Volume 23, Number 2, June 2022 ISSN 1234-7655

STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association and Statistics Poland

IN THIS ISSUE:

EDITORIAL OFFICE **ISSN 1234-7655**

Scientific Secretary
Marek Cierpiał-Wolan, *Statistical Office in Rzeszów, Rzeszów, Poland, e-mail: m.cierpial-wolan@stat.gov.pl*

Managing Editor
Adriana Nowakowska, *Statistics Poland, Warsaw, e-mail: a.nowakowska3@stat.gov.pl*

Secretary
Patryk Barszcz, *Statistics Poland, Warsaw, Poland, e-mail: p.barszcz@stat.gov.pl, phone number +48 22 – 608 33 66*

Technical Assistant
Rajmund Litkowiec, *Statistical Office in Rzeszów, Rzeszów, Poland, e-mail: r.litkowiec@stat.gov.pl*

Address for correspondence

Statistics Poland, al. Niepodległości 208, 00-925 Warsaw, Poland, tel./fax: +48 22 – 825 03 95

CONTENTS

Research articles

Other articles

39th Multivariate Statistical Analysis 2021, Lodz. Conference Papers

Research Communicates and Letters

3rd Congress of Polish Statistics

Conference reports

Volume 23, Number 2, June 2022

STATISTICS IN TRANSITION new series, June 2022
Vol. 23, No. 2, pp. III–VIII

From the Editor

The passing quarter was full of a series of events of great importance for the scientific life of the global community of statisticians. Two of them – the meeting of IAOS/International Association for Official Statistics and the 3rd Congress of Polish Statistics on the occasion of the 110th anniversary of the Polish Statistical Association/PSA – took place at the same time and in the same venue, in Krakow's Convention Center, on April 25–28. Like previous congresses, this one also gave the opportunity to award the highest distinction, which the Polish Statistical Association awards to distinguished persons for an extraordinary contribution to the development of statistical sciences, which is the medal of Jerzy Spława-Neyman. This time, the Neyman medal was awarded to Prof. Partha Lahiri, Danny Pfeffermann and Wlodzimierz Okrasa. Along with congratulations to the laureates, their short bionotes are included in the first section of this issue.

The June issue presents a set of 12 articles – there are 8 manuscripts in Research papers section, 3 conference papers from the 39th Multivariate Statistical Analysis 2021, which took place in November 2021 in Lodz, Poland provided as others articles, and 1 paper published in the Research Communciates&Letters part. Our authors come from Egypt, Saudi Arabia, Pakistan, USA, Poland, India, Nigeria, Malesia, Algeria, and Italy. We are pleased to be recognized by such a respectable group of scientists.

Research articles

In the first paper, **Abdelfattah Mustafa A.** and **M. I. Khan** discuss ***The length-biased power hazard rate distribution: some properties and applications.*** The authors show that this distribution reports an extension of several probability distributions, namely: exponential, Rayleigh, Weibull, and linear hazard rate. The procedure of maximum likelihood estimation was taken for parameters and derived. The applicability of the model was explored by three real data sets. Also, to examine the performance of the technique, a simulation study is extracted. The superiority of the new model has been exhibited by some real data sets. It has been seen that Power Hazard Rate Distribution can adequately provide better fits than other models.

The article entitled ***Jackknife winsorized variance estimator under imputed data*** prepared by **Muhammad Umair SohaiL, Fariha Sohil, Javid Shabbir,** and **Sat Gupta** show the problem of missing and extreme values for the estimation of population variance. The presence of extreme values either in the study variable, or the auxiliary variable, or in both of them, can adversely affect the performance of the estimation procedure. The authors have considered three different situations for the presence of extreme values and also have considered jackknife variance estimators for the population variance by handling these extreme values under stratified random sampling. Bootstrap technique ABB was carried out to understand the relative relationship more precisely. The authors also modified the linearized version of the jackknife variance estimator suggested by Rao (1996) for the precise estimation of winsorized variance, which is helpful with computer programs that use linearized methods for the estimation of variance. The stratified sampling scheme was discussed as it is commonly used in large scale socio-economic surveys.

Maciej Jewczak and **Magdalena Brudz** in their manuscript ***Socio-economic development and quality of life of Nuts-2 units in the European Union*** examine the level of socio-economic development and quality of life in the European Union in the years 2004 and 2018. The analyses were conducted for a rarely used level of spatial data aggregation, i.e. for NUTS-2 units, but they cover only those European regions that were EU members in 2004. As the primary research tool, the two-dimensional development matrix was adopted, which enabled the verification of the hypothesis regarding the convergence of synthetic measures that indicate the levels of socio-economic development and quality of life in the EU regions. For these indices, the development matrix was also used to identify the strengths and weaknesses as well as the opportunities and threats for selected spatial units, and, at the same time, to estimate the rates of change of the socio-economic development and quality of life levels. Depending on the criteria considered, the most common methods for determining the degree of the advancement of life quality or socio-economic development include taxonomical techniques and analyses of potential, which are based mainly on objective data sourced from official registers. A very important fact from this study is that the scientific analysis covered data at the regional level, while most studies focus only on quality of life or socio-economic development at the macro level.

Arora S., Mahajan K. K., and **Jangra V.** present ***A Bayesian estimation of the Gini index and the Bonferroni index for the Dagum distribution with the application of different priors.*** The Bayesian estimators and highest posterior density credible intervals were obtained for two popular inequality measures, viz. the Gini index and the Bonferroni index in the case of the Dagum distribution. The study has considered the informative and non-informative priors, i.e. the Mukherjee-Islam prior and the extension of Jeffrey's prior, respectively, under the presumption of the Linear Exponential (LINEX) loss function. The authors have carried out a Monte Carlo simulation study in order to obtain the relative efficiency of both the Gini and Bonferroni indices while taking into consideration different priors and loss functions.

It was observed that Mukherjee-Islam prior performs better than the extension of Jeffrey's prior in terms of having smaller estimated loss. It was also observed that the LINEX loss function results in smaller loss as compared to squared error loss function (SELF) for small, medium and large sample sizes irrespective of the choice of prior. The expected loss decreases as the sample size increases.

The paper by **Sanusi Alhaji Jibrin** and **Rosmanjawati Abdul Rahman** entitled ***ARFURIMA models: simulations of their properties and applications*** defines the Autoregressive Fractional Unit Root Integrated Moving Average (ARFURIMA) model for modelling ILM time series with fractional difference value in the interval of $1<dd<2$. The performance of the ARFURIMA model was examined through a Monte Carlo simulation. Also, some applications were presented using the energy series, bitcoin exchange rates and some financial data to compare the performance of the ARFURIMA and the Semiparametric Fractional Autoregressive Moving Average (SEMIFARMA) models. The presented simulations studies confirmed superiority of the ARFURIMA over the ARIMA in simulating nonstationary and the FURI series and thus proved the ILM properties of the ARFURIMA model and its large sample properties too. Some applications of the model were presented and further confirmed a better fit of the ARFURIMA compared to the SEMIFARMA model.

Abdelmalek Gagui's and **Abdelhak Chouaf's** article ***On the nonparametric estimation of conditional hazard estimator in the single functional index*** characterises the conditional hazard estimator of a real response where the variable is given a functional random variable (i.e. it takes values in an infinite-dimensional space). The authors focus on the functional index model as a good compromise between nonparametric and parametric models to prove the asymptotic normality of the proposed estimator under general conditions and in cases where the variables satisfy the strong mixing dependency. The means of the kernel estimator method, based on a single-index structure, were used. A simulation of the proposed methodology has shown that it is efficient for large sample sizes. It was also shown that the estimator provides good predictions under this model. In non-parametric functional statistics, the semi-metric of the projection type is very important for increasing the concentration property. The functional index model is a special case of this family of semi-metrics because it is based on the projection on a functional direction, which is important for the implementation of the method in practice.

In the next paper **Mateusz Borkowski** focuses on ***Institutional equilibrium in EU economies in 2008 and 2018: SEM-PLS models*** to identify the strength and direction of the development of the relationship between formal and informal institutions and to assess the institutional equilibrium of modern economies. The article presents a comprehensive model of the institutional structure and a unique method of measuring institutional equilibrium. The structural equations modelling based on partial least squares (SEM-PLS) was applied. The study included 27 EU economies and the research period covered the years 2008 and 2018. The results of the study demonstrate that the quality of informal institutions strongly, positively determines the quality of formal institutions. The conducted analyses indicate that modern economies are diversified in terms of the quality of informal and formal institutions and, consequently, in institutional equilibrium.

Sakshi Kaushik, Alka Sabharwal, and **Gurprit Grover** present the manuscript entitled ***Extracting relevant predictors of the severity of mental illnesses from clinical information using regularisation regression models.*** The authors describe the relevant predictors of the severity of mental illnesses (measured by psychiatric rating scales) from a wide range of clinical variables consisting of information on both laboratory test results and psychiatric factors. The laboratory test results collectively indicate the measurements of 23 components derived from vital signs and blood tests results for the evaluation of the complete blood count. The 8 psychiatric factors known to affect the severity of mental illnesses are considered, viz. the family history, course and onset of an illness, etc. Retrospective data of 78 patients diagnosed with mental and behavioural disorders were collected from the Lady Hardinge Medical College & Smt. S.K, Hospital in New Delhi, India. The observations missing in the data were imputed using the non-parametric random forest algorithm. This paper adds to the literature of medical research aimed at identifying the biomarkers for diagnosis and predictors of the severity status of mental disorders, and should be helpful in developing valid and efficient approaches to diagnose the disorders at an early stage The clinicians can use the relevant factors to build a profile of the patient and his needs, and also effective strategies for treatment planning.

Other articles

39th Multivariate Statistical Analysis 2021, Lodz. Conference Papers

The section starts with the paper prepared by **Czesław Domański** and **Robert Kubacki** entitled ***Regression model of water demand for the city of Lodz as a function of atmospheric factors.*** The authors presented the results of the work on a statistical model which determined the influence of individual atmospheric factors on the demand for water in the city of Lodz, Poland, in 2010-2019. In order to build the model, the study used data from the Water Supply and Sewage System Company (Zakład Wodociągów i Kanalizacji Sp. z o.o.) in the city of Lodz complemented with data on weather conditions in the studied period. The analysis showed that the constructed models make it possible to perform a forecast of water demand depending on the expected weather conditions. The relation between daily weather variables and water use in the city of Lodz, Poland were examined. It was confirmed that the maximum daily temperature is a good predictor of water demand, and that holidays are significant in decreasing the water demand. Moreover, wind speed is a good predictor of water demand. It is likely that higher wind speed increases evaporation of water, which induces a cooling effect and thus decreases daily water consumption. Together, all these variables explain between 65% of the variations in the city of Lodz.

Stefano Bonnini's and **Getnet Melak Assegie's** article evaluates ***Advances on permutation multivariate analysis of variance for big data.*** Due to the gap in the literature about combined permutation tests, in particular for big data with a large number of variables**,** a Monte Carlo simulation study was carried out to investigate the power behaviour of the tests, and the application to a real case study was performed to show the utility of the method. It was provided that among the distribution free solutions to the multivariate analysis of variance in the family of combined permutation tests, the method based on the Tippet combination is in general preferable, especially if there is no preventive information about the possible percentage of variables (or marginal distributions) under the alternative hypothesis. Instead of the Tippett combination, the Fisher rule can be applied when the percentage is close to 100%. The Liptak combination seems to be non-convenient in general. This methodological tool is an important and useful solution of testing problems for big data, especially when the number of variables is very large and the sample sizes are small. The usefulness and the effectiveness of the method is confirmed by the application to the case study concerning the survey on the organizational well-being at the University of Ferrara.

The paper by **Tadeusz Bednarski, Piotr B. Nowak,** and **Magdalena Skolimowska-Kulig** examines ***Scaled Fisher consistency for the partial likelihood estimation in various extensions of the Cox model.*** The Cox proportional hazards model has become the most widely used procedure in survival analysis, and the theoretical basis of the original model has been developed in various extensions. The authors have investigated the accuracy of inference based on the primer Cox model in the existence of unobserved heterogeneity, that is, when the data generating mechanism is more complex than presumed and described by the kind of an extension of the Cox model with undefined frailty. It was shown that the conventional partial likelihood estimator under the considered extension is Fisher-consistent up to a scaling factor, provided symmetry-type distributional assumptions on covariates. The results of simulation experiments that reveal an exemplary behaviour of the estimators were presented.

Research Communicates and Letters

The Research Communicates & Letters section presents a paper prepared by **Hemlata Joshi, S., Azarudheen, M. S. Nagaraja,** and **C. Singh** entitled ***On the quick estimation of probability of recovery from COVID-19 during first wave of epidemic***

in India: a logistic regression approach. Due to the fact that the COVID-19 became a threat all across the world with the new cases every day, and there is still a difficult situation with no effective medicine, it is very important to know if a patient with COVID-19 is going to recover or die. The study is based on the situation in India and the data published by the Ministry of Health and Family Welfare of India were used for the empirical analysis. The manuscript shows a model that has been developed to estimate the probability of recovery of a patient based on the demographic characteristics, as most of the Indian population is living in poor hygienic conditions. The probability model is developed using the indirect method of estimation based on some demographic factors, and it was found that the probability of recovery from coronavirus disease is statistically the same in both males and females.

Włodzimierz Okrasa
Editor

STATISTICS IN TRANSITION new series, June 2022
Vol. 23, No. 2, pp. IX

Submission information for Authors

Statistics in Transition new series (SiT) is an international journal published jointly by the Polish Statistical Association (PTS) and Statistics Poland, on a quarterly basis (during 1993–2006 it was issued twice and since 2006 three times a year). Also, it has extended its scope of interest beyond its originally primary focus on statistical issues pertinent to transition from centrally planned to a market-oriented economy through embracing questions related to systemic transformations of and within the national statistical systems, world-wide.

The SiT-*ns* seeks contributors that address the full range of problems involved in data production, data dissemination and utilization, providing international community of statisticians and users – including researchers, teachers, policy makers and the general public – with a platform for exchange of ideas and for sharing best practices in all areas of the development of statistics.

Accordingly, articles dealing with any topics of statistics and its advancement – as either a scientific domain (new research and data analysis methods) or as a domain of informational infrastructure of the economy, society and the state – are appropriate for *Statistics in Transition new series.*

Demonstration of the role played by statistical research and data in economic growth and social progress (both locally and globally), including better-informed decisions and greater participation of citizens, are of particular interest.

Each paper submitted by prospective authors are peer reviewed by internationally recognized experts, who are guided in their decisions about the publication by criteria of originality and overall quality, including its content and form, and of potential interest to readers (esp. professionals).

Manuscript should be submitted electronically to the Editor:
sit@stat.gov.pl,
GUS/Statistics Poland,
Al. Niepodległości 208, R. 296, 00-925 Warsaw, Poland

It is assumed, that the submitted manuscript has not been published previously and that it is not under review elsewhere. It should include an abstract (of not more than 1600 characters, including spaces). Inquiries concerning the submitted manuscript, its current status etc., should be directed to the Editor by email, address above, or w.okrasa@stat.gov.pl.

For other aspects of editorial policies and procedures see the SiT Guidelines on its Web site: http://stat.gov.pl/en/sit-en/guidelines-for-authors/

STATISTICS IN TRANSITION new series, June 2022
Vol. 23, No. 2, pp. XI–XII

Editorial Policy

The broad objective of *Statistics in Transition new series* is to advance the statistical and associated methods used primarily by statistical agencies and other research institutions. To meet that objective, the journal encompasses a wide range of topics in statistical design and analysis, including survey methodology and survey sampling, census methodology, statistical uses of administrative data sources, estimation methods, economic and demographic studies, and novel methods of analysis of socio-economic and population data. With its focus on innovative methods that address practical problems, the journal favours papers that report new methods accompanied by real-life applications. Authoritative review papers on important problems faced by statisticians in agencies and academia also fall within the journal's scope.

***

Abstracting and Indexing Databases

Statistics in Transition new series is currently covered in:

Databases indexing the journal:

- BASE – Bielefeld Academic Search Engine
- CEEOL – Central and Eastern European Online Library
- CEJSH (The Central European Journal of Social Sciences and Humanities)
- CNKI Scholar (China National Knowledge Infrastructure)
- CNPIEC – cnpLINKer
- CORE
- Current Index to Statistics
- Dimensions
- DOAJ (Directory of Open Access Journals)
- EconPapers
- EconStore
- Electronic Journals Library
- Elsevier – Scopus
- ERIH PLUS (European Reference Index for the Humanities and Social Sciences)
- Genamics JournalSeek
- Google Scholar
- Index Copernicus
- J-Gate
- JournalGuide
- JournalTOCs
- Keepers Registry
- MIAR
- Microsoft Academic
- OpenAIRE
- ProQuest – Summon
- Publons
- QOAM (Quality Open Access Market)
- ReadCube
- RePec
- SCImago Journal & Country Rank
- TDNet
- Technische Informationsbibliothek (TIB) - German National Library of Science and Technology
- Ulrichsweb & Ulrich's Periodicals Directory
- WanFang Data
- WorldCat (OCLC)
- Zenodo

STATISTICS IN TRANSITION new series, June 2022
Vol. 23, No. 2, pp. 1–16, DOI 10.2478/stattrans-2022-0013
Received – 09.02.2021; accepted – 30.07.2021

The length-biased power hazard rate distribution: Some properties and applications

Abdelfattah Mustafa[1,2], **M. I. Khan**[2]

ABSTRACT

In this article, the length-biased power hazard rate distribution has introduced and investigated several statistical properties. This distribution reports an extension of several probability distributions, namely: exponential, Rayleigh, Weibull, and linear hazard rate. The procedure of maximum likelihood estimation is taken for parameters. Finally, the applicability of the model is explored by three real data sets. To examine, the performance of the technique, a simulation study is extracted.

Key words: length-biased, power hazard rate distribution, maximum likelihood estimation.

1. Introduction

Importance of the statistical distributions in different fields of studies, researchers have shown their curiosity to suggest a new distribution via numerous methods. In pioneering work, Cox (1962) proposed a model dealing with the unequal probability of sample observation termed as length-biased technique. This concept has many applications in biomedical sciences, Lawless (2003).

Several papers have been arisen to investigate the performance of length-biased distributions. For instance see Gupta and Keating (1985), Khattree (1989), Gupta and Tripathi (1990), Oluyede (1999), Das and Roy (2011a,b), Ratnaparkhi and Nimbalkar (2012), Al-Khadim and Hussain (2014), Nanuwong and Bodhisuwan (2014), Seenoi et al. (2014) , Modi (2015), Saghir et al. (2016), Saghir et al. (2017), Mudasir and Ahmad (2018) and Parveen and Ahmad (2018), among others.

The lifetime distributions are always characterized by selecting a specific hazard rate function (HRF). The power HRF is one of them. The HRF is used in many fields of study (reliability analysis, actuarial sciences, demography, and economics). The inference on hazard function for lifetime data has become a prevalent tool for researchers.

The power hazard function (PHF) was introduced by Mugdadi (2005).

$$h(x) = \lambda x^{\nu}, \quad x > 0, \lambda > 0, \nu > -1. \tag{1}$$

In view of (1) cumulative distribution function (cdf) is given

$$F(x) = 1 - e^{-\frac{\lambda}{\nu+1}x^{\nu+1}}, \quad x > 0, \lambda > 0, \nu > -1, \tag{2}$$

[1] Department of Mathematics, Faculty of Science, Mansoura University, Mansoura 35516, Egypt. E-mail: amelsayed@mans.edu.eg. ORCID: https://orcid.org/0000-0002-8551-6115.

[2] Department of Mathematics, Faculty of Science, Islamic University of Madinah, KSA. E-mail: izhar.stats@gmail.com. ORCID: https://orcid.org/0000-0002-5793-9786.

and the probability density function (pdf) is

$$f(x) = \lambda x^{\nu} e^{-\frac{\lambda}{\nu+1}x^{\nu+1}}, \quad x > 0,\ \lambda > 0,\ \nu > -1. \tag{3}$$

If X has pdf (3), we denote it by $X \sim \text{PHRD}(\lambda, \nu)$.

The PHF is very simple, and it could be increasing, decreasing, or constant. Therefore, the PHR distribution contributes a better fit over two-parameter distributions when modelling monotone hazard rates. More explorationon the PHR distribution can be seen in Ismail (2014), Mugdadi and Min (2009), Tarvirdizade and Nematollahi (2016) and Tarvirdizade and Nematollahi (2020). It is important to note that some familiar distributions are special case of (3) reported in Section 2.1.

The paper is organized as follows. The formulation of length-biased PHR distribution (LBPHRD) and its structured properties are discussed in Section 2. Section 3 is devoted to estimating the parameters via the maximum likelihood method. Section 4 reveals the usefulness of the new model and, also simulation study is evaluated to examine the performance of MLEs. The conclusion is presented in Section 5.

2. Length-Biased Power Hazard Rate Distribution

The LBPHR distribution is proposed in this section. The shape of the pdf, hazard rate and some sub-models are established also.

Definition 1. If the random variable X has a pdf $f(x)$ and expected value $E(X) < 0$ then the pdf of the length- biased distribution of X can be formulated as

$$g(x) = \frac{x f(x)}{E(X)}. \tag{4}$$

From (3) and (4), the LBPHR distribution with two parameters λ (scale) and ν (shape) can be obtained as follows

$$g(x) = \frac{\lambda x^{\nu+1} e^{-\frac{\lambda}{\nu+1}x^{\nu+1}}}{\left(\frac{\nu+1}{\lambda}\right)^{\frac{1}{\nu+1}} \Gamma\left(\frac{\nu+2}{\nu+1}\right)}, \quad x > 0, \tag{5}$$

where $\Gamma(n) = \int_0^{\infty} u^{n-1} e^{-u} du$ is gamma function.

The graph of the pdf of LBPHRD is shown in Figure 1, for various values of λ and ν.

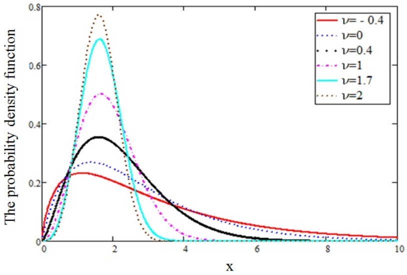


Figure 1. The plot of $g_{LBPHR}(x)$ for $\lambda = 0.73$ and various values of ν.

From Figure 1, the pdf of LBPHRD has one peak, so there is one mode.

The cdf of LBPHR distribution has the form

$$G(x) = \frac{\gamma\left(\frac{\nu+2}{\nu+1}, \frac{\lambda}{\nu+1}x^{\nu+1}\right)}{\Gamma\left(\frac{\nu+2}{\nu+1}\right)}, \quad x > 0, \tag{6}$$

where $\gamma(a,x) = \int_0^x t^{a-1}e^{-t}dt$ is an upper incomplete gamma function.

The survival (reliability) function of LBPHRD is given as

$$\bar{G}(x) = \frac{\Gamma\left(\frac{\nu+2}{\nu+1}, \frac{\lambda}{\nu+1}x^{\nu+1}\right)}{\Gamma\left(\frac{\nu+2}{\nu+1}\right)}, \quad x > 0, \tag{7}$$

where $\Gamma(a,x) = \int_x^\infty t^{a-1}e^{-t}dt$ is an incomplete gamma function.

The hazard rate of LBPHRD takes the form

$$h(x) = \frac{\lambda x^{\nu+1}e^{-\frac{\lambda}{\nu+1}x^{\nu+1}}}{\left(\frac{\nu+1}{\lambda}\right)^{\frac{1}{\nu+1}}\Gamma\left(\frac{\nu+2}{\nu+1}\right)}, \quad x > 0. \tag{8}$$

Derivative the $h(x)$, w.r.t. x,

$$h^{'}(x) = \frac{1}{\left(\frac{\nu+1}{\lambda}\right)^{\frac{1}{\nu+1}}\Gamma\left(\frac{\nu+2}{\nu+1}\right)}\left[\lambda(\nu+1) - \lambda^2 x^{\nu+1}\right]x^{\nu}e^{-\frac{\lambda}{\nu+1}x^{\nu+1}},$$

by equating $h^{'}(x)$ by zero, we find $x = 0$ and $x = \left(\frac{\nu+1}{\lambda}\right)^{\frac{1}{\nu+1}}$ are the critical points for $h(x)$.

By using the second derivetives test, we can find

$$h^{''}(x) = \frac{\lambda}{\left(\frac{\nu+1}{\lambda}\right)^{\frac{1}{\nu+1}}\Gamma\left(\frac{\nu+2}{\nu+1}\right)}\left[\nu(\nu+1) - \lambda(3\nu+2)x^{\nu+1} + \lambda^2 x^{2(\nu+1)}\right]x^{\nu-1}e^{-\frac{\lambda}{\nu+1}x^{\nu+1}}.$$

- At $x = 0$, $h^{''}(x) = 0$, then $x = 0$ is the inflection point.

- At $x = \left(\frac{\nu+1}{\lambda}\right)^{\frac{1}{\nu+1}}$,

$$h^{''}(x) = -\frac{\lambda(\nu+1)^2 x^{\nu-1}}{\left(\frac{\nu+1}{\lambda}\right)^{\frac{1}{\nu+1}}\Gamma\left(\frac{\nu+2}{\nu+1}\right)}e^{-\frac{\lambda}{\nu+1}x^{\nu+1}} < 0,$$

  then $h(x)$ has a local maximum at $x_0 = \left(\frac{\nu+1}{\lambda}\right)^{\frac{1}{\nu+1}}$.

Some hazard rate plots of the LBPHR distribution with specific parameter values are given in Figure 2.

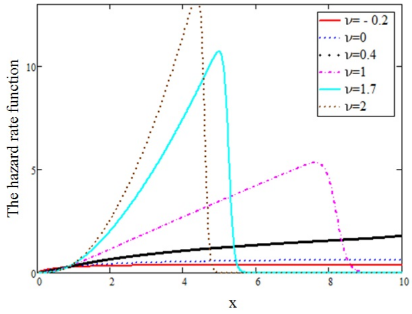


Figure 2. The HRF of the LBPHRD for $\lambda = 0.73$ and various values of ν.

Therefore, the function $h(x) \uparrow$ on the interval $(0, x_0)$ and $h(x) \downarrow$ on the interval (x_0, ∞). From the Figure 2, the hazard function exhibits that proposed model becomes a major tool to fit many lifetime data in (reliability, survival analysis, finance and economics).

2.1. Special cases of LBPHRD

The LBPHRD is very versatile distribution. It covers many noted distribution as follows.

1. Setting $\nu = \lambda - 1$, we obtain the length-biased Weibull (LBW) distribution as obtained by Shaban and Boudrissa (2007).
2. Setting $\nu = 1$, we obtain the length-biased Rayleigh (LBR) distribution with parameter $\frac{1}{\lambda}$ as obtained by Ajami and Jahanshahi (2017).
3. Setting $\nu = 0$, we obtain the length-biased exponential (LBE) distribution as obtained by Mir et al. (2013).
4. Setting $\nu = 1$, we obtain the length-biased linear failure rate (LBLFR) distribution.

The results obtained in this paper can be valid for these distributions and the other distributions which have a power hazard function.

2.2. Statistical properties

Some statistical properties of the LBPHRD are discussed in this section.
Theorem 1. If $X \sim \text{LBPHRD}(\lambda, \nu)$ then the rth moment is given as

$$E(X^r) = \frac{\left(\frac{\nu+1}{\lambda}\right)^{\frac{r}{\nu+1}} \Gamma\left(\frac{r+\nu+2}{\nu+1}\right)}{\Gamma\left(\frac{\nu+2}{\nu+1}\right)}. \tag{9}$$

Proof. The rth moments of LBPHRD can be attained by

$$E(X^r) = \int_0^\infty x^r g(x) dx,$$

from (5), then

$$E(X^r) = \int_0^\infty \frac{\lambda x^{r+\nu+1} e^{-\frac{\lambda}{\nu+1}x^{\nu+1}}}{\left(\frac{\nu+1}{\lambda}\right)^{\frac{1}{\nu+1}} \Gamma\left(\frac{\nu+2}{\nu+1}\right)} dx. \tag{10}$$

Let $u = \frac{\lambda}{\nu+1}x^{\nu+1}$, $du = \lambda x^\nu dx$. Upon simplification, (10) leads to

$$E(X^r) = \frac{\left(\frac{\nu+1}{\lambda}\right)^{\frac{r}{\nu+1}} \Gamma\left(\frac{r+\nu+2}{\nu+1}\right)}{\Gamma\left(\frac{\nu+2}{\nu+1}\right)}. \tag{11}$$

The mean and variance for LBPHRD can be calculated from (11) as follows. Setting $r = 1$, in (11),

$$E(X) = \frac{\left(\frac{\nu+1}{\lambda}\right)^{\frac{1}{\nu+1}} \Gamma\left(\frac{\nu+3}{\nu+1}\right)}{\Gamma\left(\frac{\nu+2}{\nu+1}\right)}. \tag{12}$$

Putting $r = 2$, in (11),

$$E(X^2) = \frac{\left(\frac{\nu+1}{\lambda}\right)^{\frac{2}{\nu+1}} \Gamma\left(\frac{\nu+4}{\nu+1}\right)}{\Gamma\left(\frac{\nu+2}{\nu+1}\right)}. \tag{13}$$

Therefore, variance of LBPHRD is

$$Var(X) = \frac{\left(\frac{\nu+1}{\lambda}\right)^{\frac{2}{\nu+1}} \Gamma\left(\frac{\nu+4}{\nu+1}\right)}{\Gamma\left(\frac{\nu+2}{\nu+1}\right)} - \left[\frac{\left(\frac{\nu+1}{\lambda}\right)^{\frac{1}{\nu+1}} \Gamma\left(\frac{\nu+3}{\nu+1}\right)}{\Gamma\left(\frac{\nu+2}{\nu+1}\right)}\right]^2. \tag{14}$$

The shape characteristics of the probability distribution, skewness and kurtosis play an important role.These can be derived from Theorem 1, using the following relations.

$$S_k = \frac{\mu_3^{'} - 3\mu_1^{'}\mu_2^{'} + 2\mu_1^{'3}}{(\mu_2^{'} - \mu_1^{'})^{3/2}}, \qquad K_u = \frac{\mu_4^{'} - 4\mu_1^{'}\mu_3^{'} + 6\mu_1^{'2}\mu_2^{'} - 3\mu_1^{'4}}{(\mu_2^{'} - \mu_1^{'})^2},$$

where $\mu_r^{'} = E(X^r)$.

The mode of the LBPHRD:

Taking the logarithm of (5), we have

$$\ln g(x) = \ln(\lambda) + (\nu+1)\ln(x) - \frac{\lambda}{\nu+1}x^{\nu+1} - \ln\left[\left(\frac{\nu+1}{\lambda}\right)^{\frac{1}{\nu+1}} \Gamma\left(\frac{\nu+2}{\nu+1}\right)\right]. \tag{15}$$

Differentiate (15) w.r.t. x and equating it zero,

$$\frac{d}{dx}\ln g(x) = \frac{\nu+1}{x} - \lambda x^\nu = 0, \tag{16}$$

therefore

$$x = \left(\frac{\nu+1}{\lambda}\right)^{\frac{1}{\nu+1}}.$$

Again differentiate (16),

$$\frac{d^2}{dx^2}\ln g(x) = -\frac{\nu+1}{x^2} - \lambda\nu x^{\nu-1} = -\frac{(\nu+1)+\lambda\nu x^{\nu+1}}{x^2},$$

at $x = \left(\frac{\nu+1}{\lambda}\right)^{\frac{1}{\nu+1}}$, then

$$\frac{d^2}{dx^2}\ln g(x) = -\frac{(\nu+1)^2}{x^2} < 0.$$

Therefore, the mode is $x = \left(\frac{\nu+1}{\lambda}\right)^{\frac{1}{\nu+1}}$.

Using the following relation, pth percentile can be obtained

$$G(x_p;\lambda,\nu) = p. \tag{17}$$

Substituting from (6) into (17), x_p satisfies the equation

$$\Gamma\left(\frac{\nu+2}{\nu+1}, \frac{\lambda}{\nu+1}x^{\nu+1}\right) - p\Gamma\left(\frac{\nu+2}{\nu+1}\right) = 0. \tag{18}$$

The pth percentile can be calculated numerically by using Equation (18).

The median can be calculated from Equation (18), at $p = 0.5$.

For $\lambda = 0.5$, $\nu \in (0,5)$, the values of $E(X)$, mode, $Var(X)$, s_k, k_u and CV for LBPHRD and PHRD, respectively are presented in Table 1.

Table 1. Some statistical measures for $\lambda = 0.5, \nu \in (0,5)$.

	LBPHRD						PHRD					
ν	$E(X)$	Mode	$Var(X)$	S_k	K_u	CV	$E(X)$	Mode	$Var(X)$	S_k	K_u	CV
0.0	4.000	2.000	8.000	1.414	6.000	70.71	2.000	0.000	4.000	2.000	9.000	100.00
0.5	2.743	2.080	2.059	0.813	3.780	52.31	1.878	1.000	1.626	1.072	4.390	67.90
1.0	2.257	2.000	0.907	0.486	3.108	42.20	1.772	1.414	0.858	0.631	3.245	52.27
1.5	1.998	1.904	0.507	0.269	2.864	35.64	1.689	1.552	0.522	0.359	2.857	42.78
2.0	1.837	1.817	0.323	0.111	2.786	30.94	1.623	1.587	0.348	0.168	2.729	36.35
2.5	1.726	1.744	0.224	-0.012	2.784	27.42	1.569	1.584	0.246	0.025	2.713	31.61
3.0	1.644	1.682	0.164	-0.110	2.819	24.63	1.524	1.565	0.183	-0.087	2.748	28.07
3.5	1.582	1.629	0.125	-0.191	2.874	22.35	1.487	1.541	0.141	-0.178	2.808	25.25
4.0	1.532	1.585	0.099	-0.259	2.938	20.54	1.455	1.516	0.111	-0.254	2.880	22.90
4.5	1.491	1.546	0.080	-0.318	3.006	18.97	1.428	1.491	0.09	-0.318	2.957	21.01
5.0	1.456	1.513	0.066	-0.369	3.076	17.64	1.404	1.468	0.074	-0.373	3.035	19.38

From Table 1, we can conclude that:

1. the LBPHRD is positive skewed, for $\nu < 2.5$, while PHRD is positive skewed, for $\nu \leq 2.5$.

2. the LBPHRD is negative skewed, for $\nu \geq 2.5$, while PHRD is negative skewed, for $\nu > 2.5$

3. when $\nu = 0.0$, the LBPHRD and PHR are highly skewed, ($S_k > 1$).

4. when $\nu = 0.5$, the LBPHRD is moderately skewed, $(0.5 < S_k < 1)$, while PHRD is highly skewed.

5. when $\nu = 1$, the LBPHRD is approximately symmetric, $(-0.5 < S_k < 0.5)$, while PHRD is moderately skewed $(0.5 < s_k < 1)$.

6. when $1.5 \leq \nu \leq 5$, the LBPHRD and PHRD are approximately symmetric.

7. the dispersion for the distributions are decreasing for ν increasing.

8. for $0.0 \leq \nu \leq 1.0$ the LBPHRD and PHRD are leptokurtic $(S_k > 3)$.

9. for $1.5 \leq \nu \leq 4$, the LBRHRD and PHRD are platykurtic $(S_k < 3)$.

10. for $4.5 \leq \nu \leq 5$, the LBPHRD and PHRD are mesokurtic $(S_k \cong 3)$.

11. Since the coefficient of variation $(Cv = \frac{\sqrt{Var(X)}}{mean} \times 100)$ is larger for PHRD, the PHRD are more variable than the LBPHRD, for all values of ν.

Therefore, the LBPHR model is more flexible than PHR model.

3. Estimation of Parameters

Consider $X_1, X_2, \cdots, X_n$ be a random sample from LBPHRD, the Maximum likelihood estimation (MLE) can be applied to estimate the parameters as follows. The likelihood function is given by

$$L(\lambda, \nu; x) = \frac{\lambda^n \left(\prod_{i=1}^n x_i^{\nu+1}\right) e^{-\frac{\lambda}{\nu+1}\sum_{i=1}^n x_i^{\nu+1}}}{\left(\frac{\nu+1}{\lambda}\right)^{\frac{n}{\nu+1}} \left(\frac{1}{\nu+1}\right)^n \Gamma^n\left(\frac{1}{\nu+1}\right)}, \quad x > 0. \tag{19}$$

The log-likelihood function is

$$\begin{aligned}\mathscr{L} &= n\ln(\lambda) + (\nu+1)\sum_{i=1}^n \ln(x_i) - \frac{\lambda}{\nu+1}\sum_{i=1}^n x_i^{\nu+1} - \frac{n}{\nu+1}\ln\left(\frac{\nu+1}{\lambda}\right) \\ &\quad + n\ln(\nu+1) - n\ln\left[\Gamma\left(\frac{1}{\nu+1}\right)\right]. \end{aligned} \tag{20}$$

Differentiate Equation (20) w.r.t. λ and ν. Equating the derivatives to zero, we get the normal equations as follows.

$$\frac{\partial \mathscr{L}}{\partial \lambda} = \frac{n}{\lambda}\left(1 + \frac{1}{\nu+1}\right) - \frac{1}{\nu+1}\sum_{i=1}^n x_i^{\nu+1} = 0, \tag{21}$$

$$\begin{aligned}\frac{\partial \mathscr{L}}{\partial \nu} &= \sum_{i=1}^n \ln(x_i) + \frac{\lambda}{(\nu+1)^2}\sum_{i=1}^n x_i^{\nu+1}\Big[1 - (\nu+1)\ln(x_i)\Big] + \frac{n}{(\nu+1)^2}\ln\left(\frac{\nu+1}{\lambda}\right) \\ &\quad + \frac{n\nu}{(\nu+1)^2} - n\psi\left(\frac{1}{\nu+1}\right) = 0, \end{aligned} \tag{22}$$

where $\psi(x) = \frac{d}{dx}\ln\Gamma(x)$ be a digamma function.

The asymptotic normality of the MLEs can be applied to compute the confidence interval (C.I.) for the parameters. The observed variance and covariance matrix of $\Theta = (\lambda, \nu)$ is

$$I^{-1}(\Theta) = \begin{bmatrix} -\frac{\partial^2 \mathscr{L}}{\partial \lambda^2} & -\frac{\partial^2 \mathscr{L}}{\partial \lambda \partial \nu} \\ -\frac{\partial^2 \mathscr{L}}{\partial \nu \partial \lambda} & -\frac{\partial^2 \mathscr{L}}{\partial \nu^2} \end{bmatrix}^{-1} = \begin{bmatrix} -I_{11} & -I_{12} \\ -I_{21} & -I_{22} \end{bmatrix}^{-1},$$

where

$$I_{11} = -\frac{n}{\lambda^2}\left(1 + \frac{1}{\nu+1}\right), \tag{23}$$

$$I_{12} = -\frac{n}{\lambda(\nu+1)^2} + \frac{1}{(\nu+1)^2}\sum_{i=1}^{n} x_i^{\nu+1} - \frac{1}{\nu+1}\sum_{i=1}^{n} x_i^{\nu+1}\ln(x_i), \tag{24}$$

$$I_{21} = I_{12}, \tag{25}$$

$$\begin{aligned} I_{22} = {} & -\frac{\lambda}{(\nu+1)^3}\sum_{i=1}^{n} x_i^{\nu+1}\Big[1-(\nu+1)\ln(x_i)\Big]\Big[2-(\nu+1)\ln(x_i)\Big] - \frac{\lambda}{(\nu+1)^2} \times \\ & \sum_{i=1}^{n} x_i^{\nu+1}\ln(x_i) - \frac{2n}{(\nu+1)^3}\ln\left(\frac{\nu+1}{\lambda}\right) + \frac{(2-\nu)n}{(\nu+1)^3} - n\psi'\left(\frac{1}{\nu+1}\right), \end{aligned} \tag{26}$$

and $\psi'(x) = \frac{d^2}{dx^2}\ln\Gamma(x)$.

Asymptotic confidence interval can be derived by using observed variance and covariance matrix. A $100(1-\alpha)\%$ C.I.s of $\Theta = (\lambda, \nu)$ have the form $\hat{\lambda} \pm z_{\alpha/2}\sqrt{Var(\hat{\lambda})}$ and $\hat{\nu} \pm z_{\alpha/2}\sqrt{Var(\hat{\nu})}$. The $z_{\alpha/2}$ is upper $(\alpha/2)$th percentile of the standard normal distribution.

4. Applications

4.1. Real data

In this section, an application of LBPHR distribution using three real data sets to illustrate that it provides significant improvements over its sub-model.

Example 4.1. The data of fatigue cycle of 6061–T6 aluminum coupons cut in the horizontal direction of rolling, which is oscillated 18 rounds per second reported by Birnbaum and Saunders (1969). The data set includes 100 observations having an optional stress per round 31×10^3 psi which is reported after reducing 65 as follows.

5	25	31	32	34	35	38	39	39	40	42	43	43	43	44	44
47	47	48	49	49	49	51	54	55	55	55	56	56	56	58	59
59	59	59	59	63	63	64	64	65	65	65	66	66	66	66	66
67	67	67	68	69	69	69	69	71	71	72	73	73	73	74	74
76	76	77	77	77	77	77	77	79	79	80	81	83	83	84	86
86	87	90	91	92	92	92	92	93	94	97	98	98	99	101	103
105	109	136	147												

In Table 2, MLEs of the unknown parameters of LBR, LBW, PHR and LBPHR distributions are given along with criterion log-likelihood, AIC (Akaike's information criterion) and BIC (Bayesian information criterion).

Table 2. MLEs, $\mathscr{L}$, AIC and BIC.

Model	θ	λ	v	$\mathscr{L}$	AIC	BIC
LBR	1.722×10^{3}	–	–	-874.485	1.751×10^{3}	1.754×10^{3}
LBW	–	0.342	–	-553.06	1.108×10^{3}	1.111×10^{3}
PHR	–	1.303×10^{-5}	1.85	-475.692	955.384	960.594
LBPHR	–	1.028×10^{-4}	1.425	-454.493	912.986	918.197

Table 2 indicates that the LBPHR is best than LBR, LBW and PHR distributions in terms of model fitting for this data.

The variance and covariance matrix is given as

$$I^{-1} = \begin{bmatrix} 8.494\times10^{-9} & -1.991\times10^{-5} \\ -1.991\times10^{-5} & 0.047 \end{bmatrix}$$

Then the 95% C.I. for λ and v for LBPHRD are $(0, 2.83481\times10^{-4})$ and $(0.9993, 1.84998)$, respectively.

Figure 3 shows that the likelihood function has unique solution.

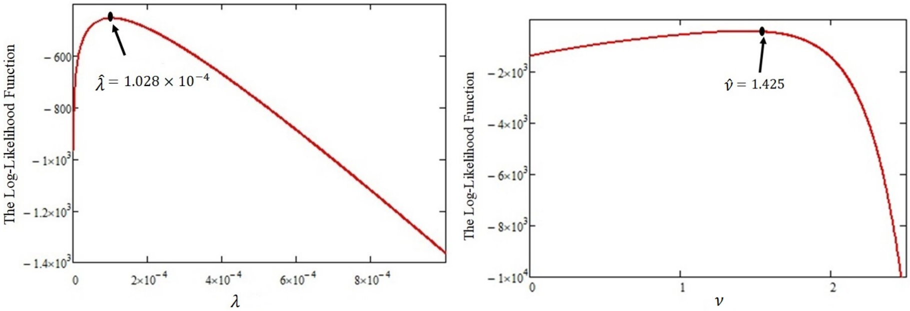


Figure 3. The outline of the $\mathscr{L}$ of λ and v.

For $\hat{\lambda} = 1.028\times10^{-4}$ and $\hat{v} = 1.425$, some statistical measures can be calculated, see Table 3.

Table 3. Some statistical measures for LBPHR at $\hat{\lambda}$ and $\hat{v}$.

Mean	Mode	Variance	Skewness	Kurtosis
67.283	63.585	602.188	0.297	2.887

From Table 3, the LBPHR distribution has,

1. the distribution is right skewed ($S_k > 0$) and it is approximately symmetric($-0.5 < S_k < 0.5$).
2. the distribution is platykurtic ($K_u < 3$).

Example 4.2. We use data collected by Balakrishnan et al. (l2010). The behavioral and emotional issues of children are scaled by GRASP (general rating of affective symptoms for preschoolers). The data (with frequency in parenthesis is the score of GRASP measurement of children) are:

19(16)	20(15)	21(14)	22(9)	23(12)	24(10)	25(6)	26(9)	27(8)	28(5)	29(6)
30(4)	31(3)	32(4)	33	34	35(4)	36(2)	37(2)	39	42	44

The MLEs and $\mathscr{L}$, AIC and BIC are reported in Table 4.

Table 4. MLEs and $\mathscr{L}$, AIC and BIC.

Model	θ	λ	ν	$\mathscr{L}$	AIC	BIC
LBR	217.216	–	–	-884.464	1.771×10^3	1.774×10^3
LBW	–	0.411	–	-594.469	1.191×10^3	1.194×10^3
PHR	–	8.275×10^{-5}	2.234	-436.482	876.963	882.759
LBPHR	–	1.216×10^{-5}	2.929	-420.866	845.731	851.527

Table 4 indicates that the LBPHR is best than LBR, LBW and PHR distributions in terms of model fitting for this data.

The variance and covariance matrix is given as

$$I^{-1} = \begin{bmatrix} 1.091\times10^{-10} & -2.797\times10^{-6} \\ -2.797\times10^{-6} & 0.072 \end{bmatrix}$$

Then the 95% C.I. for λ and ν are $(0, 3.26336\times10^{-5})$ and $(2.40217, 3.45613)$, respectively.

Figure 4 shows that the likelihood function has unique solution.

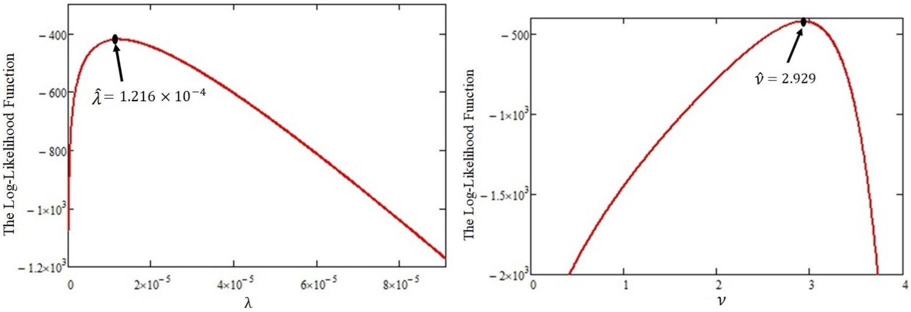


Figure 4. The sketch of the log-likelihood function of λ and ν.

For $\hat{\lambda} = 1.216\times10^{-5}$ and $\hat{\nu} = 2.929$, some statistical measures can be calculated, see Table 5.

Table 5. Some statistical measures, for LBPHR at $\hat{\lambda}$ and $\hat{\nu}$.

Mean	Mode	Variance	Skewness	Kurtosis
24.716	25.244	38.138	-0.097	2.813

From Table 5,

1. the distribution is left skewed ($S_k < 0$) and it is approximately symmetric ($-0.5 < S_k < 0.5$).

2. the distribution is platykurtic ($K_u < 3$).

Example 4.3. The following uncensored data is taken from Mahmoud and Mandouh (2013), which comprises 100 observations(breaking the stress of carbon fibers in Gba) are:

0.92	0.928	0.997	0.9971	1.061	1.117	1.162	1.183	1.187	1.192	1.196
1.213	1.215	1.2199	1.22	1.224	1.225	1.228	1.237	1.24	1.244	1.259
1.261	1.263	1.276	1.31	1.321	1.329	1.331	1.337	1.351	1.359	1.388
1.408	1.449	1.4497	1.45	1.459	1.471	1.475	1.477	1.48	1.489	1.501
1.507	1.515	1.53	1.5304	1.533	1.544	1.5443	1.552	1.556	1.562	1.566
1.585	1.586	1.599	1.602	1.614	1.616	1.617	1.628	1.684	1.711	1.718
1.733	1.738	1.743	1.759	1.777	1.794	1.799	1.806	1.814	1.816	1.828
1.830	1.884	1.892	1.944	1.972	1.984	1.987	2.020	2.0304	2.029	2.035
2.037	2.043	2.046	2.059	2.111	2.165	2.686	2.778	2.972	3.504	3.863
5.306										

The MLES, $\mathscr{L}$, AIC and BIC are given in Table 6.

Table 6. MLEs of the parameters and $\mathscr{L}$, AIC and BIC.

Model	θ	λ	ν	$\mathscr{L}$	AIC	BIC
LBR	1.035	–	–	-131.653	265.306	267.911
LBW	–	1.406	–	-101.918	205.835	208.44
PHR	–	0.521	1.632	-90.149	184.298	189.509
LBPHR	–	0.877	1.237	-84.566	173.132	178.342

Table 6 indicates that the LBPHR is best than LBR, LBW and PHR distributions in terms of model fitting for this data.

The variance and covariance matrix is given as

$$I^{-1} = \begin{bmatrix} 9.448 \times 10^{-3} & -0.011 \\ -0.011 & 0.027 \end{bmatrix}$$

Then the 95% C.I. for λ and ν are $(0.68651, 1.06753)$ and $(0.91431, 1.55924)$, respectively.

Figure 5 shows that the likelihood function has unique solution.

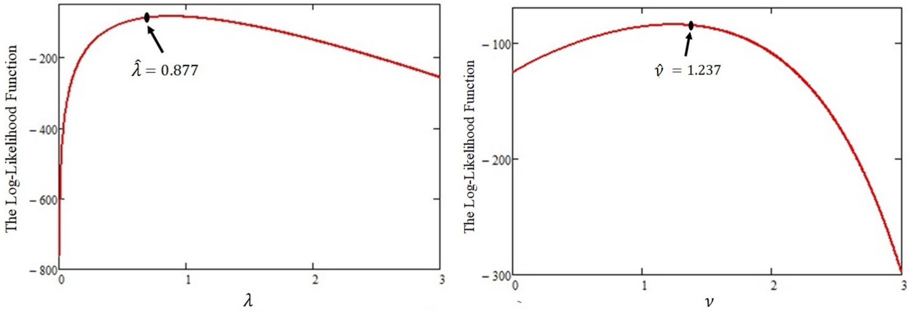


Figure 5. The shape of the log-likelihood function of λ and ν.

For $\hat{\lambda} = 0.877$ and $\hat{\nu} = 1.237$, some statistical measures can be calculated, see Table 7.

Table 7. Some statistical measures, for LBPHRD at $\hat{\lambda}$ and $\hat{\nu}$.

Mean	Mode	Variance	Skewness	Kurtosis
1.647	1.52	0.408	0.374	2.962

From Table 7, we observe that

1. the distribution is right skewed ($S_k > 0$) and it approximately symmetric skewed ($-0.5 < S_k < 0.5$).

2. the distribution is platykurtic. ($K_u < 3$).

4.2. Simulation study

We evaluate the performance of MLE of the model through Monte-Carlo simulation. The simulation's steps are as follows.

1. Fix the vector of parameters $\Theta = (\lambda, \nu)$, and sample of size n.

2. From LBPHR(λ, ν) distribution generate random observation with size n. Since CDF for LBPHR has no closed form, the random observation can be generated by using the Newton's Raphson method.

$$x_{i+1} = x_i - \frac{F(x_i, \Theta) - u_i}{f(x_i, \Theta)}, \quad i = 0, 1, \cdots, n-1, \tag{27}$$

where, $u \sim uniform(0,1)$.

3. Using step 2, estimate $\hat{\Theta}$ through MLE scheme.

4. Steps 2 and 3, repeated N times.

5. To enumerate MREs (mean relative estimates) and MSEs (mean square errors) using $\hat{\Theta}$ and Θ through the following equations.

$$MRE = \frac{1}{N}\sum_{j=1}^{N}\frac{\hat{\Theta}_{i,j}}{\Theta_i}, \qquad MSE = \frac{1}{N}\sum_{j=1}^{N}\left(\hat{\Theta}_{i,j}-\Theta_i\right)^2,$$

$$Bias = \frac{1}{N}\sum_{j=1}^{N}\hat{\Theta}_{ij}-\Theta_i, \quad i=1,2.$$

Simulation results are obtained via MATHCAD 2007. The selected parameter values are $\Theta=(0.5,2)$, $N=10000$ and $n=(10,20,30,40,50,75,100,150,200,250,300,400,500)$.

Table 8 contains the MLEs, Bias, MREs, and MSEs, for the estimators $\hat{\Theta}_i, i=1,2$, for different values of n.

Table 8. The MLEs, MREs and MSEs, for different values of n.

	λ				ν			
n	MLE	Bias	MRE	MSE	MLE	Bias	MRE	MSE
10	0.22357	-0.27643	0.44714	0.08764	3.56421	1.56421	1.78211	3.32282
20	0.49478	-0.00522	0.98956	0.00040	2.30885	0.30885	1.15443	0.15088
30	0.64100	0.14100	1.28200	0.02070	1.85722	-0.14278	0.92861	0.02979
40	0.49451	-0.00549	0.98902	0.00021	2.33661	0.33661	1.16830	0.17403
50	0.56267	0.06267	1.12535	0.00423	2.37851	0.37851	1.18926	0.14668
75	0.50153	0.00153	1.00306	0.00018	2.12308	0.12308	1.06154	0.01756
100	0.35535	-0.14465	0.71069	0.02103	2.42129	0.42129	1.21064	0.18009
150	0.39449	-0.10551	0.78898	0.01120	2.53190	0.53190	1.26595	0.30598
200	0.55981	0.05981	1.11963	0.00361	2.06537	0.06537	1.03269	0.00489
250	0.50780	0.00780	1.01559	0.00008	2.07485	0.07485	1.03743	0.00591
300	0.42144	-0.07856	0.84289	0.00618	2.16489	0.16489	1.08244	0.02738
400	0.54172	0.04172	1.08344	0.00175	1.90172	-0.09828	0.95086	0.00976
500	0.54186	0.04186	1.08372	0.00176	1.89278	-0.10722	0.94639	0.01154
Average	0.48004	-0.01996	0.96008	0.01223	2.27856	0.27856	1.13928	0.33749

MREs approximate to one when MSEs approaches to zero. Figures 6 – 9 display the estimated MLs,Bias, MREs and MSEs.

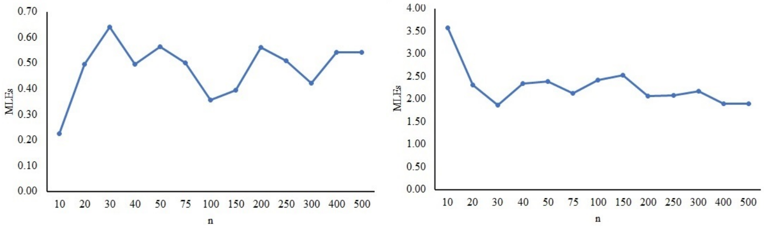


Figure 6. The MLEs for λ and ν.

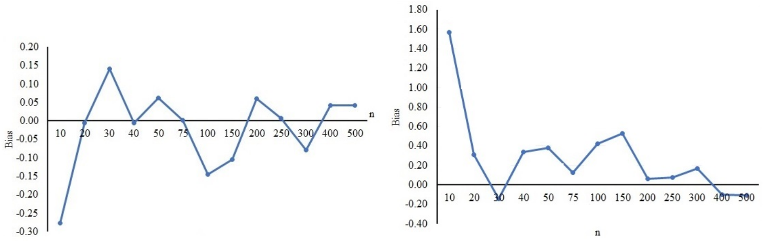


Figure 7. The Bias for λ and ν.

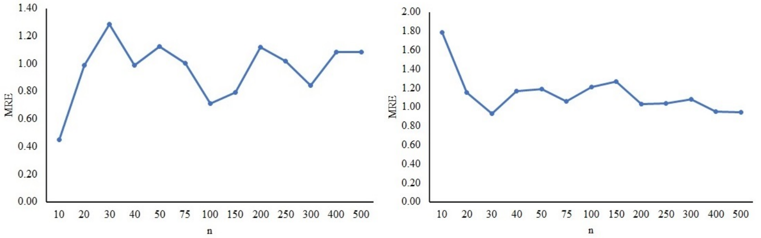


Figure 8. The MREs for λ and ν.

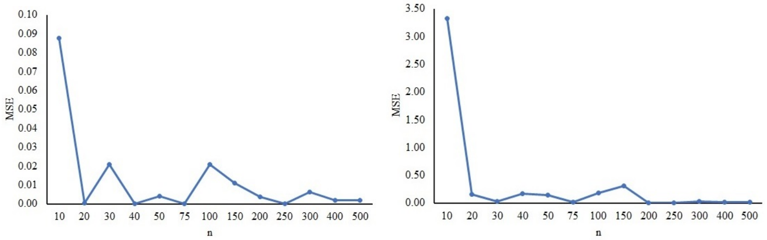


Figure 9. The MSEs for λ and ν,

We notice from Figures 6–9 as follows.

1. For large same size: (i) Estimate of MSE → 0, (ii) Expected (MRS) → 1, (iii) Biases of $(\lambda, \nu) \rightarrow 0$.
2. Biases of λ are positive/negative.
3. Biases of ν are approximately positive.
4. Estimates of parameters are asymptotically unbiased.

Therefore, the MLE is an suitable for estimating parameters of LBPHR distribution. Similar results can be obtained for different parameters.

5. Conclusions

We propose the length-biased power hazard rate distribution and study its various characteristics. The maximum likelihood estimate for parameters is derived. The superiority of the new model has been exhibited by some real data sets. It has been seen that PHRD can adequately provide better fits over other models.

References

Ajami, M., Jahanshahi S. M. A., (2017). Parameter estimation in weighted Rayleigh distribution. Journal of Modern Applied Statistical Methods, 16(2), pp. 256–276.

Al-Khadim, A. K., Hussein A. N., (2014). New proposed length-biased weighted Exponential and Rayleigh distribution with application. *Mathematical Theory and Modeling*, 4, pp. 2224–5804.

Balakrishnan, N., Victor, L., Antonio, S., (2010). *A mixture model based on Birnhaum-Saunders Distributions, A study conducted by Authors regarding the Scores of the GRASP (General Rating of Affective Symptoms for Preschoolers), in a city located at South Part of the Chile*.

Birnbaum, Z. W., Saunders, S. C., (1969). Estimation for a family of life distribution with applications to fatigue. *Journal of Applied Probability*, 6(2), pp. 328–347.

Cox, D. R., (1962). *Renewal Theory*. New York, NY: Barnes & Noble.

Das, K. K., Roy, T. D., (2011). Applicability of length-biased generalized Rayleigh distribution. *Advances in Applied Science Research*, 2, pp. 320–327.

Das, K. K., Roy, T. D., (2011). On some length-biased weighted Weibull distribution. *Advances in Applied Science Research*, 2, pp. 465–475.

Gupta, P. L., Tripathi, R. C., (1990). Effect of length-biased sampling on the modeling error. *Communications Statistics - Theory and Methods*, 19, pp. 1483–1491.

Gupta, R. C., Keating, J. P., (1985). Relations for reliability measures under length-biased sampling. *Scandanavian Journal of Statistics*, 13, pp. 49–56.

Ismail, K., (2014). Estimation of for distribution having power hazard function. *Pakistan Journal of Statistics*, 30, pp. 57–70.

Lawless, J. F., (2003). Statistical Models and Methods for Lifetime Data, 2nd Edition, Weiley, Canada.

Mir, K. A., Ahmed, A., Reshi, J. A., (2013). On Size-biased Exponential Distribution. *Journal of Modern Mathematics and Statistics*, 7(2), pp. 21–25.

Khattree, R., (1989). Characterization of inverse-Gaussian and gamma distributions through their length-biased distributions. *IEEE Transactions on Reliability*, 38, pp. 610–611.

Mahmoud, M. R., Mandouh, R. M., (2013). On the Transmuted Fréchet Distribution. *Journal of Applied Sciences Research*, 9(10), pp. 5553–5561.

Modi, K., (2015). Length-biased Weighted Maxwell distribution. *Pakistan Journal of Statistics and Operation Research*, 11(4), pp. 465–472.

Mudasir, S., Ahmad, S. P., (2018). Characterization and estimation of the length-biased Nakagami distribution. *Pakistan Journal of Statistics and Operation Research*, 14(3), pp. 697–715.

Mugdadi, A. R., (2005). The least squares type estimation of the parameters in the power hazard function. *Applied Mathematics Computation*, 169, pp. 737–748.

Mugdadi, A. R., Min, A., (2009). Bayes estimation of the power hazard function. *Journal of Interdisciplinary Mathematics*, 12, pp. 675–689.

Nanuwong, N., Bodhisuwan, W., (2014). Length-biased nu Pareto distribution and its structural properties with application. *Journal of Mathematics and Statistics*, 10, pp. 49–57.

Oluyede, B. O., (1999). On inequalities and selection of experiments for length-biased distributions. *Probability in the Engineering and Informational Sciences*, 13, pp. 169–185.

Praveen, Z., Ahmad, M., (2018). Some properties of size- biased weighted Weibull distribution. *International Journal of Advanced and Applied Sciences*, 5(5), pp. 92–98.

Ratnaparkhi, M. V., Naik-Nimbalkar, U. V., (2012). The length-biased lognormal distribution and its application in the analysis of data from oil field exploration studies. *Journal of Modern Applied Statistical Methods*, 11, pp. 225–260.

Saghir, A., Khadim, A., Lin, Z., (2017). The Maxwell length-biased distribution: Properties and Estimation. *Journal of Statistical Theory and Practice*, 11, pp. 26–40.

Saghir, A., Tazeem, S., Ahmad, I., (2016). The length-biased weighted exponentiated inverted Weibull distribution. *Cogent Mathematics*, 3(1), DOI: 10.1080/23311835.2016. 1267299.

Seenoi, P., Supapa, K.T., Bodhisuwan, W., (2014). The length-biased exponentiated inverted Weibull Distribution. *International Journal of Pure and Applied Mathematics*, 92, pp. 191–206.

Shaban, S.A., Boudrissa, N. A., (2007). The Weibull length-biased distribution: Properties and estimation. *Interstat*,
http://interstat.statjournals.net/YEAR/2007/articles/0701002.pdf

Tarvirdizade, B., Nematollahi, N., (2016). Parameter estimation based on record data from power hazard rate distribution. *13th Iranian Statistical Conference*, Kerman, Iran.

Tarvirdizade, B., Nematollahi, N., (2020). Inference on $P(X > Y)$ based on record values from power hazard rate distribution. *Journal of Computational Statistics and Modelling*, 1(1), pp. 59–76.

STATISTICS IN TRANSITION new series, June 2022
Vol. 23 No. 2, pp. 17–32, DOI 10.2478/stattrans-2022-0014
Received – 12.12.2018; accepted – 25.06.2021

Jackknife winsorized variance estimator under imputed data

Fariha Sohil[1], Muhammad Umair Sohail[2], Javid Shabbir[3], Sat Gupta[4]

ABSTRACT

In the present study, we consider the problem of missing and extreme values for the estimation of population variance. The presence of extreme values either in the study variable, or the auxiliary variable, or in both of them, can adversely affect the performance of the estimation procedure. We consider three different situations for the presence of extreme values and also consider jackknife variance estimators for the population variance by handling these extreme values under stratified random sampling. Bootstrap technique ABB is carried out to understand the relative relationship more precisely.

Key words: adjusted imputation, jackknife variance estimators, linearized jackknife, missing values, winsorized variance

2000 AMS Classification: 62D05

1. Introduction

In most social science studies, researchers often face the problem of non-response due to sensitive or embarrassing issues. For example, in the case of student grade point surveys, the students may be reluctant to provide the information about grade point average. Basically, non-response is classified into two basic complete categories: (1) Unit non-response, which occurs when either the interviewee refuses to provide the response regarding the variable of interest or the interviewee is not available. (2) Item none-response, which occurs mainly due to the sensitive or embarrassing nature of the study variable (Y). Muhamad (2016) studied the imputation of missing responses by using the higher order moments of the auxiliary variable.

[1] Department of Education, The Women University, Multan, Pakistan. E-mail: s.fariha@gmail.com.

[2] Department of Statistics, University of Narowal, Narowal, Pakistan. E-mail: umair.sohail@uon.edu.pk. ORCID: https://orcid.org/0000-0002-5440-126X.

[3] Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan. E-mail: js@qau.edu.pk. ORCID: https://orcid.org/0000-0002-0035-7072.

[4] Department of Mathematics and Statistics, University of North Carolina, Greensboro, USA.

The main goal of our current work is to consider the problem of missing at random (MAR) values in the estimation of finite population variance. When the item non-response occurs the missing values of the non-respondent class can be imputed by utilizing the available information from the respondent class. Many methods use the auxiliary information for imputing the missing value.

Rubin (1976) gave a comprehensive concept of missing values by defining terms such as missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR) values. Rubin (1978) considered the problem of inflation in estimated variance by discussing the idea of multiple imputation (MI). The suggested procedure obtains $\lambda\ (\geq 2)$ data sets by imputing the missing values under the same imputation procedure of λ times. To define the Multiple Imputation (MI) methodology by $\bar{y}_{I1}, \bar{y}_{I2}, \cdots, \bar{y}_{I\lambda}$, the λ imputed estimators for the population mean. The final imputed estimator for population mean is given by $\bar{y}_{I.} = \frac{1}{\lambda}\sum_{l=1}^{\lambda}\bar{y}_{Il}$, with estimated variance

$$v\left(\bar{y}_{I.}\right) = \frac{1}{\lambda}\sum_{l=1}^{\lambda}\left(\frac{1}{n}-\frac{1}{N}\right)s_{Il}^{2} + \frac{\lambda+1}{\lambda}\left\{\frac{1}{\lambda-1}\sum_{l=1}^{\lambda}\left(\bar{y}_{Il}-\bar{y}_{I.}\right)^{2}\right\}, \tag{1.1}$$

where s_{Il}^{2} is the sample variance for the $l-th$ imputed data set having n sample and N population size respectively. The variance estimator leads to valid inference about the parameter of interest, when the number of imputation is large, provided the imputation is proper in the sense that imputed values for the non-respondent group are obtained from the posterior distribution of the study variable (Rubin and Schenker, 1986; Mujtaba et al. 2014). The traditional imputation methods like hot deck (HD) may give the underestimate variance of $\bar{y}_{I.}$. Rubin and Schenker (1986) provided the Approximate Bayesian Bootstrapping (ABB) approach for proper variance estimation. For $l = 1, 2, 3, \cdots, \lambda$; we draw r values randomly with replacement from the r observed values and then obtain $(n-r)$ missing values from the r bootstrap donors. The resultant estimates based on the g reference distribution performed well in terms of large sample selection probability, even for λ as small as 2 or 3.

MI is a proper tool to handle the missing data but some of the major limitations are: (1) Cost for handling the multiple data sets is high as compared to single imputation, especially in complex surveys. (2) The general ABB approach for imputing the non-response, that has some issues regarding the clustering, stratification, unequal probabilities of selection, is not currently taken into account. (3) Sometimes the imputation is deterministic, missing values are obtained by the sample of donor set and the auxiliary data. (4) For smaller values of λ, we may attain a low level of precision for the multiple imputation variance estimator (MIVE), because the last term in (1.1) approaches to zero for a small value of λ.

The main focus of this investigation is to consider the univariate statistics such as mean and total under imputation and provide some recent work on jackknife variance estimation to adjust the imputed values in the presence of extreme observations. We consider the problem of extreme values in the study variable, the auxiliary variable, or in both of them, before imputing the missing values. We consider the stratified random sampling design with commonly used imputation methods such as traditional ratio, classical linear regression method and hot deck within imputation. These imputation strategies are not proper in the sense of (Rubin and Schenker, 1986), but all of them would have the valid design based on inference about the suggested variance estimator. Recently, Chen et al. (2107) suggested an approaches to improving survey-weighted estimates by precisely weighting the survey estimates.
The aim of the study is to consider the problem of extreme values either at the upper or lower end for the precise imputation of missing responses. In present study, we proposed a jacknified Winsorized variance estimator under imputed data by discussing three different cases for the occurrence of extreme values in the field of survey sampling. These are given below:

Case I: Extreme values in study variable

Let the extreme values occur only in the study variable (Y) but not in the auxiliary variable (X). These extreme values should lead to the low correlation with the auxiliary variable which will affect the performance of the estimation procedure.

Case II: Extreme values in the auxiliary variable

Let $X_1, X_2, \cdots, X_N$ be the values of the auxiliary variable having a population mean $(\bar{X})$. Suppose the characteristics of the auxiliary information are not available but we have some relevant information. We want to utilize the auxiliary information in a significant manner for better inference. The one of the possible ways to handle this situation is to use the idea of winsorization for the valid inference about the population parameters.

Case III: Extreme values in both variables

Suppose that, extreme values occur both in the study and the auxiliary variable due to some natural or unnatural disturbance in the experiment. This irregular behaviour of the study and the auxiliary variables may lead to underestimation or overestimation of population parameters. Under such circumstances, we need to use some standard procedures for the valid inference. So, we truncate both variables by using some specified standard procedures and the results by using a truncated set of values are more reliable as compared to the irregular observed values.

We follow the standard truncation process, where low extreme values are truncated at the first quartile and high extreme values are truncated at the third quartile.

2. Proposed procedure

Motivated by Rao (1996), we consider the problem of extreme values in both the study and the auxiliary variables under stratified random sampling for the estimation of winsorized variance.

2.1. Complete response case

Most of the daily life surveys based on well established frames are often used in stratified sampling. Let n_h be the number of sampled units selected from the $h-th$ stratum $(h=1,\ 2,\ 3,\ \cdots\ ,L)$ such that $\sum_{h=1}^{L} n_h = n.$ For the complete response case, the usual unbiased estimator of $\bar{Y}$ is given by $\bar{\Upsilon} = \sum_{h=1}^{L} W_h \bar{\Upsilon}_h$, where W_h is the stratum weight and $\bar{\Upsilon}_h$ is the sample mean of the $h-th$ stratum after truncation. The variance of the winsorized set of values is

$$v\left(\bar{\Upsilon}\right) = \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{n_h N_h}\right) s^2_{\Upsilon h}, \tag{2.1}$$

where $s^2_{\Upsilon h} = \frac{1}{n_h - 1}\sum_{j=1}^{n_h}\left(\Upsilon_{j_h} - \bar{\Upsilon}_h\right)^2$ and Υ_{j_h} is the truncated response of the $j-th$ respondent in the $h-th$ stratum.

The jackknife variance estimator of $\bar{\Upsilon}$ after deleting the extreme values, is given by

$$v_J\left(\bar{\Upsilon}\right) = \sum_{h=1}^{L} (n_h - 1)\left(\frac{N_h - n_h}{n_h N_h}\right) s^{2J}_{\Upsilon h}, \tag{2.2}$$

where $s^{2J}_{\Upsilon h} = \frac{1}{n_h}\sum_{j=1}^{n_h}\left\{\bar{\Upsilon}(hj) - \bar{\Upsilon}\right\}^2$ and $\bar{\Upsilon}(hj)$ is sample mean obtained after deleting the $j-th$ response from the $h-th$ stratum.

2.2. Adjusted imputed value

In the case of missing values in Υ, suppose s_h be the sample of size n_h is selected from the $h-th$ stratum having Ω_h sampled units, let r_h be the respondents and r'_h be the non-respondents units who refuse to provide the response regarding the variable of interest. So, $s_h = s_{r_h} \cup s_{r'_h}$. Let $\overline{\Upsilon}_{r_h}$ be the winsorized sample mean of s_{r_h} in $h-th$ stratum. Suppose $\Upsilon^{\#}$ is the imputed value for the $j-th$ unit in $s_{r'_h}$.

The estimator for the population mean is then given by

$$\overline{\Upsilon}_I = \sum_{h=1}^{L} \frac{W_h}{n_h} \left\{ \sum_{j_h \in s_{r_h}} \Upsilon_{j_h} + \sum_{j_h \in s_{r'_h}} \Upsilon^{\#}_{j_h} \right\}, \tag{2.3}$$

With deterministic approach, the jackknife variance estimator of $\overline{\Upsilon}_I$ is obtained in the usual way by deleting the respondents s_{r_h}, each of the imputed value in the $h-th$ stratum is adjusted in magnitude as $\left\{ \Upsilon^{\#}_{j_h}(hJ) - \Upsilon^{\#}_{j_h} \right\}$, where $\Upsilon^{\#}_{j_h}(hJ)$ is the imputed value for the $j-th$ non-respondent unit in $h-th$ stratum, when hJ respondent is deleted from s_h. Then, the adjusted imputed missing value is equal the "*correct*" value $\Upsilon^{\#}_{j_h}(hJ)$ if $hJ \in s_{r_h}$ and remaining values are unchanged, if the non-respondent hJ is deleted. In the case of stochastic imputation, each of the imputed value is adjusted by $E^{\#}_{hJ}\Upsilon^{\#}_{j_h} - E^{\#}\Upsilon^{\#}_{j_h}$, when $hJ \in s_{r_h}$, where $E^{\#}$ denotes the expectation with respect to the imputation procedure given the donor class and $E^{\#}_{hJ}$ is the expectation when the donor values are adjusted by removing hJ units. Note that the adjusted imputation values reflect that the donor set is changed, when the respondent set is deleted from the sample.

The imputed estimator based on the original and imputed values of the $j-th$ sample units in the $h-th$ stratum is expressed as $\overline{\Upsilon}^{@}_I(hJ)$, after deleting the hJ units. Then, the jackknife winsorized variance estimator, after ignoring the finite population correction factor, is given by (2.4)

$$v_J\left(\overline{\Upsilon}_I\right) = \sum_{h=1}^{L} (n_h - 1) s^{2@}_{\Upsilon h}, \tag{2.4}$$

where

$$s^{2@}_{\Upsilon h} = \frac{1}{n_h} \sum_{j=1}^{n_h} \left\{ \overline{\Upsilon}^{@}_I(hj) - \overline{\Upsilon}_I \right\}^2 .$$

2.3. Ratio imputation

Suppose the auxiliary information is available on all the sampled units in s_h . The traditional ratio imputation procedure with winsorized data is defined as:

$$\Upsilon^{\#}_{j_h} = \left(\frac{\overline{\Upsilon}_{r_h}}{\overline{\eta}_{r_h}} \right) \eta_{j_h}, \tag{2.5}$$

for $j_h \in s_{r'_h}$. Where $\overline{\Upsilon}_{r_h}$ and $\overline{\eta}_{r_h}$ are the sample mean from the s_{r_h} respondent class in the $h-th$ stratum respectively. This imputation procedure is motivated by the fact that $\Upsilon^{\#}_{j_h}$ is the best predictor of the units which are in $s_{r'_h}$ group, under the following ratio super population model, which is given by:

$$E\left(\Upsilon_{j_h}\right) = b_h \eta_{j_h}, \quad V\left(\Upsilon_{j_h}\right) = \sigma_h^2 \eta_{j_h} \quad \text{and} \quad \mathrm{Cov}\left(\Upsilon_{j_h} \Upsilon_{k_h}\right) = 0, \tag{2.6}$$

This model also holds for s_{r_h} , if there is no selection bias. The probability of response depends upon the η_{j_h}·Sarndal (1992) has shown (2.6) as an imputation model.

Under such an approach, (2.3) will be written as:

$$\overline{\Upsilon}_{Ih} = \left(\frac{\overline{\Upsilon}_{r_h}}{\overline{\eta}_{r_h}} \right) \overline{\eta}_h \tag{2.7}$$

and hence

$$\overline{\Upsilon}_I = \sum_{h=1}^{L} W_h \left(\frac{\overline{\Upsilon}_{r_h}}{\overline{\eta}_{r_h}} \right) \overline{\eta}_h, \tag{2.8}$$

Under (2.6), the estimator $\overline{\Upsilon}_I$ is a design based model, $E\left(\overline{\Upsilon}_I\right) = \overline{Y}$, provide that the model also holds for the respondent units. For the uniform response from all strata, the (2.8) has the same properties as the two-phase separate ratio estimators.

It is readily seen that, $\Upsilon^{\#}_{j_h}(hJ) = \dfrac{\overline{\Upsilon}_{r_h}(hJ)}{\overline{\eta}_{r_h}(hJ)} \eta_{j_h}$, with the ratio imputation method hJ respondent units are deleted, where $\overline{\Upsilon}_{r_h}(hJ) = \dfrac{r\overline{\Upsilon}_{r_h} - \Upsilon_{j_h}}{r_h - 1}$ and similar for η as $\overline{\eta}_{r_h}(hJ) = \dfrac{r\overline{\eta}_{r_h} - \eta_{j_h}}{r_h - 1}$. Using the values, the jackknife variance estimator for $\overline{\Upsilon}_I$ is obtained from (2.4). The linearized version of the jackknife variance estimator is obtained under the model (2.6).

The linearized version of the jackknife variance estimator is helpful in estimating the variance through a computer program. This suggested method is helpful in obtaining the valid jackknife estimator under the uniform response from all strata. Let $\overline{\Upsilon}_I^{@}\left(hJ\right)-\overline{\Upsilon}_I=W_h\left\{\overline{\Upsilon}_{I_h}^{@}(hJ)-\overline{\Upsilon}_{I_h}\right\}$ be the adjusted imputed estimator for $\overline{Y}_h$. If hJ units are deleted from h^{th} stratum, we have

$$v_j\left(\overline{\Upsilon}_I\right)=\sum_{h=1}^{L}W_h^2 v_j\left(\overline{\Upsilon}_{Ih}\right) \tag{2.9}$$

where
$$v_j\left(\overline{\Upsilon}_{Ih}\right)=\sum_{j=1}^{n_h}\frac{n_h-1}{n_h}\left\{\overline{\Upsilon}_{Ih}^{@}(hJ)-\overline{\Upsilon}_{Ih}\right\}^2 .$$

A linearized version of $v_J(\overline{\Upsilon}_I)$ is given by $v_{L.rat.}\left(\overline{\Upsilon}_I\right)=\sum_h W_h^2 v_J\left(\overline{\Upsilon}_{Ih}\right)$ with

$$v_J\left(\overline{\Upsilon}_{Ih}\right)=\left(\frac{\overline{\eta}_h}{\overline{\eta}_{r_h}}\right)^2\frac{\vartheta_h}{r_h}+2\left(\frac{\overline{\eta}_h}{\overline{\eta}_{r_h}}\right)\frac{\delta_h}{n_h}+\frac{\Delta_h}{n_h}, \tag{2.10}$$

where $\vartheta_h=\sum_{j_h\in s_{r_h}}\frac{\zeta_{j_h}^2}{(r_h-1)}$, $\delta_h=\left(\frac{\overline{\Upsilon}_{r_h}}{\overline{\eta}_{r_h}}\right)\sum_{j_h\in s_{r_h}}\frac{\zeta_{j_h}\eta_{j_h}}{(r_h-1)}$ and

$$\Delta_h=\left(\frac{\overline{\Upsilon}_{r_h}}{\overline{\eta}_{r_h}}\right)^2\sum_{j_h\in s_{r_h}}\frac{(\eta_{j_h}-\overline{\eta}_h)}{(n_h-1)} \text{ with } \varsigma_{j_h}=\Upsilon_{j_h}-\frac{\overline{\Upsilon}_{r_h}}{\overline{\eta}_{r_h}}\eta_{j_h} .$$

Formula in (2.10) is obtained by using the following expression

$$\overline{\Upsilon}_{Ih}^{@}-\overline{\Upsilon}_{Ih}=-\frac{\overline{\Upsilon}_{r_h}(\eta_{j_h}-\overline{\eta}_h)}{\overline{\eta}_{r_h}(n_h-1)}-\left\{\frac{\overline{\eta}_h(hJ)}{\overline{\eta}_{r_h}(hJ)}\right\}\frac{\zeta_{j_h}}{(r_h-1)}$$

if $hJ\in s_{r_h}$ and equals $-\frac{\overline{\Upsilon}_{r_h}(\eta_{j_h}-\overline{\eta}_h)}{\overline{\eta}_{r_h}(n_h-1)}$ if $hJ\in s_{r_h'}$, and noting that $\frac{\overline{\eta}_h(hJ)}{\overline{\eta}_{r_h}(hJ)}=\frac{\overline{\eta}_h}{\overline{\eta}_{r_h}}$ for large r_h, where $\overline{\eta}_h\left(hJ\right)=\frac{n_h\overline{\eta}-\eta_{j_h}}{n_h-1}$.

Rao and Sitter (1992) have shown that (2.10) is a design consistent variance estimator under two phase sampling design. It shows that the jackknife variance estimator in (2.4) and linearized version of (2.4) are effective under the uniform response within each stratum.

Moreover Sarndal (1992) provided the following approximation under the model (2.4)

$$v_s(\bar{\Upsilon}_{Ih})=\left(\frac{\bar{\eta}_h}{\eta_{r_h}}\right)^2\frac{\vartheta_h}{r_h}+2\left(\frac{r_h}{n_h}\right)\frac{\delta_h}{n_h}+\frac{\Delta_h}{n_h}. \tag{2.11}$$

After the relative comparison of (2.10) and (2.11), it is noted that $E\left(\delta_h\right)=0$ for the large value of r_h, which could lead the estimators in (2.4) and (2.10) to be unbiased under the model of (2.6). Moreover, it is observed that

$$v_s\left(\bar{\Upsilon}_I\right)=\sum_h W_h^2 v_s\left(\bar{\Upsilon}_{I_h}\right) \tag{2.12}$$

is not the consistent estimator under the uniform response within each stratum. The adjustment of the finite population correction (fpc) is shown by Rao (1996) comprehensively. There is no simple relation to adjust the finite correction in (2.4), but Rao and Sitter (1995) use some internal relationships to recover the finite population correction (fpc). The modified estimator in (2.4) can be used in two-phase sampling within strata, when imputation is used; that is, when the response of non-sampled units of Υ is imputed using the first phase auxiliary information. Whitridge and Kovar (1990) considered the importance of mass imputation by utilizing it on business data. Kovar and Chen (1994) discussed the finite sample properties of (2.4) by the real life application to business survey data.

Here, we discuss the stochastic counterpart of the traditional ratio imputation. In this approach the first donor is selected from $h_0 i_0$ by using the simple random sample with replacement for s_h. Then, the ratio residual $\left(\zeta^*_{j_h}\right)$ is added to (2.5) to get the random imputation value as

$$\Upsilon^{\#}_{j_h}=\left(\frac{\bar{\Upsilon}_{r_h}}{\bar{\eta}_{r_h}}\right)\eta_{j_h}+\zeta^*_{j_h} \tag{2.13}$$

Noting that $E^{\#}\left(\zeta^*_{j_h}\right)=0$, the resultant ratio estimator is unbiased for $\bar{Y}$ under the model (2.6) and uniform response from all the strata.

With this ratio imputation procedure, we have

$$E^{\#}_{hJ}(\Upsilon^{\#}_{j_h})=\left\{\frac{\bar{\Upsilon}_{r_h}(hJ)}{\bar{\eta}_{r_h}(hJ)}\right\}\eta_{j_h},\quad \text{if } hJ \text{ respondents are deleted.} \tag{2.14}$$

and

$$E^{\#}\left(\Upsilon^{\#}_{j_h}\right)=\left(\frac{\overline{\Upsilon}_{r_h}}{\overline{\eta}_{r_h}}\right)\eta_{j_h} \tag{2.15}$$

Thus, the adjusted imputed values under the model (2.4) are, as:

$$y^{\#}_{j_h}+\left\{\frac{\overline{\Upsilon}_{r_h}(hJ)}{\overline{\eta}_{r_h}(hJ)}\right\}\eta_{j_h}-\left\{\frac{\overline{\Upsilon}_{r_h}}{\overline{\eta}_{r_h}}\right\}\eta_{j_h},$$

If hJ units are deleted and remaining units are unchanged. It is easy to express the linear version of jackknife variance estimator, is given by

$$v_L\left(\overline{\Upsilon}_I\right)=\sum_h W_h^2 v_L\left(\overline{\Upsilon}_{I_h}\right), \tag{2.16}$$

where $v_L\left(\overline{\Upsilon}_{I_h}\right)$ is simply obtained by adding a term which is obtained due the random selection from a donor set to (2.10) under the ratio estimator. The extra terms are given by

$$\left(\frac{m_h}{n_h^2}\right)\left\{2\left(\frac{\overline{\Upsilon}_{r_h}}{\overline{\eta}_{r_h}}\right)s^*_{\zeta\eta_h}+s^{*2}_{\zeta h}+\left(\frac{r_h}{n_h}\right)\overline{\zeta}^{*2}_h\right\},$$

where $s^*_{\zeta\eta_h}=\sum_{j_h\in s_h}\zeta^*_{j_h}\left(\eta_{j_h}-\overline{\eta}_h\right)\Big/n_h$, $s^{*2}_{\zeta_h}=\sum_{j_h\in s_h}\left(\zeta^*_{j_h}-\overline{\zeta}^*_h\right)\Big/n_h$ and $\overline{\zeta}^*_h=\sum_{j_h\in s_h}\zeta^*_{j_h}\Big/n_h$.

If auxiliary information regarding the variable of interest is unavailable then the traditional ratio imputation is reduced to the simple random imputation within each stratum. For these type of situations, Little and Rubin (1987) considered the approximate Bayesian Bootstrapping for handling it and discussed it in detail in their text in the chapter MI.

2.4. Regression imputation

Let η be observed on all the sampled units in s_h. The classical linear regression estimator is defined as:

$$\Upsilon^{\#}_{j_h}=\overline{\Upsilon}_{r_h}+\hat{\beta}_{r_h}\left(\eta_{j_h}-\overline{\eta}_{r_h}\right)\quad\text{for } j_h\in s_h \tag{2.17}$$

where $\hat{\beta}_{r_h}$ is a linear simple regression coefficient based on the respondent units in the h^{th} stratum. The imputed values $\Upsilon^{\#}_{j_h}$ are the best predictor for the unobserved units of Υ_{j_h} under the following super population model:

$$E\left(\Upsilon_{j_h}\right) = \alpha_h + \beta_h \eta_{j_h}, \quad V\left(\Upsilon_{jh}\right) = \sigma_h^2, \text{ and } \quad \operatorname{cov}\left(\Upsilon_{j_h} \Upsilon_{k_h}\right) = 0, \tag{2.18}$$

provided that the model holds s_{r_h} , if there is no selection bias.

Under regression imputation, (2.3) can be written as:

$$\overline{\Upsilon}_I = \sum_{h=1}^{L} W_h \left\{ \overline{\Upsilon}_{r_h} + \hat{\beta}_{r_h}\left(\overline{\eta}_h - \overline{\eta}_{r_h}\right)\right\} \tag{2.19}$$

The given estimator in (2.19) is $E(\overline{\Upsilon}_I) = \overline{Y}$ with a uniform response from strata. When the hJ item is deleted, then, the estimator $\Upsilon^{\#}_{j_h}$ is written as:

$$\Upsilon^{\#}_{j_h}(hJ) = \overline{\Upsilon}_{r_h}(hJ) + \hat{\beta}_{r_h}(hJ)\left\{\eta_{j_h} - \overline{\eta}_{r_h}(hJ)\right\} \tag{2.20}$$

where $\hat{\beta}_{r_h}(hJ)$ is the linear regression coefficient, obtained after deleting the hJ units.

Using (2.4), the linear version of $v_J(\overline{\Upsilon}_I)$ is given by $v_{L.reg.}\left(\overline{\Upsilon}_I\right) = \sum_{h=1}^{L} W_h^2 v_L\left(\overline{\Upsilon}_h\right)$ with

$$v_L(\overline{\Upsilon}_h) = \frac{1}{r_h^2}\sum_{j_h \in s_{r_h}} \zeta_{j_h}^2 \left\{1 + \frac{\left(\eta_{j_h} - \overline{\eta}_{r_h}\right)\left(\overline{\eta}_h - \overline{\eta}_{r_h}\right)}{\tilde{s}^2_{\eta r_h}}\right\}^2 + 2\hat{\beta}_{r_h}\frac{1}{n_h r_h}\sum_{j_h \in s_{r_h}} \zeta_{j_h}$$

$$\left(\eta_{j_h} - \overline{\eta}_h\right)\left\{1 + \frac{\left(\eta_{j_h} - \overline{\eta}_{r_h}\right)\left(\overline{\eta}_h - \overline{\eta}_{r_h}\right)}{\tilde{s}^2_{\eta r_h}}\right\} + \frac{\hat{\beta}^2_{r_h}}{n_h}\tilde{s}^2_{\eta_h}, \tag{2.21}$$

where $\zeta_{j_h} = \left(\Upsilon_{j_h} - \overline{\Upsilon}_{r_h}\right) - \hat{\beta}_{r_h}\left(\eta_{j_h} - \overline{\eta}_{r_h}\right)$, $\tilde{s}^2_{\eta r_h} = \frac{1}{r_h}\sum_{j_h \in s_{r_h}}\left(\eta_{j_h} - \overline{\eta}_{r_h}\right)^2$ and $\tilde{s}^2_{\eta_h} = \frac{1}{n_h}\sum_{j_h \in s_{n_h}}\left(\eta_{j_h} - \overline{\eta}_h\right)^2$. Following Rao and Shao (1992), we can say that both the jackknife and its linear version for variance are approximately unbiased under the model (2.18). Sitter (1997) extended the results under multiple linear regression imputation.

Now, we have to consider the stochastic counterpart of the regression imputation for the winsorized variance estimator. Under this approach, the first donor set $h_0 j_0$ is selected by simple random sample with replacement from the s_{r_h}, independently from each stratum. Then, the regression residual $\zeta^*_{j_h} = \zeta_{h_0 j_0}$ are added to $\hat{\Upsilon}_{j_h} = \overline{\Upsilon}_{r_h} + \hat{\beta}_{r_h r_h}\left(\eta_{j_h} - \overline{\eta}_{r_h}\right)$ to get random imputed missing value $\Upsilon^{\#} = \hat{\Upsilon}_{j_h} + \zeta^*_{j_h}$, $hJ \in s_{n_h}$. Noting that $E^{\#}\left(\zeta^*_{j_h}\right) = 0$, the resultant imputed estimator would be $\overline{\Upsilon}_I$ is unbiased for $\overline{Y}$ under model (2.18), as well as it is also assumed that the probability of response is the same in all strata. So, we have

$$E^{\#}_{hJ}\Upsilon^{\#}_{j_h} = \hat{\Upsilon}_{jh}(hJ) = \overline{\Upsilon}_{r_h}(hJ) + \hat{\beta}_{r_h}(hJ)\left\{\eta_{j_h} - \overline{\eta}_{r_h}(hJ)\right\} \tag{2.22}$$

and $E^{\#}\Upsilon^{\#}_{j_h} = \hat{\Upsilon}_{j_h}$. Thus the adjusted imputed values used (2.4) for variance estimator by $\Upsilon^{\#}_{j_h} + \hat{\Upsilon}_{j_h}(hJ) - \hat{\Upsilon}_{j_h}$. If the hJ respondent units are deleted, $\hat{\Upsilon}_{j_h}(hJ)$ is given by (2.22).

A linear version of (2.4) under stochastic regression imputation is defined as:

$$v_L\left(\overline{\Upsilon}_I\right) = \sum_h W_h^2 v_L\left(\overline{\Upsilon}_{Ih}\right), \tag{2.23}$$

where $v_L(\overline{\Upsilon}_{Ih})$ is obtained by adding a term due to hot-deck imputation from the given formula in (2.21) under linear regression imputation. The extra term is given by $\frac{m_h}{n_h^2}\left\{2\hat{\beta}_{r_h} s^*_{\zeta\eta_h} + s^{*2}_{\zeta} + \frac{r_h}{n_h}\overline{\zeta}^{*2}_h\right\}$, where $s^*_{\zeta\eta_h}, s^{*2}_{\zeta_h}$ and $\overline{\zeta}^{*2}_h$ are the regression residuals.

If $V\left(\Upsilon_{j_h}\right)$ is not the same in each stratum, say $V\left(\Upsilon_{j_h}\right) = \sigma_h^2 \eta_{j_h}$ as in the ratio model (2.6), then the weighted linear regression is appropriate as compared to others. The resultant imputation estimator is unbiased for $\overline{Y}$ but it is not consistent under the uniform response within strata.

3. Numerical study

In addition to our study, here, we discuss numerical results by using bootstrap technique ABB on a real life data set. We obtained the data set from Rudolf et al. (2006), and modified the data using ABB technique and then applied the jackknife technique on the modified data set.

Data Set: In FEV.DAT.csv, the strata are created using the age group of the patients. There are three strata, which are used as imputation classes. Two variables FEV status (Y) and height (X) in inches of the patients are considered. The summary statistics of Y and X are given in Table 1.

Table 1. Summary statistics of the sample data set

Stratum (i)	N_h	$\overline{Y}_h$	$\overline{X}_h$	C_{yh}^2	C_{xh}^2	C_{yxh}^2	ρ_{yxh}^2
Stratum 1	300	2.0335	56.9610	0.0642	0.0060	0.8280	0.8280
Stratum 2	300	3.0530	64.2746	0.0536	0.0037	0.0108	0.7556
Stratum 3	54	3.6667	69.0909	0.0587	0.0026	0.0004	0.7795

For the truncation of the available data set, the procedure is defined as follow:

$$\Upsilon_{jh} = \begin{cases} Q_{1h} & \text{if}\, y_{jh} < Q_{1h} \\ y_{jh} & \text{if}\, Q_{1h} < y_{jh} < Q_{3h}, \\ Q_{3h} & \text{if}\, y_{jh} > Q_{3h} \end{cases} \quad \eta_{jh} = \begin{cases} Q_{1h} & \text{if}\, x_{jh} < Q_{1h} \\ x_{jh} & \text{if}\, Q_{1h} < x_{jh} < Q_{3h} \\ Q_{3h} & \text{if}\, x_{jh} > Q_{jh} \end{cases} \tag{3.1}$$

In Figure 1 we illustrate the original (O) behaviour of the study and the auxiliary variables respectively within each stratum. In the second row, the truncated (T) behaviour of the target study variable w.r.t the auxiliary variable is expressed. After applying the above mentioned truncation procedure, we observed that the correlation coefficient in the first two strata is decreased but in the third stratum it improved significantly.

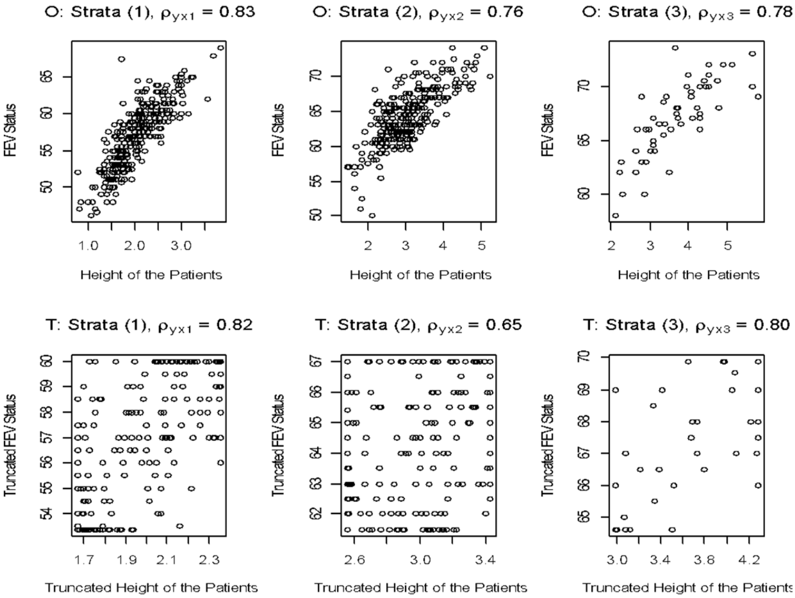


Figure 1. Data illustration within strata

Table 2. Variance of the suggested estimators

Response Rate									Variances				
Strata (i)													
1			2			3							
n_1	r_1		n_2	r_2		n_3	r_3		$v\left(\bar{y}_{I.}\right)$	$v_j\left(\bar{\Upsilon}_I\right)$	$v_{L.rat.}\left(\bar{\Upsilon}_I\right)$	$v_s\left(\bar{\Upsilon}_I\right)$	$v_{L.reg.}\left(\bar{\Upsilon}_I\right)$
80	20		80	20		20	5		0.000000049	0.015270	0.002447	0.000271	0.000026
	40			40			10		0.000000046	0.015039	0.000245	0.000063	0.000022
	60			60			15		0.000000026	0.015019	0.000066	0.000037	0.000018
160	40		160	40		30	10		0.000000180	0.010799	0.000246	0.000031	0.000017
	80			80			15		0.000000159	0.010702	0.000063	0.000016	0.000011
	120			120			20		0.000000145	0.010324	0.000024	0.000011	0.000008
240	60		240	60		40	15		0.000000051	0.009220	0.000069	0.000010	0.000007
	120			120			20		0.000000046	0.008899	0.000025	0.000006	0.000004
	180			180			25		0.000000025	0.008031	0.000012	0.000005	0.000002

Based on the sampled information with hot deck imputation for non-respondent, we consider jackknife winsorized variance estimation with imputed data by adjusting the imputed values. The winsorized variance of a given data set is 0.045130.

A different jackknife and its linearized version of estimators is considered under the ABB approach. In Table 2, the variance of the different versions of the jackknife estimator is given under a different response rate. It is clearly noticed that, the linearized version of the regression estimator under the jackknife technique outperforms as compared to others.

4. Discussion

In our present study, we discussed jackknife winsorized variance estimation based on the single imputed value. We also modified the linearized version of the jackknife variance estimator suggested by Rao (1996) for the precise estimation of winsorized variance, which is helpful for computer programs that use linearized methods for the estimation of variance. We discussed the stratified sampling scheme, because it is commonly used in large scale socio-economic surveys. We used the traditional ratio, classical linear regression and weighted hot deck imputation procedure within the classes. As we know, these imputation procedures are not reliable under multiple imputation but they could provide the valid design-based inference about the stated problem. For the practical application of this procedure, the available complete data set has information of the response status for each item and for the imputation group. The current existing computer algorithms are modified to implement these variance estimators without the permanent retention of the supplementary data.

For the stratified random sampling the imputed estimator for the population characteristics (say mean) is unbiased, under the ratio estimators with the same probability of response from all strata and the design model is also unbiased under the ratio super population model. Similarly, our modified procedure under the guideline of Rao (1996) is also consistent for the estimation of winsorized variance under uniform response from all the strata as well as the unbiased estimator under the ratio model. In this study, we estimate the approximate unbiasedness of the jackknife estimator, when the weighted or hot deck imputation is used to impute the missing values.

Our study is concentrated around the univariate estimation of the population parameters like mean and total under marginal imputation. For some complex population parameters like regression and correlation coefficient, marginal imputation is considered the association between variables. For the common donor hot deck imputation, we have the same donor set for this joint imputing of the non-response values by handling those problems which are being faced in the marginal imputation.

The current work can be extended under some modern methods like the Gibbs sampling for drawing the imputation values from the posterior distribution of the non-observed values instead of common donor hot deck imputation, but we have mentioned earlier the modern methods for obtaining the significant imputation that take into account the design features which are currently under consideration.

Acknowledgement

Thanks to the editorial team and reviewers for their the expert opinion, which improve the quality of the manuscript.

References

Chen, Q., Elliott, M. R., Haziza, D., Yang, Y., Ghosh, M., Little, R. J., and Thompson, M., (2017). Approaches to improving survey-weighted estimates. *Statistical Science*, 32(2), pp. 227–248.

Fay, R. E., (1993). Valid inferences from imputed survey data, "in proceedings of the survey research methods". *Journal of the American Statistical Association*, 1, pp. 227–232.

Korn, E. L., Graubard, B. I., (2011). *Analysis of health surveys* (Vol. 323), John Wiley & Sons.

Kovar, J. G., Chen, E. J., (1994). Jackknife variance estimation of imputed survey data. *Survey Methodology*, 20, pp. 45–52.

Little, R., Rubin, D. B., (1987). *Statistical Analysis With Missing Data*, New York: Wiley.

Mujtaba, A., Ali, M. and Kohli, K., (2014). Statistical optimization and characterization of pH-independent extended-release drug delivery of cefpodoxime proxetil using Box–Behnken design. *Chemical Engineering Research and Design*, 92(1), pp. 156–165.

Mohamed, C., Sedory, S. A. and Singh, S., (2016). *Imputation using higher order moments of an auxiliary variable. Communications in Statistics-Simulation and Computation*, 46(8), pp. 6588–6617.

Rao, J., (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, 91(434), pp. 499–506.

Rao, J., Sitter, R. R., (1992). Jackknife variance estimation under imputation for missing survey data. *Technical Report 214 Carleton University*, Laboratory for Research in Statistics and Probability.

Rao, J. N. and Shao, J., (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79(4), pp. 811–822.

Rao, J. N., Sitter, R., (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82(2), pp. 453–460.

Rao, J. N. K., (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, 91(434), pp. 499–506.

Little, R. J., Rubin, D. B., (2019). *Statistical analysis with missing data,* Vol. 793, John Wiley & Sons.

Rudolf, F. J., William, W. J. and Ping, S., (2006). Regression analysis: statistical modeling of a response variable. Elsevier.

Rubin, D. B., (1976). Inference and Missing Data. *Biometrika*, 63(3), pp. 581–592.

Rubin, D. B., (1978). Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. *Journal of the American Statistical Association*, 1, pp. 20–34.

Rubin, D. B., Schenker, N., (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81(394), pp. 366–374.

Sarndal, C.-E., (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, 18(2), pp. 241–252.

Sitter, R., (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*, 92(438), pp. 780–787.

Whitridge, P., Kovar, J., (1990). *Use of mass imputation to estimate for subsample variables*, 1, pp. 132–137.

Wolter, K., (1985). *Introduction to Variance Estimation*, New York: Springer-Verlag.

STATISTICS IN TRANSITION new series, June 2022
Vol. 23 No. 2, pp. 33–48, DOI 10.2478/stattrans-2022-0015
Received – 15.05.2021; accepted – 10.01.2022

Socio-economic development and quality of life of NUTS-2 units in the European Union

Maciej Jewczak[1], **Magdalena Brudz**[2]

ABSTRACT

Analyses regarding socio-economic development and quality of life are an important aspect of research and discussion for many international organisations, states and local authorities. Due to the complexity and multidimensionality of these issues, conducting research can be problematic. The conclusions of various analytical centres indicate that there are many paths towards establishing a set of factors which affect quality of life and ways of assessing socio-economic development levels. Depending on the criteria considered, the most common methods for determining the degree of the advancement of life quality or socio-economic development include taxonomical techniques and analyses of potential, which are based mainly on objective data sourced from official registers.

The main purpose of the paper is to investigate the level of socio-economic development and quality of life in the European Union in the years 2004 and 2018. The analyses were conducted for a rarely used level of spatial data aggregation, i.e. for NUTS-2 units. The analysis covers only those European regions that were *EU* members in 2004. As the primary research tool, the two-dimensional development matrix was adopted, which enabled the verification of the hypothesis regarding the convergence of synthetic measures that indicate the levels of socio-economic development and quality of life in the *EU* regions. For these indices, the development matrix is also used to identify the strengths and weaknesses as well as the opportunities and threats for selected spatial units, and, at the same time, to estimate the rates of change of the socio-economic development and quality of life levels.

Key words: quality of life, development matrix, taxonomic techniques, regional analyses.

[1] Corresponding author: University of Lodz, Department of Operational Research, Łódź, Poland. E-mail: maciej.jewczak@uni.lodz.pl. ORCID: https://orcid.org/0000-0002-7837-3486.

[2] University of Lodz, Department of Operational Research, Łódź, Poland. E-mail: magdalena.brudz@uni.lodz.pl. ORCID: https://orcid.org/0000-0003-4994-9682.

1. Introduction

Dynamic economic and social progress forces people living in modern societies to attach great importance not only to a better/higher quality of life, but also to the socio-economic development of their inhabitants. In the European Union (*EU)* and worldwide, it is possible to identify clusters of better/less developed regions, and when following the tendencies, their arrangement may change spatio-temporally.

To make an objective analysis of both socio-economic development (*SE*) and the level of quality of life (*QoL*), it is necessary to use appropriate tools. As part of development research, some challenges may arise related to various aspects of everyday life that should be covered by the study. As a result, it is possible to identify better/weaker developed areas. Development is a term strictly connected with the issue of progress (PWN, 2021) and it is usually defined as a long-term process of directional change or as the transformation from simple, lower, less perfect forms to more complex and advanced solutions. In the socio-economic sciences, development is generally defined as the overall change or the transformations that affect both society and the economy. It is a multi-faceted and long-term process (Schumpeter, 1960; Cyrson, 1997; Begg et al., 2007; Samuelson and Nordhaus, 2012). Therefore, it should not be assigned only with direct economic progress; it should also include important social, cultural or environmental factors. For this reason, various indicators become analytically useful, although all aspects are rarely, if ever, developed equally.

Focusing solely on the assessment of the advancement of economic development, the literature most often uses gross domestic product (*GDP*) *per capita* as a development measure (Stiglitz, Sen and Fitoussi, 2010). An indicator of socio-economic progress that is frequently used due to the provision of information about the health of the surveyed population is infant mortality *per* 1 000 live births (Robine, Romieu and Cambois, 1999), or the percentage of girls attending school. In addition to these indicators, the level of development is also estimated with energy consumption *per person*, research and development (*R&D*) expenditure, educational attainment or gender wage comparisons (Stiglitz, Sen and Fitoussi, 2018).

By analysing the socio-economic development and the quality of life simultaneously, from the point of view of a single social unit, these phenomena are characterised by a subjective assessment and are not clearly defined or comprehended. In order to live better and/or happier, it is essential to consider many aspects of daily activity and discuss the issue from a broader perspective. Currently, a decent wage and a reliable occupation or a good socio-economic background are no longer sufficient (Tomkiewicz, 2018). For this reason, this type of research also involves qualitative indicators such as opinions that reflect an immeasurable element of development. However, these indications are still subjective because each respondent has personal assessment criteria.

The study analysed the regional results for 262 NUTS-2 regions of the European Union Member States according to the *EU*'s members composition in 2004. Reducing the research to a lower level of spatial data aggregation is justified by the increase in regional heterogeneity, which represents the statistical significance of the variability level (Đurović, Bigović and Milović, 2017). Considering the conditions of local economies makes it possible to identify inequalities in regional development (Annoni, Dominicis and Khabirpour, 2019). Ertur, Le Gallo and Baumont (2006) claim that the spatial distribution of areas characterised by high/low economic development tends to show constant decomposition over time.

The data on both socio-economic development and quality of life were sourced from the Eurostat and the Organisation for Economic Co-operation and Development (*OECD*) databases. Several detailed statistics were unavailable for selected spatial data aggregation. Thus, to make the research database complete, comparable and reliable, the missing information was supported with data from the local Central Statistical Offices (*CSOs*). Based on the collected data, comparative research was conducted for 2004 and 2018, when the socio-economic differentiation and quality of life levels in the NUTS-2 units were assessed. The study estimated the synthetic indicators of socio-economic differentiation and the quality of life in each of the 262 analysed *EU* regions in order to obtain information on the quality of life and socio-economic condition of the spatial units. Additionally, as a result of the research, in the empirical part of the article, the analysed objects were further classified in the development matrix. Applying a combined and multidimensional approach to the analysed issue allowed to achieve the research goal concerning changes in the socio-economic development in relation to the quality of life of the population at the provincial level. This approach allowed for the verification of the overall hypothesis of permanent positive changes in both spheres of life of every human being.

2. Criteria for building life quality and socio-economic development indices

The aforementioned indicators illustrate only a fragment of reality. For this reason, the United Nations Development Agency (*UNDP*) annually publishes the collective Human Development Index (*HDI*). It analyses the level of development of countries based on a long and healthy life, knowledge level and also standard of living. The highest value of the synthetic index for each analysed country is a unity, with zero as the lowest value. Since it was first developed, this indicator has been modified many times. In the 30th edition in 2020, a factor measuring the impact on the natural environment was also taken into account, which analysed countries and their inhabitants' impact on nature. As a result, the real situation of the analysed states has become much more realistic. This modification significantly influenced the classification of countries in the ranking. Some of them, previously considered worth following, fell to lower positions in the development hierarchy in 2020 (UNDP, 2020).

In 1994, the World Health Organization (*WHO*) constituted its Quality of Life Department (identified by the acronym *WHOQOL*), which characterised the term quality of life as a subjective perception of individuals' life. It takes into consideration cultural background and values assigned to personal ambitions, possibilities, rules and different obstacles (WHO, 1997). Quality of life has a huge impact on physical and mental health and relationships with others, which was assumed as a reference point in the study of theoretical and empirical considerations. Hawthorne and Osborne (2005) indicated that while constructing the quality-of-life measure, indications should always be explained from a personal perspective. However, it should be noted that each social unit makes a global assessment of the quality of its life differently, which is often influenced by the place of residence or position in the social structure.

The selection of the most important criteria that allowed to determine the indices of life quality and socio-economic development in 262 *EU* spatial units were organised in accordance with the "Better Life Index" proposed by the *OECD,* and the applied methodology corresponds with the latest recommendations of the *OECD* and the Joint Research Centre (*JRC*) of the European Commission (*EC*), *i.e.* the 10-step system for constructing indicators (OECD, 2008). Some of the determinants specified by the *OECD* were not included, or they were replaced by other characteristics due to difficulties related to the data availability at the regional level. Only objective and accessible data sources were selected. While the quality of life is a multidimensional phenomenon when constructing the life quality (*LQ)* measure, indisputable intangible factors such as education, the state of the environment or digital and information development were taken into account, following Agénor and Lim (2018).

The characteristics were initially compiled into subgroups, consistent with the classification of European statistics. The construction of the synthetic *LQ* indicator included 16 quantitative determinants that express various quality of life aspects. The measure representing the socio-economic background (*SE* index) was composed by analogy and based on 17 quantitative characteristics that illustrate numerous aspects of socio-economic development. Stanickova (2015) defined the main factors of socioeconomic development, listing six groups of characteristics that are crucial for *EU* economies. She focused on economic growth, infrastructure level, and everyday human life and education, although the interest of her research was national economies. After collecting all the studies conducted thus far, a list of summary factors was created and further applied in the analysis (Tab. 1).

Table 1. Factors applied in the construction of the synthetic *LQ* and *SE*

Life Quality index		Socioeconomic Development index	
Subgroup	**Determinants**	**Subgroup**	**Determinants**
Education	Participation in education, and additional training rates by attainment level	**Economic accounts**	GDP in constant prices
Health protection, environment and social welfare	Health care conditioning; healthy life years and life expectancy; usage of resources, qualified medical staff; efficiency of health care; mortality rate; air pollution level of 2.5PM; subjective life satisfaction	**Labour market**	Employment level; other labour assets
		Poverty and social exclusion	Poverty rate by type; households at risk of poverty
Income	household accounts	**Science and technology**	R&D expenditure
Digital economy and information society	Internet access; use of IT tools and solutions	**Transportation**	Public roads and railroads; vehicle stock; road safety – victims of road accidents by type and severity

Source: own elaboration.

2.1. Method of building the life quality and socio-economic development indices, including the development matrix

The variables listed in the previous section were initially standardised and used to estimate the indicators of life quality and socio-economic development for each NUTS-2 region and the selected period separately. As a result, it was possible to compare the variability of the quality-of-life indicators with the results of socio-economic differentiation, *i.e.* with tendencies and indicators calculated for the analysed spatial objects.

Many paths were considered regarding how these indicators should be estimated (pattern and non-pattern methods). One crucial condition for the construction of synthetic measures is data comparability (the additivity postulate). The normalisation process also includes the elimination of negative values from the calculations and the stability of the level of variability – the postulate of constant range or stability of extreme values. To maintain data comparability, the standardisation transformation with mean and standard deviation values was used according to the following formula:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \quad (1)$$

where: x_{ij} – the factor's value, $\bar{x}_j$ – the factor's value, S_j – the factor's standard deviation. Ultimately, the calculation was based on the hierarchical taxonomic measure of development proposed by Hellwig (1968):

$$m_i = 1 - \frac{d_{i0}}{d_0}, \quad (i = 1,2, \dots, n) \tag{2}$$

where: $\boldsymbol{d_{i0}} = \sqrt{\sum_{j=1}^{m}\left(z_{ij} - z_{0j}\right)^2}$ – the Euclidean distance between the i-th observation from the pattern of development, $\boldsymbol{d_0} = \sqrt{\sum_{j=1}^{m}\left(z_{0j} - z_{-0j}\right)^2}$ – the Euclidean distance between the pattern $\boldsymbol{z_{0j}}$ and $\boldsymbol{z_{-0j}}$ anti-pattern of development, which are implemented in the multivariant analysis in accordance with the character of each individual variable. The method considers the stimulative or destimulative impact of a characteristic for the overall level of the mi taxonomic measure. In the case of model values, when a given factor has been defined as stimulating for the general level of a complex phenomenon, its maximum value is adopted as $\boldsymbol{z_{0j}}$; in the case of a destimulating effect, by contrast, it is adopted as its minimum value. For anti-model values $\boldsymbol{z_{-0j}}$ the completely opposite situation occurs, and the minimum value for the stimulating factor is accepted, and the maximum value in the case of a destimulant (Suchecki, 2010).

The combined procedures make it possible to determine two objects (most often hypothetical) that represent the best and the worst possible alternatives. The pattern and anti-pattern perform two functions in the analysis. The first one is to assess the individual level of the phenomenon in the given i-th object; the second one is to provide a certain standardisation point of the size of the phenomenon.

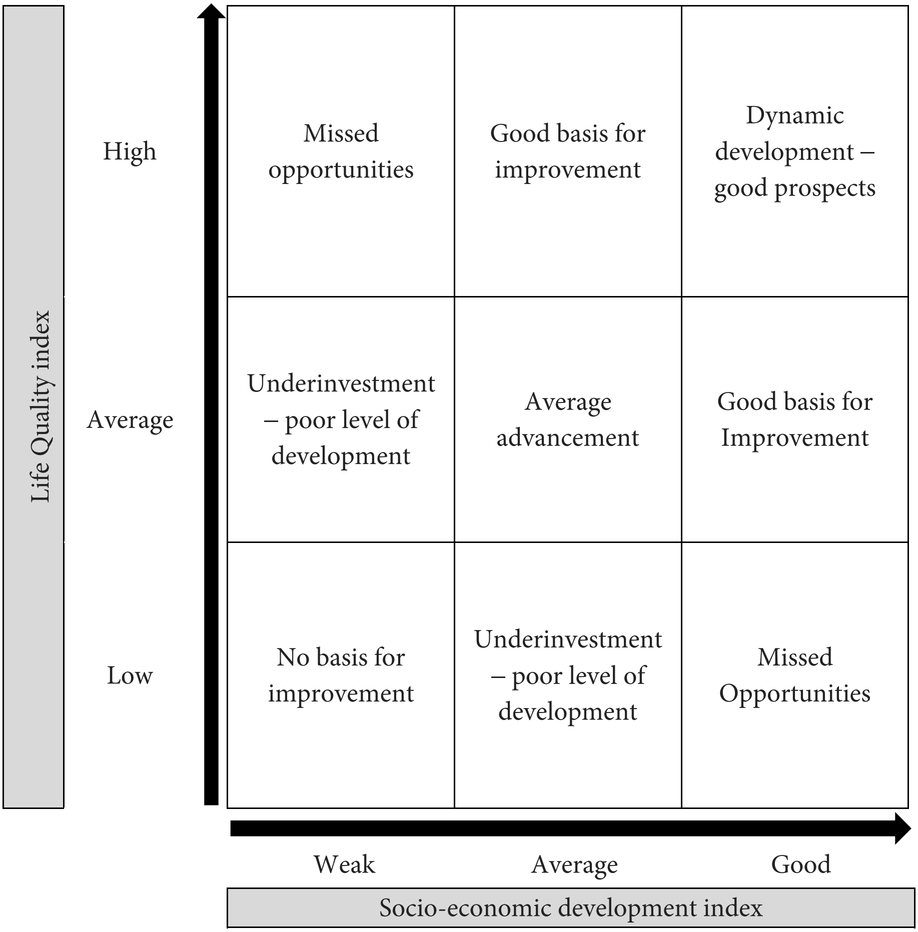
It is known from the properties of the taxonomic measure that the higher the level of a complex phenomenon, the higher the level of the mi development measure. The measure assumes values in the range [0,1]; for the model, it takes the value of unity, and for the anti-model, it takes the zero value.

Due to the comparability of objects ordered inside the measure, it becomes intuitive to interpret, especially for assessing the development or deterioration in the quality of life or the socio-economic development in the NUTS-2 regions. The use of the uniform set of diagnostics features makes it possible to compare the tendencies of changes in the levels of the two indices for local populations for two periods (or moments). Therefore, it is possible to introduce a graphic summary of the numerical results based on the formula of a relative increase (rate of change), calculated as follows (Hydzik, 2012):

$$\boldsymbol{rLQI}_{\frac{t}{t-1}} = \frac{\boldsymbol{LQI_t - LQI_{t-1}}}{\boldsymbol{LQI_{t-1}}}, \boldsymbol{rSE}_{\frac{t}{t-1}} = \frac{\boldsymbol{SE_t - SE_{t-1}}}{\boldsymbol{SE_{t-1}}} \tag{3}$$

where: $\boldsymbol{rLQI}_{\frac{t}{t-1}}$ – regional rate of change of the mi indicator of life quality, and $\boldsymbol{rSE}_{\frac{t}{t-1}}$ – the regional rate of change of the mi indicator of socio-economic development.

The indications that result from the change rates allowed for an additional interpretation of observed tendencies, indicating regions of improvement and/or deterioration of the development of the analysed phenomena. This comparison also allowed for the assessment of the pace of the changes noted, which is important for the implementation of *EU* policy goals and for discussions about equalising opportunities for regions considered to be weakly developed compared to Western Europe or Scandinavian countries.



Life Quality index	Weak	Average	Good
High	Missed opportunities	Good basis for improvement	Dynamic development – good prospects
Average	Underinvestment – poor level of development	Average advancement	Good basis for Improvement
Low	No basis for improvement	Underinvestment – poor level of development	Missed Opportunities
	Socio-economic development index		

Figure 1. Development matrix design

Source: based on Jewczak and Korczak (2020).

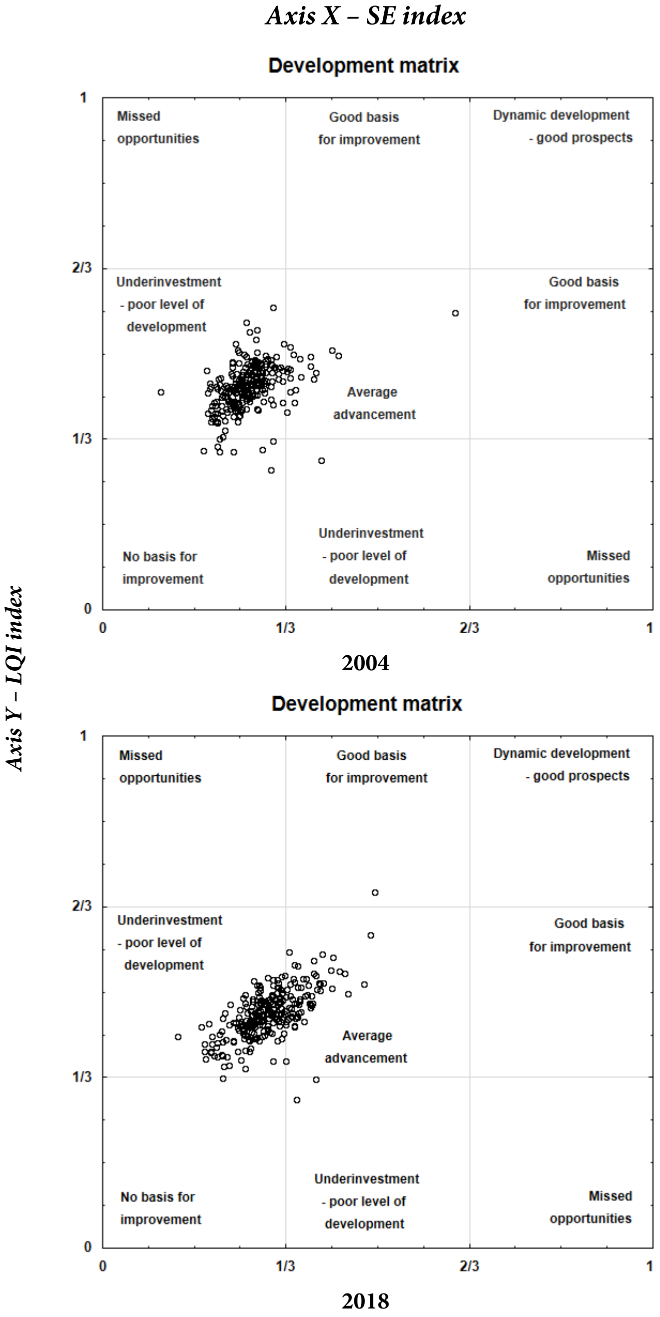
The overall summary of the analysis carried out in the paper is based on the construction of a proposed development matrix, which is a classification technique that makes it possible to position two objects in a two-dimensional format that describes the relationship between two analysed phenomena, here: quality of life and the socio-economic development (Jewczak and Korczak, 2020).

The development matrix consists of rows and columns that present the level of individual features that differentiate the positions of objects on a scale from 0 to 1. The matrix is divided into nine equal fields, each representing the characteristics of the phenomena's development level, and they should be interpreted in accordance with the strategic field definition. From the interpretational perspective, the scatter plot design has the obvious advantage, which is connected with indicating the relationship between phenomena, the intensity and, when making temporal comparisons, indicating tendencies of change.

3. Empirical analysis results for *EU* NUTS-2 regions

To establish the relationship between the quality of life of inhabitants and the situation of socio-economic development of NUTS-2 regions in the selected *EU* countries, the quality of life *LQI* and socio-economic development *SE* indices were assessed. The spatio-temporal analysis for 2004 and 2018 adopted the reference object approach and further comparison in the ordered development matrix. On this basis, spatial objects were classified as illustrated in Fig. 2. The cloud image of the analysed objects in the 2004 development matrix allowed us to illustrate positive trends in quality of life. Meanwhile, the image for 2018 is more dispersed, and a significant part of the regions shifted to more positive strategic fields of greater development.

The impact of the socio-economic development changes was less explicit. As the research tool indicated, for both of the analysed periods, most of the spatial units were counted as objects with "no basis for improvement". Only one of the objects (the Île-de-France region) recorded an improvement in *SE* development in 2018 compared to 2014, moving to an average level. In both time points, the best situation in terms of quality of life and socio-economic development was recorded in the Swedish East Middle region, as shown by the highest coordinate in the development matrix.



2004

2018

Figure 2. Results of the development matrices in 2004 and 2018

Source: developed by the Authors based on EUROSTAT, *OECD* and local *CSO*'s *LFS* data.

From the properties of the proposed analytical tool of the development matrix and the positioning of individual strategic fields, it follows quite intuitively that the closer the object is to the origin of the coordinate system, the worse the situation of the object is, i.e. "no basis for improvement". The objects' movement over time towards strategic fields of higher values of complex phenomena should be assessed as positive changes that result from the improvement of one or both phenomena simultaneously in the direction of the (1,1) coordinate. Additionally, by analysing the shape of the cloud image of the distribution of points in the scatterplot in the development matrix, it is also possible to identify the relationship that occurs between the analysed phenomena.

In 2004, the vast majority of spatial objects were characterised by an average level of intensity in the quality of life and a low level of socio-economic development. The result of such a two-dimensional classification is the concentration of coordinates within the "underinvestment and poor level of development" field. One object with the opposite relationship between the quality of life and socio-economic development was the Lithuania region. This unit recorded an average socio-economic development level with a low intensity level of the quality of life. The best positioned spatial unit in 2004 was Île de France, which noted the highest level of socio-economic development and quality of life; however, the intensity of the phenomena is considered to be an "average advancement".

When analysing the *EU* NUTS-2 regions in the final year of the analysis, it is possible to conclude that the situation of the objects generally improved over time. There was only one spatial unit (Lithuania) which was classified as "no basis for improvement" – it was also the worse-positioned object in 2004. Most objects were classified as "poor level of development", characterised by an average level of quality of life and low levels of socio-economic development. By reversing the direction of the analysis, only two objects of "poor level of development" were identified (Pays de la Loire, Calabria), which were characterised by an average level of socio-economic development and a low intensity of quality of life.

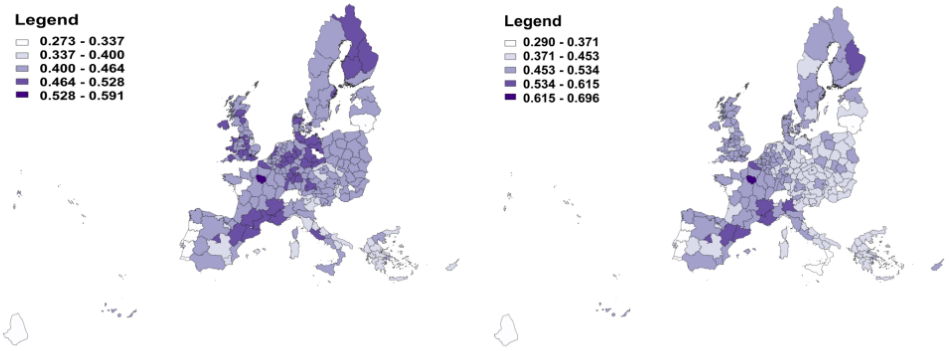
In 2018, objects indicating an "average advancement" in both quality of life and socio-economic development constituted quite a large group (more numerous than in 2004). However, the distribution of the coordinates clearly indicates that the position of the NUTS-2 objects is more stimulated by the intensity of the quality of life than by the socio-economic development. In the analysed period, only one object with a "good basis for improvement" was specified, and its position results from the noted high level of quality of life. Again, this was the region of Île de France.

By analysing the overall perception of development, a positive tendency should be emphasised, as a significant number of regions mostly positioned in the field of "underinvestment – poor level of development" in 2004 shifted towards fields of better assessment in 2018. The direction of the tendency of the frequency distribution of the measures indicated that the observable changes should be perceived as favourable – pushing objects towards the strategic field of an average level of advancement.

When comparing the cloud images for the differentiation of distribution of points in the development matrix scatterplots, it can be easily noticed that the advancement of the condition of NUTS-2 regions is the result of changes in their levels of socio-economic development more than an improvement in the quality of life. However, the analysis demonstrated that the life quality also advanced.

This relationship between quality of life and socioeconomic development showed a positive empowerment relationship with the passage of time. In 2004, the strength of this association was established by the Rxy Spearman's coefficient at (+0.503) level and advanced in 2018 to (+0.642) – this relationship was significant at p <0.05.

The research indicated that there was an increase in the overall intensity of the life quality – this is noticeable in the extreme, minimum and maximum values of the complex phenomena. The highest levels of the recorded quality of life were observed in the Scandinavian area and Western Europe (Fig. 3).



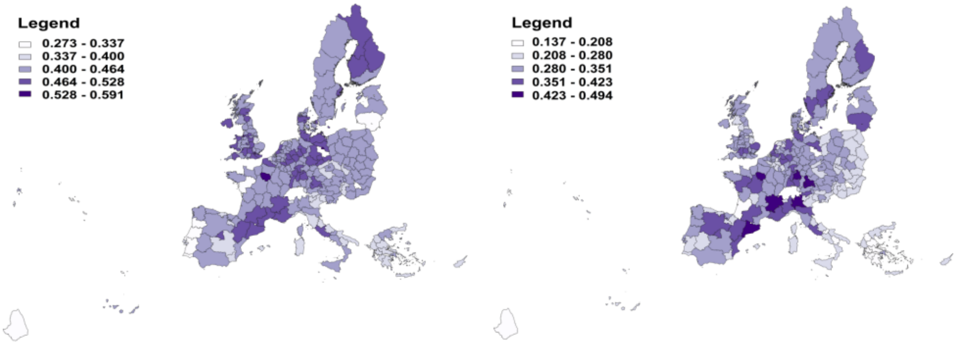
2004 **2018**

Figure 3. Values of synthetic measures for *Quality of Life* in 2004 and 2018

Source: developed by the Authors based on EUROSTAT, *OECD* and local *CSO's LFS* data.

This could be summarised by the statement that the regions located within the borders of the founding members of the *EU* were characterised by a high quality of life. These tendencies were convergent for both periods; however, it is clearly visible that the quality of life deteriorated in the regions of the United Kingdom, Germany, for most regions of Poland, Slovakia, the Czech Republic, and Romania, as well as the Balkan area.

In contrast to the LQ indicators, for the socio-economic development index, the intensity of the phenomenon decreased between 2004 and 2018. Again, this could be summarised by the extreme level of the maximum value of the synthetic measures (Fig. 4). The spatial arrangement in the selected periods is quite convergent, and the improvement in the intensity of the spatial distribution mostly concerned objects that noted a higher level in the first place.



2004 **2018**

Figure 4. Values of synthetic measures for socio-economic development in 2004 and 2018

Source: developed by the Authors based on EUROSTAT, *OECD* and local *CSO*'s *LFS* data.

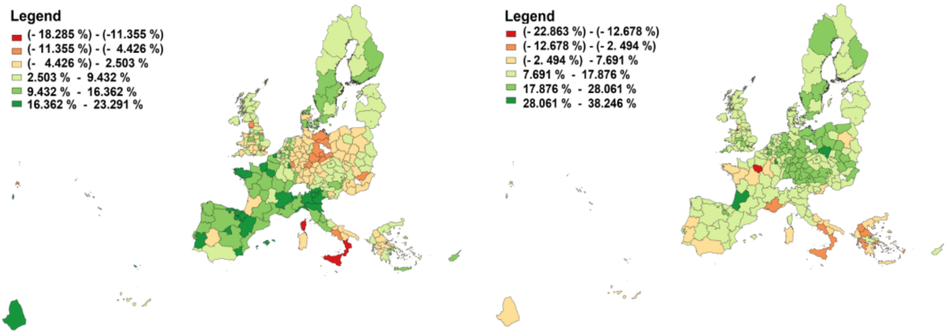
This situation may suggest a general improvement in the socio-economic situation in most of the NUTS-2 regions (which was previously observed by the greater dispersion within the strategic fields in the development matrix). The socio-economic development level improved in regions of the Iberian peninsula, France and Northern Italy, whereas in Poland, the level of socio-economic development for most of the analysed spatial units in 2018 was assessed at a lower intensity level compared to 2004. A positive conclusion of the analysis is the improvement in the situation of the Lithuania region, which advanced to a higher intensity group of socio-economic development in 2018. However, it should be remembered that this region was assessed as the worst for quality of life in both periods.

4. Conclusions

The biggest advantage of the research is that the scientific analysis covered data at the regional level, while most studies focus only on quality of life or socio-economic development at the macro level. It is hard to find studies that consider differences at the regional or even local level that do not focus on the internal differentiation within national borders. The study carried out in the article provided information on the quality of life and socio-economic development in 262 NUTS-2 regions in the *EU*

Member States. The research goal was achieved by measuring the improvement in the quality of life and the socio-economic development using a taxonomic measure of development. This allowed not only to evaluate each of the *i*-th objects in relation to the reference object, but also to obtain results taking into account the intensity of variability of phenomena. The proposed procedure solves the problem of the impact of individual components that were taken into account when constructing the indices to reflect more accurately both the quality of life and the socio-economic situation of the regions. However, due to the configuration of the assessed measures (although they are based on reliable information sourced in official registers), the resulting quality of life and socio-economic indicators should be treated as an information point. This is due to the limitations related to data availability at the lower spatial data aggregation level.

One of the positive conclusions from the study is that the relationship between the two measures of *LQ* and *SE* was significant, in terms of non-parametrical Spearman's correlation coefficient – for 2004, the R_{xy} amounted to 0.503 (significant at p <0.05), while in 2018, the R_{xy} was 0.642 (also significant). This connection indicates that the relationship between the phenomena is positive and strengthens over time. Considering the results based on the applied development matrices, the conclusion (supported by the graphical presentation of change rate tendencies (Fig. 5)) indicates that the arrangement of objects in the coordinate system shifted towards a more positive assessment, defined as an "average advancement" in both phenomena.



rLQI ***rSE***

Figure 5. Rates of change in *LQI* and *SE* measures

Source: developed by Authors based on EUROSTAT, *OECD* and local CSO's *LFS* data.

For the life quality (*rLQI*), in particular, the improvement was observed within countries considered to be "more developed", with the exception of Germany. The graphics of change rates also indicate that for most of the analysed regions, there was a positive change in terms of socio-economic development (*rSE*). Again looking at the

German NUTS-2 units, although they recorded an unfavourable change in terms of the quality of life, there was an improvement in socio-economic development.

To summarise the results of the multivariate analysis, there was a positive relationship between the quality of life and socio-economic development, which strengthened over the analysed period. One should evaluate positively the regions that recorded favourable change rates in the levels of synthetic measures, which is consistent with the previously noted trends. For most of the NUTS-2 areas, the quality of life improved, except for the areas of Germany and Poland and neighbouring countries, which share a similar socio-economic background. For the regions identified with negative (unfavourable) rates of change, although the changes were not spectacularly low/high, these results might be a consequence of their migration policies of opening borders to residents of the countries admitted to the *EU* in the analysed period.

The same may be true for some regions of France, Italy and Germany, which are seen as a constant target of migration movements in Europe. Changes in the levels of socio-economic development, which accelerated in Central and Eastern Europe, should be assessed positively, with simultaneous downward trends recorded in the regions of the "old Union" countries. However, this finding may support the previously-mentioned concept of underdeveloped countries catching up to highly developed countries rather than it being the case that, overall, the quality of life or socio-economic development in well-developed economies deteriorated significantly.

References

Agénor, P. R., Lim, K. Y., (2018). Unemployment, growth and welfare effects of labor market reforms. *Journal of Macroeconomics*, 58, pp. 19–38.

Annoni, P., De Dominicis, L. and Khabirpour, N., (2019). Location matters: A spatial econometric analysis of regional resilience in the European Union. *Growth and Change*, 50(3), pp. 824–855.

Begg, D., Fischer, S. and Dornbusch, R., (2007). *Makroekonomia*, PWE, Warszawa.

Brudz, M., Jewczak M., (2019). Quality of life and labour market in Poland, In M. Papież and S. Śmiech (Eds.), *The 13th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio-Economic Phenomena, Conference Proceedings*, Wydawnictwo C.H. Beck, Warszawa, pp. 17–25.

Cyrson, E. (ed.), (1997). *Kompendium wiedzy o gospodarce*, PWN, Warszawa–Poznań.

Đurović, G., Bigović, M. and Milović, N., (2017). Support for further enlargement of the EU: Statistical analysis of regional differences. *Journal of Balkan and Near Eastern Studies*, 19(3), pp. 243–258.

Ertur, C., Le Gallo, J. and Baumont, C., (2006). The European regional convergence process, 1980-1995: Do spatial regimes and spatial dependence matter?. *International Regional Science Review*, 29(1), pp. 3–34.

European Commission, (2019). *Constructing a composite indicator*, https://ec.europa.eu/jrc/en/coin/10-step-guide/overview (accessed on: 10.01.2019).

Fitouss, J. P., Sen, A. K. and Stiglitz, J. E., (2011). *Mismeasuring Our Lives: Why GDP Doesn't Add Up*. ReadHowYouWant.com.

Hawthorne, G., Osborne, R., (2005). Population norms and meaningful differences for the Assessment of Quality of Life (AQoL) measure. *Australian and New Zealand journal of public health*, 29(2), pp. 136–142.

Hellwig, Z., (1968). Zastosowanie metody taksonomicznej do typologicznego podziału krajów ze względu na poziom ich rozwoju oraz zasoby i strukturę wykwalifikowanych kadr. *Przegląd statystyczny*, 4(1968), pp. 307–326.

Hydzik P., (2012). Zastosowanie metod taksonomicznych do oceny poziomu rozwoju społeczno-ekonomicznego powiatów województwa podkarpackiego. *Zeszyty Naukowe Politechniki Rzeszowskiej, Ekonomia i Nauki Humanistyczne*, z. 19 (2/2012), pp. 17–32.

Jewczak, M., Korczak, K., (2020). Poverty and Health State in Poland: Evidence from a regional perspective. *Management Issues*, 18, 3(89), Warsaw, pp. 49–66.

Nermend K., (2017). *Metody analizy wielokryterialnej i wielowymiarowej wspomaganiu decyzji*, PWN, Warszawa.

OECD, JRC European Commission, (2008). *Handbook on constructing composite indicators: methodology and user guide*, OECD Publishing.

PWN, (2021). Rozwój, https://encyklopedia.pwn.pl/haslo/rozwoj;4009883.html (accessed on: 10.02.2021).

Robine, J. M., Romieu, I. and Cambois, E., (1999). Health expectancy indicators. *Bulletin of the World Health Organisation*, 77(2), p. 181.

Samuelson, P. A., Nordhaus, W. D., (2012). Ekonomia. *Rebis*, Warszawa.

Schumpeter, J.A., (1960). *Teoria rozwoju gospodarczego*, PWN, Warszawa.

Stanickova, M., (2015). Classifying the EU competitiveness factors using multivariate statistical methods. *Procedia Economics and Finance*, 23, pp. 313–320.

Stiglitz, J., Sen, A. and Fitoussi, J., (2018). Beyond GDP. *Project Syndicate*.

Suchecki, B. (ed.), (2010). *Ekonometria przestrzenna: metody i modele analizy danych przestrzennych,* Wydawnictwo CH Beck.

Tomkiewicz, J., (2018). The labour market and income distribution in postsocialist economies – Non-obvious regularities. *Communist and Post-Communist Studies*, 51(4), pp. 315–324.

UNDP, (2020). Human Development Report 2020. *The next frontier. Human development and the Anthropocene*. The United Nations Development Programme Publishing.

WHO, (1997). *Division of Mental Health and Prevention of Substance Abuse*. WHOQOL: measuring quality of life. World Health Organization, https://apps.who.int/iris/handle/10665/63482 (accessed on: 10.09.2020).

STATISTICS IN TRANSITION new series, June 2022
Vol. 23 No. 2, pp. 49–68, DOI 10.2478/stattrans-2022-0016
Received – 20.07.2020; accepted – 01.03.2021

A Bayesian estimation of the Gini index and the Bonferroni index for the Dagum distribution with the application of different priors

Sangeeta Arora[1], **Kalpana K. Mahajan**[2], **Vikas Jangra**[3]

ABSTRACT

Bayesian estimators and highest posterior density credible intervals are obtained for two popular inequality measures, viz. the Gini index and the Bonferroni index in the case of the Dagum distribution. The study considers informative and non-informative priors, i.e. the Mukherjee-Islam prior and the extension of Jeffrey's prior, respectively, under the presumption of the Linear Exponential (LINEX) loss function. A Monte Carlo simulation study is carried out in order to obtain the relative efficiency of both the Gini and Bonferroni indices while taking into consideration different priors and loss functions. The estimated loss proves lower when using the Mukherjee-Islam prior in comparison to the extension of Jeffrey's prior and the LINEX loss function outperforms the squared error loss function (SELF) in terms of the estimated loss. Highest posterior density credible intervals are also obtained for both these measures. The study used real-life data sets for illustration purposes.

Key words: Inequality measures, Bayes estimator, credible interval, LINEX loss function.

1. Introduction

The Dagum distribution (also called the inverse Burr distribution; Dagum called it a generalized Logistic-Burr distribution (Kleiber and Kotz, 2003) is a well-known distribution popularly used to model income distribution. Camilo Dagum proposed the Dagum distribution in 1970, which is a skewed and heavy tailed distribution and is appropriate to model the distribution of financial, income as well as wealth distribution. The Dagum distribution was developed as an alternative to the Pareto distribution and lognormal distribution and it performs better than other two/three

[1] Department of Statistics, Panjab University, Chandigarh, India. E-mail: sarora131@gmail.com. https://orcid.org/0000-0002-2052-4798.

[2] Department of Statistics, Panjab University, Chandigarh, India. E-mail: mahajan_kr@pu.ac.in. https://orcid.org/0000-0001-5274-2418.

[3] Department of Statistics, Panjab University, Chandigarh, India. E-mail: vjangra49@gmail.com. ORCID: https://orcid.org/0000-0002-4228-1782.

parameters income/wealth distribution models when applied to empirical data (Chotikapanich and Griffiths, 2006). One of the special cases of the Dagum distribution appeared for the first time in Burr (1942) as the third example (Burr III) of solutions of the Burr distribution system. The three parameter Dagum Type I distribution evolved from Dagum's experimentation with a shifted log-logistic distribution (Chotikapanich, 2008).

The probability density function of the Dagum distribution is given as

$$f(x;a,b,p)=\begin{cases}\frac{ap}{x}\left(\frac{\left(\frac{x}{b}\right)^{ap}}{\left(\left(\frac{x}{b}\right)^{a}+1\right)^{p+1}}\right), & x>0;\ \ a,b,p>0,\\ 0, & othwerwise.\end{cases} \tag{1}$$

The plot of probability density function of the Dagum distribution for various $p = 3.5$,4.5, 7.8 with $a = 2.5, b = 1.5$ is shown in Figure 1.

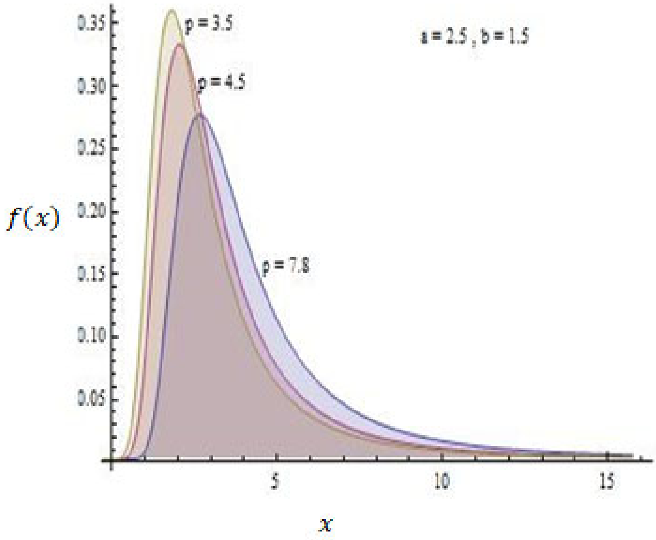


Figure 1. Probability density function of the Dagum distribution

The cumulative distribution function of the Dagum distribution is given by

$$F(x;a,b,p)=\begin{cases}\left(1+\left(\frac{x}{b}\right)^{-a}\right)^{-p}, & x>0;\quad a,b,p>0,\\ 0, & otherwise,\end{cases} \tag{2}$$

where a and p are shape parameters and b is a scale parameter. For $p = 1$, the Dagum distribution is also referred to as log-logistic distribution (Dagum, 1975).

Inequality is a vital characteristic of non-negative distribution. It is used to analyze data in socio-economic sciences, in the context of income distribution. In the context of income inequality, the Gini index (Gini, 1912; Foster et al., 1984) is generally defined as

$$G = 1 - 2\int_0^1 L(p)dp\,, \quad 0 \le p \le 1, \tag{3}$$

where $L(p) = (1/\mu)\int_0^p F^{-1}(t)dt$ is the equation of the Lorenz curve and $\mu = \int_0^1 F^{-1}(t)dt$ is the mean of the distribution.

The Bonferroni index is defined as

$$B = 1 - \int_0^1 B(p)dp, \qquad 0 \le p \le 1, \tag{4}$$

where $B(p) = \left(\frac{1}{p\mu}\right)\int_0^p F^{-1}(t)dt$ is the equation of the Bonferroni curve.

The curve was introduced by Bonferroni (1930) and has been analysed and studied by various authors: see for instance De Vergottini (1940), Tarsitano (1990), Giorgi & Crescenzi (2001) and Zenga (2013).

In the case of the Dagum distribution, the Gini index (G) and the Bonferroni index (B) are given by

$$G = \frac{\Gamma p\, \Gamma(2p+\frac{1}{a})}{\Gamma 2p\, \Gamma(p+\frac{1}{a})} - 1, \qquad a, p > 0, \tag{5}$$

where $\Gamma(.)$ is the Gamma function,

and $$B = p\left[\varphi\left(p + \frac{1}{a}\right) + \varphi(p)\right], \quad a, p > 0, \tag{6}$$

where $\varphi(x) = \frac{d}{dx} ln\sqrt{x} = \frac{\Gamma(x)'}{\Gamma(x)}$, is the Digamma function.

Note that both values are independent of the scale-parameter b.

A huge literature exists on the estimation of the Gini index and inequality measures using classical approach, i.e. parametric and non-parametric (Moothathu, 1985; Sen, 1988; Dixon, 1987; Bansal et al., 2011). But in the case of Bayesian set up, a lot of work still needs attention (Sathar et al., 2009; Bhattacharya and Chaturvedi, 1999) particularly in the context of income inequality. In the case of Pareto distribution Bayesian estimators of the Gini index (Kaur et al., 2015) are obtained using different priors under LINEX loss function. Some work regarding Bayesian estimation of the shape parameter p of the Dagum distribution is available under different loss functions using informative and non-informative priors (Naqash et al., 2017) while Layla et al. (2020) discussed the Bayesian estimation of the survival function using Gamma as informative and Jeffrey as non-informative prior, but the income inequality field still awaits the attention of researchers. In the present paper, Bayesian estimators for two famous inequality indices, viz. the Gini index and the Bonferroni index will be obtained for the Dagum distribution along with their credible intervals. These inequality indices are not only used in the economic set up but have applications in other fields such as survival analysis, reliability and bio-statistics.

When the Bayesian approach is used, the selection of a suitable prior distribution plays a major role. Basically, priors can be divided into informative (an informative prior depends on elicitation of prior distribution based on pre-existing scientific knowledge in the area of investigation), non-informative (a non-informative prior is usually improper, $i.e.$ it does not have a proper density function but the resulting posterior distribution is a proper density function), and conjugate prior (if the posterior distribution $p(\theta|x)$ is from the same family of probability distributions as the prior probability distribution $p(\theta)$) (Kass and Wasserman, 1996; Berger, 2006). In the Bayesian estimation, the benchmark for quality (good) estimator for the parameters of interest is the selection of the proper loss function. A squared error loss function is the simplest loss function among all the loss functions. It is also known as a quadratic loss function, defined as

$$L(\theta) = (\hat{\theta} - \theta)^2, \tag{7}$$

where $\hat{\theta}$ is the estimator of θ.

The squared error loss function (SELF) is symmetrical and shows equal importance to losses due to overestimation and underestimation of equal magnitude. One disadvantage of using the squared error loss function is that it penalizes overestimation or underestimation. Overestimation of a parameter can lead to more severe or less severe consequences than underestimation, or vice versa. In the case of income inequality under-estimation is more serious as compared to overestimation (Kaur et al., 2015). For this reason, the use of an asymmetrical loss function, which can provide greater importance to overestimation or underestimation, can be considered for the estimation of the parameters. Many asymmetrical loss functions are available in the statistical literature and one such Linear exponential loss function (LINEX) has been proposed by Varian (1975) as

$$L\big(\hat{\theta} - \theta\big) = e^{b(\hat{\theta}-\theta)} - b(\hat{\theta} - \theta) - 1,\ b \neq 0. \tag{8}$$

The posterior expectation of the LINEX loss function is

$$E(L\big(\hat{\theta} - \theta\big)) = e^{b\hat{\theta}}E(e^{-b\theta}) - b\left(\hat{\theta} - E(\theta)\right) - 1,$$

where $E(.)$ denotes posterior expectation with respect to the posterior density of θ.

By a result of Zellner (1986) the Bayes estimator of θ denoted by $\hat{\theta}$ under the LINEX loss function is the value which minimizes posterior expectation and is given by

$$\hat{\theta} = -\tfrac{1}{b}\ln\big[E\big(e^{-b\theta}\big)\big], \tag{9}$$

provided the expectation $E(e^{-b\theta})$ exists and is finite.

The LINEX loss function is approximately equal to the squared error loss function for the small values of b.

In this paper, the LINEX loss function is used for obtaining Bayesian estimators for two popular inequality indices, i.e. the Gini index and the Bonferroni index in the case of the Dagum distribution using Mukherjee-Islam prior (informative prior) and the extension of Jeffrey's prior (non-informative prior). The plan of the paper is as follows. In Section 2, prior and posterior distributions are discussed in the case of the Dagum distribution. In Section 3, Bayesian estimators are obtained for the Gini index and the Bonferroni index for the Dagum distribution under the assumption of the LINEX loss function. In Section 4, using simulation, relative efficiency of Bayesian estimates is obtained for both the Gini and Bonferroni index taking into consideration different priors and two loss functions, LINEX and SELF. In Section 5, the credible intervals are defined and highest posterior density credible intervals are carried out for both the Gini index and the Bonferroni index. Two real life examples to illustrate the method of Bayesian setup are given in Section 6.

2. Prior and posterior distribution

2.1. Case 1: Shape parameter p is unkown and a, b are known

Let $X = (x_1, x_2, \ldots, x_n)$ be a random sample from the Dagum distribution with shape parameters p and a and scale parameter b, i.e. $X \sim D(a, b, p)$, then the likelihood function for the Dagum distribution as a function of p (keeping a and b fixed) is given by

$$L = \left(\frac{ap}{b^{ap}}\right)^n \prod_{i=1}^n {x_i}^{ap-1} \prod_{i=1}^n \left(1 + \left(\frac{x_i}{b}\right)^a\right)^{-(p+1)} \tag{10}$$

$$L(a,b,p) \propto p^n \prod_{i=1}^n \left(\frac{x_i}{b^{ap}}\right) \prod_{i=1}^n \left[1 + \left(\frac{x_i}{b}\right)^a\right]^{-p}$$

$$= p^n e^{p \sum_{i=1}^n \ln\left(\frac{x_i}{b}\right)^a} e^{-p \sum_{i=1}^n \ln\left[1+\left(\frac{x_i}{b}\right)^a\right]}$$

$$\Rightarrow \; L(p|x) \propto \; p^n e^{-p \sum_{i=1}^n \ln\left[1+\left(\frac{x_i}{b}\right)^{-a}\right]}$$

$$= p^n e^{-pT}$$

where

$$T = \sum_{i=1}^n \ln\left[1 + \left(\frac{x_i}{b}\right)^{-a}\right]. \tag{11}$$

Posterior distribution under Mukherjee-Islam prior

Mukherjee Islam (1983) is a well-known probability distribution used by many researchers to model a failure distribution for the purpose of reliability and Bayesian analysis.

Assume that p has a Mukherjee-Islam prior with hyper parameters $(\alpha,\sigma) > 0$, defined by

$$\pi(p) = \alpha\sigma^{-\alpha}p^{\alpha-1} \ ; \ p > 0, \alpha > 0, \sigma > 0. \tag{12}$$

Then, the posterior distribution of p under Mukherjee-Islam prior is given by

$$\pi_M(p|x) = \frac{L(p)*\pi(p)}{\int_0^\infty L(p)*\pi(p)\,dp}$$

$$\pi_M(p|x) \propto p^{n+\alpha-1}e^{-pT}$$

$$= hp^{n+\alpha-1}e^{-pT}$$

where h is the normalized constant given by

$$h = \int_0^\infty p^{n+\alpha-1}\, e^{-pT}dp$$

$$= \Gamma(n+\alpha)/T^{n+\alpha}\,.$$

Thus, the posterior distribution of p is given by

$$\pi_M(p|x) = \frac{T^{n+\alpha}}{\Gamma(n+\alpha)}p^{n+\alpha-1}e^{-pT}, \tag{13}$$

which is gamma density with parameters $T\ and\ \beta_1 = n + \alpha$.

Posterior distribution under extension of Jeffreys' prior

Jeffreys' prior is a particular case of the extension of Jefferys' prior proposed by Kutubi and Ibrahim (2009). The extension of Jeffreys' prior is defined as

$$\pi(p) \propto [I(p)]^m \,; m > 0,$$

where $[I(p)]$ is the Fisher Information given by

$$[I(p)] = -E\left[\frac{\partial^2 l}{\partial p^2}\right] = \frac{n}{p^2},$$

where l is the log-likelihood function. For $m = \ 0.5$, it reduces to Jeffreys' prior. Thus, the extension of Jeffreys' prior is given by

$$\pi(p) \propto \frac{1}{p^{2m}}, m > 0. \tag{14}$$

The posterior distribution is defined by

$$\pi_{EJ}(p|x) \propto p^{n-2m}e^{-pT} = Kp^{n-2m}e^{-pT},$$

where k is the normalized constant given by

$$K = \int_0^\infty p^{n-2m}\, e^{-pT}dp = \frac{\Gamma(n-2m+1)}{T^{n-2m+1}}.$$

Thus, the posterior distribution of p/x is given by

$$\pi_{EJ}(p|x) = \frac{T^{n-2m+1}}{\Gamma(n-2m+1)}p^{n-2m}e^{-pT}, \tag{15}$$

which is a gamma density with parameters $T and\ \beta_2 = n - 2m + 1$.

2.2. Case 2: Shape parameter $\boldsymbol{a}$ is unkown and $\boldsymbol{p, b}$ are known

Let $X = (x_1,x_2,\ldots,x_n)$ be a random sample from $D(a, b, p)$ Dagum distribution. Then, the likelihood function of the scale parameter a (keeping p and b fixed) is given by

$$L = \left(\frac{ap}{b^{ap}}\right)^n \prod_{i=1}^{n} x_i{}^{ap-1} \prod_{i=1}^{n} \left(1 + \left(\frac{x_i}{b}\right)^a\right)^{-(p+1)}.$$

Posterior distribution under Mukherjee-Islam prior

Assume that a has a Mukherjee-Islam prior with hyper parameters $(\alpha, \sigma) > 0$ defined by

$$g(a) = \alpha\sigma^{-\alpha}a^{\alpha-1} \ ; \alpha > 0, \sigma > 0. \tag{16}$$

The posterior distribution of a is

$$\pi_M(a|x) = \frac{(a^{n+\alpha-1})(\frac{1}{b^{ap}})^n \prod_{i=1}^{n} x_i{}^{ap-1}\left(1+\left(\frac{x_i}{b}\right)^a\right)^{-(p+1)}}{\int_0^\infty (a^{n+\alpha-1})(\frac{1}{b^{ap}})^n \prod_{i=1}^{n} x_i{}^{ap-1}\left(1+\left(\frac{x_i}{b}\right)^a\right)^{-(p+1)} da} \tag{17}$$

Posterior distribution under extension of Jeffreys' prior

The extension of Jeffreys' prior is given by

$$g(a) \propto \frac{1}{a^{2m}}, m > 0. \tag{18}$$

The posterior distribution of a is

$$\pi_{EJ}(a|x) = \frac{(a^{n-2m})(\frac{1}{b^{ap}})^n \prod_{i=1}^{n} x_i{}^{ap-1}\left(1+\left(\frac{x_i}{b}\right)^a\right)^{-(p+1)}}{\int_0^\infty (a^{n-2m})(\frac{1}{b^{ap}})^n \prod_{i=1}^{n} x_i{}^{ap-1}\left(1+\left(\frac{x_i}{b}\right)^a\right)^{-(p+1)} da} \tag{19}$$

3. Bayesian estimation under Linear Exponential (LINEX) loss function using different priors

3.1. Case 1: Shape parameter $\boldsymbol{p}$ is unkown and $\boldsymbol{a, b}$ are known

Bayesian estimators using Mukherjee-Islam prior

Using the posterior distribution given in (13) the Bayesian estimator $\widehat{G_{ML}}$ of the Gini index G using Mukherjee-Islam prior is

$$\widehat{G_{ML_1}} = \frac{-1}{b}\log\ E[e^{-bG}]$$

$$= \frac{-1}{b}\log\left[\int_0^\infty e^{-b\left(\frac{\Gamma p\, \Gamma\left(2p+\frac{1}{a}\right)}{\Gamma 2p\, \Gamma\left(p+\frac{1}{a}\right)}-1\right)} \frac{T^{n+\alpha}}{\Gamma(n+\alpha)} p^{n+\alpha-1} e^{-pT} dp\right]$$

$$= \frac{-1}{b} \log \left[\frac{T^{n+\alpha}}{\Gamma(n+\alpha)} \int_0^\infty e^{-\left(b \frac{\Gamma p\, \Gamma\left(2p+\frac{1}{a}\right)}{\Gamma 2p\, \Gamma\left(p+\frac{1}{a}\right)} - b - pT \right)} p^{n+\alpha-1} dp \right]. \tag{20}$$

The Bayesian estimator $\widehat{B_{ML}}$ of the Bonferroni index B using Mukherjee-Islam prior is

$$\widehat{B_{ML_1}} = \frac{-1}{b} \log\ E[e^{-bB}]$$
$$= \frac{-1}{b} \log \left[\int_0^\infty e^{-bp\left[\varphi\left(p+\frac{1}{a}\right)+\varphi(p)\right]} \frac{T^{n+\alpha}}{\Gamma(n+\alpha)} p^{n+\alpha-1} e^{-pT} dp \right]$$
$$= \frac{-1}{b} \log \left[\frac{T^{n+\alpha}}{\Gamma(n+\alpha)} \int_0^\infty e^{-bp\left[\varphi\left(p+\frac{1}{a}\right)+\varphi(p)\right]-pT} p^{n+\alpha-1} dp \right]. \tag{21}$$

Bayesian estimators using extension of Jeffreys' Prior

Using the posterior distribution given in (15) the Bayesian estimator $\widehat{G_{EL}}$ of the Gini index G using the extension of Jeffreys' prior is

$$\widehat{G_{EL_1}} = \frac{-1}{b} \log\ E[e^{-bG}]$$
$$= \frac{-1}{b} \log \left[\int_0^\infty e^{-b\left(\frac{\Gamma p\, \Gamma\left(2p+\frac{1}{a}\right)}{\Gamma 2p\, \Gamma\left(p+\frac{1}{a}\right)} - 1 \right)} \frac{T^{n-2m+1}}{\Gamma(n-2m+1)} p^{n-2m} e^{-pT} dp \right]$$
$$= \frac{-1}{b} \log \left[\frac{T^{n-2m+1}}{\Gamma(n-2m+1)} \int_0^\infty e^{-\left(b \frac{\Gamma p\, \Gamma\left(2p+\frac{1}{a}\right)}{\Gamma 2p\, \Gamma\left(p+\frac{1}{a}\right)} - b - pT \right)} p^{n-2m} dp \right]. \tag{22}$$

The Bayesian estimator $\widehat{B_{EL}}$ of the Bonferroni index B using the extension of Jeffreys' prior is

$$\widehat{B_{EL_1}} = \frac{-1}{b} \log\ E[e^{-bB}]$$
$$= \frac{-1}{b} \log \left[\int_0^\infty e^{-bp\left[\varphi\left(p+\frac{1}{a}\right)+\varphi(p)\right]} \frac{T^{n-2m+1}}{\Gamma(n-2m+1)} p^{n-2m} e^{-pT} dp \right]$$
$$= \frac{-1}{b} \log \left[\frac{T^{n-2m+1}}{\Gamma(n-2m+1)} \int_0^\infty e^{-bp\left[\varphi\left(p+\frac{1}{a}\right)+\varphi(p)\right]-pT} p^{n-2m} dp \right]. \tag{23}$$

3.2. Case 2: Shape parameter a is unkown and p, b are known

Bayesian estimators using Mukherjee-Islam prior

Using the posterior distribution given in (17) the Bayes estimator $\widehat{G_{ML}}$ of the Gini index G using Mukherjee-Islam prior is

$$\widehat{G_{ML_2}} = \frac{-1}{b} \log\ E[e^{-bG}]$$
$$= \frac{-1}{b} \log \left[\int_0^\infty e^{-b\left(\frac{\Gamma p\, \Gamma\left(2p+\frac{1}{a}\right)}{\Gamma 2p\, \Gamma\left(p+\frac{1}{a}\right)} - 1 \right)} \frac{(a^{n+\alpha-1})(\frac{1}{bap})^n \prod_{i=1}^n x_i{}^{ap-1}\left(1+\left(\frac{x_i}{b}\right)^a\right)^{-(p+1)}}{\int_0^\infty (a^{n+\alpha-1})(\frac{1}{bap})^n \prod_{i=1}^n x_i{}^{ap-1}\left(1+\left(\frac{x_i}{b}\right)^a\right)^{-(p+1)} da} \right]. \tag{24}$$

The Bayes estimator $\widehat{B_{ML}}$ of the Bonferroni index B using Mukherjee-Islam prior is

$$\widehat{B_{ML_2}} = \frac{-1}{b}\log\ E[e^{-bB}]$$
$$= \frac{-1}{b}\log\left[\int_0^\infty e^{-bp\left[\varphi\left(p+\frac{1}{a}\right)+\varphi(p)\right]} \frac{(a^{n+\alpha-1})(\frac{1}{b^{ap}})^n \prod_{i=1}^n x_i{}^{ap-1}\left(1+\left(\frac{x_i}{b}\right)^a\right)^{-(p+1)}}{\int_0^\infty (a^{n+\alpha-1})(\frac{1}{b^{ap}})^n \prod_{i=1}^n x_i{}^{ap-1}\left(1+\left(\frac{x_i}{b}\right)^a\right)^{-(p+1)} da}\right]. \tag{25}$$

Bayesian estimators under extension of Jeffreys' prior

The Bayes estimator $\widehat{G_{EL}}$ of the Gini index G using the extension of Jeffreys' prior is

$$\widehat{G_{EL_2}} = \frac{-1}{b}\log\ E[e^{-bG}]$$
$$= \frac{-1}{b}\log\left[\int_0^\infty e^{-b\left(\frac{\Gamma p\,\Gamma\left(2p+\frac{1}{a}\right)}{\Gamma 2p\,\Gamma\left(p+\frac{1}{a}\right)}-1\right)} \frac{(a^{n-2m})(\frac{1}{b^{ap}})^n \prod_{i=1}^n x_i{}^{ap-1}\left(1+\left(\frac{x_i}{b}\right)^a\right)^{-(p+1)}}{\int_0^\infty (a^{n-2m})(\frac{1}{b^{ap}})^n \prod_{i=1}^n x_i{}^{ap-1}\left(1+\left(\frac{x_i}{b}\right)^a\right)^{-(p+1)} da}\right]. \tag{26}$$

The Bayes estimator $\widehat{B_{EL}}$ of the Bonferroni index B using the extension of Jeffreys' prior is

$$\widehat{B_{EL_2}} = \frac{-1}{b}\log\ E[e^{-bB}]$$
$$= \frac{-1}{b}\log\left[\int_0^\infty e^{-bp\left[\varphi\left(p+\frac{1}{a}\right)+\varphi(p)\right]} \frac{(a^{n-2m})(\frac{1}{b^{ap}})^n \prod_{i=1}^n x_i{}^{ap-1}\left(1+\left(\frac{x_i}{b}\right)^a\right)^{-(p+1)}}{\int_0^\infty (a^{n-2m})(\frac{1}{b^{ap}})^n \prod_{i=1}^n x_i{}^{ap-1}\left(1+\left(\frac{x_i}{b}\right)^a\right)^{-(p+1)} da}\right] \tag{27}$$

Remark As all these expressions cannot be simplified further, the Bayesian estimators have been obtained using simulation techniques in R software.

4. Simulation study

In order to assess the statistical performance of these estimators for the Gini index and the Bonferroni index, a simulation study is conducted. The $VGAM$ package in R software is used to draw the sample from the Dagum distribution and using simulation the Bayes estimates and their corresponding losses are computed. Theprocess is replicated 10000 times and the average of the results has been presented in the tables below (Tables 1-4). The estimated losses are computed for both LINEX and the squared error loss function (SELF) using generated random samples from the Dagum distribution and by considering three sample sizes, (i) small sample size $n = 25$, (ii) moderate sample size $n = 50$, (iii) large sample size $n = 100$. The estimated losses are repeated for Mukherjee-Islam prior and the extension of Jeffreys' prior using different configuration of scale and shape parameters, viz. $b(known)$ =1.5, 1.3, 1.2, $a(known)$ =2.5, 1.75, 1.6, $p(unkown)$ =7.8, 4.5, 3.5 and $b(known)$ =1.2, 2.2, 3.2, $p(known)$ =1.4, 1.7, 1.9, $a(unkown)$ =5.8, 2.5, 1.5. The hyper parameters values are α = 2.2, 3.5 and m = 1, 1.7 are chosen using MLE values in R software.

Table 1. Bayesian Estimates under LINEX loss function and Estimated loss (in parenthesis) for Gini Index and Bonferroni Index under Mukherjee-Islam prior (when p is unknown, a and b are known)

n	b	a	p	α =2.2		α =3.5	
				$\widehat{G}_{ML_1}$	$\widehat{B}_{ML_1}$	$\widehat{G}_{ML_1}$	$\widehat{B}_{ML_1}$
25	1.5	2.5	7.8	0.14119	0.17720	0.14053	0.17590
				(0.03647, 0.03848)	(0.05690,0.05817)	(0.03672,0.03825)	(0.05748,0.05950)
	1.3	1.75	4.5	0.141006	0.25390	0.14177	0.25368
				(0.03654, 0.03789)	(0.08675,0.08836)	(0.03626,0.03798)	(0.08785,0.08837)
	1.2	1.6	3.5	0.14341	0.27770	0.14179	0.27523
				(0.03566, 0.036148)	(0.06531,0.06661)	(0.03625,0.03775)	(0.06627,0.06819)
50	1.5	2.5	7.8	0.14139	0.17759	0.14063	0.17684
				(0.03640, 0.03715)	(0.05686,0.05739)	(0.03668,0.03799)	(0.05706,0.05874)
	1.3	1.75	4.5	0.14121	0.25448	0.14205	0.25394
				(0.03647, 0.03718)	(0.08648,0.08739)	(0.03616,0.03701)	(0.08673,0.08735)
	1.2	1.6	3.5	0.14404	0.27925	0.14376	0.27860
				(0.03543, 0.03608)	(0.06472,0.06597)	(0.03553,0.03691)	(0.06497,0.06691)
100	1.5	2.5	7.8	0.14152	0.17746	0.14088	0.17722
				(0.03638, 0.03699)	(0.05522,0.05669)	(0.03660,0.03709)	(0.05688,0.05797)
	1.3	1.75	4.5	0.14292	0.25450	0.14235	0.25447
				(0.03584,0.03615)	(0.08531,0.08649)	(0.03605,0.03700)	(0.08548,0.08646)
	1.2	1.6	3.5	0.14587	0.27879	0.14519	0.27933
				(0.03481,0.03513)	(0.06390,0.06459)	(0.03501,0.03611)	(0.06369,0.06479)

Notation used: Estimated loss (under LINEX, under SELF)

Table 2. Bayesian Estimates under LINEX loss function and Estimated loss (in parenthesis) for Gini and Bonferroni index under the extension of Jeffreys' prior (when p is unknown, a and b are known)

n	b	a	p	m =1		m =1.7	
				$\widehat{G}_{ML_1}$	$\widehat{B}_{ML_1}$	$\widehat{G}_{ML_1}$	$\widehat{B}_{ML_1}$
25	1.5	2.5	7.8	0.14072	0.17732	0.14007	0.17682
				(0.036684,0.03801)	(0.05741,0.05964)	(0.03689,0.03816)	(0.05707,0.05936)
	1.3	1.75	4.5	0.21439	0.25300	0.21493	0.25368
				(0.06362,0.06560)	(0.08727,0.08902)	(0.06339,0.06549)	(0.08685,0.08894)
	1.2	1.6	3.5	0.24123	0.17982	0.23999	0.18100
				(0.06760,0.06996)	(0.20488,0.22827)	(0.06808,0.06971)	(0.20398,0.22873)
50	1.5	2.5	7.8	0.14098	0.17735	0.14070	0.17737
				(0.03655,0.03715)	(0.05683,0.05806)	(0.03665,0.03793)	(0.05682,0.05815)
	1.3	1.75	4.5	0.21519	0.25428	0.21505	0.25449
				(0.06366,0.06499)	(0.08657,0.08748)	(0.06335,0.06499)	(0.08647,0.08764)
	1.2	1.6	3.5	0.24176	0.18305	0.24153	0.18294
				(0.06739,0.06861)	(0.20241,0.21953)	(0.06748,0.06879)	(0.20149,0.22238)
100	1.5	2.5	7.8	0.14103	0.17806	0.14088	0.17727
				(0.03643,0.03709)	(0.05583,0.05696)	(0.03659,0.03704)	(0.05586,0.05619)
	1.3	1.75	4.5	0.21478	0.25431	0.21471	0.25471
				(0.06346,0.064185)	(0.08556,0.08602)	(0.06320,0.06435)	(0.08546,0.08675)
	1.2	1.6	3.5	0.24193	0.18424	0.24098	0.18328
				(0.06720,0.06818)	(0.20006,0.21174)	(0.06700,0.06861)	(0.20106,0.21352)

Notation used: Estimated loss (under LINEX, under SELF)

Table 3. Bayesian Estimates under LINEX loss function and Estimated loss (in parenthesis) for Gini Index and Bonferroni Index under Mukherjee-Islam prior (when α is unknown, p and b are known)

n	b	p	a	α =2.2		α =3.5	
				$\widehat{G}_{ML_2}$	$\widehat{B}_{ML_2}$	$\widehat{G}_{ML_2}$	$\widehat{B}_{ML_2}$
25	1.2	1.4	5.8	0.41101	0.37029	0.41738	0.37936
				(0.00324,0.00349)	(0.00260,0.00277)	(0.00325,0.00344)	(0.00261,0.00285)
	2.2	1.7	2.5	0.48465	0.53917	0.49700	0.53748
				(0.00558,0.00582)	(0.00362,0.00382)	(0.00564,0.00595)	(0.00377,0.00390)
	3.2	1.9	1.5	0.50321	0.55357	0.50791	0.55153
				(0.00665,0.00683)	(0.00161,0.00187)	(0.00655,0.00689)	(0.00167,0.00185)
50	1.2	1.4	5.8	0.41186	0.37684	0.41071	0.37579
				(0.00324,0.00341)	(0.00251,0.00268)	(0.00322,0.00340)	(0.00261,0.00275)
	2.2	1.7	2.5	0.49196	0.53931	0.49994	0.53254
				(0.00557,0.00579)	(0.00361,0.00379)	(0.00556,0.00590)	(0.00362,0.00382)
	3.2	1.9	1.5	0.50789	0.55070	0.51258	0.55503
				(0.00654,0.00680)	(0.00160,0.00179)	(0.00654,0.00681)	(0.00161,0.00176)
100	1.2	1.4	5.8	0.41284	0.38057	0.41004	0.38513
				(0.00320,0.00333)	(0.00244,0.00255)	(0.00322,0.00331)	(0.00241,0.00263)
	2.2	1.7	2.5	0.49897	0.54934	0.49148	0.54864
				(0.00555,0.00569)	(0.00358,0.00362)	(0.00555,0.00586)	(0.00350,0.00371)
	3.2	1.9	1.5	0.51124	0.56258	0.51245	0.56856
				(0.00641,0.00679)	(0.00157,0.00165)	(0.00644,0.00677)	(0.00158,0.00160)

Notation used: Estimated loss (under LINEX, under SELF)

Table 4. Bayesian Estimates under LINEX loss function and Estimated loss (in parenthesis) for Gini and Bonferroni index under the extension of Jeffreys' prior (when α is unknown, p and b are known)

n	b	p	a	m =1		m =1.7	
				$\widehat{G}_{ML_2}$	$\widehat{B}_{ML_2}$	$\widehat{G}_{ML_2}$	$\widehat{B}_{ML_2}$
25	1.2	1.4	5.8	0.38574	0.35028	0.38974	0.35525
				(0.00347,0.00365)	(0.00275,0.00288)	(0.00351,0.00369)	(0.00268,0.00279)
	2.2	1.7	2.5	0.46251	0.50124	0.46925	0.502173
				(0.00562,0.00581)	(0.00367,0.00379)	(0.00573,0.00588)	(0.00368,0.00379)
	3.2	1.9	1.5	0.48196	0.45218	0.48202	0.45869
				(0.00667,0.00679)	(0.00179,0.00183)	(0.00660,0.00671)	(0.00172,0.00189)
50	1.2	1.4	5.8	0.38159	0.35585	0.38874	0.35968
				(0.00349,0.00350)	(0.00267,0.00271)	(0.00342,0.00359)	(0.00258,0.00266)
	2.2	1.7	2.5	0.46095	0.50143	0.46748	0.50147
				(0.00560,0.00579)	(0.00366,0.00370)	(0.00563,0.00571)	(0.00367,0.00372)
	3.2	1.9	1.5	0.47259	0.45748	0.47179	0.45321
				(0.00659,0.00661)	(0.00168,0.00176)	(0.00659,0.00669)	(0.00167,0.00170)
100	1.2	1.4	5.8	0.39561	0.35249	0.39095	0.35648
				(0.00335,0.00341)	(0.00259,0.00266)	(0.00335,0.00349)	(0.00249,0.00251)
	2.2	1.7	2.5	0.47259	0.50852	0.47125	0.50357
				(0.00558,0.00561)	(0.00359,0.00362)	(0.00559,0.00569)	(0.00359,0.00368)
	3.2	1.9	1.5	0.50014	0.45354	0.50934	0.45258
				(0.00649,0.00656)	(0.00158,0.00163)	(0.00649,0.00655)	(0.00158,0.00166)

Notation used: Estimated loss (under LINEX, under SELF)

Comments: One can observe that

1) The estimated loss in each case decreases as sample size n increases for all the configurations of various parameters.
2) The estimated loss using the LINEX loss function is smaller as compared with the squared error loss function (SELF) for both Mukherjee-Islam prior and the extension of Jeffrey's prior.
3) The estimated loss is also lower using Mukherjee-Islam prior than the extension of Jeffrey's prior.

5. Credible interval

According to Eberly and Casella (2003), the $100(1-\gamma)\%$ equal tail credible interval for the exact posterior distribution can be defined as

$$P(\theta < L) = \int_{-\infty}^{L} \pi(\theta/x)d\theta = \frac{\gamma}{2}, \qquad P(\theta < U) = \int_{U}^{\infty} \pi(\theta/x)d\theta = \frac{\gamma}{2}$$

where $\pi(\theta/x)$ is the posterior distribution of θ and (L, U) are the lower and upper limits of the credible interval respectively for the specified value of γ level of significance.

Highest Posterior Density (HPD) Credible Intervals

Chen and Shao (1999) introduced the algorithm to find the HPD credible intervals. $100(1-\gamma)\%$ HPD credible interval is the $100(1-\gamma)\%$ credible interval with smallest width among all possible $100(1-\gamma)\%$ credible intervals. Once the posterior sample is generated for parameter θ_i $\left(i = 1,2,\dots,(N-N_0)\right)$, then $\theta_{(1)} \le \theta_{(2)} \le \cdots \le \theta_{(N-N_0)}$ denote the ordered values of $\theta_1, \theta_2, \dots, \theta_{(N-N_0)}$. The $100(1-\gamma)\%$ HPD interval for θ is defined by $\left(\theta_{(j)}, \theta_{(j+[(1-\gamma)(N-N_0)])}\right)$, where j is chosen such that

$$\theta_{(j+[(1-\gamma)(N-N_0)])} - \theta_{(j)} = \min_{1\le j\le M}\left(\theta_{(j+[(1-\gamma)(N-N_0)])} - \theta_{(j)}\right), \qquad j = 1,2,\dots,(N-N_0),$$

where $[x]$ denotes to greatest integer less than or equal to x.

Table 5. 95% HPD Credible Intervals, width of the interval and Bayesian estimates (in 2nd row) for Gini Index under Mukherjee–Islam Prior

n	b	a	p	α =2.2	α =3.5
				(Credible interval) (width) (Bayes Estimate)	(Credible interval) (width) (Bayes Estimate)
25	1.5	2.5	7.8	(0.08764,1.159667) (1.072027) (0.141195)	(0.098866,1.418709) (1.319843)(0.140532)
	1.3	1.75	4.5	(0.034647,1.065941) (1.031294) (0.1410067)	(0.090902,1.285003) (1.194101)(0.141778)
	1.2	1.6	3.5	(0.086689,1.750493) (1.663804)(0.1434128)	(0.029834,1.230873) (1.201039)(0.141793)

Table 5. 95% HPD Credible Intervals, width of the interval and Bayesian estimates (in 2nd row) for Gini Index under Mukherjee–Islam Prior (cont.)

n	b	a	p	α =2.2	α =3.5
50	1.5	2.5	7.8	(0.04379,1.016813)	(0.079860,1.239483)
				(0.973023)(0.1413926)	(1.159623) (0.1406394)
	1.3	1.75	4.5	(0.010254,1.01777)	(0.006467,1.069109)
				(1.007516)(0.1412134)	(1.062642)(0.1420578)
	1.2	1.6	3.5	(0.072973,1.210653)	(0.094263,1.058927)
				(1.13768)(0.1440414)	(0.964664)(0.1437612)
100	1.5	2.5	7.8	(0.050407,1.00031)	(0.046396,1.091680)
				(0.949903) (0.1415200)	(1.045284) (0.1408886)
	1.3	1.75	4.5	(0.259496,1.00071)	(0.001569,1.031742)
				(0.741214)(0.1429269)	(1.030173) (0.1423524)
	1.2	1.6	3.5	(0.005600,1.000285)	(0.047262,1.007538)
				(0.994685) (0.1458737)	(0.960276) (0.1451941)

Table 6. 95% Credible Intervals, width of the interval and Bayesian estimates for Bonferroni Index under Mukherjee–Islam Prior

n	b	a	p	α =2.2	α =3.5
				(Credible interval)	(Credible interval)
				(width) (Bayes Estimate)	(width) (Bayes Estimate)
25	1.5	2.5	7.8	(0.155791,1.996669)	(0.118866, 1.907854)
				(1.840878)(0.1772007)	(1.788988)(0.1759073)
	1.3	1.75	4.5	(0.134647,1.901148)	(0.110902,1.903905)
				(1.766501)(0.2539019)	(1.793003)(0.2536816)
	1.2	1.6	3.5	(0.116897,1.537517)	(0.129834,1.986688)
				(1.42062)(0.2777098)	(1.856854)(0.2752306)
50	1.5	2.5	7.8	(0.11965,1.743861)	(0.109801,1.618079)
				(1.624211)(0.1775938)	(1.508278)(0.1768471)
	1.3	1.75	4.5	(0.103105,1.349774)	(0.106467,1.322476)
				(1.246669)(0.2544812)	(1.216009)(0.2539437)
	1.2	1.6	3.5	(0.102973,1.182072)	(0.194236,1.460026)
				(1.079099)(0.2792508)	(1.26579)(0.2786031)
100	1.5	2.5	7.8	(0.045070,1.109478)	(0.054969,1.535899)
				(1.064408)(0.1774652)	(1.48093)(0.1772273)
	1.3	1.75	4.5	(0.059796,1.179001)	(0.043519,1.017423)
				(1.119205)(0.2545054)	(0.973904)(0.2544711)
	1.2	1.6	3.5	(0.006575,1.000118)	(0.004072,1.087538)
				(0.993543)(0.2787902)	(1.083466)(0.279333)

Table 7. 95% HPD Credible Intervals, width of the interval and Bayesian estimates for the Gini index under the extension of Jeffreys' Prior

n	**b**	**a**	**p**	**m** =1	**m** =1.7
				(Credible interval) (width) (Bayes Estimate)	(Credible interval) (width) (Bayes Estimate)
25	1.5	2.5	7.8	(0.102227,1.930575) (1.828348)(0.1407223)	(0.17144,1.901145) (1.729705)(0.1400771)
	1.3	1.75	4.5	(0.140647,1.674105) (1.533458)(0.2143965)	(0.16113,1.693517) (1.532387)(0.2149375)
	1.2	1.6	3.5	(0.166116,1.744702) (1.578586)(0.2412356)	(0.147876,1.842412) (1.694536)(0.239996)
50	1.5	2.5	7.8	(0.113269,1.459546) (1.346277)(0.1409815)	(0.154915,1.654938) (1.500023)(0.1407081)
	1.3	1.75	4.5	(0.10900,1.310285) (1.201285)(0.2151987)	(0.174812,1.450959) (1.276147)(0.2150537)
	1.2	1.6	3.5	(0.131304,1.105901) (0.974597)(0.2417637)	(0.135988,1.437119) (1.301131)(0.2415304)
100	1.5	2.5	7.8	(0.176421,1.111614) (0.935193)(0.1410336)	(0.13556,1.147504) (1.011944)(0.140881)
	1.3	1.75	4.5	(0.157634,1.133661) (0.976027)(0.2147873)	(0.139321,1.037455) (0.898134)(0.2147116)
	1.2	1.6	3.5	(0.103214,1.057451) (0.954237)(0.2419366)	(0.035624,1.098726) (1.063102)(0.240987)

Table 8. 95% HPD Credible Intervals, width of the interval and Bayesian estimates for Bonferroni Index under the extension of Jeffreys' Prior

n	b	a	p	m =1	m =1.7
				(Credible interval) (width) (Bayes Estimate)	(Credible interval) (width) (Bayes Estimate)
25	1.5	2.5	7.8	(0.115489,1.799312) (1.683823)(0.1773219)	(0.109144,1.86549) (1.756346)(0.1768241)
	1.3	1.75	4.5	(0.180611,1.841856) (1.661245)(0.2530068)	(0.150197,1.699720) (1.549523)(0.2536889)
	1.2	1.6	3.5	(0.13116,1.971017) (1.839857)(0.1798252)	(0.132148,1.360320) (1.228172)(0.1810069)
50	1.5	2.5	7.8	(0.113269,1.272277) (1.159008)(0.1773576)	(0.134027,1.590231) (1.456204)(0.1773717)
	1.3	1.75	4.5	(0.150900,1.479691) (1.328791)(0.2542884)	(0.168629,1.265633) (1.097004)(0.2544995)
	1.2	1.6	3.5	(0.113309,1.698651) (1.585342)(0.1830509)	(0.145988,1.243351) (1.097363)(0.1829449)
100	1.5	2.5	7.8	(0.169611,1.029145) (0.859534)(0.1780651)	(0.16900,1.080353) (0.911353)(0.1772798)
	1.3	1.75	4.5	(0.163091,1.154255) (0.991164)(0.2543103)	(0.122073,1.105629) (0.983556)(0.2547158)
	1.2	1.6	3.5	(0.136478,1.175373) (1.038895)(0.1842456)	(0.118012,1.006210) (0.888198)(0.1832878)

Comment: One can further infer that as sample sizes increases, the width of the credible interval decreases for 95% credible intervals for both Mukherjee-Islam prior and the extension of Jeffreys' prior. The width of HPD is smaller in the case of Mukherjee-Islam prior.

6. For illustration, two real data sets are taken up in this section

Example 1. A real data is considered for the illustration of the proposed study. This data (Daren et al. (2014)) set represents the degree of reading power (DRP) scores for a sample of 30 third grade students.

40, 26, 39, 14, 42, 18, 25, 43, 46, 27, 19, 47, 19, 26, 35, 34, 15, 44, 40, 38, 31, 46, 52, 25, 35, 35, 33, 29, 34, 41. By using easy fit software, it is seen that data fit well to the Dagum distribution and p-value for the Kolmogorov-Smirnov test is 0.87284 at 5% level of significance. The value of shape parameters and scale parameter $p = 0.14, a = 18.5,\ b = 45.7$ are obtained using easy fit software and the Bayes estimates are obtained along with HPD credible intervals for the Gini and Bonferroni Index using both Mukherjee and the extension of Jeffrey' Priors. The results have been presented in the table below (Table 9).

Table 9: Bayesian estimates along with 95% HPD Credible Intervals under LINEX loss function and Estimated loss under LINEX and SELF (in parenthesis) for Gini index and Bonferroni index under Mukherjee-Islam prior and the extension of Jeffrey's Prior

Priors		$\hat{\mathbf{G}}_{\mathbf{ML}}$	$\hat{\mathbf{B}}_{\mathbf{ML}}$
Mukherjee-Islam prior	α = 2.2	0.11480 (0.04757,0.06558)	0.06611 (0.0026,0.00277)
	95% HPD Credible Intervals (width)	(0.07145,1.54261) (1.47115)	(0.00483,0.01642) (0.01159)
Extension of Jeffrey's prior	m =1	0.12146 (0.08620,0.1982)	0.06919 (0.07093,0.14253)
	95% HPD Credible Intervals (width)	(0.0531,1.8642) (1.8111)	(0.0015,0.9631) (0.9616)

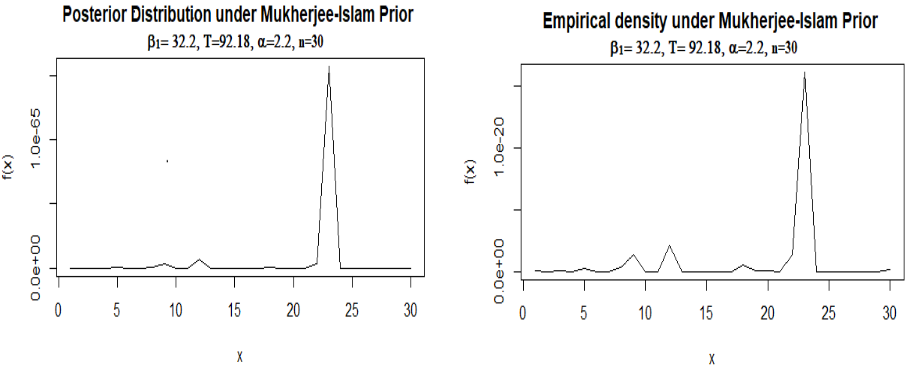


Figure 2. Comparison of Posterior density with Empirical density under Mukherjee-Islam Prior

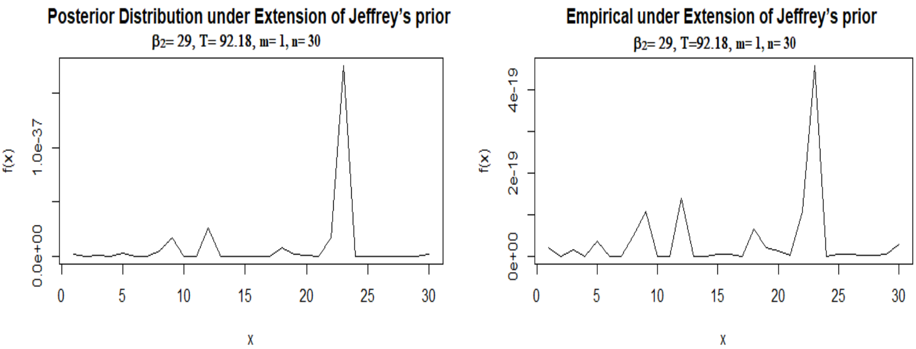


Figure 3. Comparison of Posterior density with Empirical density under Extension of Jeffreys's Prior

From the above findings of graph, we can see the posterior density and empirical density under Mukherjee-Islam prior and the extension of Jeffreys' prior are nearly the same.

Example 2. The data (Sanku et al. (2017)) set consists of 30 observations on breaking stress of carbon fibres (in Gba). The data are: 3.7, 2.74, 2.73, 3.11, 3.27, 2.87, 4.42, 2.41, 3.19, 3.28, 3.09, 1.87, 3.75, 2.43, 2.95, 2.96, 2.3, 2.67, 3.39, 2.81, 4.2, 3.31, 3.31, 2.85, 3.15, 2.35, 2.55, 2.81, 2.77, 2.17. By using easy fit software, it is seen that data fit well to the Dagum distribution and the p-value for the Kolmogorov-Smirnov test is 0.99668 at 5% level of significance. The values of shape parameters and scale parameter $p = 0.97, a = 9.7,\ b = 2.9$ are obtained using easy fit software and the Bayes estimates are obtained along with HPD credible intervals for the Gini and Bonferroni Index using both Mukherjee and Uniform Prior. The results have been presented in the table below (Table 10).

Table 10. Bayesian estimates along with 95% HPD Credible Intervals under LINEX loss function and Estimated loss under LINEX and SELF (in parenthesis) for Gini index and Bonferroni index.

Priors		$\widehat{\mathbf{G}}_{\mathbf{ML}}$	$\widehat{\mathbf{B}}_{\mathbf{ML}}$
Mukherjee-Islam prior	α = 1.2	0.02634 (0.03926,0.07219)	0.01152 (0.0011,0.00161)
	95% HPD Credible Intervals (width)	(0.00926,1.12018) (1.11092)	(0.00916,0.01849) (0.00933)
Extension of Jeffrey's prior	m =1	0.18001 (0.04165,0.21013)	0.04629 (0.06282,0.09932)
	95% HPD Credible Intervals (width)	(0.01406,1.50674) (1.49268)	(0.00132,1.73965) (1.73833)

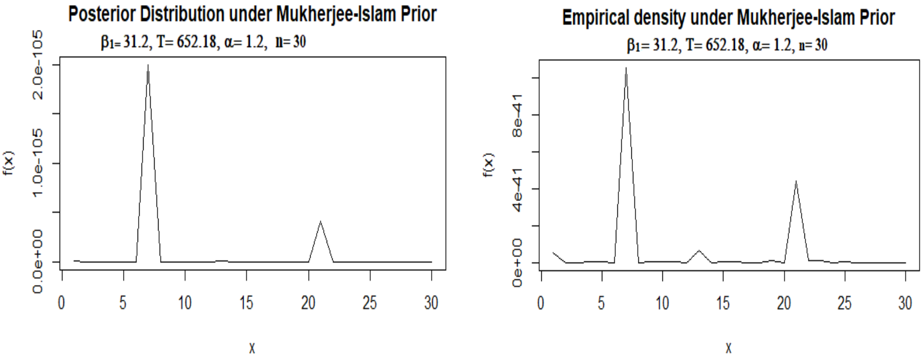


Figure 4. Comparison of Posterior density with Empirical density under Mukherjee-Islam Prior.

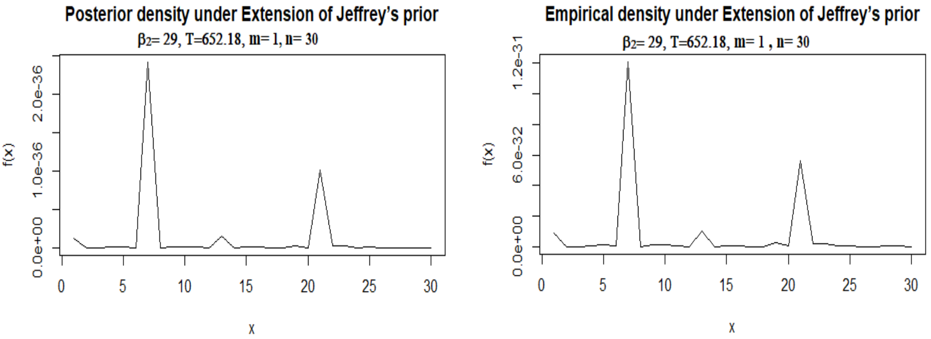


Figure 5. Comparison of Posterior density with Empirical density under Extension of Jeffrey's Prior.

From the above findings of the graph, we can see the posterior density and empirical density under Mukherjee-Islam prior and the extension of Jeffrey's prior are nearly the same.

As seen above, the findings from real life examples are in accordance with those of the simulation study. One can see that in the case of the real data set also Mukherjee-Islam prior results in smaller estimated loss in comparison with the extension of Jeffrey's prior. Even the width of HPD credible interval is smaller in the case of Mukherjee prior. The estimated loss is also smaller in the case of LINEX than SELF irrespective of the prior being used. The findings from the real life example are in accordance with those of the simulation study.

7. Conclusion

Bayes estimates of two inequality indices are obtained in the case of the Dagum distribution, an important income distribution. As seen from the simulation study it is observed that Mukherjee-Islam prior performs better than the extension of Jeffrey's prior in terms of having smaller estimated loss. It is also observed that the LINEX loss function results in smaller loss as compared to SELF for small, medium and large sample sizes irrespective of the choice of prior. One can further see that the expected loss decreases as the sample size increases. The real data set is also in conformity with above results.

Acknowledgement

The authors acknowledge with thanks the valuable suggestions/comments made by the anonymous reviewers, which has led to a considerable improvement of the manuscript.

References

Abdul-Sathar, E. I., Jeevanand, E. S. and Nair, K. R. M., (2009). Bayes estimation of Lorenz curve and Gini-index for classical Pareto distribution in some real data situation, Journal of Applied Statistical Science, 17(2), pp. 315–329.

Al-Kutubi, H. S., Ibrahim, N. A, (2009). Bayes estimator for exponential distribution with extension of Jefferys' prior information, Malays. J. Math. Sci., 3(2), pp. 297–313.

Bansal, P., Arora, S. and Mahajan, K. K., (2011). Testing homogeneity of Gini indices against simple-ordered alternative, Communications in Statistics: Simulation and Computation, 40(2), pp. 185–198.

Berger J., (2006). The case for objective Bayesian analysis, Bayesian Analysis, 1(3), pp. 385–402.

Bhattacharya, S. K., Chaturvedi, A. and Singh, N. K., (1999). Bayesian estimation for the Pareto income distribution, Statistical Papers, 40(3), pp. 247–262.

Bonferroni, C. E., (1930). Statistics Elements Generated, Seber Library, Florence.

Burr, I. W., (1942). Cumulative Frequency Function Annals of Mathematical Statistics, 13, pp. 215–232.

Chotikapanich, D., Griffiths, W. E., (2006). Bayesian Assessment of Lorenz and Stochastic Dominance in Income Distributions, in J. Creedy and G. Kalb (eds.) Research on Economic Inequality, 13: Dynamics of Inequality and Poverty, pp. 297–321, Elsevier, Amsterdam.

Dagum, C., (1975). A Model of Income Distribution and the Conditions of Existence of Moments of Finite Order, Proceedings of the 40th session of the International Statistical Institute, XLVI, Book 3, Warsaw, pp. 196–202.

Dagum, C., (1977a). A New Model of Income Distribution, the Lorenz Curve and the Gini Concentration Ratio, 30(3), pp. 413–436.

Daren S, Starnes, Daniel S. Yates and David S. Moore., (2014). The Practice of Statistics for AP*4th edition (2014), Macmillan Learning

De Vergottini., (1940). On the meaning of some concentration indices, Journal of Economics and Annuals of Economic, 11, pp. 317–347.

Dixon, P. M., Weiner, J., Mitchell-Olds, T. and Woodley, R., (1987). Bootstrapping the Gini coefficient of inequality, Ecology, 68(5), pp. 1548–1561.

Foster, J., Greer, J. and Thorbecke, E., (1984). A class of decomposable poverty measures, Econometrica, 52(3), 761–766.

Gini, C., (1912). Variability and Mutabiltity, C. Cuppini, Bologna, Italy.

Giorgi, G. M., and Crescenzi, M., (2001). A proposal of poverty measures based on the Bonferroni inequality index, Metron, 59 (3–4), pp. 3–16.

Kass, R., Wasserman, L., (1996). The selection of prior distributions by formal rules, Journal of American Statistical Association, 91(431), pp. 1343–1370.

Kaur, K., Arora, S. and Mahajan, K. K., (2015). Bayesian Estimation of Inequality and Poverty Indices in Case of Pareto Distribution using different priors under LINEX loss function, Journal of advances in statistics, Vol. 2015.

Kleiber, C., Kotz, S., (2003). Statistical Size Distributions in Economics and Actuarial Sciences, John Wiley, Hoboken, NJ.

Layla M. Nassir and Nathier A. Ibrahim, (2020). The Bayesian Estimator for Probabilistic Dagum Distribution, International Journal of Innovation, Creativity and Change, 12(5).

Moothathu, T. S., (1985). Sampling distributions of Lorenz curve and Gini index of the Pareto distribution, Sankhya (Statistics), Series B, 47(2), pp. 247–258.

Naqash, S., Ahmad S. P. and Ahmed, A., (2017). Bayesian analysis of Dagum Distribution, Journal of Reliability and Statistical Studies, 10, pp. 123–136.

Sanku Dey, Bander Al-Zahrani and Samerah Basloom, (2017). Dagum Distribution: Properties and Different Methods of Estimation, International Journal of Statistics and Probability; Vol. 6, No. 2; March 2017.

Sathar, E. I., Jeevanand, E. S. and Nair, K. R. M., (2005). Bayesian estimation of Lorenz curve, Gini-index and variance of logarithms in a Pareto distribution, Statistica, 65(2), pp. 193–205.

Sen, P. K., (1988). The harmonic Gini coefficient and affluence indexes, Mathematical Social Sciences, 16(1), pp. 65–76.

Tarsitano, A., (1990). The Bonferroni Index of Income Inequality. In Dagum C., Zenga M. (eds) Income and Wealth Distribution, Inequality and Poverty, Studies in Contemporary Economics, Springer, Berlin, Heidelberg, pp. 228–242.

Varian, H. R., (1975). A Bayesian approach to real estate assessment, in Studies in Bayesian Econometrics and Statistics in Honor of Leonard J. Savage, S. E. Fienberg and A. Zellner, Eds., 195–208, North-Holland, Amsterdam, The Netherlands.

Zellner, A., (1986). Bayesian estimation and prediction using asymmetric loss functions, Journal of the American Statistical Association, 81(394), pp. 446–451.

Zenga, M. M., (2013). Decomposition by Sources of the Gini, Bonferroni and Zenga Inequality Indexes., Statistica & Applicazioni, 11(2), pp. 133–161.

STATISTICS IN TRANSITION new series, June 2022
Vol. 23 No. 2, pp. 69–87, DOI 10.2478/stattrans-2022-0017
Received – 28.02.2020; accepted – 26.11.2021

ARFURIMA models: simulations of their properties and application

Sanusi Alhaji Jibrin[1], **Rosmanjawati Abdul Rahman**[2]

ABSTRACT

This article defines the Autoregressive Fractional Unit Root Integrated Moving Average (ARFURIMA) model for modelling ILM time series with fractional difference value in the interval of $1 < d < 2$. The performance of the ARFURIMA model is examined through a Monte Carlo simulation. Also, some applications were presented using the energy series, bitcoin exchange rates and some financial data to compare the performance of the ARFURIMA and the Semiparametric Fractional Autoregressive Moving Average (SEMIFARMA) models. Findings showed that the ARFURIMA outperformed the SEMIFARMA model. The study's conclusion provides another perspective in analysing large time series data for modelling and forecasting, and the findings suggest that the ARFURIMA model should be applied if the studied data show a type of ILM process with a degree of fractional difference in the interval of $1 < d < 2$.

Key words: interminable long memory, autocorrelation, fractional unit root integrated series, fractional unit root differencing, ARFURIMA model.

1. Introduction

Long Memory (LM) is a statistical property that may arise in time series data. The information of its occurrence in financial and economic variables can be exploited by investors and policy makers to predict equity prices, quantify market's risk, inflation and economic growth of the country. There are past and recent works that have discovered degree of long memory also called fractional differencing value in the interval of $0 < d < 1$ (see Granger and Joyeux (1980), Hosking (1981), Ballie et al., (2014), Boubaker et al. (2016) and Pumi et al., (2019)).

[1] Department of Statistics, Kano University of Science and Technology, Wudil. Nigeria. E-mail: sanusijibrin46@gmail.com. ORCID: https://orcid.org/0000-0003-1276-7965.

[2] School of Mathematical Sciences, Universiti Sains Malaysia. Malaysia. E-mail: rosmanjawati@usm.my. ORCID: https://orcid.org/0000-0002-5384-0674.

Furthermore, according to Rahman and Jibrin (2018), when the ACF of time series exhibits decays more slowly and have fractional difference value in the interval of $1 < d < 2$, the series are said to be having an Interminable LM (ILM) process. In view of this, we need a family of models that can simulate a very strong dependent relationship (autocorrelation) between distance observations, at the same time being flexible enough to model both the integrated, $I(1)$ and Fractional Unit Root Integrated (FURI), I$(1 < d < 2)$ process.

The Autoregressive Fractional Unit Root Integrated Moving Average (ARFURIMA) model, suggested by Rahman and Jibrin (2019), provides a method for modelling FURI time series with fractional difference value in the interval of $1 < d < 2$. Therefore, in this paper, some of the basic properties of the ARFURIMA model were derived and presented as follows. In Section 2 we state the basic properties of the ARFURIMA*(0,d,0)* process. This is followed by the general ARFURIMA(p,d,q) family, its properties, and some special cases of ARFURIMA(p,d,q) in Section 3. A short introduction of SEMIFARMA(p,d,q) model is given in Section 4. Then some simulations by using the ARFURIMA are carried out in Section 5 to assess its properties. Finally, its applications by using the energy series, Bitcoin exchange rate and financial data in are presented in Section 6.

2. Methodology

Dolado and Marmol (1997) named the Data Generating Process (DGP) of y_t to be Nonstationary Fractionally Integrated (NFI) process and defined it as:

$$(1-L)^d y_t = \varepsilon_t, \tag{1}$$

where d$\geq \frac{1}{2}$, $\varepsilon_t \sim iid(0, \sigma^2)$ and d is decomposed as $d = \propto + \delta$, where $\propto = 1,2,3, \dots \dots \dots$ and $|\delta| < \frac{1}{2}$. In this case, d can be in the range of $1 < d < \infty$ while most time series usually have a fractional difference value, d, in the interval of $1 < d < 2$ (see Gil-Alana et al., (2018) and Sabzikar et al. (2019)). However, Hurvich and Chen (2000) and Erfani and Samimi (2009) have highlighted the repercussion of over-differencing including loss of information, negative values of differenced series, $d \leq -0.5$ and estimation of complex models. In view of these, Rahman and Jibrin (2019) resolved that the possible highest value of fractional difference is in the interval of $1 < d < 2$. Also, this type of time series exhibits very slow decaying ACF, which is slower than usual decay seen in the literature of time series and LM analysis. Having said this, Rahman and Jibrin (2019) named the DGP of y_t to be the FURI process and defined its operator as:

$$\{(1-L)(1-d^*)(1+L)\} y_t = \varepsilon_t, \tag{2}$$

where $d^* = d - 1, 0 < d^* < 1$ and $1 < d < 2$.

Details of the derivation and its R algorithm can be found in Rahman and Jibrin (2019).

2.1. The ARFURIMA(p,d,q) models

In a similar way on how Granger and Joyeux (1980) and Hosking (1981) introduced ARFIMA model due to FI(d) process, Porte-Hudak (1990) introduced SARFIMA model due to seasonal FI(d) process and ARTFIMA model of Meerschaert et al. (2014) was introduced due to tempered FI(d) process. Rahman and Jibrin (2019) introduced the ARFURIMA model due to the FURI(d) processes. In order to obtain the ARFURIMA model, the lag representation of the proposed non power operator is incorporated as:

$$\varphi(L)\{(1-L)\big(1-d^*(1+L)\big)\}Y_t = \theta(L)\varepsilon_t, \tag{3}$$

where $\varphi(L)$ and $\theta(L)$ are stationary and invertible. L is the backward shift operator, ε_t represents a white noise process and $\{(1-L)\big(1-d^*(1+L)\big)\}$ is the proposed non-power operator. The operator fractionally differenced is a process that exhibits a very slow decaying (unusual decay) ACF. Here, $d^* = d-1$ such that $0 < d^* < 1$ and $1 < d < 2$ and both d^* and d are the LM and ILM parameters respectively. The identification of the ARFURIMA(*p,d,q*) model followed the Box and Jenkins approach and was discussed in detail in Rahman and Jibrin (2019).

2.2. The ARFURIMA(*0,d,0*) process

The ARFURIMA(*0,d,0*) process was defined to be a discrete time series $\{Y_t\}$, which was presented as:

$$\big(\nabla - d^*\nabla(1+L)\big)Y_t = \varepsilon_t \tag{4}$$

where $\nabla = (1-L)$, L is the backward-shift operator and Y_t represents the FURI series. The fractional differencing parameter d was estimated by applying GPH (1983) semi-parametric method defined by:

$$\ln\{I(\varphi_j)\} = a - d\ln\left\{4\sin^2\left(\frac{\varphi_j}{2}\right)\right\} + \varepsilon_t, \tag{5}$$

where $j = 1, \dots, n$, and $I(\varphi_j) = \left(\frac{1}{2\pi}\right)\left|\sum_{i=1}^{T} y_t exp(i\varphi_j t)\right|^2$was the periodogram at the frequency $\varphi_j = \frac{2\pi j}{T}$. The following theorems were some of the derived properties of ε_t.

Theorem 1

For $0 < d^* < 1$ such that $d^* = d-1$ and $1 < d < 2$, $\{Y_t\}$ is

a. stationary, causal and has infinite Moving Average (MA) function written as

$$Y_t = \psi(L)\varepsilon_t = \sum_{n=0}^{\infty}\psi_n L^n \varepsilon_t = \sum_{n=0}^{\infty}\psi_n \varepsilon_{t-n}. \tag{6}$$

The fractional differencing operator $(1-L)^d$, $0 < d < 1$ in Granger and Joyeux (1980), Hosking (1981), Dolado and Marmol (1997), Meerschaert et al., (2014), Boubaker et al., (2016) and Pumi et al., (2019), is defined as an infinite binomial series expansion in powers of the backward-shift operator

$$(1-L)^d Y_t = \sum_{i=0}^{\infty} \psi_i L^i Y_t = \sum_{i=0}^{\infty} \psi_i Y_{t-i}, \tag{7}$$

where the coefficient, ψ_i, is expanded by:

$$\psi_i = \prod_{k=1}^{i} \frac{k+d-1}{k} = \frac{\Gamma \mathrm{i} + \mathrm{d}}{\Gamma \mathrm{d}\, \Gamma \mathrm{i} + 1}, i = 1,2, \dots \tag{8}$$

However, ψ_n can be expanded as:

$$\psi_n = \left((L^n - L^{1+n}) - d^*(L^n - L^{2+n})\right). \tag{9}$$

So,

$$\psi_n \sim (1-L)(1-d^*(1+L), \tag{10}$$

and hence, (6) can be re-written as:

$$Y_t = \psi(L)\varepsilon_t = \sum_{n=0}^{\infty} \psi_n L^n \varepsilon_t = \sum_{n=0}^{\infty} \left((L^n - L^{1+n}) - d^*(L^n - \ L^{2+n})\right)\varepsilon_t, \tag{11}$$

where L, ε_t and $d^*$is as defined in (4). Also,

$$\textstyle\sum_{n=0}^{\infty} |\psi_n| < \infty \tag{12}$$

satisfied the causality condition, where it stated that $\{Y_t\}$ depended on past residuals ε_t and the dependency was gradually decreasing asymptotically for a long time.

Proof.

Using $Y_t = \psi(L)\varepsilon_t$, we have $\psi(L) = \left((L^n - L^{1+n}) - d^*(L^n - L^{2+n})\right)^{-1}$. When $1 < d < 2$, the expansion of $\psi(L)$ converged for $|L| \leq 1$ and so $\{Y_t\}$ is stationary. The expansion of $(L^n - L^{1+n}) - d^*(L^n - L^{2+n})$ resulted in (10) when $n \to \infty$, that was $(L^n - L^{1+n}) - d^*(L^n - L^{2+n}) \sim \left((1-L)(1-d^*(1+L)\right)$.

b. $\{Y_t\}$ is invertible and has infinite AR function writen as:

$$\Phi(L)Y_t = \textstyle\sum_{n=0}^{\infty} \Phi_n L^n Y_t = \sum_{n=0}^{\infty} \Phi_n Y_{t-n} = \varepsilon_t, \tag{13}$$

where Φ_n is defined similar to ψ_n in (9) with the invertibility

$$\textstyle\sum_{n=0}^{\infty} |\Phi_n| < \infty \tag{14}$$

Proof.

The proof was similar to (a).

c. The spectral density function of $\{Y_t\}$ is

$$f(\lambda) = \sum_{k=0}^{\infty} e^{i\lambda k}\,\gamma(k) \tag{15}$$

where $\gamma(k)$ was the autocovariance function of $\{Y_t\}$. $f_{(\lambda)} \sim |\lambda|^{-d} C_f$ described the pole at the zero frequency of the spectral density as $C_f > 0$ and $1 < d < 2$.

d. The autocovariance function of $\{Y_t\}$ is

$$\gamma(k) = E(Y_t Y_{t-k}) = \frac{d^*\gamma_{k-2} - \gamma_{k-1}}{d^*}, \tag{16}$$

where $\gamma_{(k)} \sim K^{d-1}$, $1 < d < 2$ described the very slow decay in the autocorrelation function of Y_t as $k \to \infty$.

Proof.

Using

$$\begin{aligned}\gamma_k = E(Y_t Y_{t-k}) &= Y_{t-k}\big(Y_t - Y_{t-1} - d^*(Y_t - Y_{t-2})\big)\\ &= \gamma_k - \gamma_{k-1} - d^*(\gamma_k - \gamma_{k-2})\\ &= \gamma_k(1 - d^*) - \gamma_{k-1} + d^*\gamma_{k-2},\end{aligned}$$

and re-arranging the equation as $-\gamma_k(1 - d^*) = d^*\gamma_{k-2} - \gamma_{k-1}$, we get $\gamma_k d^* = d^*\gamma_{k-2} - \gamma_{k-1}$ and therefore,

$$\gamma_k = \frac{d^*\gamma_{k-2} - \gamma_{k-1}}{d^*}.$$

e. The autocorrelation function of $\{Y_t\}$ is

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{d^*\gamma_{k-2} - \gamma_{k-1}}{d^*\gamma_2 - \gamma_1} \tag{17}$$

where d^* is as defined in (4).

Proof.

Using

$$\begin{aligned}\gamma_0 = E(Y_t Y_t) &= Y_t\big(Y_t - Y_{t-1} - d^*(Y_t - Y_{t-2})\big)\\ &= \gamma_0 - \gamma_1 - d^*(\gamma_0 - \gamma_2)\\ &= \gamma_0(1 - d^*) - \gamma_1 + d^*\gamma_2.\end{aligned}$$

Re-arranging the equation will get $\gamma_0 - \gamma_0(1 - d^*) = d^*\gamma_2 - \gamma_1$.
Therefore,

$$\gamma_0 = \frac{d^*\gamma_2 - \gamma_1}{d^*}. \tag{18}$$

Therefore, (17) is resulted from substituting (16) and (18).

2.3. The Nonstationary ARFURIMA(*p,d,q*) model

Consider the stationary ARFURIMA(*p,d,q*) model, written as:

$$\varphi(L)\left((1-L)\left(1-d^{*}(1+L)\right)\right)(Y_t-\mu)=\theta(L)\varepsilon_t, \tag{19}$$

where $\varphi(L)=1-\varphi_1 L-\varphi_2 L^2-\cdots-\varphi_p L^p$ and $\theta(L)=1-\theta_1 L-\theta_2 L^2-\cdots-\theta_q L^q$. For (19) to be stationary and invertible, each zero of $\varphi(L)$ and $\theta(L)$ must be outside the unit circle respectively. Noticed that when $d^{*}=0$, the ARFURIMA is reduced to the ARIMA model.

Theorem 2

The ARFURIMA(*p,d,q*) model as mentioned by (19) is

a. stationary if

$$Y_t=\left((1-L)\left(1-d^{*}(1+L)\right)\right)^{-1}\varphi(L)^{-1}\theta(L)\varepsilon_t. \tag{20}$$

b. It is invertible when

$$\varphi(L)\theta(L)^{-1}\left((1-L)\left(1-d^{*}(1+L)\right)\right)Y_t=\varepsilon_t. \tag{21}$$

c. Its spectral density function is given by

$$f(\omega)=\frac{\sigma_e^2}{2\pi}\frac{\left|\theta(e^{-i\omega})\right|^2}{\left|\varphi(e^{-i\omega})\right|^2}\left((1-L)\left(1-d^{*}(1+L)\right)\right)\left(e^{i2\pi\omega}\right). \tag{22}$$

d. Then, the non-stationary ARFURIMA model can be represented as

$$\varphi(L)\left((1-L)\left(1-d^{*}(1+L)\right)\right)Y_t=\upsilon+\theta(L)\varepsilon_t \tag{23}$$

where υ is a constant.

2.4. Maximum Likelihood Estimation Method for ILM Model and Its Hybrid

Consider series $Y=(y_1,\dots,y_t)'$, where $y_1,\dots,y_t$was the FURI process. In order to obtain the estimates of the ARFURIMA, the series $Y_t$was filtered by the non-power operator, ε_t, where

$$\varepsilon_t=\left\{\left\{(1-L)\left(1-d^{*}(1+L)\right)\right\}Y_t\right\}. \tag{24}$$

Following Kang and Yoon (2013), ε_t in equation (24) was assumed to be normal. The parameters of the ARFURIMA (*p,d,q*) model were estimated by using the Maximum Likelihood Estimation (MLE) and nonlinear optimization procedures. The maximized of the logarithm of the normal likelihood function was given in equation (25).

$$\ln\{L(\mu,d,\varphi,\theta,\sigma^2)\}=-\tfrac{n}{2}\ln(2\pi)-\tfrac{1}{2}\ln|\Sigma|-\tfrac{1}{2}Y'\Sigma^{-1}Y, \tag{25}$$

where n is the number of observations, Σ represents the n x n covariance matrix of Y dependent on $\mu, d, \varphi, \theta, \sigma^2$ and $|\Sigma|$is the determinant of Σ.

2.5. The SEMIFARMA*(p,d,q)* model

With reference to the Semiparametric Fractional Autoregressive Moving Average (SEMIFARMA) model by Beran and Feng (2002), we used the definition of the SEMIFARMA*(p,d,q)* model as:

$$\varphi(L)(1-L)^d\{(1-L)^m Y_t - \mu\} = \theta(L)\varepsilon_t \tag{26}$$

where μ is the mean of Y_t, $\varphi(L) = 1 - \varphi_1 L - \varphi_2 L^2 - \cdots \varphi_p L^p$, $\theta(L) = 1 - \theta_1 L - \theta_2 L^2 - \cdots \theta_q L^q$, m and d defined as $\delta = m + d$ such that $d \in (-0.5, 0.5)$ and $m \in \{0,1\}$.

3. Simulation properties of ARFURIMA (p,d,q) model

This section discusses the simulation to assess the ILM, large sample, conceptual and unbiased properties of the ARFURIMA models. The simulated models are summarized in Table 1.

Table 1. Different models with their different selection of d,$\varphi_1\theta_1$ and n

Model	d	φ_1	θ_1	Sample size, n
ARFURIMA(1,d,0)	1.1	0.5 to 0.9	-	6000
ARIMA(1,1,0)	1		-	6000
ARFURIMA(1,d,0)	1.1,1.5,1.9	-0.9,0.7,0,0.7,0.9	-	6000
ARFURIMA(1,d,0)			-	375
ARFURIMA(1,d,1)		0.4	0.6	375, 750, 1500, 3000, 6000

Referring to Figure 1, all the ACF indicate a very strong hyperbolic decay, implying evidence of ILM. Therefore, on average, all the simulated series are not stationary. Also, the degree of dependency between observations may produce fractional differencing value in interval of $1 < d < 2$.

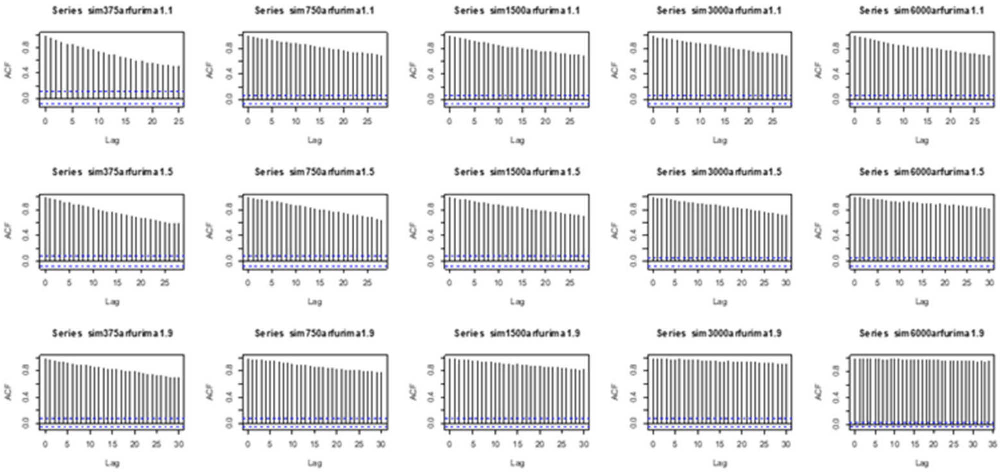


Figure 1. ACF for simulated series using ARFURIMA *(1,d,1)* based on $\varphi = 0.4$, $\theta = 0.6$, $d = \{1.1, 1.5, 1.9\}$ and sample size $n = \{375, 750, 1500, 3000, 6000\}$.

Table 2. Autocorrelation of ARFURIMA (1,1.1,0) and ARIMA (1,1,0) process for different values of φ with n=6000

	ρ_k **of ARFURIMA(1, 1.1, 0)**					ρ_k **of ARIMA(1, 1, 0)**				
k	$\varphi = 0.5$	$\varphi = 0.6$	$\varphi = 0.7$	$\varphi = 0.8$	$\varphi = 0.9$	$\varphi = 0.5$	$\varphi = 0.6$	$\varphi = 0.7$	$\varphi = 0.8$	$\varphi = 0.9$
1	1	1	1	1	1	0.999	0.999	1	1	1
2	0.999	0.999	1	1	1	0.998	0.998	0.999	1	1
3	0.999	0.999	0.997	0.999	1	0.997	0.997	0.998	0.999	0.999
4	0.999	0.999	0.996	0.999	0.999	0.995	0.995	0.997	0.999	0.999
5	0.999	0.999	0.994	0.999	0.999	0.994	0.994	0.996	0.999	0.998

The autocorrelation values of ARFURIMA (1,d,0) and ARIMA (1,1,0), as shown in Table 2, indicated a perfect relationship and strong dependency between observations. On the average, the dependence degrees captured by ARFURIMA was higher compared to the ARIMA models. Therefore, the simulations have provided adequate explanations about the quality of the proposed ARFURIMA model in simulating ILM and FURI series and thus proved the ILM properties of the model.

Meanwhile, for $k \leq 3$, the autocorrelation values of ARFURIMA (1,1.1,0) when $\varphi_1 = 0.9$, as shown, indicate that both the large theoretical fractional difference and φ parameter value have influenced the degree of dependence among the simulated series. Also, for $k \leq 3$, the autocorrelation values of ARFURIMA (1,d,0) for $d = 1.9$ and $\varphi_1 = 0.9$, as shown in Table 3, was perfect indicating that the large theoretical fractional difference and φ parameter value has influenced the degree of dependence among the simulated series. Meanwhile, by comparing Table 3 and 4, the occurrence of perfect autocorrelations among simulated series with $n = 6000$ was found higher compared to $n = 375$. This implied the existence of large sample size properties of the ARFURIMA model.

Table 3. Autocorrelation of the ARFURIMA (1,d,0) process for various values of d and φ , with n=6000

k	d	$\varphi = -0.9$	$\varphi = -0.7$	$\varphi = 0$	$\varphi = 0.7$	$\varphi = 0.9$
1	1.1	0.999	0.999	0.999	0.999	0.999
2		0.999	0.999	0.999	0.999	0.999
3		0.999	0.999	0.997	0.999	0.999
4		0.998	0.999	0.996	0.999	0.999
5		0.997	0.999	0.994	0.999	0.999
1	1.5	0.999	0.999	0.999	1	1
2		0.999	0.999	0.999	0.999	1
3		0.999	0.999	0.999	0.999	1
4		0.999	0.999	0.999	0.999	0.999
5		0.998	0.999	0.999	0.999	0.999
1	1.9	0.999	0.999	0.999	1	1
2		0.999	0.999	0.999	1	1
3		0.999	0.999	0.999	0.999	1
4		0.998	0.999	0.999	0.999	0.999
5		0.998	0.998	0.999	0.999	0.999

The results of the simulation for ARFURIMA(p,d,q) with various settings mentioned in Table 1 showed that means and variances of all the estimated ARFURIMA models confirmed and supported the assumption that the residuals are normally distributed since all the means are zero with variances in the interval of $0.5 < \sigma_e^2 < 1.2$ specifically for $n \geq 1000$. Again, this proves the large sample and also the conceptual properties of the proposed ARFURIMA model. The authors can be contacted for a complete result of these simulations.

4. The application

This section presents the application of the proposed ARFURIMA model by using data of energy series, bitcoin exchange rates and some financial data.

4.1. Data

The description of nine series of data consisted of energy prices series, bitcoin exchanged rates, a financial index and few currencies exchange rates, which are displayed in Table 4. As shown in Figure 1-3, the time series plots of the studied series exhibited nonlinearity deterministic trends. All the ACF showed a very slow decay in the long term with positive autocorrelations, which provided evidence of the LM process. In view of this, there exists LM in the studied series, and it can be described as an ILM. On average, all the nine series are not stationary.

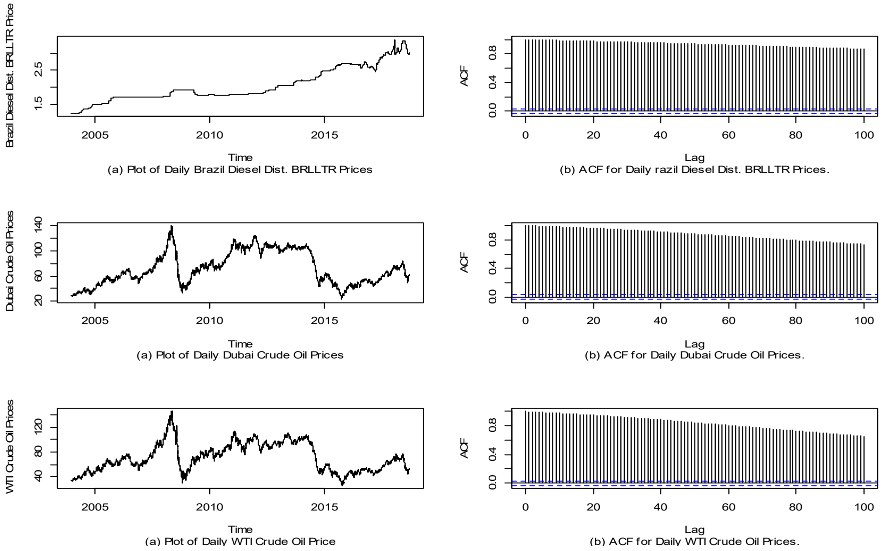
Table 4. Daily Time Series Used for Analysis

S/No.	Type of Data	Abbreviation	Sample Size	Date
1	Brazil Diesel Distributors BRLLTR Prices	BDDP	3915	26/01/04 - 25/01/19
2	Dubai Crude Oil Prices	DBCP	3896	26/01/04 - 25/01/19
3	WTI Crude Oil Prices	WTCP	3896	26/01/04 - 25/01/19
4	Bitcoin to 1000 Euro Exchange Rate	BEUR	1056	15/12/14 - 31/12/18
5	Bitcoin to 1000 Pound Exchange Rate	BPOU	1056	15/12/14 - 31/12/18
6	Bitcoin to 1000 US Dollar Exchange Rate	BDOL	1056	15/12/14 - 31/12/18
7	ATHEX Composite Index	ATIN	7891	03/10/98 - 31/12/18
8	Kuwait Dinar to US Dollar Exchange Rate	KUSD	5196	01/02/99 - 31/12/18
9	Uruguay Peso to UK Pound Exchange Rate	UUKP	5196	01/02/99 - 31/12/18

Source: datastream of Thomson Reuters and Morgan Stanley Capital International (MSCI).

Table 5 presented the descriptive statistics, serial correlation, and normality test for the nine series. It showed that the mean for the Brazil Diesel, Dubai and WTI price each is 2.05, 71.97 and 71.06 respectively. Meanwhile, the standard deviation, which measured the variability or volatility of bitcoin exchange rate for each Euro, British pound sterling, United State (U.S) dollar and Japan Yen is 1.68, 2.32 and 1.50 respectively. The skewness and kurtosis for all the series indicated non-normality. Similarly, the Ljung-Box Q-statistic at lag 50 and Jarque-Bera statistic showed that for all the studied series, the null hypotheses of no serial correlation and normality were rejected at the 0.05 significance level respectively.

In testing and estimating the LM, the bandwidth was chosen between 0 and 1 (GPH, 1983). Hurvich *et al.* (1998) suggests that the best bandwidth (bw) is 0.8, however, Baillie and Morana (2012) uses 0.6 and as for this study, we considered the bw as 0.5, the average of all possible fractional values in interval $0 < d < 1$. For the purpose of comparison, we also produced the estimations based on the bw suggested in Baillie and Morana (2012) and Hurvich et al. (1998) and results were presented in Table 6. The GPH and LWE confirmed the incidence of ILM at level among the studied series. The null hypothesis of no ILM was rejected due to the p-values are less than the significant level of 0.05. Besides, notice that the two estimators, GPH and LWE with bandwidth of 0.5 and 0.8 respectively, produced inconsistent results at level series, with $0 < d < 1$ and $1 < d < 2$. Also, on average, the GPH produced higher fractional differencing values that can adequately eliminate the unwanted noise signals across the nine series.



(a) Plot of Daily Brazil Diesel Dist. BRLLTR Prices

(b) ACF for Daily razil Diesel Dist. BRLLTR Prices.

(a) Plot of Daily Dubai Crude Oil Prices

(b) ACF for Daily Dubai Crude Oil Prices.

(a) Plot of Daily WTI Crude Oil Price

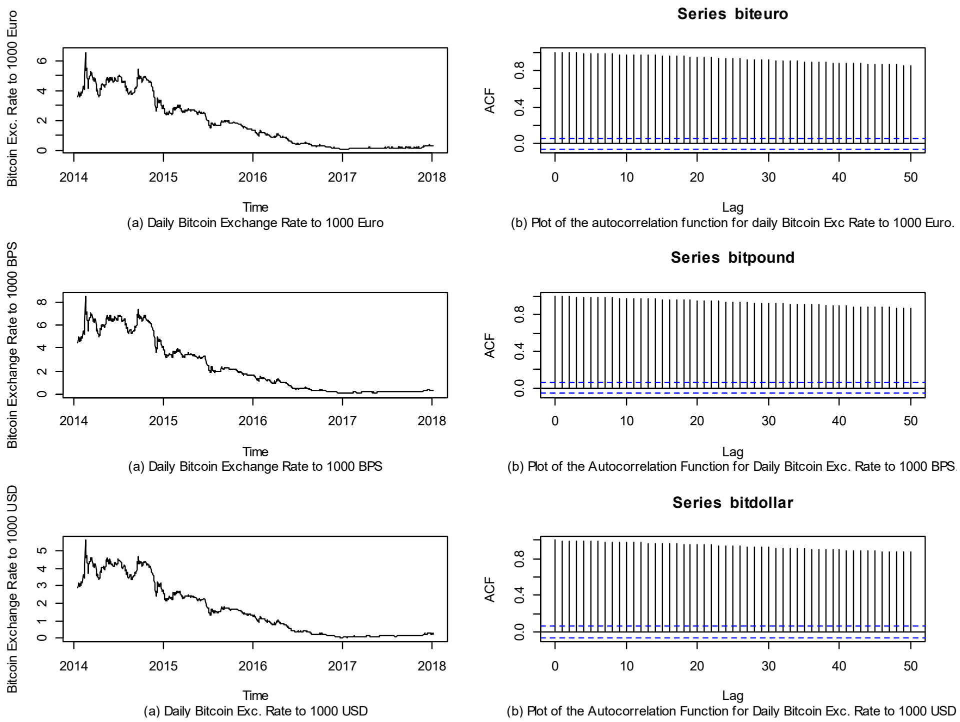
(b) ACF for Daily WTI Crude Oil Prices.

Figure 1. Time Series Plot and ACF for Brazil Diesel, Dubai and WTI Crude Oil Prices (in Dollar per barrel)

Consequently, we suggest that using GPH estimator with a bandwidth equal to 0.5 will produce an adequate fractional differencing value. The adequate fractional differencing value would eliminate the deterministic trend and help in producing a series with less variability. Table 7 presents standard errors of the means of the series. The series were differenced using the $\{(1-L)(1-d^*(1+L))\}Y_t$, $(1-L)^dY_t$ and $(1-L)^{\propto+\delta}Y_t$ operators or fractional filters of Rahman and Jibrin (2019), Granger and Joyuex (1981) and Dolado and Marmol (1997) respectively.

Note that the filters of Dolado and Marmol (1997) and Beran and Feng (2002), shown in the last two columns respectively, produced almost the same standard errors and can be considered to be similar to the current operator used in this study.

A comparison of the standard errors of the mean produced by these three differenced series have shown evidence of a better performance of the fractional unit root difference filter, in which it gave the most minimum standard errors of mean compared to the other two filters. Although the KUSD series indicated that the three filters produced the same standard error, there is a reason to believe that the fractional unit root differenced filter procedure used in this study for fractionally differencing FURI time series was the most appropriate among the three filters because it has reduced the volatility, dependency and linearity structures in all the considered series.



(a) Daily Bitcoin Exchange Rate to 1000 Euro

(b) Plot of the autocorrelation function for daily Bitcoin Exc Rate to 1000 Euro.

(a) Daily Bitcoin Exchange Rate to 1000 BPS

(b) Plot of the Autocorrelation Function for Daily Bitcoin Exc. Rate to 1000 BPS

(a) Daily Bitcoin Exc. Rate to 1000 USD

(b) Plot of the Autocorrelation Function for Daily Bitcoin Exc. Rate to 1000 USD

Figure 2. Time Series Plot and ACF for Daily Bitcoin Exchange Rate to 1000 Euro, USD and USP

4.2. Models identification

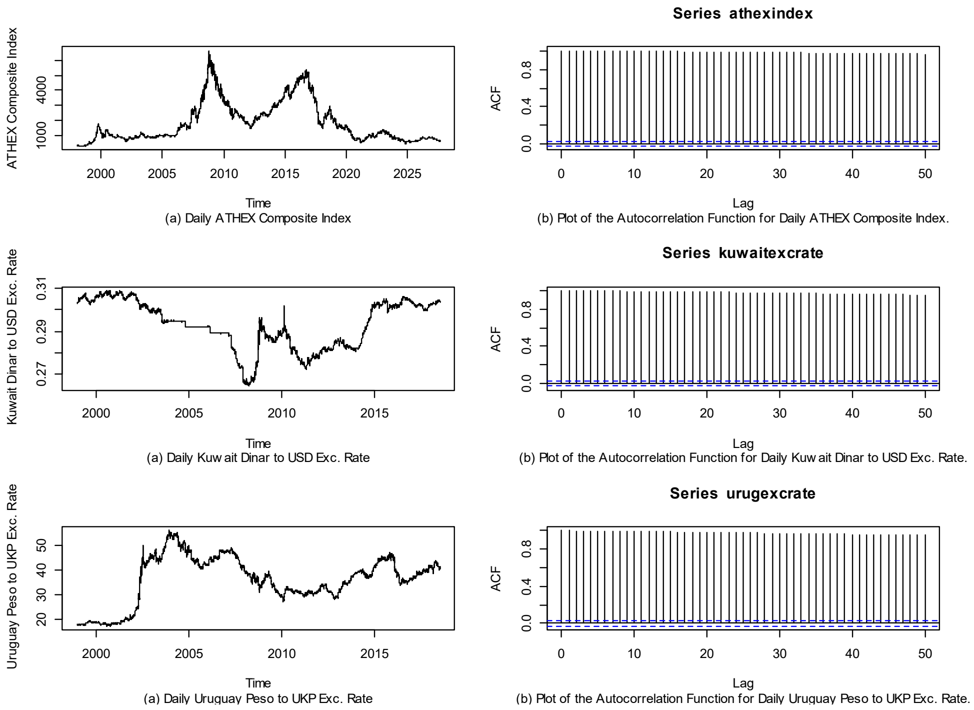
The AIC values for ARFURIMA and SEMIFARMA models were presented in Table 8 and 9 respectively. The best model according to the least AIC value (the values were bold) was identified for each series among the candidate models of ARFURIMA and SEMIFARMA.

4.3. Estimation, diagnostic tests and forecast

In this section, the estimated parameters of the mean model ARFURIMA and SEMIFARMA for each studied series are presented.

4.3.1. Estimation of the ARFURIMA and SEMIFARMA Model

The results of the estimated parameters of both models ARFURIMA(1,d,1) and SEMIFARMA(1,d,1) for each series and their log-likelihood values are reported in Table 10. All the parameters of the ARFURIMA models were found significant due to their minimum standard errors. Furthermore, the ARFURIMA had larger log-likelihood values compared to the SEMIFARMA model, implying that the ARFURIMA have fitted the data well. Also, the proposed non power operator in ARFURIMA had successfully eliminated large inherent noise signals in all the considered series.



(a) Daily ATHEX Composite Index

(b) Plot of the Autocorrelation Function for Daily ATHEX Composite Index.

(a) Daily Kuwait Dinar to USD Exc. Rate

(b) Plot of the Autocorrelation Function for Daily Kuwait Dinar to USD Exc. Rate.

(a) Daily Uruguay Peso to UKP Exc. Rate

(b) Plot of the Autocorrelation Function for Daily Uruguay Peso to UKP Exc. Rate.

Figure 3. Time Series Plot and ACF for Daily ATHEX Index, Kuwait, and Uruguay Exchange Rate to USD and UKP

Table 5. Descriptive Statistics

Variables	Minimum	Maximum	Mean	SD	Skewness	Kurtosis	Q-Test (50)	JB Test
BDDP	1.22	3.40	2.05	0.49	0.76	-0.30	185894.69***	387.62
DBCP	22.79	140.56	71.97	26.17	0.31	-1.08	177008.13***	250.34
WTCP	26.19	145.31	71.06	22.88	0.38	-0.66	170140.41***	162.49
BEUR	0.06	6.58	1.78	1.68	0.67	-0.93	46627.64***	117.96
BPOU	0.07	8.51	2.34	2.32	0.75	-0.89	47436.33***	132.84
BDOL	0.05	5.59	1.60	1.50	0.65	-0.98	47423.24***	115.83
ATIN	269.45	6633.90	1825.10	1334	1.15	0.40	382594.01***	1793.57
KUSD	0.26	0.31	0.29	0.01	-0.48	-0.70	250067.39***	310.36
UUKP	17.36	56.03	36.10	9.71	-0.43	-0.59	248247.15***	237.57

Notes: SD=Standard Deviation, the Jarque–Bera test corresponds to the test statistic for the null hypothesis of normality in the distribution of sample data. The Ljung–Box statistic, Q(n), check for serial correlation of the series up to the nth order.

Table 6. Tests and Estimation of ILM and LM

	Average bw, bw=0.5	BM (2012), bw=0.6	H (1998), bw=0.8	Average of bw, bw=0.5	BM (2012), bw=0.6	H (1998), bw=0.8
Energy Data	LWE			GPH		
BDDP	1.1202(0.000)	1.1202(0.000)	0.9945(0.000)	1.1182(0.000)	1.1182(0.000)	1.0077(0.000)
DBCP	1.2667(0.000)	1.0885(0.000)	1.0170(0.000)	1.2613(0.000)	1.0983(0.000)	1.0469(0.000)
WTCP	1.2348(0.000)	1.0635(0.000)	1.0095(0.000)	1.2294(0.000)	1.0582(0.000)	1.0220(0.000)
BEUR	1.1463(0.000)	0.9809(0.000)	1.0011(0.000)	1.2872(0.000)	1.0002(0.000)	1.0257(0.000)
BPOU	1.1756(0.000)	0.9940(0.000)	1.0061(0.000)	1.3156(0.000)	1.0229(0.000)	1.0187(0.000)
BDOL	1.1832(0.000)	1.0019(0.000)	1.0074(0.000)	1.3150(0.000)	1.0275(0.000)	1.0215(0.000)
ATIN	1.1695(0.000)	1.0750(0.000)	1.0775(0.000)	1.1738(0.000)	1.0592(0.000)	1.0229(0.000)
KUSD	1.2580(0.000)	1.1486(0.000)	0.9875(0.000)	1.2523(0.000)	1.1376(0.000)	0.9858(0.000)
UUKP	1.1184(0.000)	1.0682(0.000)	0.9816(0.000)	1.1020(0.000)	1.1112(0.000)	0.9751(0.000)

Note: p-values are in parenthesis (.), bw denotes the bandwidth for the LWE and GPH tests. BM is Bailie and Morana, meanwhile H is Hurvich)

Table 7. Standard Errors of the Mean for Fractional Unit Root and Fractional Differenced of the Studied Series

Variables	$\{(1-L)(1-d^{*}(1+L))\}Y_t$	$(1-L)^{d}Y_t$	$(1-L)^{\propto+\delta}Y_t$
BDDP	0.00015	0.00027	0.00028
DBCP	0.01926	0.02633	0.02842
WTCP	0.02036	0.02657	0.02830
BEUR	0.00226	0.00343	0.00366
BPOU	0.00291	0.00442	0.00476
BDOL	0.00191	0.00284	0.00306
ATIN	0.36372	0.47519	0.48515
KUSD	0.00001	0.00001	0.00001
UUKP	0.00405	0.00573	0.00582

Table 8. AIC Values for ARFURIMA(*p,d,q*) Models

Variables	ARFURIMA(1,d,0)	ARFURIMA(1,d,1)	ARFURIMA(2,d,0)
BDDP	-25616.84	**-25618.43**	-25616.48
DBCP	11716.44	**11584.88**	11607.83
WTCP	12444.65	12379.17	**12378.06**
BEUR	-2618.18	**-2639.72**	-2627.75
BPOU	-2111.12	**-2150.71**	-2130.78
BDOL	-2998.07	**-3038.86**	-3019.29
ATIN	77220.03	**77208.76**	77212.13
KUSD	-66159.66	-66206.75	**-66369.59**
UUKP	1915.799	**1905.627**	1916.047

Table 9. AIC Values for SEMIFARMA(*p,d,q*) Models

Variables	SEMIFARMA(1,d,0)	SEMIFARMA(1,d,1)	SEMIFARMA(2,d,0)
BDDP	-24550.76	**-24573.65**	-24564.02
DBCP	14170.81	**14007.84**	14071.67
WTCP	14589.55	**14473.07**	14507.58
BEUR	-1870.80	**-1918.11**	-1880.35
BPOU	-1283.53	**-1345.28**	-1297.33
BDOL	-2173.91	**-2237.29**	-2189.36
ATIN	80076.43	**79966.76**	80018.9
KUSD	-65401.79	**-65787.28**	-65691.27
UUKP	3066.39	**3017.78**	3050.02

Table 10. Estimation of ARFURIMA(*p,q*) and SEMIFARMA(*p,q*) with their Log-likelihood Values

Variables	Candidate Models	φ_1	φ_2	θ_1	Log-Likelihood
BDDP	ARFURIMA(1,d,1)	0.09(0.0003)	-------	0.27(0.0005)	12813.22
	SEMIFARMA(1,d,1)	0.41(0.0821)	-------	-0.54(0.0757)	12290.81
DBCP	ARFURIMA(1,d,1)	-0.09(0.0003)	-------	0.53(0.0007)	-5788.442
	SEMIFARMA(1,d,1)	0.24(0.0412)	-------	-0.62(0.0345)	-7031.430
WTCP	ARFURIMA(2,d,0)	-0.37(0.0006)	-0.14(0.0004)	-------	-6185.031
	SEMIFARMA(1,d,1)	0.40(0.0532)	-------	-0.68(0.0443)	−7263.33
BEUR	ARFURIMA(1,d,1)	0.13(0.0004)	-------	0.48(0.0007)	1323.859
	SEMIFARMA(1,d,1)	0.65(0.0568)	-------	-0.87(0.0391)	963.440
BPOU	ARFURIMA(1,d,1)	0.13(0.0003)	-------	0.55(0.0007)	1079.354
	SEMIFARMA(1,d,1)	0.63(0.0531)	-------	-0.88(0.0352)	676.159

Table 10. Estimation of ARFURIMA(p,q) and SEMIFARMA(p,q) with their Log-likelihood Values (cont.)

Variables	Candidate Models	φ_1	φ_2	θ_1	Log-Likelihood
BDOL	ARFURIMA(1,d,1)	0.12(0.0004)	-------	0.54(0.0007)	1523.427
	SEMIFARMA(1,d,1)	0.61(0.0562)	-------	-0.87(0.0383)	1121.097
ATIN	ARFURIMA(1,d,1)	0.59(0.0008)	-------	0.69(0.0008)	-38600.380
	SEMIFARMA(1,d,1)	0.84(0.0202)	-------	-0.90(0.0157)	−39979.420
KUSD	ARFURIMA(2,d,0)	-0.12(0.0003)	0.2(0.0004)	-------	33188.790
	SEMIFARMA(1,d,1)	0.12(0.0265)	-------	-0.62(0.0213)	32897.460
UUKP	ARFURIMA(1,d,1)	0.43(0.0007)	-------	0.59(0.0007)	-948.814
	SEMIFARMA(1,d,1)	0.68(0.0521)	-------	-0.77(0.0454)	−1504.901

Note: standard errors are in (·) except in the second column

4.3.2. The Diagnostic Test

Tests based on the residual’s normality test of Jarque-Bera, the Ljung-Box and ARCH-LM tests were applied, and the results showed evidence of non-normality, serial correlation, and heteroscedasticity in both the ARFURIMA and SEMIFARMA models due to large statistic and p-values less than 0.05. However, a comparison of the statistics from the three tests showed that the ARFURIMA model performed better due to larger test statistic for each test. A table of this analysis can be provided by the author on request.

4.3.3. Forecasting accuracy

The Mean Absolute Error (MAE), Mean Percentage Error (MPE) and Mean Absolute Percentage Error (MAPE) were used to evaluate the forecast performance. The results are presented in Table 11 and showed that the ARFURIMA model produced a better forecast with minimum MAE, MPE and MAPE.

Table 11. Forecasts Accuracy Values of ARFURIMA and SEMIFARMA Model

Variables	Candidate Models	MAE	MPE	MAPE
BDDP	ARFURIMA(1,d,1)	0.001542	0.021271	0.063209
	SEMIFARMA(1,d,1)	0.001953	123.7131	149.1917
DBCP	ARFURIMA(1,d,1)	1.045593	-0.028338	1.598621
	SEMIFARMA(1,d,1)	1.046859	59.90192	193.3132
WTCP	ARFURIMA(2,d,0)	1.079942	-0.043681	1.625366
	SEMIFARMA(1,d,1)	1.084338	117.8703	221.0361
BEUR	ARFURIMA(1,d,1)	0.040129	1.028172	3.097527
	SEMIFARMA(1,d,1)	0.049917	108.1391	397.9863

Table 11. Forecasts Accuracy Values of ARFURIMA and SEMIFARMA Model (cont.)

Variables	Candidate Models	MAE	MPE	MAPE
BPOU	ARFURIMA(1,d,1)	0.032087	1.044535	3.116316
	SEMIFARMA(1,d,1)	0.054838	49.77582	222.7510
BDOL	ARFURIMA(1,d,1)	0.006603	-0.923786	3.006184
	SEMIFARMA(1,d,1)	0.036540	89.33419	188.2536
ATIN	ARFURIMA(1,d,1)	11.91682	-0.039669	1.227970
	SEMIFARMA(1,d,1)	19.86253	72.46209	153.1615
KUSD	ARFURIMA(2,d,0)	0.000223	-0.000138	0.076933
	SEMIFARMA(1,d,1)	0.137226	800.8632	1001.293
UUKP	ARFURIMA(1,d,1)	0.007133	0.002391	0.568356
	SEMIFARMA(1,d,1)	0.296838	122.9570	154.4095

Similarly, Diebold and Mariano (1995) accuracy tests indicated that the ARFURIMA was better in forecasting all the series at 0.05 level of significance. A table of this analysis can be provided by the author on request.

5. Conclusions

In this work, we defined the family of the ARFURIMA (p,d,q) model and the stationarity, invertibility and basic properties of the models were derived and presented. The presented simulations studies confirmed superiority of the ARFURIMA over the ARIMA in simulating nonstationary and the FURI series and thus proved the ILM properties of the ARFURIMA model and its large sample properties too. Besides, some applications of the model were presented and further confirmed a better fit of the ARFURIMA compared to the SEMIFARMA model.

In conclusion, this study provided another perspective in analysing large time series data for modelling and forecasting, and the findings suggested that the ARFURIMA model should be considered if the data show a type of the ILM process with a degree of fractional difference in the interval of $1 < d < 2$.

Acknowledgement

This work was sponsored by Universiti Sains Malaysia, under Bridging Grant 304/PMATHS/6316286.

References

Baillie, R. T., Morana, C., (2012). Adaptive ARFIMA models with application to inflation, *Economic Modelling*, Vol. 29, pp. 2451–2459.

Beran, J., Feng, Y., (2002). SEMIFAR Models-a semiparametric approach to modelling trends, long-range dependence and nonstationarity, *Computational Statistics and Data analysis*, Vol. 40, pp. 393–419.

Boubaker, H., Canarella, G., Gupta, R. and Miller, M. S., (2016). Time-varying persistence of inflation: evidence from a wavelet-based approach. Working Paper Series, *University of Connecticut, Department of Economics.*

Diebold, F. X., Mariano, R. S. (1995). Comparing predictive accuracy, *Journal of Business and Economic Statistics*, Vol. 13, pp. 253–263.

Dolado, J. J., Marmol, F., (1997). On the properties of the Dickey Pantula test against fractional alternatives, *Economics Letters*, Vol. 57, pp. 11–16

Erfani, A., Samimi, A. J., (2009). Long memory forecasting of stock price index using a fractionally differenced ARMA model, *Journal of Applied Sciences Research*, Vol. 5, pp. 1721–1731.

Geweke, J., Porter-Hudak, S., (1983). The estimation and application of long memory time series models, *Journal of Time Series Analysis*, Vol. 4, pp. 221–238.

Gil-Alana, L. A., Gupta, R, Shittu, O. I. and Yaya, O. S., (2018). Market efficiency of Baltic Stock Markets: A fractional integration approach, *Physica A: Statistical Mechanics and Its Applications*, Vol. 511(1), pp 251–262.

Granger, C. W. J., Joyeux, R., (1980). An Introduction to long memory time series models and fractional differencing, *Journal of Time Series Analysis*, Vol. 1, pp. 15–39.

Geweke, J., Porter-Hudak, S., (1983). The estimation and application of long memory time series models, *Journal of Time Series Analysis*, Vol. 4, pp. 221–238.

Hurvich, C. M., Deo, R. and Brodsky, J., (1998). The mean squared error of Geweke and Porter-Hudak's estimator of the memory parameter of a long memory time series, *Journal of Time Series Analysis*, Vol. 19, pp. 19–46.

Hurvich, C. M., Chen, W. W., (2000). An efficient taper for potentially overdifferenced longmemory time series, *Journal of Time Series Analysis*, 21(2), pp. 155–180.

Hosking, J. R. M., (1981). Fractional differencing, *Biometrika*, Vol. 68, pp. 165–176.

Kang, S. H., Yoon, S., (2013). Modeling and forecasting the volatility of petroleum futures prices, *Energy Economics*, Vol. 36, pp. 354–362.

Meerschaert, M. M. Sabzikar, F. Phanikumar, M. S. and Zeleke, A., (2014). Tempered fractional time series model for turbulence in geophysical flows, *Journal of Statistical Mechanics*: *Theory and Experiment*, Vol. P09023, pp. 1–13.

Porter-Hudak, S., (1990). An application of the seasonal fractionally differenced model to the monetary aggregates, *Journal of the American Statistical Association*, Vol. 45 No. 410, pp. 338–344.

Pumi, G., Valk, M., Bisognin, C., Bayer, F. M. and Prass, T. S., (2019). Beta Autoregressive Fractionally Integrated Moving Average models, *Journal of Statistical Planning and Inference*, Vol. 200, pp. 196–212.

Rahman, R. A., Jibrin, S. A., (2018). A fractional difference returns for stylized fact studies, *Journal of Physics: Conference Series*, Vol. 113, pp. 012074

Rahman, R. A., Jibrin S. A., (2019). Modeling and forecasting tapis crude oil price: A long memory approach, *AIP Conference Proceedings* 2184, 050005-1–050005-8.

Sabzikar, F., Mcleod, A. I. and Meerschaert, M. M., (2019). Parameter estimation for ARTFIMA time series, *Journal of Statistical Planning and Inference*, Vol. 200, pp.129–145, https://doi.org/10.1016/j.jspi.2018.09.010.

STATISTICS IN TRANSITION new series, June 2022
Vol. 23, No. 2, pp. 89–105, DOI 10.2478/stattrans-2022-0018
Received – 30.01.2021; accepted – 18.01.2022

On the nonparametric estimation of the conditional hazard estimator in a single functional index

Abdelmalek Gagui[1], **Abdelhak Chouaf**[2]

ABSTRACT

This paper deals with the conditional hazard estimator of a real response where the variable is given a functional random variable (i.e it takes values in an infinite-dimensional space). Specifically, we focus on the functional index model. This approach offers a good compromise between nonparametric and parametric models. The principle aim is to prove the asymptotic normality of the proposed estimator under general conditions and in cases where the variables satisfy the strong mixing dependency. This was achieved by means of the kernel estimator method, based on a single-index structure. Finally, a simulation of our methodology shows that it is efficient for large sample sizes.

Key words: single functional index, conditional hazard function, nonparametric estimation, α-mixing dependency, asymptotic normality, functional data.

1. Introduction

The nonparametric estimation of the hazard function plays a crucial role in statistical analyses. This subject can be approached from multiple perspectives depending on the complexity of the problem. Many techniques have been studied in the literature to treat these various situations but all treat only real or multidimensional explanatory random variables. We refer to Watson and Leadbetter (1964), who were the first to study the nonparametric estimation of the hazard function. In the sequel, many authors have been interested in the study of such a function (see, for example Tanner and Wong (1983), Delecroix and Yazourh (1992), Collomb et al. (1985) and Youndjé et al. (1996)).

Focusing on functional data, the first results on the nonparametric estimate of this model, were achieved by Ferraty et al. (2000). They have studied the almost complete convergence of an estimator with kernel for the function of a chance of a real random variable conditioned by a functional explanatory variable. For instance, Masry (2005) showed the asymptotic normality of the estimator for the function of regression, Ferraty et al. (2007) studied the mean squared convergence, Burba et al. (2008) are interested in the estimate of the function of regression by using the method of k-nearest neighbours, Quintela-del-Rio (2008) obtained the asymptotic normality of the non-parametric estimation of the conditional hazard function. Ferraty et al. (2010) they etablished the almost complete convergence uniform on the functional component of this nonparametric model.

[1] Djillali Liabes University, Algeria. E-mail: gagui.abdelmalek@gmail.com. ORCID: https://orcid.org/0000-0002-2715-4304.

[2] Djillali Liabes University, Algeria. E-mail: abdo_stat@yahoo.fr.

The modelling of the spatial data was also considered in nonparametric estimation for functional data. On this subject, Dabo-Niang et al. (2012) studied the almost convergence of an estimator with kernel for the function of regression. Laksaci et al. (2009) treated the almost complete convergence of the estimator with a kernel of the function of conditional distribution and the conditional quantiles. Li and Tran (2007) obtained the asymptotic normality of a kernel estimator of the hazard function. The study of the kernel estimator of the conditional hazard function when the covariates take values in functional statistic was treated by Lakssaci et al. (2010).

Our goal in this work is devoted to the study of the single functional index model. This approach consists of making a projection between the explanatory variable Y on the functional response variable X to the non-parametric context on a function directly θ. In the finite-dimensional, random variables have been widely studied, see for example Hardle et al. (1993), Hristache et al. (2001). Furthermore, when the case is infinite dimensions or when the explanatory variable is functional, the first work which was interested in the single-index model for the nonparametric estimation is Ferraty et al. (2003). They stated for i.i.d. variables and obtained the almost complete convergence under some conditions. In the same context Ait Saidi et al. (2005) studied the dependent case of these estimators, Ait Saidi et al. (2008) proposed cross-validated estimation where the functional index is an unknown, Attaoui et al. (2011) obtained the uniform almost complete convergence of conditional density in the functional single index. More recently Tabti et al. (2017) obtained the pointwise almost complete convergence and the uniform almost complete convergence of a kernel estimator of the hazard function with the quasi-association condition in a single-index approach.

In the present paper, we obtain, under some conditions, the asymptotic normality of the conditional hazard function estimator. This result enables us to obtain the confidence intervals of this estimator. In practice, this study has great importance because it permits us to construct a prediction method based on the maximum risk estimation with a single functional index.

In Section 2, we introduce the estimator of our model in the single-functional index. Section 3 we introduce assumptions and asymptotic properties are given. Practical aspects are discussed in Section 4. Simulations are given in Section 5. Finally, Section 6 is devoted to the proofs of the results.

2. The model

Let $\{(X_i,Y_i),\ 1 \leq i \leq n\}$ be n random variables, identically distributed as the random pair (X,Y) with values in $\mathbb{H} \times \mathbb{R}$, where $\mathbb{H}$ is a separable real Hilbert space with the norm $\|.\|$ generated by an inner product $< .,. >$. We consider the semi-metric d_θ associated with the single index $\theta \in \mathbb{H}$ defined by $\forall x_1,x_2 \in \mathbb{H} : d_\theta(x_1,x_2) :=|< x_1 - x_2, \theta >|$. Assume that the explanation of Y given X is done through a fixed functional index θ in $\mathbb{H}$. In the sense that there exists a θ in $\mathbb{H}$ (unique up to a scale normalization factor) such that: $\mathbb{E}[Y|X] = \mathbb{E}[Y| < \theta, X >]$. The conditional cumulative distribution function of Y given $< X, \theta >$ is denoted by

$$F^x(\theta,y) := F(y| < \theta,x >) = \mathbb{P}(Y \leq y \,|< X,\theta >=< \theta,x >),\ \forall y \in \mathbb{R}$$

Clearly we have for all $x \in \mathbb{H}$,

$$F_1(.| < x,\theta_1 >) = F_2(.| < x,\theta_2 >) \Rightarrow F_1 \equiv F_2 \text{ and } \theta_1 = \theta_2.$$

The natural kernel estimator of $F(\theta,y,x)$ is defined as

$$\widehat{F}(\theta,y,x) = \frac{\sum_{i=1}^{n} K(h_K^{-1} d_\theta(x,X_i))H(h_H^{-1}(y-Y_i))}{\sum_{i=1}^{n} K(h_K^{-1} d_\theta(x,X_i))},\ \forall y \in \mathbb{R} \tag{1}$$

We suppose that the conditional density of Y given $X = x$ denoted by $f(.|x)$ exists and is given by $\forall y \in \mathbb{R},\ f_\theta(y|x) := f(y| < x,\theta >)$. In the following, we denote by $f(\theta,.,x)$, the conditional density of Y given $< x,\theta >$ and we define the kernel estimator $\widehat{f}(\theta,.,x)$ of $f(\theta,.,x)$ by:

$$\widehat{f}(\theta,y,x) = \frac{h_H^{-1}\sum_{i=1}^{n} K(h_K^{-1} d_\theta(x,X_i))H'(h_H^{-1}(y-Y_i))}{\sum_{i=1}^{n} K(h_K^{-1} d_\theta(x,X_i))},\ \forall y \in \mathbb{R} \tag{2}$$

with the convention $0/0 = 0$, where K and H are kernels function (H' is the derivate of H) and $h_K := h_{n,K}$ (resp $h_H := h_{n,H}$) is a sequence of bandwidths that decrease to zero as n goes to infinity.

We are interested in estimating non parametrically the conditional hazard function λ defined by:

$$\hat{\lambda}(\theta,y,x) = \frac{\widehat{f}(\theta,y,x)}{1-\widehat{F}(\theta,y,x)},\ \forall y \in \mathbb{R}.$$

3. Main results

We begin with introducing some notations. Let $(X_i,Y_i)_{i=1}^{\infty}$ be a sequence of random variables and $\alpha(n)$ be a sequence of real numbers. A stationary process $(X_i,Y_i)_{i=1}^{\infty}$ is called α-mixing or strongly mixing, if
$\alpha(n) = \sup_{A\in\mathscr{A}_1^k} \sup_{B\in\mathscr{A}_{n+k}^{\infty}} |\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)| \to 0$, as $n\to\infty$, where $\mathscr{F}_a^b$ is the σ-algebra generated by $(X_j,Y_j)_{j=a}^{b}$.

In this section, we give some obtained results on the asymptotic normality of the estimator $\widehat{\lambda}(\theta,y,x)$, which require the following additional hypotheses. All along the paper, when no confusion is possible, we will denote by C and C' some strictly positive generic constants. We put, for any $x \in \mathbb{H}$, and $i = 1,...,n$ $K_i(\theta,x) := K(h_K^{-1} d_\theta(x,X_i))$ and, for all $y \in \mathbb{R}$, $H_i^j := H^j(h_H^{-1}(y-Y_i))$ for $j = 0,1$ In the following, for any $x \in \mathbb{H}$ and $y \in \mathbb{R}$, let $\mathscr{N}_x$ be a fixed neighbourhood of x in $\mathbb{H}$, $S_{\mathbb{R}}$ will be a fixed compact subset of $\mathbb{R}$,and we will use the notation $B_\theta(x,h) := \{x_1 \in \mathbb{H} :\ 0 < | < x-x_1,\theta > | < h\}$, the ball centered at x, with radius h. All along the paper, when no confusion will be possible, we will denote by C, C'

and $C_{\theta,x}$ some generic constant in $\mathbb{R}_+^*$

(H1) $\mathbb{P}(X \in B_{\theta,x}(h)) = \phi_{\theta,x}(h) > 0$. Moreover, there exists a function $\beta_{\theta,x}(.)$ such that: $\forall s \in [-1,1]$, $\lim\limits_{n \longrightarrow \infty} \dfrac{\beta_{\theta,x}(sh_K)}{\beta_{\theta,x}(h_K)} = \beta_{\theta,x}(s)$.

(H2) For $l \in \{0,2\}$, the functions $\psi_l(s) = \mathbb{E}\left[\dfrac{\partial^l f(\theta,y,X)}{\partial y^l} - \dfrac{\partial^l f(\theta,y,x)}{\partial y^l} | d_\theta(x,X=s)\right]$ and $\Psi_l(s) = \mathbb{E}\left[\dfrac{\partial^l F(\theta,y,X)}{\partial y^l} - \dfrac{\partial^l F(\theta,y,x)}{\partial y^l} | d_\theta(x,X=s)\right]$ are differentiable at $s=0$.

(H3) The kernel K is a differentiable function and its derivative K' exists and is such that there exist two constants C and C' with $-\infty < C < K'(t) < C' < 0$, for $t \in [0,1]$.

(H4) The kernels K and H are an even bounded function .

(H5) The bandwidths h_K and h_H satisfy
(1*) $\lim\limits_{n \longrightarrow \infty} \dfrac{1}{nh_H \phi_{\theta,x}(h_K)} = 0$,
(2*) $\lim\limits_{n \longrightarrow \infty} nh_H^5 \phi_{\theta,x}(h_K) = 0$ and $\lim\limits_{n \longrightarrow \infty} nh_H h_k^2 \phi_{\theta,x}(h_K) = 0$,
(3*) $\lim\limits_{n \longrightarrow \infty} h_H = 0$, $\lim\limits_{n \longrightarrow \infty} h_K = 0$, *and* $\lim\limits_{n \longrightarrow \infty} \dfrac{\log n}{n\phi_{\theta,x}(h_K)} = 0$,
(4*) $\lim\limits_{n \longrightarrow \infty} h_K^{2b_1} \phi_{\theta,x}(h_K) = 0$, *and* $\lim\limits_{n \longrightarrow \infty} h_H^{2b_1} \phi_{\theta,x}(h_K) = 0$.

(H6) $(X_i,Y_i)_{i\in\mathbb{N}}$ is a strongly mixing sequence, whose mixing coefficient $\alpha(n)$ satisfies $\exists a > (5+\sqrt{17})/2$, $\exists C > 0 : \forall n \in \mathbb{N}$, $\alpha(n) \leq Cn^{-a}$.

(H7) $0 < \sup\limits_{i \neq j} \mathbb{P}\left((X_i,X_j) \in B_\theta(x,h_K) \times B_\theta(x,h_K)\right) = O\left(\dfrac{\phi_{\theta,x}(h_K)^{(a+1)/a}}{n^{1/a}}\right)$.

(H8) $\exists \beta_0 > 0$, C_1, $C_2 > 0$, such that: $C_1 n^{\frac{3-a}{a+1}+\beta_0} \leq \phi_{\theta,x}(h_K) \leq C_2 n^{\frac{1}{1-a}}$.

Comments on the assumptions

Assumptions (H1)-(H4) are technicals and permit to give an explicit asymptotic variance. The function $\beta_{\theta,x}(.)$ will play a major role in our results, it intervenes to compute the exact constant terms involved in our asymptotic expansions (for more of this assumptions, see Ferraty et al. 2007). Finally (H5)-(H8) permits to remove the bias term in the asymptotic normality result.
Now, we give our main result.

Theorem 3.1. *Assume that (H1)-(H5) hold, and (H6)-(H8) hold, as n goes to infinity, we have*

$$(nh_H \phi_{\theta,x}(h_K))^{1/2}(\widehat{\lambda}(\theta,y,x) - \lambda(\theta,y,x) - B_n(\theta,y,x)) \xrightarrow{\mathscr{D}} \mathscr{N}(0,\sigma_h^2(\theta,y,x)),$$

where

$$B_n(\theta,y,x) = \frac{1}{1-F(\theta,y,x)}\Big((B_H^f - \lambda(\theta,y,x)B_H^F)h_H^2 + ((B_K^f - \lambda(\theta,y,x)B_K^F)h_K)\Big)$$

with

$$\sigma_h^2(\theta,y,x) = \frac{M_2\lambda(\theta,y,x)}{M_1^2(1-F(\theta,y,x))}$$

$M_0 = K(1) - \int_0^1 sK'(s)\beta_{\theta,x}(s)ds, \quad M_j = K^j(1) - \int_0^1 (K^j)'(s)\beta_{\theta,x}(s)ds$
for $j = 1,2$

and

$$B_H^f(\theta,y,x) = \frac{1}{2}\frac{\partial^2 f(\theta,y,x)}{\partial y^2}\int t^2 H'(t)dt,$$

$$B_K^f(\theta,y,x) = h_k\psi_0'(0)\frac{M_0}{M_1}h_K.$$

$$B_H^F(\theta,y,x) = \frac{1}{2}\frac{\partial^2 F(\theta,y,x)}{\partial y^2}\int t^2 H'(t)dt,$$

$$B_K^F(\theta,y,x) = h_k\Psi_0'(0)\frac{M_0}{M_1}h_K.$$

and $\mathscr{D}$ means the convergence in distribution.

Corollary 3.1. *Under the hypotheses of Theorem 3.1,and if the bandwidth parameters (h_K and h_H) satisfies (H5) and if the function $\phi_{\theta,x}(h_K)$ satisfies :*

$$\lim_{n\longrightarrow\infty}(h_H^2 + h_K)(n\phi_{\theta,x}(h_K))^{1/2} = 0,$$

we have

$$(nh_H\phi_{\theta,x}(h_K))^{1/2}(\widehat{\lambda}(\theta,y,x) - \lambda(\theta,y,x)) \xrightarrow{\mathscr{D}} \mathscr{N}(0,\sigma_h^2(\theta,y,x)),$$

The proof of Theorem 3.1 is based on the following decomposition:

$$\begin{aligned}
\widehat{\lambda}(\theta,y,x) - \lambda(\theta,y,x) \;&=\; \frac{1}{\widehat{F}_D(\theta,x) - \widehat{F}_N(\theta,y,x)}\Big(\widehat{f}_N(\theta,y,x) - \mathbb{E}[\widehat{f}_N(\theta,y,x)]\Big)\\
&+\; \frac{1}{\widehat{F}_D(\theta,x) - \widehat{F}_N(\theta,y,x)}\Big\{\lambda(\theta,y,x)\Big(\mathbb{E}[\widehat{F}_N(\theta,y,x)] - F(\theta,y,x)\Big)\\
&+\; \Big(\mathbb{E}[\widehat{f}_N(\theta,y,x)] - f(\theta,y,x)\Big)\Big\}\\
&+\; \frac{\widehat{h}(\theta,y,x)}{\widehat{F}_D(\theta,x) - \widehat{F}_N(\theta,y,x)}\Big\{1 - \mathbb{E}[\widehat{F}_N(\theta,y,x)]\\
&-\; \Big(\widehat{F}_D(\theta,x) - \widehat{F}_N(\theta,y,x)\Big)\Big\}
\end{aligned}$$

Lemma 3.1. *Under the Assumptions of Theorem 3.1, as n goes to infinity, we have*

$$(nh_H\phi_{\theta,x}(h_K))^{1/2}(\widehat{f_N}(\theta,y,x)-\mathbb{E}[\widehat{f_N}(\theta,y,x)])\xrightarrow{\mathscr{D}}\mathcal{N}(0,\sigma_f^2(\theta,y,x)).$$

Proof of lemma 3.1 First, we define that

$$Z_i(\theta,y,x)=\frac{\sqrt{\phi_{\theta,x}(h_K)}}{\sqrt{nh_H}\mathbb{E}[K_1(\theta,x)]}(\zeta_i(\theta,y,x)-\mathbb{E}[\zeta_i(\theta,y,x)]),$$

and

$$T_n:=\sum_{i=1}^{n}Z_i(\theta,y,x).$$

where $\zeta_i(\theta,y,x)=H_i'(\theta,x)K_i(\theta,x)$,
Thus,

$$T_n=\sqrt{nh_H\phi_{\theta,x}(h_K)}(\widehat{f_N}(\theta,y,x)-\mathbb{E}[\widehat{f_N}(\theta,y,x)]).$$

So, our claimed result is now

$$T_n\longrightarrow\mathcal{N}(0,\sigma_f^2(\theta,x)). \tag{3}$$

Therefore, we have

$$\begin{aligned}Var(T_n)&=nh_H\phi_{\theta,x}(h_K)Var(\widehat{f_N}(\theta,y,x)-\mathbb{E}[\widehat{f_N}(\theta,y,x)])\\&=nh_H\phi_{\theta,x}(h_K)Var(\widehat{f_N}(\theta,y,x))\end{aligned} \tag{4}$$

Now, we need to evaluate the variance of $\widehat{f_N}(\theta,y,x)$. For this we have for all $1\le i\le n$, :

$$\begin{aligned}Var(\widehat{f_N}(\theta,y,x))&=\frac{1}{(nh_H\mathbb{E}[K_1(\theta,x)])^2}\sum_{i=1}^{n}\sum_{j=1}^{n}Cov(\zeta_i(\theta,y,x),\zeta_j(\theta,y,x))\\&=I_{1,n}+I_{2,n}.\end{aligned}$$

where

$$\begin{aligned}I_{1,n}&=\frac{1}{n(h_H\mathbb{E}[K_1(\theta,x)])^2}Var(\zeta_1(\theta,y,x))\\I_{2,n}&=\frac{1}{(nh_H\mathbb{E}[K_1(\theta,x)])^2}\sum_{i=1}^{n}\sum_{j=1_{i\neq j}}^{n}Cov(\zeta_i(\theta,y,x),\zeta_j(\theta,y,x)).\end{aligned}$$

First, for the quantity $I_{1,n}$, we have

$$\begin{aligned} Var(\zeta_1(\theta,y,x)) &\leq \mathbb{E}\left[H'^2_1(y)K_1^2(\theta,x)\right] \\ &\leq \mathbb{E}\left[K_1^2(\theta,x)\mathbb{E}\left[H'^2_1(y)| < \theta,X_1 >\right]\right]. \end{aligned}$$

$$\begin{aligned} |\mathbb{E}\left[H'^2_1(y)| < \theta,X_1 >\right]| &= \left|\int_{\mathbb{R}} H'^2(h_H^{-1}(y-z))f(\theta,z,x)dz\right| \\ &\leq h_H \int_{\mathbb{R}} H'^2|f(\theta,y-h_Ht,x)f(\theta,y,x)|dt \\ &+ h_H f(\theta,y,x)\int_{\mathbb{R}} H'^2 dt \\ &\leq h_H^{1+b_2}\int_{\mathbb{R}} |t|^{b_2}H'^2dt + h_H f(\theta,y,x)\int_{\mathbb{R}} H'^2 dt \\ &= h_H\left(o(1)+f(\theta,y,x)\left(\int_{\mathbb{R}} H'^2dt\right)\right). \end{aligned}$$

As $n \longrightarrow \infty,\ \mathbb{E}[K_1^2(\theta,x)] \longrightarrow M_2\phi_{\theta,x}(h_K)$, one gets

$$Var(\zeta_1(\theta,y,x)) = M_2\phi_{\theta,x}(h_K)h_H\left(o(1)+f(\theta,y,x)\left(\int_{\mathbb{R}} H'^2dt\right)\right).$$

So, using (H5-1*), we get

$$\begin{aligned} I_{1,n} &= \frac{M_2\phi_{\theta,x}(h_K)}{n(M_1h_H\phi_{\theta,x}(h_K))^2}h_H\left(o(1)+f(\theta,y,x)\left(\int_{\mathbb{R}} H'^2dt\right)\right) \\ &= o\left(\frac{1}{nh_H\phi_{\theta,x}(h_K)}\right)+\frac{M_2f(\theta,y,x)}{M_1^2nh_H\phi_{\theta,x}(h_K)}\left(\int_{\mathbb{R}} H'^2dt\right) \\ &\longrightarrow \frac{M_2f(\theta,y,x)(\int_{\mathbb{R}} H'^2dt)}{M_1^2nh_H\phi_{\theta,x}(h_K)},\ as\ n \longrightarrow \infty. \end{aligned} \tag{5}$$

Second, for the quantity $I_{2,n}$, we will use the following decomposition:

$$I_{2,n} = \sum_{i=1}^{n}\sum_{\substack{j=1 \\ 0<|i-j|\leq m_n}}^{n} Cov(\zeta_i(\theta,y,x),\zeta_j(\theta,y,x)) + \sum_{i=1}^{n}\sum_{\substack{j=1 \\ |i-j|>m_n}}^{n} Cov(\zeta_i(\theta,y,x),\zeta_j(\theta,y,x)).$$

Similarly to Attaoui said (2014), we can easily write

$$I_{2,n} = O(nh_H^2\phi_{\theta,x}(h_K)).$$

It yields,

$$\frac{1}{nh_H\phi_{\theta,x}(h_K)}I_{2,n} \longrightarrow 0,\ as\ n \longrightarrow \infty. \tag{6}$$

Finally, the proof of the Lemma is completed, to get

$$Var(T_n) \longrightarrow \frac{M_2 f(\theta,y,x)}{M_1^2}\left(\int_{\mathbb{R}} H'^2 dt\right) =: \sigma_f^2(\theta,x).$$

Lemma 3.2. *Under the Assumptions (H1)-(H4), and (H7), as n goes to infinity, we have*

$$\mathbb{E}[\widehat{F}_N(\theta,y,x)] - F(\theta,y,x) = B_H^F(\theta,y,x)h_H^2 + B_K^F(\theta,y,x)h_K + o(h_H^2) + o(h_K)$$

Proof of lemma 3.2 First, for $\mathbb{E}[\widehat{F}(\theta,y,x)]$, we start by writing

$$\mathbb{E}[\widehat{F}_N(\theta,y,x)] = \frac{1}{\mathbb{E}[K_1(\theta,x)]}\mathbb{E}\left[K_1(\theta,x)\mathbb{E}[h_H^{-1}H_1'(y)|X]\right]$$

with

$$h_H^{-1}\mathbb{E}[H_1'(y)|X] = \int_{\mathbb{R}} H'(t)F(\theta,y-h_H t,X)dt$$

The latter can be re-written, using a Taylor expansion under (H4), as follows:

$$h_H^{-1}\mathbb{E}[H_1'(y)|X] = F(\theta,y,X) + \frac{h_H^2}{2}\left(\int t^2 H'(t)dt\right)\frac{\partial^2 F(\theta,y,X)}{\partial^2 y} + o(h_H^2).$$

Thus, we get

$$\begin{aligned}\mathbb{E}[\widehat{F}_N(\theta,y,x)] &= \frac{1}{\mathbb{E}[K_1(\theta,x)]}\left(\mathbb{E}[K_1(\theta,x)F(\theta,y,X)] + \left(\int t^2 H'(t)dt\right)\right.\\ &\times \left.\mathbb{E}\left[K_1(\theta,x)\frac{\partial^2 F(\theta,y,X)}{\partial^2 y}\right] + o(h_H^2)\right).\end{aligned}$$

Let $\Psi_l(.,y) := \frac{\partial^l F(.,y,.)}{\partial^l y}$: for $l \in \{0,2\}$, since $\Psi_l(0) = 0$, we have

$$\begin{aligned}\mathbb{E}[K_1(\theta,x)\psi(X,y)] &= \Psi_l(x,y)\mathbb{E}[K_1(\theta,x)] + \mathbb{E}[K_1(\theta,x)(\Psi_l(X,y) - \Psi(x,y))]\\ &= \Psi(x,y)\mathbb{E}[K_1(\theta,x)] + \mathbb{E}[K_1(\theta,x)(\Psi_l(d_\theta(x,X)))]\\ &= \Psi_l(x,y)\mathbb{E}[K_1(\theta,x)] + \Psi_l'(0)\mathbb{E}[d_\theta(x,X)K_1(\theta,x)]\\ &+ o(\mathbb{E}[d_\theta(x,X)K_1(\theta,x)]).\end{aligned}$$

So

$$\begin{aligned}\mathbb{E}[\widehat{F}_N(\theta,y,x)] &= F(\theta,y,x)+\frac{h_H^2}{2}\frac{\partial^2 F(\theta,y,X)}{\partial^2 y}\int t^2 H'(t)dt+o\left(h_H^2\frac{\mathbb{E}[d_\theta(x,X)K_1(\theta,x)]}{\mathbb{E}[K_1(\theta,x)]}\right)\\ &+ \Psi_0'(0)\frac{\mathbb{E}[d_\theta(x,X)K_1(\theta,x)]}{\mathbb{E}[K_1(\theta,x)]}+o\left(\frac{\mathbb{E}[d_\theta(x,X)K_1(\theta,x)]}{\mathbb{E}[K_1(\theta,x)]}\right).\end{aligned}$$

Similarly to Ferraty et al. (2007), we show that

$$\frac{1}{\phi_{\theta,x}(h_K)}\mathbb{E}[d_\theta(x,X)K_1(\theta,x)] = M_0 h_K + o(h_K)$$

and

$$\frac{1}{\phi_{\theta,x}(h_K)}\mathbb{E}[K_1(\theta,x)] \longrightarrow M_1.$$

Hence,

$$\mathbb{E}[\widehat{F}_N(\theta,y,x)] = F(\theta,y,x)+\frac{h_H^2}{2}\frac{\partial^2 F(\theta,y,X)}{\partial^2 y}\int t^2 H'(t)dt+\Psi_o'(0)\frac{M_0}{M_1}h_K+o(h_H^2)+o(h_K)$$

Lemma 3.3. *Under the Assumptions (H1)-(H4), and (H7), as n goes to infinity, we have*

$$\mathbb{E}[\widehat{f}_N(\theta,y,x)] - f(\theta,y,x) = B_H^f(\theta,y,x)h_H^2 + B_K^f(\theta,y,x)h_K + o(h_H^2) + o(h_K)$$

Proof of lemma 3.3. The proof of this lemma follows the steps as for proving lemma 3.2, to study $\mathbb{E}[\widehat{f}_N(\theta,y,x)]$ it suffices to write by an integration by part

$$\mathbb{E}[\widehat{f}_N(\theta,y,x)] = \frac{1}{\mathbb{E}[K_1]}\mathbb{E}[K_1\mathbb{E}[H_1 \mid X]] \text{ with } \mathbb{E}[K_1\mathbb{E}[H_1 \mid X]] = \int_{\mathbb{R}} H'(t)f^X(y-h_H t)dt$$

Then we can follow to prove that

$$\mathbb{E}[\widehat{f}_N(\theta,y,x)] = f(\theta,y,x)+\frac{h_H^2}{2}\frac{\partial^2 f(\theta,y,X)}{\partial^2 y}\int t^2 H'(t)dt+\psi_0'(0)\frac{M_0}{M_1}h_K+o(h_H^2)+o(h_K)$$

Lemma 3.4. *Under the hypotheses of Theorem 3.1*

$$\widehat{F}_D(\theta,x) - \widehat{F}_N(\theta,y,x) \longrightarrow 1 - F(\theta,y,x), \text{ in probability.}$$

And

$$\left(\frac{nh_H\phi_{\theta,x}(h_K)}{\sigma_h^2(\theta,y,x)}\right)^{1/2}\left(\widehat{F}_D(\theta,x) - \widehat{F}_N(\theta,y,x) - 1 + \mathbb{E}[\widehat{F}_N(\theta,y,x)\right) = o_p(1).$$

Proof of lemma 3.4. It is clear that

$$\mathbb{E}\left[\widehat{F}_D(\theta,x)-\widehat{f}(\theta,y,x)-1+F(\theta,y,x)\right]\longrightarrow 0,$$

and

$$Var\left[\widehat{F}_D(\theta,x)-\widehat{f}(\theta,y,x)-1+F(\theta,y,x)\right]\longrightarrow 0,$$

then

$$\widehat{F}_D(\theta,x)-\widehat{f}(\theta,y,x)-1+F(\theta,y,x)\xrightarrow{\mathbb{P}} 0.$$

Moreover, the asymptotic variance of $\widehat{F}_D(\theta,x)-\widehat{f}_N(\theta,y,x)$ (see Djebbouri et al.(2015)), allows to obtain

$$\frac{nh_H\phi_{\theta,x}(h_K)}{\sigma_h^2(\theta,y,x)}Var\left[\widehat{F}_D(\theta,x)-\widehat{f}_N(\theta,y,x)-1+\mathbb{E}[\widehat{f}_N(\theta,y,x)]\right]\longrightarrow 0.$$

By combining the result with the fact that

$$\mathbb{E}\left[\widehat{F}_D(\theta,x)-\widehat{f}_N(\theta,y,x)-1+\mathbb{E}[\widehat{f}_N(\theta,y,x)]\right]\longrightarrow 0,$$

we obtain the claimed result.

4. Simulation study

We first construct the simulation of the explanatory functional variables. In the second part, we focus on the ability of the nonparametric functional regression to predict responses variable from functional predictors. Finally we illustrated the Monte Carlo methodology and we will test the efficiency of the asymptotic normality results in parallel with the practical experiment.

For this purpose, we consider the following process explanatory functional variables for $n=350$:

$$X_i(t)=1-\sin(2\Omega_i t)\alpha_i+\Omega_i t\,,\ \forall t\in[0,\pi]$$

where α_i and Ω_i are n independent real random variables (r.r.v.) uniformly distributed over $[0.3;2]$ $(resp.[1;3])$, t is assumed that these curves are observed on a discretization grid of 100 points in the interval. These functional variables are represented in the Figure 1

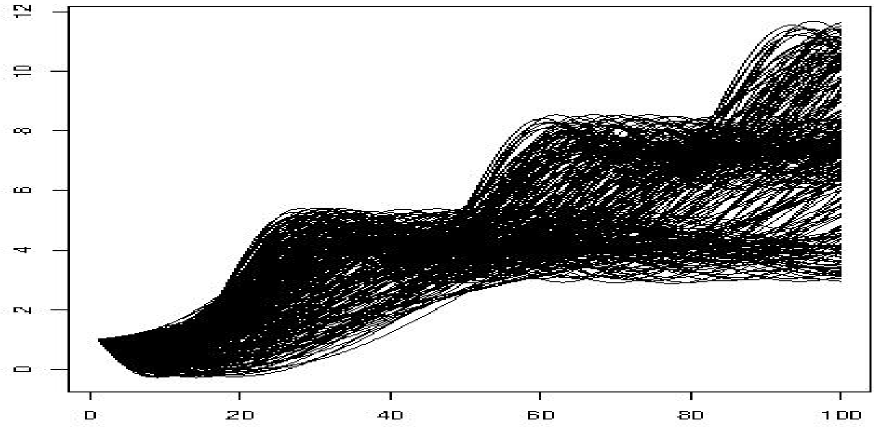


Figure 1. The curves $X_{i=1,\ldots,200}$

For response variables Y_i, we consider the following model for all $i = 1,\ldots n$ and $l = 1,\ldots n$:

$$Y = \lambda(< X_i, \theta_l >) + \varepsilon$$

where $\lambda(\mathscr{X}) = \int_0^{t_j} \frac{1}{1-\mathscr{X}_i(v)^2} dv$ and ε is a centred normal variable and it is assumed to be independent of $(X_i)_i$. Our goal in this illustration is to show the usefulness of conditional density in the context of forecasting.

Now, we precise the different parameters of our estimators. Indeed, first of all, it is clear that the shape of the curves allows us to use

$$d(x_1,x_2) = \sqrt{\int_0^1 (x_1(t) - x_2(t))^2} \text{ ; } \forall x_1, x_2 \in \mathscr{H} \text{ where H is semi-metric}$$

We choose particularly the quadratic kernels defined by

$$K(x) = \frac{3}{2}(1-x^2)\; x \in [0,1]\;;\; K_0(x) = \frac{3}{4}(1-x^2)\; x \in [-1,1] \text{ and } H(x) = \int_{-\infty}^{x} K_0(u)du.$$

In this illustration, we select the functional index θ on the set of eigenvectors of the empirical covariance operator.

$$\frac{1}{200}\sum_{i=1}^{200}(X_i-\bar{X})^t((X_i-\bar{X})).$$

Indeed, we recall that the ideas of Aitsaidi (2007) can be adapted to find a method of practical selection for θ. However, this adaptation in the case of the conditional density requires tools and additional preliminary results (see the discussion Attaoui *et al.* (2010) and Attaoui (2014)).

For this purpose, we divide our observations into two packets: learning sample $(X_i,Y_i)_{i=1,...200}$ and test sample and $(X_i,Y_i)_{i=201,...250}$(see, Ferraty et al. (2006)). For the choice of smoothing parameters h_K and h_H, we will adopt the selection criterion used by Ferraty and Vieu (2006) in the case of the kernel method for which h_K and h_H are obtained. by minimizing the next criterion

$$\text{for each} \quad X_i \quad \text{in the sample of the test} \qquad err(h_K,h_H) = |Y_{i^*} - \theta(X_{i^*})| \qquad (7)$$

where i^* denotes the index of the nearest curve X_i from all the curves of the learning sample.

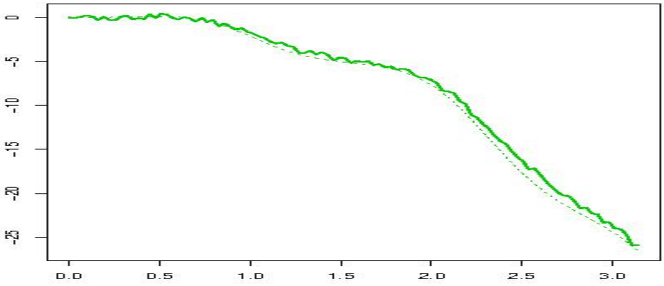


Figure 2. Predicted functional responses (solid lines); observed functional responses (dashed lines).

In this simulation study, we assume the quality of prediction by comparing the predicted functional responses (i.e. $\widehat{\lambda}(\theta,y,x)$ for any X in the testing sample) and the true functional operator (i.e. $\lambda(\theta,y,x)$) as in Figure. 2. However, if one wishes to assess the quality of prediction for the whole testing sample, it is much better to see what happens direction by direction. In other words, displaying the predictions onto the direction θ_l amounts to plotting the 50 points $(\lambda(<X_i,\theta_l>),\widehat{\lambda}(<X_i,\theta_l>))_{i=201,\ldots,250}$. Figure. 3 proposes a componentwise prediction graph for the two first components $(i.e. l = 1,2)$. The quality of componentwise predictions is quite good for each component.

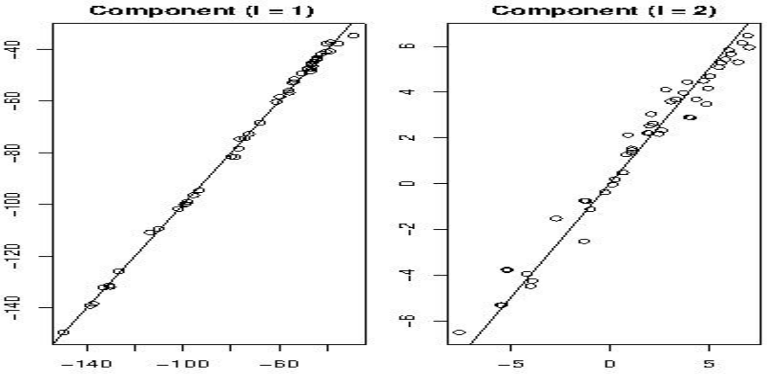


Figure 3. Representation of the prediction quality for each component.

For the next simulation algorithm we used:

- Simulate a sample of size n
- Calculate the smoothing parameters h_K and h_H that are varied over an interval [0,1] and which minimizes in 7
- We compute the quantities

$$(nh_H\phi_{\theta,x})^{1/2}(\widehat{\lambda}(\theta,y,x)-\lambda(\theta,y,x))$$

where $\widehat{\lambda}(\theta, y, x)$ is the functional hazard kernel estimator from the sample $(X_i, Y_i)_{i=1,\ldots,200}$,

- compute a standard hazard function estimator by the kernel method .
- compare the estimated $\widehat{\lambda}(\theta, y, x)$ with the corresponding estimated $\lambda(\theta, y, x)$.

The obtained results are shown in Figure. 4.

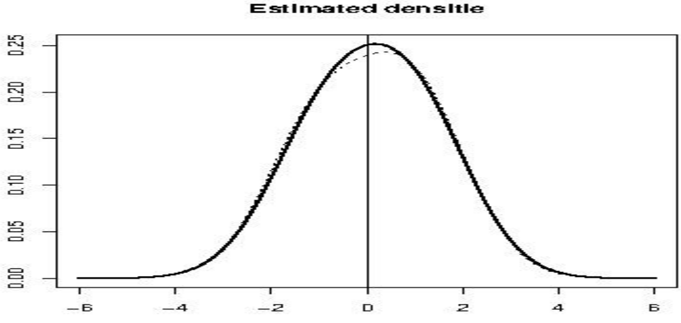


Figure 4. Representation of the asymptotic distribution of the hazard function estimator.

It can be seen that, both are very close and have good behaviours with respect to the standard normal distribution.

5. Conclusions

In this paper, we are mainly interested in the nonparametric estimation of the conditional hazard function estimator for a variable explanatory functionally conditioned to an actual response variable via a functional single index model. We show that the estimator provides good predictions under this model. One of the main contributions of this work is the choice of a semi-metric. Indeed, it is well known that, in non-parametric functional statistics, the semi-metric of the projection type is very important for increasing the concentration property. The functional index model is a special case of this family of semi-metrics because it is based on the projection on a functional direction which is important for the implementation of our method in practice.

Acknowledgements

The authors thank the associate editor and the anonymous reviewers for their valuable comments and suggestions, which substantially improved the quality of an earlier version of this article.

References

Ait Saidi, A., Ferraty, F., Kassa, P. and Vieu, P., (2005). Single functional index model for a time series. *Rev. Roumaine Math. Pures Appl.* 50, pp. 321–330.

Ait Saidi, A., Ferraty, F., Kassa, P. and Vieu, P., (2008). Cross-validated estimations in the single functional index model. *Statistics.* Vol. 42, No. 6, pp. 475–494.

Ait Saidi, A. and Mecheri, K., (2016). The conditional cumulative distribution function in single functional index model. *Comm. Statist. Theory Methods.* 45, pp. 4896–4911.

Arfi, M., (2013). Nonparametric Estimation for the Hazard Function. *Communications in Statistics - Theory and Methods* 42, pp. 2543–2550.

Attaoui, S., (2014). Strong uniform consistency rates and asymptotic normality of conditional density estimator in the single functional index modeling for time series data. *AStA - Advances in Statistical Analysis* 98, pp. 257–286.

Attaoui, S., laksaci, A. and Ould Said, F., (2011). A note on the conditional density estimate in the single functional index model. *Statist. Probab. Lett.,* 81, pp. 45–53.

Bagkavos, D., (2011). Local linear hazard rate estimation and bandwidth selection . *Annals of the Institute of Statistical Mathematics* 63, pp. 1019–1046.

Bouraine, M., Ait Saidi, A., Ferraty, F. and Vieu, P., (2010). Choix optimal de l'indice Multi-fonctionnel: Methode de validation croisée. *Rev. Roumaine Math.Pures Appl.*. 55, pp. 355–367.

Burba, F., Ferraty, F. and Vieu, P., (1996). Convergence of *k* nearest neighbor kernel estimator in nonparametric functional regression. *Comptes Rendus Mathematique*. 346, pp. 339–342.

Collomb, G., Hassani, S., Sarda, P. and Vieu, P., (1985). Estimation non parametrique de la fonction de hasard pour des observations dependentes. *Statistique et Analyse des Données* 10, pp. 42–49.

Dabo-Niang, S., Kaid, Z. and Laksaci, A., (2012). On spatial conditional mode estimation for a functional regressor. *Statistics and probability letters,*. 82, pp. 1413–1421.

Delecroix, M. , Yazourh , O., (1992). Estimation de la fonction de hazard en présence de censure droite. Méthode des fonctions orthogonales. *Statistique et Analyse des Données* 16, pp. 39–62 .

Estevèz-Pérez, G., Quintela-del-R, A. and Vieu, P., (2002). Convergence rate for cross-validatory bandwith in kernel hazard estimation from dependent samples. *J. Statist. Plann. Inference* 104, pp. 1–30.

Ferraty, F. and Vieu, P., (2006). *Nonparametric functional data analysis. Theory and Practice*. Springer Series in Statistics. New York.

Ferraty, F., Laksaci , A. and Vieu, P., (2005). Functional times series prediction via conditional mode, *C. R., Math. Acad. Sci. Paris* 340(5), pp. 389–392.

Ferraty, F., Laksaci , A. and Vieu, P., (2006). Estimating some characteristics of the conditional distribution in nonparametric functional models, *Stat. Inf. Stoch. Proc.* 9, pp. 47–76.

Ferraty, F., Laksaci, A., Tadj, A., and Vieu, P.,(2010). Rate of uniform consistency for nonparametric estimates with functional variables, *Journal of Statistical Planning and Inference*. 140, pp. 335–352.

Härdle, W., Hall, P. and Ichumira, H., (1993). Optimal smoothing in single-index models *Ann. Statist.*. 27, pp. 157–178.

Hristache, M., Juditsky, A., and Spokoiny, V., (2001). Direct estimation of the index coefficient in the single-index model. *Ann. Statist.*, 29 , pp. 595–623.

Laksaci, A. and Mechab, B., (2010). Estimation nonparamétrique de la fonction de hasard avec variable explicative fonctionnelle : cas des données spatiales. *Rev. Roumaine Math. Pures Appl.* 55, pp. 35–51.

Laksaci, A. and Maref, F., (2009). Nonparametric estimation of conditional quantiles for functional and spatial dependent variables. *Comptes Rendus Mathematique*. 347, pp. 1075–1080.

Lecoutre, J. P. and Ould-Said, E.,(1992). Estimation de la densité et de la fonction de hasard conditionnelle pour un processus fortement mélangeant avec censure. *C.R. Math. Acad. Sci.Paris*. 314, pp. 295–300.

Li, J. and Tran, L. T., (2007). Hazard rate estimation on random fields, *J. Multivariate Anal.* 98 , pp. 1337–1355.

Masry, E., (2005). Nonparametric regression estimation for dependent functional data: asymptotic normality, *Stoch. Process. Appl.* 115, pp. 155–177.

Patil, P. N. , Wells M. T. and Marron J. S., (1993). Some heuristics of kernel based estimators of ratio functions. *Journal of Nonparametric Statistics* 4, pp. 203–209.

Quintela-Del-Río A., (2008). Hazard function given a functional variable: Non-parametric estimation under strong mixing conditions. *Journal of Nonparametric Statistics* 5, pp. 413–430.

Roussas, G. G., (1968). On some properties of nonparametric estimates of probability density functions *Bull. Soc. Math. Greece (N. S.)* 9, pp. 29–43 .

Tabti, H. and Ait Saidi, A., (2017). Estimation and simulation of conditional hazard function in the quasi-associated framework when the observations are linked via a functional single-index structure, *Communications in Statistics - Theory and Methods* 47, pp. 816–838 .

Tanner, M. A. and Wong, W. H., (1983). The Estimation of the Hazard Function from Randomly Censored Data by the Kernel Method *The Annals of Statistics* 11, pp. 989–993 .

Watson G. S. and Leadbetter M. R., (1964). Hazard Analysis. I *Biometrika* 51, pp. 175–184 .

Youndjé, É., Sarda, P. and Vieu, P., (1996). Optimal smooth hazard estimates. *Test.* 5, pp. 379–394.

STATISTICS IN TRANSITION new series, June 2022
Vol. 23, No. 2 pp. 107–127, DOI 10.2478/stattrans-2022-0019
Received – 22.07.2021; accepted – 21.03.2022

Institutional equilibrium in EU economies in 2008 and 2018: SEM-PLS models

Mateusz Borkowski[1]

ABSTRACT

The aim of the research is to identify the strength and direction of the development of the relationship between formal and informal institutions and to assess the institutional equilibrium of modern economies. The structural equations modelling based on partial least squares (SEM-PLS) is applied to achieve the purpose of the article. It is an econometric method that allows the measurement and analysis of the dependencies between latent variables (measures that cannot be directly observed). The study included 27 EU economies and the research period covered the years 2008 and 2018. The results of the study demonstrate that the quality of informal institutions strongly, positively determines the quality of formal institutions. The conducted analyses indicate that modern economies are diversified in terms of the quality of informal and formal institutions and, consequently, in institutional equilibrium. Considerable institutional disparities also translate into a large diversification in economic development. The article proposes a different meaning of institutional equilibrium, understood as the achieved state of institutional structure characterised by high quality informal institutions which interact with each other to improve the efficiency of formal institutions. The article presents a comprehensive model of the institutional structure and a unique method of measuring institutional equilibrium.

Key words: institutional equilibrium, SEM-PLS, economic growth and development.

1. Introduction

The institutional approach is gaining popularity today. For many years, institutions in macroeconomic models have been covered by the ceteris paribus assumption, or treated as an undoubted pro-development factor. However, an increasing number of researchers have taken up the topic of institutional structure in search of the sources of economic failures. It turns out that the inefficiency of the system may be the cause of development disparities and their increase in the world.

[1] Doctoral School in the Social Sciences (economics and finance), University of Bialystok, Bialystok, Poland, E-mail: m.borkowski@uwb.edu.pl, ORCID: https://orcid.org/0000-0003-0644-4764.

In modern economic theory there is a noticeable gap in the modelling of institutions. Existing models that take into account institutional variables are based mainly on simple correlation and regression analyses. Most often they concern the quality of only one selected institution. There is a noticeable lack of econometric models of the entire institutional system in the social literature. Moreover, the measurement of institutional equilibrium is rare. This article is an attempt to complement institutional theory with tools measuring institutional quality and levels of institutional equilibrium.

The problem of institutional equilibrium is gaining interest among scholars from all over the world. Interestingly, the understanding of institutional equilibrium varies. The most common assumption is that institutions themselves are a kind of equilibrium in a game (Hindriks & Guala, 2015). This paper proposes that institutional equilibrium can be understood as an achieved state of institutional structure that is characterized by high quality informal institutions that interact to improve the efficiency of formal institutions.

The purpose of the research is to identify the strength and direction of the relationship between formal and informal institutions and to assess the institutional equilibrium of modern economies. The problem addressed is the differentiation of EU economies in terms of the quality of institutional systems. The paper adopts three research hypotheses:

H_1: Informal institutions positively and strongly influence the quality of formal institutions in the EU countries.

H_2: The relationship between informal and formal institutions is getting weaker over the time (from 2008 to 2018).

H_3: Countries of a higher level institutional equilibrium feature economies with a higher level of GDP per capita.

This paper applies structural equation modelling using the partial least squares method (SEM-PLS). The years 2008 and 2018 were selected as the period of research, as these are the most recent statistical data available. The study covered 27 EU economies.

2. Literature review

Defining institutions is not a simple task. The reason for the difficulty in conceptualizing this term is its multidimensional and interdisciplinary nature. Differences in explaining the meaning of institutions arise not only in different social science disciplines, but also within those disciplines (Godłów-Legiędź, 2010, p. 65). Within the economic sciences there are three main approaches to defining institutions (Gancarczyk, 2002, p. 82). The first assumes that institutions are norms or customs that are embedded in the economy (processes). Second, institutions are identified with organizations. The third one equates institutions with a state of equilibrium in a game – a strictly model-based approach.

This article uses the definition by G. M. Hodgson. Institutions are a system of embedded and well-established both formal and informal norms, rules, customs, which influence economic, social and political interactions among individuals in the economy (Hodgson, 2006, p. 18). The work uses a process approach, which means that institutions (processes) and organizations (entities) are related concepts, although not identical.

Institutions are characterized by the following features:

- universality (Vitola & Šenfelde, 2015, p. 278) – they are universal in nature, affecting all relations in the economy,
- variability over time – they change, evolve; changes depend on the type of institution and the elasticity of the institutional system; change can take the form of: complete displacement, layering, drift or conversion (Mahoney & Thelen, 2010, p. 16),
- immateriality and direct immeasurability – the quality of institutions cannot be directly observed, institutions cannot be seen (Ostrom, 2008, p. 822),
- heterogeneity – each institution is unique, original,
- endogenous nature – they arise within the society/economy – either created by people consciously or unconsciously,
- internal complexity – the institutional system consists of many institutions, which also have components, and components have elements and so on,
- internal interdependence, which can take the form of:
  a) complementary relationships – institutions function in the environment of other institutions, they can complement and strengthen each other (Höpner, 2005, p. 333),
  b) mutual exclusion, competitive relationships (Amable, 2016, p. 79) – institutions can also be an obstacle for the functioning of other institutions, they can mutually limit each other, weaken incentives for interaction (Helmke & Levitsky, 2004, p. 729),
  c) relationships of substitutability – outdated institutions are replaced by new ones, better suited to the conditions of the present (Gruszewska, 2011, p. 55),
- dependence on the past – new institutions are the product of past socio-economic processes, they are ideally suited to past conditions, but will never be in line with the conditions of the present (Veblen, 2016, s. 88).

Institutions are of undeniable importance in the economy. All relations, whether economic, social or political, are regulated by institutions. They give a sense of action to all units in the economy, create a safe area for functioning, and thus contribute to increasing the predictability of participants in socio-economic processes. It would seem that the most important task of institutions in modern economies is to determine the possible solutions, create opportunities, and also to set the rules for all units in the economy (Gruszewska, 2013, p. 136).

When studying the institutions in modern world economies, one should focus on the analysis of the institutional system. In institutional theory there are many divisions of the institutional system of the economy. The most widely used in the literature is the division proposed by D. C. North (1992; 1994), according to which the institutional system consists of three elements: formal institutions, informal institutions and the mechanisms for their enforcement. It is this view of the institutional system that was applied in this paper. Special attention has been paid especially to formal and informal institutions.

Formal institutions have a statutory character, and are the result of the activities of the governance. Most often they are written down in the form of normative acts. They can also take the form of actions – for example, markets' regulations. Their specificity makes their variability over time much greater than in the case of informal constructions (Fuentelsaz et al., 2019, pp. 6–8). Their boundaries of change are determined by informal institutions, which are the core of the entire institutional system. The components of the formal institutional environment include the institutions of (Rodrik, 2007, pp. 150–161): legal order, property rights, macroeconomic stability, regulation, social security, and conflict management.

Informal institutions are the second main component of the institutional system. In contrast to formal ones, they arise spontaneously, endogenously (Seidler, 2011). They are not written down, but deeply rooted in the mentality of society. They change very slowly, thus conditioning changes in the entire institutional system (Mohmand, 2015, p. 7). Changes in formal rules, which can be introduced by the governance in a relatively short time, are limited by informal institutions. New formal norms are not immediately aligned with social norms. There is a dissonance between formal and informal institutions. The community, only after some time, adapts to the new formal structures (Gruszewska, 2017, p. 41). The informal institutions include (Fiedor, 2015, p. 94): culture (including economic culture), attitudes towards religion, behavioural patterns, social trust, and the so-called "mental models", i.e. established behavioural patterns.

The continuous adjustment processes of formal institutions to informal ones show that the institutional system is in a constant disequilibrium. The degree of institutional disequilibrium varies. As J. Wilkin points out, institutional equilibrium is a state, not a point, at which: various needs of the members of society are balanced; there is an inclination of the members of society to follow the established rules of conduct, which have been considered socially beneficial, with the possibility of choosing to achieve their goals; the continuity of prevailing rules and social mechanisms is guaranteed and a high degree of predictability of other members of society is ensured (Wilkin, 2011, p. 32).

The relationship between informal and formal institutions and the enforcement mechanisms that support them can be the basis for defining institutional equilibrium. B. Fiedor (2019, p. 176) defines institutional equilibrium as a state in which informal institutions strengthen and positively influence the enforcement of formal rules, and strengthen the enforcement mechanisms. Institutional equilibrium is of a higher importance than other equilibrium found in the economy. Institutions are considered to be the foundations of the economy, they are a form of security and a stabilizer for the economy (Wilkin et al., 2019, p. 662–663). J. Platje distinguishes five levels of institutional equilibrium depending on the efficiency of formal and informal institutions and enforcement mechanisms (Table 1).

Table 1. Levels of institutional equilibrium according to J. Platje

No.	Efficiency of*:			Level of institutional equilibrium
	formal institutions	informal institutions	institutional governance	
1.	+	+	+	ideal institutional equilibrium
2.	+	+	–	weak institutional equilibrium
3.	+	–	+	institutional disequilibrium
4.	+	–	–	
5.	–	+	+	
6.	–	+	–	
7.	–	–	+	strong institutional disequilibrium
8.	–	–	–	ideal institutional disequilibrium

* "+" – high, "–" – low.

Source: own work on the basis of: (Platfje, 2008, p. 147).

This paper applies the institutional equilibrium matrix, proposed and then empirically used by C. R. Williamson (2009, p. 373), to assess the institutional balance of modern economies. The institutional system is divided into formal and informal institutions. Their quality determines the level of institutional equilibrium. Strong informal and formal rules create conditions, which allow obtaining benefits to be obtained by all individuals functioning in society.

This paper assumes that institutional equilibrium is defined by the quality of both formal and informal institutions. The state of institutional equilibrium can take three forms: strong institutional equilibrium (high quality of both formal and informal rules); weak formal institutional equilibrium (high quality of formal institutions, low of informal ones) and weak informal institutional balance (low quality of formal institutions, high of formal ones). When both formal and informal institutions are characterized by low quality, this implies institutional disequilibrium.

3. Research method – SEM-PLS

The assumptions of the paper were met using partial least squares structural equation modelling (SEM-PLS or PLS-SEM), which was created by H. Wold (1980). SEM-PLS is an econometric method for studying phenomena that are not directly observable

(Ciborowski & Skrodzka, 2020, p. 1355). SEM-PLS is one of two structural equation modelling techniques – the other, a more restrictive one, is covariance-based structural equation modelling (CB-SEM). SEM modelling strongly combines empirics with theory (Skrodzka, 2016, p. 283). The use of SEM-PLS, rather than CB-SEM, seems appropriate for the topic under study. Several arguments support this, including (Hair et al., 2011, pp. 139–141): (1) institutions do not have an elaborated theory of econometric modelling, so the aim of the study is not to test theory but to create a new one; (2) the number of observations is rather small (27 EU countries); (3) the data do not follow a normal distribution (characteristics of macroeconomic data); (4) it is planned to use the values of latent variables to linearly order the objects in terms of the level of directly unobservable phenomena.

Each SEM-PLS model consists of two sub-models: an internal (structural) one and an external (measurement) one (Skrodzka, 2016, pp. 282–283). The first one describes the relationships between latent variables, while the second one presents the relationships between latent variables and their diagnostic variables. The general form of the internal model is presented in Formula 1.

$$\xi_j = \alpha_{0j} + \sum_{j\to p} \alpha_{qj}\xi_q + \varepsilon_j \tag{1}$$

where: ξ_j –j-th endogenous latent variable; ξ_q –q-th exogenous latent variable; α_{0j} – location parameter of the internal relationship for the endogenous variable; α_{qj} – structural parameter of the internal model showing the link between the q-th exogenous variable and the j-th endogenous variable; ε_j – random error of the internal relation for j-th endogenous variable.

There are two types of relationships between latent structures and their explanatory variables in the external model: weighting (2) and reflective (3). The first one assumes that the latent variables are linear combinations of their explanatory indicators. Reflective relations represent the strength of the "reflection" of an unobservable feature by its explanatory variables (Rogowski, 1990, pp. 36–37).

$$\xi_{tj} = \sum_{i=1} w_{ij}x_{tij} \tag{2}$$

where: ξ_{jt} – t-th value of the j-th latent variable; x_{ijt} – t-th value of i-th indicator explaining j-th latent variable; w_{ij} – weight of i-th indicator explaining j-th latent variable.

$$x_{ij} = \pi_{ij0} + \pi_{ij}\xi_j + \mu_{ij} \tag{3}$$

where: π_{ij0} – location parameter of reflective relationship; π_{ij} – factor loading, the relationship of reflecting the j-th latent variable by the i-th indicator; μ_{ij} – random element whose expected value is equal to zero.

Latent variables can be determined in two ways: deductively and inductively. In the deductive approach, the explanatory indicators are reflective, whereas in the inductive analysis they are formative. The reflective indicators should be highly correlated with each other, while the formative ones are not (Hair et. al., 2014, pp. 46–47). Depending on the approach used, different measures of statistical validation are used. SEM-PLS proceeds in steps (Lohmöller, 1989, pp. 30–31): (1) First, the values of the weights are estimated. The estimation of weights is iterative. Estimation of internal values of weights can be done using the centroid, factorial or path scheme (preferred, used in this article). (2) Next, the values of the latent variables are calculated according to Formula 2. (3) The next step is to calculate the values of factor loadings for the external model and the parameters of the internal model using OLS. (4) The final step is to determine the location parameters for the reflective and internal relationships (optional step in cross-sectional models).

The estimated SEM-PLS model needs to be verified. The validation starts with the substantive analysis. It is assessed whether the model is consistent with the initial assumptions and theory. It is also necessary to check the signs of the model parameters. Statistical verification involves the use of appropriate measures to assess specific properties of the model. Table 2 presents the measures and verification criteria divided into those appropriate for a structural model, an external model defined inductively (formative indicators) and an external model defined deductively (reflective indicators).

Table 2. Verification measures and criterions of SEM-PLS model

Versification measure	**Brief description**	**Verification criterion**
validation of structural model		
variance inflation factor (VIF)	By using the VIF measure, collinearity of exogenous variables is checked.	VIF < 5.00
coefficient of determination (R^2)	A classic measure of econometrics, it determines how much of the variation in an endogenous latent variable is explained by exogenous latent structures.	lack of standard
standard deviation of parameter (S_α)	The standard errors of the parameters are obtained using the bootstrapping procedure. The full evaluation of the significance of the parameters proceeds as in classical econometrics - t-student test. Alternative measure: standard deviations calculated using Tukey's Jackknifing method - the "2s" rule for significance testing.	p-value < significance level
Stone-Gaisser test value (S-G)	Assessment of predictive ability. The S-G test value is obtained from the blindfolding procedure. Data for the model are blindfolded L times. Every L-th element is blindfolded and replaced by, for example, the arithmetic mean of the others. Based on the substitution relationships, predictions are determined from the SEM-PLS model, which can be used to calculate S-G test value. (L should belong to the interval <5,10>).	S-G ≥ 0.00

Table 2. Verification measures and criterions of SEM-PLS model (cont.)

Versification measure	Brief description	Verification criterion
validation of outer model (formative approach)		
variance inflation factor (VIF)	In formative outer models indicators forming a latent variable should not be highly correlated with each other.	VIF < 5.00
standard deviation of weight (S_{α})	Same as for testing the significance of the internal relationship parameter.	
validation of outer model (reflective approach)		
Cronbach's α composite reliability (pc)	Internal consistency verification. Reflective indicators should be highly correlated with each other.	0,95 > Cb's α and pc > 0.70
π_{ij} – factor loading value average variance extracted (AVE)	Convergent reliability validation. Variables that have less than 0.40 strength of correlation with the latent variable should be removed. Latent construct should extract more than 50% of total variability.	$\pi_{ij} \geq 0.40$ AVE ≥ 0.50
standard deviation of factor loading (S_{α})	Same as for testing the significance of the internal relationship parameter and weights.	
cross loadings analysis	Discriminatory validity assessment. Indicators of a given latent variable should be the ones that correlate most strongly with that variable. Alternatives: Fornell-Larcker criterion or Heterotrait-monotrait ratio (HTMT).	-

Source: own work on the basis of: (Hair et al., 2014; Rogowski, 1990).

Two computational packages from the R environment will be used to estimate the SEM-PLS model: cSEM (Rademaker & Schuberth, 2021) and SEMpls (Monecke & Leisch, 2012).

4. Data

A precise quantitative analysis of the quality of institutions is, and probably will always be, impossible. This is mainly because institutions are deeply embedded in society. Contemporary attempts to assess the quality of institutions are based on measures prepared by inter-national statistical organizations. Many institutional researchers deny the use of such indicators. They believe that the study of institutions can only have a qualitative dimension (Skarbek, 2020, p. 409).

Table 3. Selected measures of quality of institutions

Statistical organization	Report/ roup of measures	Formal or informal	Range values
World Bank	The Worldwide Governance Indicators (WGI)	formal	<-2.5;2.5>
	Doing Business (DB)	formal	varied
Heritage Foundation	Index of Economic Freedom	mainly informal	<0;100>
Fraser Institute	Economic Freedom in the World (EFW)	formal	<0;10>
Fraser Institute & Cato Institute	Human Freedom in the World (HFW)	informal	<0;10>
Freedom House	Freedom in the World (FIW)	mainly informal	varied

Source: own work.

Doubts about the use of these types of metrics seem justified. The greatest objections arise for methodological reasons. Institutional indicators are more often created on the basis of surveys or experts' opinions rather than on the basis of "hard" data. Although such measures do not reflect the reality in a one-to-one ratio, they give some general approximation of the quality of institutions. However, in the opinion of many researchers of institutions (Balcerzak, 2020; Miłaszewicz & Nermend, 2020; Nifo & Vecchione, 2015), such measures can be used to assess the quality of institutions. Nevertheless, the interpretation of the results should be approached carefully. Table 3 presents a brief description of the indicators used.

5. Specification of the SEM-PLS model

Figure 1 shows the specification of the SEM-PLS model that will be estimated in this paper. The model consists of two latent variables: quality of informal institutions (INF) and quality of formal institutions (FOR). The explanatory variables of the latent constructs are defined deductively (reflective indicators).

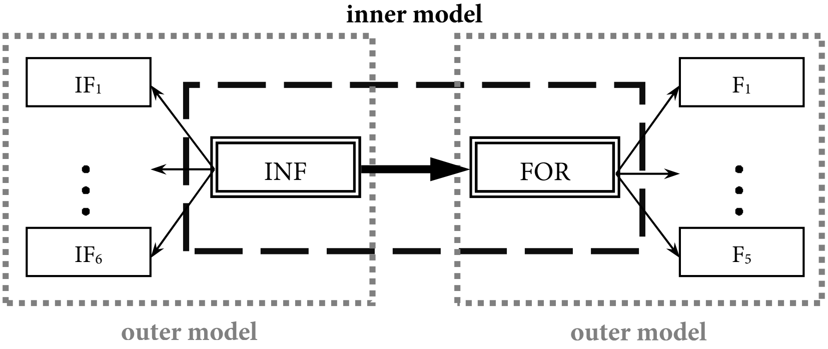


Figure 1. Specification of SEM-PLS model applied in the article

Source: own work.

The selected explanatory indicators for each latent variable are presented in Table 4. The FOR variable (quality of formal institutions) is reflected by five variables, which represent the quality of law system, property rights, regulatory institutions and institutions for macroeconomic stabilization. While INF (quality of informal institutions) is explained by six variables pertaining to freedom (personal, political and economic) and culture (religion, social behaviour).

Table 4. Measures of the quality of institutions (outer model specification)

Symbol	Variable	Source of data
the quality of formal institutions (FOR)		
F_1	Rule of Law	World Bank (WGI)
F_2	Legal enforcement of contracts	Fraser Institute (EFW)
F_3	Business regulations	
F_4	Regulation	
F_5	Property Rights	Heritage Foundation
the quality of informal institutions (INF)		
IF_1	Media Freedom	Fraser Institute & Cato Institute (HFW)
IF_2	Expression & Information	
IF_3	Association, Assembly, & Civil Society	
IF_4	Freedom of Expression and Belief	Freedom House (FIW)
IF_5	Personal Autonomy and Individual Rights	
IF_6	Business Freedom	Heritage Foundation

Source: own work.

The presented set of diagnostic variables was selected on the basis of substantive and statistical (classical coefficient of variation higher than 5% and positively verified SEM-PLS model) evaluation. Variables: IF_2 and IF_3 are characterized by a slightly lower coefficient of variation than 5%. Nevertheless, the variables remained in the study because of their substantive relevance.

The internal sub-model is in the form of a single equation (Formula 4). The formula represents the dependence of informal institutions (INF) on formal ones (FOR). The relationship was determined on the basis of theoretical analysis. It is the informal norms that are of fundamental importance in the economy, they affect the entire institutional system, but also are the basis for the establishment of formal institutions.

$$\mathrm{FOR_t} = \alpha_1 \mathrm{INF_t} + \alpha_2 + \varepsilon_t \tag{4}$$

The values of the latent variables will be used to construct an institutional equilibrium matrix to divide economies into four typological groups of equilibrium levels (Figure 2). Countries will be divided into those with: institutional equilibrium, weak formal equilibrium, weak informal equilibrium and institutional disequilibrium.

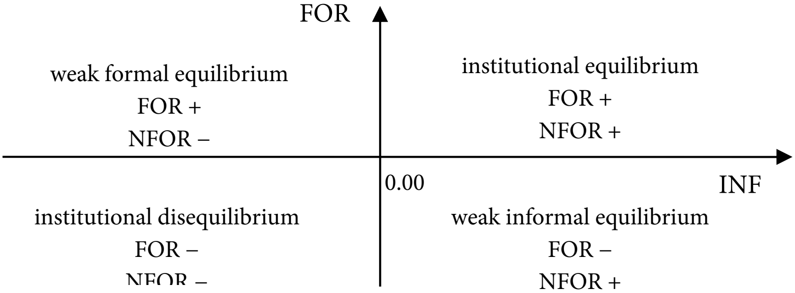

FOR
weak formal equilibrium
FOR +
NFOR –
institutional equilibrium
FOR +
NFOR +
0.00
INF
institutional disequilibrium
FOR –
NFOR –
weak informal equilibrium
FOR –
NFOR +

Figure 2. Institutional equilibrium matrix

Source: own work.

6. Research findings and discussion

6.1. Institutional equilibrium in the EU countries in 2008 – SEM-PLS results

Table 5 presents the estimates of the external sub-model of SEM-PLS model for 2008. The significance of the factor loadings was checked using the bootstrapping procedure. The number of samples was set at 5 000. At a significance level of 5% ($p < 5\%$), it can be concluded that all parameters are significantly different from zero. All indicators, both of the INF and FOR latent variables, are consistent in sign – they are all stimulants, which is consistent with the initial assumptions and economic theory.

Post-measurement convergent reliability is also observed – the values of factor loadings are greater than 0.4000. In addition, variables with a loading factor value of less than 0.7000 were examined in detail. Moreover, latent variables explain over 50% of total variability of unobservable phenomena. Based on the results, the internal consistency of the latent variables can be concluded (internal consistency measures takes values above 0.7000 and under 0.9500).

The strongest correlated indicator with the latent variable INF is IF_1 (0.9212), which is the media freedom variable. The least correlated is IF_3 (0.5530) – an indicator describing Association, Assembly & Civil Society in the economy. The values of the FOR variable are most strongly reflected by F_1 (0.9412) – a synthetic measure of the rule of law. The lowest factor loading of the FOR variable is found with F_4 (0.4027), a variable describing the general quality of regulation.

Table 5. Parameters of the outer sub-model (SEM-PLS model for 2008)

Symbol	Factor loading *(st. dev.)*	t stat	p-value	AVE	α-Cb	pc
	the quality of formal institutions (FOR)					
F_1	0.9412 *(0.0231)*	40.6819	0.0000	0.6437	0.8485	0.8947
F_2	0.8134 *(0.0697)*	11.6706	0.0000			
F_3	0.8005 *(0.0880)*	9.0936	0.0000			
F_4	0.4027 *(0.1963)*	2.0513	0.0402			
F_5	0.9318 *(0.0187)*	49.8499	0.0000			
	the quality of informal institutions (INF)					
IF_1	0.9212 *(0.0237)*	38.9112	0.0000	0.5716	0.8440	0.8859
IF_2	0.6619 *(0.1040)*	6.3661	0.0000			
IF_3	0.5530 *(0.1932)*	2.8620	0.0042			
IF_4	0.8049 *(0.0679)*	11.8502	0.0000			
IF_5	0.8721 *(0.0427)*	20.4331	0.0000			
IF_6	0.6545 *(0.1198)*	5.4634	0.0000			

Source: own work.

Table 6 presents cross loadings between FOR and INF variables in SEM-PLS model for 2008. The model has good discriminative abilities - the indicators were properly assigned to the latent structures in the model. The measurement model is considered to be positively validated.

Table 6. Cross loadings between latent variables in SEM-PLS model (2008)

Symbol	FOR	INF
F_1	**0.9412**	0.8833
F_2	**0.8134**	0.6162
F_3	**0.8005**	0.6182
F_4	**0.4028**	0.3077
F_5	**0.9317**	0.9003

Symbol	FOR	INF
IF_1	0.8371	**0.9212**
IF_2	0.4458	**0.6619**
IF_3	0.3873	**0.5531**
IF_4	0.6765	**0.8049**
IF_5	0.7886	**0.8721**
IF_6*	0.6834	0.6546

* Variable IF6 – Business Freedom – correlates a bit stronger with FOR than INF. Nevertheless, this variable remained in the modelling due to its substantive relevance

Source: own work.

The quality of informal institutions strongly, positively determine (0.8771) the quality of formal institutions (Formula 5). This is consistent with theory. Informal institutions are the core of every institutional system. The parameter at the latent variable

INF is statistically significant at the 1% level (p < 1%). The variability of FOR is explained in more than 77% by the variability of INF – the result should be considered satisfactory. The SEM-PLS (2008) model also has fairly good predictive ability (S-G test for the FOR variables at 10 folds is equal to 0.45).

$$\widehat{\mathrm{FOR}}_{2008} = \underset{(0.0376)}{0.8771^{***}}\mathrm{INF}_{2008} - 8.7850 \tag{5}$$

The SEM-PLS model estimated for data from 2008 is considered to be positively verified both substantively and statistically. The estimated SEM-PLS model allowed to estimate the values of the latent variables of the quality of formal institutions (FOR) and informal institutions (INF) for the 27 EU economies. Figure 3 presents the institutional equilibrium matrix for the EU economies in 2008. Countries were divided into four typological groups ac-cording to the level of institutional equilibrium. Institutional equilibrium was recorded in 11 economies, weak informal equilibrium in 5, institutional disequilibrium in 11. There was no countries with weak formal institutional equilibrium in 2008.

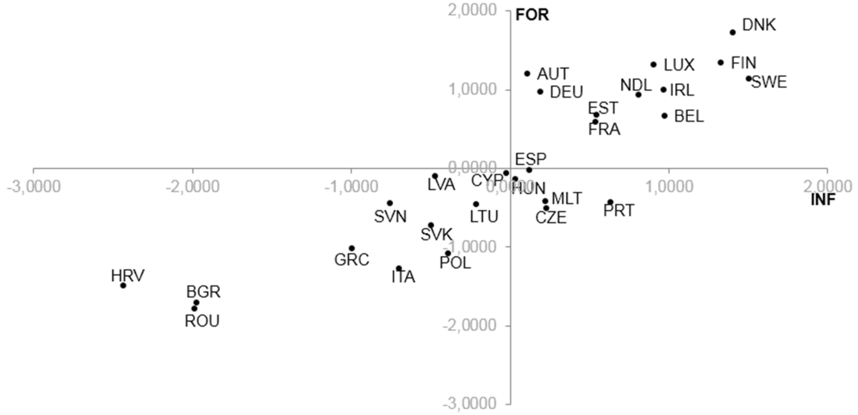


Figure 3. Institutional equilibrium matrix in 27 EU economies in 2008

Source: own work.

The results show that researched economies are diversified in terms of the quality of institutional equilibrium. In 2008, institutional equilibrium was mainly found in highly developed EU countries (e.g. Denmark, Sweden, Finland, Luxemburg, Germany), while institutional disequilibrium was recorded mainly in underdeveloped countries (Bulgaria, Romania, Greece).

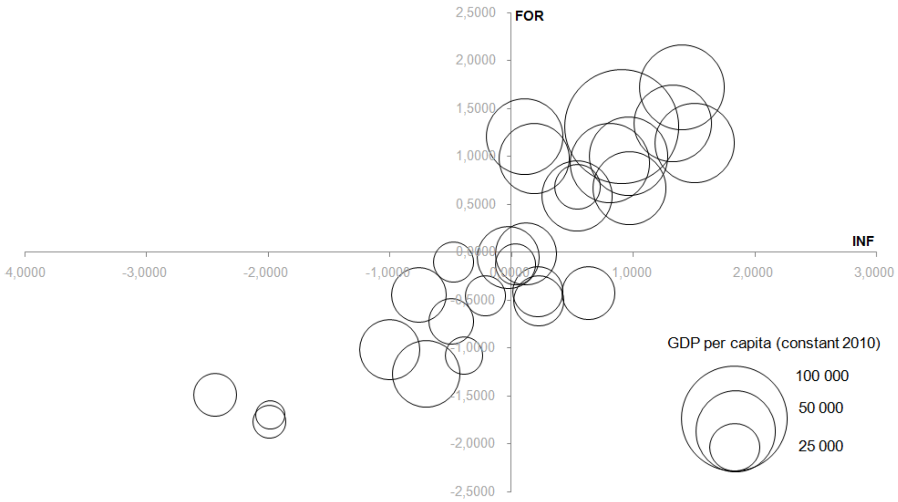


Figure 4. Institutional equilibrium and GDP per capita in 2008 (27 EU economies)

Source: own work.

Analysis of the statistical data may allow one to conclude that as institutional equilibrium improves, the level of GDP per capita in the economy rises. The average level of GDP per capita in economies with an observed institutional equilibrium is more than $51 thousand, while the average level of GDP per capita in countries with institutional imbalances is the lowest, at about $19 thousand. Figure 4 presents the institutional equilibrium matrix and GDP per capita in 2008 for 27 researched EU economies. As it turns out, institutional systems in developed economies are in institutional equilibrium.

6.2. Institutional equilibrium in the EU countries in 2018 – SEM-PLS results

The parameter estimates of the outer sub-model of the SEM-PLS model of the dependence of the quality of formal institutions on the quality of informal ones was presented in Table 7. All parameters are statistically significant at the $p < 1\%$ level. Moreover, outer sub-model is coincident. Cronbach's α and composite reliability values indicate the internal consistency of the latent variables. There is also convergent validity noted.

The strongest changes in the value of the latent variable informal institutions (INF) are reflected by the synthetic indicator representing media freedom (IF_1, 0.9066). The Association, Assembly & Civil Society (IF_3, 0.7879) variable is the least correlated with the latent variable INF. The formal institutions (FOR) variable is reflected by the rule of law measure (F_1, 0.9480) in the strongest way, while the general regulation indictor (F_4, 0.5935) has the lowest factor loading value. The results are similar compared to the sub-model estimated for data from 2008.

Table 7. Parameters of the outer sub-model (SEM-PLS model for 2018)

Symbol	Factor loading *(st. dev.)*	t stat	p-value	AVE	α-Cb	pc
the quality of formal institutions (FOR)						
F_1	0.9480 *(0.0162)*	58.4582	0.0000	0.7305	0.9031	0.9297
F_2	0.8351 *(0.0640)*	13.0567	0.0000			
F_3	0.9258 *(0.0179)*	51.8289	0.0000			
F_4	0.5935 *(0.1299)*	4.5685	0.0000			
F_5	0.9203 *(0.0234)*	39.2777	0.0000			
the quality of informal institutions (INF)						
IF_1	0.9066 *(0.0280)*	32.3694	0.0000	0.7147	0.9198	0.9375
IF_2	0.8278 *(0.0570)*	14.5295	0.0000			
IF_3	0.7879 *(0.0960)*	8.2037	0.0000			
IF_4	0.8805 *(0.0456)*	19.3152	0.0000			
IF_5	0.7911 *(0.0531)*	14.9022	0.0000			
IF_6	0.8713 *(0.0648)*	13.4453	0.0000			

Source: own work.

Table 8 contains a cross loadings between latent variables in SEM-PLS model estimated for data from 2018. Cross loadings values indicate that the variables were correctly assigned to the latent structures. The discriminant ability of the external model can be positively validated.

Table 8. Cross loadings between latent variables in SEM-PLS model (2018)

Symbol	FOR	INF
F_1	**0.9480**	0.7694
F_2	**0.8351**	0.4930
F_3	**0.9259**	0.8188
F_4	**0.5935**	0.3909
F_5	**0.9203**	0.7571

Symbol	FOR	INF
IF_1	0.7232	**0.9066**
IF_2	0.6404	**0.8278**
IF_3	0.4787	**0.7879**
IF_4	0.6494	**0.8805**
IF_5	0.7563	**0.7911**
IF_6	0.6885	**0.8713**

Source: own work.

Latent variable FOR is strongly, positively (0.7891) determined by INF latent variable (Formula 6). The relationship is statistically significant at the level of 1%. Again, the thesis that informal institutions are the core of the institutional system is confirmed. The coefficient of determination is at the level of 0.62, which indicated quite good, but satisfactory, model fit. S-G test value (10 folds) is equal to 0.43 – SEM-PLS model has fairly good abilities to predict blindfolded observations.

$$\widehat{\mathrm{FOR}}_{2018} = \underset{(0.0543)}{0.7891^{***}} \mathrm{INF}_{2018} - 1.8568 \tag{6}$$

The analysed SEM-PLS model for 2018 is considered to be positively verified both in terms of substantial and statistical criterions. The consequence is that the latent variable values can be used for the institutional equilibrium designation.

Figure 5 shows the institutional equilibrium matrix for the 27 EU economies in 2018. Institutional equilibrium was recorded in 11 economies, weak informal equilibrium in 2, weak formal equilibrium in 2. The remaining EU countries (12) were classified into the group of countries with institutional disequilibrium.

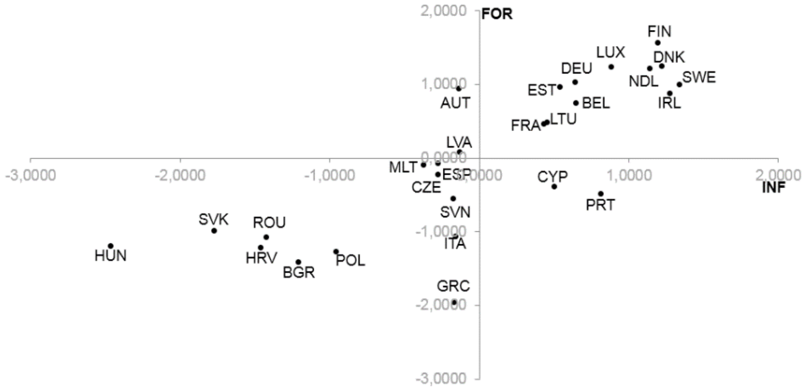


Figure 5. Institutional equilibrium matrix in 27 EU economies in 2018

Source: own work.

Institutional equilibrium is characteristic of highly developed countries (e.g. Finland, Denmark, Sweden, Ireland, Netherlands or Germany) in the European Union, while institutional disequilibrium occurs in economies of a low level of economic development (e.g. Bulgaria, Romania, Slovakia).

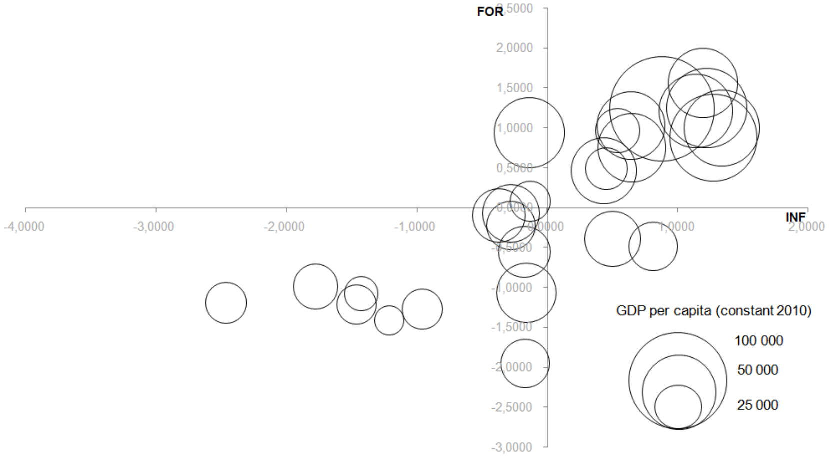


Figure 6. Institutional equilibrium and GDP per capita in 2018 (27 EU economies)

Source: own work.

The mean level of GDP per capita in the EU countries with institutional equilibrium is more than $53 thousand in 2018. As institutional equilibrium gets worse, the level of GDP per capita in the economy falls down. Countries with institutional disequilibrium achieve relatively low levels of GDP per capita on average (approximately $21 thousands in 2018). Figure 6 presents the equilibrium matrix and GDP per capita in 2018 for EU economies. Sustainable institutional systems imply higher levels of economic development. It turns out that institutional equilibrium is an important factor of economic development of modern world economies.

7. Conclusions

The main aim of the article was to identify the relationship between formal and informal institutions, as well as to measure and assess the institutional equilibrium of EU economies. The aim of the paper was achieved using SEM-PLS modelling.

Three research hypotheses are considered to be positively verified. As it turned out, the efficiency of informal institutions strongly, positively determines the quality of formal institutions. This is evidenced by the parameter of the internal relationship, which is equal to 0.8771 in 2008 and 0.7891 in 2018. The obtained results are consistent with economic and institutional theory. Informal institutions, which are the "core" of the institutional system, interact with formal ones. They strengthen their operation, but also set certain limits of their change. The strength of the relation between informal and formal rules is getting weaker over time. It seems that there is a trend in the EU economies towards disintegration rather than integration of the institutional structure. Moreover, institutional equilibrium positively influences the dynamics of economic development processes. The higher the level of institutional equilibrium, the higher, on average, the level of earned income.

The constructed models allowed for the assessment of the quality of formal and informal institutions, which enabled the construction of the institutional equilibrium matrix. In 2008, the highest efficiency of formal institutions was in Denmark and the lowest in Romania. In 2008, informal institutions were the strongest in Sweden and the weakest in Croatia. In 2018, Finland led the classification in terms of the FOR latent variable, while Greece closed the ranking. Sweden was characterised by the strongest informal institutions in 2018. The lowest quality of informal institutions in 2018 was observed in Hungary (this was also the largest fall in the ranking - from 16th place in 2008 to 27th place in 2018). Changes in the level of institutional equilibrium were not major. Noteworthy is the improvement in Lithuania, where institutional disequilibrium was in 2008 and institutional equilibrium in 2018 (the largest improvement among the EU countries in 2018, compared to 2008).

The proposed research method can be a beneficial tool for monitoring the relationship between formal and informal institutions. Moreover, the concept of measuring institutional equilibrium, admittedly very simple, can be a useful mechanism for institutional analysis.

The analyses carried out in this paper indicate that EU economies are diversified in terms of the quality of informal and formal institutions and, consequently, in institutional equilibrium. Large institutional disparities also translate into a large diversification in economic development. This problem would still appear to be still relevant and topical.

References

Amable, B., (2016). Institutional complementarities in the dynamic comparative analysis of capitalism. *Journal of Institutional Economics*, Vol. 12, No. 1, pp. 79–103, doi: 10.1017/S1744137415000211.

Balcerzak, A., (2020). Quality of Institutions in the European Union countries. Application of TOPSIS Based on Entropy Measure for Objective Weighting. *Acta Polytechnica Hungarica*, Vol. 17, No. 1, pp. 101–122, doi: 10.12700/APH. 17.1.2020.1.6.

Ciborowski, R. W., Skrodzka, I., (2020). International technology transfer and innovative changes adjustment in EU. *Empirical Economics*, Vol. 59, No. 3, pp.1351–1371. doi: 10.1007/s00181-019-01683-8.

Fiedor, B., (2015). Instytucje formalne i nieformalne w kształtowaniu trwałego rozwoju [Formal and Informal Institutions in Shaping Sustainable Development]. *Zeszyty Naukowe Uniwersytetu Szczecińskiego. Studia i Prace Wydziału Nauk Ekonomicznych i Zarządzania*, Vol. 2, No. 40, pp. 83–107, doi: 10.18276/sip.2015.40/2-07.

Fiedor, B., (2019). Proces transformacji a nowa ekonomia instytucjonalna (NEI): Koszty transakcyjne i równowaga instytucjonalna [Process of transformation vs. New Institutional Economy (NIE): Transaction Costs and Institutional Equilibrium]. *In Ekonomia i polityka. Wokół teorii Grzegorza W. Kołodko* [*Economics and Politics. Around the Theory of Grzegorz W. Kołodko*], E. Mączyńska (eds.), Warszawa: PWN, pp. 164–182.

Fuentelsaz, L., González, C., Maicas, J. P., (2019). Formal institutions and opportunity entrepreneurship. The contingent role of informal institutions. *BRQ Business Research Quarterly*, Vol. 22, No. 1. pp. 5–24. doi: 10.1016/j.brq. 2018.06.002.

Gancarczyk, M., (2002). Instytucja a organizacja w nowej ekonomii instytucjonalnej [Institution and Organisation in New Institutional Economics], *Gospodarka Narodowa. The Polish Journal of Economics*, Vol. 176, No. 5-6, pp. 78–94, doi: 10.33119/gn/113844.

Godłów-Legiędź, J., (2010). Współczesna ekonomia: Ku nowemu paradygmatowi? [*Contemporary economics: Towards a new paradigm?*], Warszawa: C. H. Beck.

Gruszewska, E., (2011). Dezintegracja w zinstytucjonalizowanym świecie [Disintegration in an Institutionalized World]. *Ekonomia i Prawo*, Vol. 7, No. 1, pp. 49–65, doi: 10.12775/eip.2011.003.

Gruszewska, E., (2013). Instytucje a proces tworzenia kapitału w Polsce [*Institutions vs. the process of capital formation in Poland*]. Białystok: Wydawnictwo Uniwersytetu w Białymstoku.

Gruszewska, E., (2017). Instytucje formalne i nieformalne. Skutki antynomii [Formal and Informal Institutions. Results of Antinomy]. *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu*, No. 493, pp. 36–50, doi: 10.15611/ pn.2017.493.03.

Hair, J. F., Hult, G. T. M., Ringle, C., Sarstedt, M., (2014). *A Primer on Partial Least Squares Structural Equation Modeling*, Los Angeles, London, New Delhi, Singapore, Washington DC, Melbourne: SAGE Publications.

Hair, J. F., Ringle, C. M., & Sarstedt, M., (2011). PLS-SEM: Indeed a Silver Bullet. *Journal of Marketing Theory and Practice*, Vol. 19, No. 2, pp. 139–152, doi: 10.2753/mtp1069-6679190202.

Helmke, G., Levitsky, S., (2004). Informal Institutions and Comparative Politics: A Research Agenda. *Perspectives on Politics*, Vol. 2, No. 4, pp. 725–740, doi: 10.1017/S1537592704040472.

Hindriks, F., Guala, F., (2015). Institutions, rules, and equilibria: A unified theory*. *Journal of Institutional Economics*, Vol. 11, No. 3, pp. 459–480, doi: 10.1017/S1744137414000496.

Hodgson, G. M., (2006). What Are Institutions? *Journal of Economic Issues*, Vol. 40, No. 1, pp. 1–25, doi: 10.1080/00213624.2006.11506879.

Höpner, M., (2005). What connects industrial relations and corporate governance? Explaining institutional complementarity. *Socio-Economic Review*, Vol. 3, No. 2, pp. 331–358, doi: 10.1093/SER/mwi014.

Lohmöller, J.-B., (1989). *Latent Variable Path Modeling with Partial Least Squares*, Verlag, Berlin, Heidelberg: Springer.

Mahoney, J., Thelen, K., (2010). *Explaining Institutional Change: Ambiguity, Agency, and Power*, Cambridge: Cambridge University Press.

Miłaszewicz, D., Nermend, K., (2020). Application of Vector Measure Construction Methods to Estimate Quality of Institutions: Nations in Transition. *European Research Studies Journal*, Vol. XXIII, No. special 2, pp. 16–29. doi: /10.35808/ersj/1805.

Mohmand, S. K., (2015). *Customary institutions and public authority: A literature review*, Bern: Swiss Agency for Development and Cooperation.

Monecke, A., Leisch, F., (2012). semPLS: Structural Equation Modeling Using Partial Least Squares. *Journal of Statistical Software*, Vol. 48, No. 1, pp. 1–32. doi: 10.18637/jss.v048.i03

Nifo, A., Vecchione, G. (2015). Measuring Institutional Quality in Italy. *Rivista Economica Del Mezzogiorno*, No. 1-2, pp. 157–182.

North, D. C., (1992). Institutions, Ideology, and Economic Performance. *Cato Journal*, Vol. 11, No. 3, pp. 477–496.

North, D. C., (1994). Economic Performance Through Time. *The American Economic Review*, Vol. 84, No. 3, pp. 359–368.

Ostrom, E., (2008). *Doing Institutional Analysis: Digging Deeper than Markets and Hierarchies*, In Handbook of New Institutional Economics, C. Ménard, M. M. Shirley (eds.), Berlin, Heidelberg: Springer, pp. 819–848, doi: 10.1007/0-387-25092-1_31.

Platje, J., (2008). "Institutional capital" as a factor of sustainable development - the importance of an institutional equilibrium. *Technological and Economic Development of Economy*, Vol. 14, No. 2, pp. 144–150. doi: 10.3846/1392-8619.2008.14.144-150

Rademaker, M., Schuberth, F., (2021). *Composite-Based Structural Equation Modeling*, Retrieved April 13, 2021 from: https://m-e-rademaker.github.io/cSEM/

Rordik, D., (2007). *One Economics, Many Recipes: Globalization*, Institutions, and Economic Growth, Princeton: Princeton University Press.

Rogowski, J., (1990). Modele miękkie. Teoria i zastosowanie w badaniach ekonomicznych [*Soft models. Theory and application in economic research*], Białystok: Wydawnictwo Filii Uniwersytetu Warszawskiego w Białymstoku.

Seidler, V., (2011). The Role of Informal Institutions in Building the Institutional Framework of an African State: The Case of the Kanuri in Nigeria. *SSRN Scholarly Paper* ID 3050608, doi:10.2139/ssrn.3050608.

Skarbek, D., (2020). Qualitative research methods for institutional analysis. *Journal of Institutional Economics*, Vol. 16, No. 4, pp. 409–422, doi: 10.1017/ S174413741900078X.

Skrodzka, I., (2016). Knowledge-Based Economy in the European Union: Cross-Country Analysis. *Statistics in Transition new series*, Vol. 17, No. 2, pp. 281-294. doi: 10.21307/stattrans-2016-019.

Veblen, T. B., (2016). *The Theory of the Leisure Class*, Scotts Valley: CreateSpace Independent Publishing Platform.

Vitola, A., Šenfelde, M., (2015). The role of institutions in economic performance. *Verslas: Teorija ir Praktika*, Vol. 16, No. 3, pp. 271–279, doi: 10.3846/ btp.2015.498.

Wilkin, J., (2011). Institutional Equilibrium: What Is it about and what Is its Role in the Economy? *Prace Naukowe Uniwersytetu Ekonomicznego We Wrocławiu. Ekonomia*, Vol. 15, No. 208, pp. 26–37.

Wilkin, J., Kargol-Wasiluk, A., Zalesko, M., (2019). Równowaga instytucjonalna - fundament równowagi gospodarczej [Institutional Equilibrium: the Foundation of Eco-nomic Equilibrium]. *Ekonomista*, No. 6, pp. 659–679.

Williamson, C. R., (2009). Informal institutions rule: Institutional arrangements and economic performance. *Public Choice*, Vol. 139, No. 3, pp. 371–387, doi: 10.1007/s11127-009-9399-x.

Wold, H., (1980). Soft modelling: Intermediate between traditional model building and data analysis. *Banach Center Publications*, Vol. 6, No. 1, pp. 333–346, doi: 10.4064/-6-1-333-346.

STATISTICS IN TRANSITION new series, June 2022
Vol. 23 No. 2, pp. 129–152, DOI 10.2478/stattrans-2022-0020
Received – 23.08.2019; accepted – 19.07.2021

Extracting relevant predictors of the severity of mental illnesses from clinical information using regularisation regression models

Sakshi Kaushik[1], Alka Sabharwal[2], Gurprit Grover[3]

ABSTRACT

Mental disorders are common non-communicable diseases whose occurrence rises at epidemic rates globally. The determination of the severity of a mental illness has important clinical implications and it serves as a prognostic factor for effective intervention planning and management. This paper aims to identify the relevant predictors of the severity of mental illnesses (measured by psychiatric rating scales) from a wide range of clinical variables consisting of information on both laboratory test results and psychiatric factors . The laboratory test results collectively indicate the measurements of 23 components derived from vital signs and blood tests results for the evaluation of the complete blood count. The 8 psychiatric factors known to affect the severity of mental illnesses are considered, viz. the family history, course and onset of an illness, etc. Retrospective data of 78 patients diagnosed with mental and behavioural disorders were collected from the Lady Hardinge Medical College & Smt. S.K, Hospital in New Delhi, India. The observations missing in the data are imputed using the non-parametric random forest algorithm. The multicollinearity is detected based on the variance inflation factor. Owing to the presence of multicollinearity, regularisation techniques such as ridge regression and extensions of the least absolute shrinkage and selection operator (LASSO), viz. adaptive and group LASSO are used for fitting the regression model. Optimal tuning parameter λ is obtained through 13-fold cross-validation. It was observed that the coefficients of the quantitative predictors extracted by the adaptive LASSO and the group of predictors extracted by the group LASSO were comparable to the coefficients obtained through ridge regression.

Key words: adaptive LASSO, group LASSO, mental disorder, multicollinearity, random forest imputation, ridge regression, severity of an illness.

[1] Quartesian, Chennai, Tamil Nadu, India. E-mail: sakshi2007@gmail.com. ORCID: https://orcid.org/0000-0002-4219-1488.

[2] Department of statistics, Kirori mal College, University of Delhi, India. E-mail: alkasabharwal@kmc.du.ac.in. ORCID: https://orcid.org/0000-0002-8252-8284.

[3] Department of statistics, Faculty of mathematical sciences, University of Delhi, India.
E-mail: gurpritgrover@yahoo.com. ORCID: https://orcid.org/0000-0003-2051-4810.

1. Introduction

Mental disorders are common non-communicable diseases rising with epidemic rates globally with over one third of people in most countries reporting sufficient criteria to be diagnosed at some point in their life (World Health Organization, 2000). The determination of the severity of mental illness has important clinical implications. Measures of severity help in the evaluation of outcome in treatment studies and may be used as a meaningful endpoint in clinical practice (Zimmerman, Morgan, & Stanton, 2018). It serves as an important prognostic factor for effective intervention planning and management.

Blood has been regarded as a source of information on illness and health since ancient times. With the emergence of experimental medical techniques in the mid-1800s, studies of blood have been carried out to identify physical characteristics that could be used to diagnose a psychiatric illness or assess the severity of its symptoms (Bahn et al. (2013)). In recent years, studies have increasingly been made on reports of blood tests such as platelets to understand psychiatric disorders, assess their impact on the severity of illness and evaluate the pharmacological properties of psychiatric drugs. Canan et al. (2012) showed that mean platelet volume (MPV) values were high in patients with major depression and decreased treatment.

Various general psychiatric aspects (such as family history, onset and course of illness, number of episodes, etc.) commonly observed across all mental disorders significantly impact the diagnosis, prognosis, severity, and remission of mental illness. Various studies in the past have identified family history as a potential risk factor for developing a mental illness and have associated it with seriousness indicators of illness such as recurrence, impairment, and age at onset (Laursen et al. (2005); Milne et al. (2009)). The number of episodes plays a cardinal role in determining the severity of illness. It has been observed that patients with a higher number of episodes have a more severe outcome (Marzo et al. (2006)). Such patients are more likely to relapse than those with fewer episodes. The onset of illness refers to how the symptoms of the disease begin to appear in a patient. The onset of symptoms in mental illness is known to be a prognostic indicator of its severity. The course of illness refers to the usual trajectory the disease follows from the onset of the first symptom until recovery or death. The course reflects the different grades of the severity of the illness. It has been observed that the chronic course of illness is associated with higher levels of depressive and somatic symptoms and greater mental dysfunction (Stegenga et al. (2010)). Studies in the past have shown that a higher amount of alcohol and tobacco consumption is found to be associated with greater severity of illness (Goldstein, Velyvis, & Parikh (2006); Krishnadas et al. (2012); Dwivedi, Chatterjee, & Singh (2017)). Further, Brådvik (2018) suggested that suicidal ideation and self-harm are related to mental illness. Insight of an illness is defined as a patient's capacity to understand the nature, significance, and severity of his or her illness. Literature suggests that insight interacts with the trajectory of the person's illness and predicts outcome in psychosis. It is found that the severity of illness increases with a progressive loss of insight (McDaniel, Edland, & Heyman (1995); Jacob (2016)). Although each mental disorder has its own complications and risks involved, a certain illness is considered to be more severe than others owing to the level of disability caused by them. These illnesses include disorders that produce psychotic symptoms, such as schizophrenia, and severe forms of other disorders, such as major depression and bipolar disorder (World Health Organization (2003)). Thus, different types of mental disorders have different severity levels. These worsen the symptoms and the course of mental illness.

Missing values are commonly encountered in medical datasets, especially mental disorders. Performing analysis with only complete patient datasets leads to a smaller sample size resulting in a loss of statistical power and bias in the estimation of parameters. Multiple imputation is a robust technique for handling missing data. In this approach, a prediction of the missing data is made using the existing data from other variables. There are several imputation methods available based on different statistical models such as regression, Random Forest, etc.

The inclusion of a large number of variables in a regression model often results in multicollinearity. Multicollinearity refers to high inter-correlations or inter-associations among the independent variables. The existence of multicollinearity affects the estimation of the model as well as the interpretation of the results. It leads to biased coefficient estimation and a loss of power. The regression models based on regularization techniques such as l_1 (Least Absolute Shrinkage and Selection Operator (LASSO) Regression; Tibshirani (1996)), l_2 (Ridge Regression; Hoerl and Kennard (1970)) and elastic net (Zou and Hastie (2005)) model, can solve this problem by adding a penalty to model parameters (except intercept) so the model generalizes the data instead of overfitting. Both ridge and LASSO regression belong to the class of penalised regression models. The key difference between these two techniques lies in the penalty that is imposed on the model. LASSO selects features that are predictive of the outcome by penalizing irrelevant features' weights to zeros while the ridge regression penalizes the irrelevant features by converging their weights to zero but never exactly equal to zero. Thus, both LASSO and ridge identify relevant predictors, however, LASSO is considered to be advantageous over ridge since it performs variable selection as well.

Many previous studies have used regularization regression models with multiply imputed data to determine relevant predictors from a class of independent variables (Jain (1985)). Brewer et al. (2009) used ridge and LASSO regression to predict an individual's score on the Unified Parkinson Disease Rating Scale based on Advanced Sensing for Assessment of Parkinson's disease (ASAP) data. Haenisch et al. (2016) identified protein analytes from a blood-based panel as potential biomarkers for diagnosing bipolar disorder using LASSO regression. Upadhya & Cheeran (2018) compared six regression techniques including ridge and LASSO to predict the Parkinson disease severity score using speech features.

Although, LASSO is an oracle procedure for simultaneously achieving consistent variable selection and optimal estimation (prediction), however, there are many solid arguments against the LASSO oracle statement (Zou (2006)). Further, Zhao and Yu (2006) showed that variable selection with LASSO could be consistent if the model satisfies some irrepresentable conditions. These conditions are restrictive and for data sets that fail to satisfy them, LASSO may not select the correct model. Therefore, to recognize relevant predictors some improvements of LASSO model have been proposed. The adaptive LASSO is a new version of the LASSO, in which adaptive weights (data driven) are used for penalizing different coefficients in the l_1 penalty. It also enjoys the oracle properties (Zou (2006)).

In some problems, when the predictors belong to pre-defined groups or factors; for example, collections of indicator (dummy) variables for representing the levels of a multiple categorical predictor such as onset and course of illness, LASSO and the adaptive LASSO are not suitable for variable selection as they are designed for selecting individual input variables. When directly applied to model they tend to select based on the strength of individual derived input variables rather than the strength of groups of input variables, often resulting in selecting more factors than necessary. In this situation it may be desirable to shrink and select the members of a group together. The group LASSO is a generalization of the LASSO for doing group-wise variable selection by introducing a suitable extension in the penalty of LASSO (Yuan & Lin (2006)).

This paper aims to identify relevant predictors for estimating the severity of mental illness (measured by psychiatric rating scales) from a wide range of clinical variables consisting of information on both laboratory test results and psychiatric aspects. The laboratory test results collectively indicate measurements on 23 components derived from vital signs and blood tests (complete blood count (CBC)) results such as diastolic and systolic blood pressure (DBP, SBP), pulse rate, haemoglobin (hb), red blood cell (RBC), etc. *Further,* 8 psychiatric factors known to affect severity of mental illness are considered, viz. family history (fh), number of episodes experienced by the patient (epi), onset and course of illness (onset), etc. The impact of covariates age and gender is also studied.

To achieve our aim, firstly missing values in the data consisting of 34 variables are imputed using the non-parametric random forest algorithm. Secondly, the problem of multicollinearity between explanatory variables is detected based on variance inflation factor (VIF). Since coefficients estimated from linear regression are biased in the presence of multicollinearity, thus, regularization techniques are used for fitting the regression model. Thirdly, prior to application of regularized regression models to the data, the dummy coding is applied to the 8 categorical variables consisting of clinical information on psychiatric factors related to mental disorders. These 8 categorical variables transform into 26 dichotomous variables with each variable representing each category. Fourthly, the ridge regression is applied to a total of 51 regressors including 25 quantitative and 26 binary variables with response variable being psychiatric rating scale score (RSS). Next, the adaptive LASSO is applied to the 25 quantitative variables including clinical variables consisting of information on vital signs and laboratory test result reports, age and number of episodes to extract the relevant predictors of RSS. Finally, the group LASSO is applied to the 26 dichotomous variables representing 8 groups of psychiatric variables to extract the relevant groups.

To the best of our knowledge, none of the previous studies has attempted to assess the relationship of such diverse and wide range of predictors with the severity of mental illness. The outline of the rest of the paper is as follows: Section 2 describes the dataset used for the application of methods discussed in Section 3. In Section 4, the application of the model to the dataset along with the results is discussed. The paper is concluded with a discussion in Section 5.

2. Data description

The retrospective data considered for this study consisted of 146 patients diagnosed with mental and behavioural disorders as per DSM-V (American Psychiatric Association (2013)) and ICD-10 (World Health Organization (1992)), collected from the Department of Psychiatry, Lady Hardinge Medical College & Smt. S.K, Hospital, New Delhi, India for the calendar year 2013-2014. The patients were diagnosed with Bipolar Affective Disorder (BPAD), schizophrenia, depression, and other disorders. The others category includes disorders, viz. Acute Transient Psychotic Disorder (ATPD), dementia, psychotic disorder: Not otherwise Specified (NOS), and alcohol abuse. Out of these 146 patients, only 78 patients could be included in the study as the clinical information on psychiatric variables as well as laboratory test result reports were available for them. The dataset of the remaining 68 patients was completely unavailable with respect to the variables considered in the study (i.e. either complete information on psychiatric variables and/or laboratory test reports were unavailable or both) and hence they were excluded.

The severity of mental disorders considered in this study is measured by various psychiatric rating scales recommended for each disorder. Since rating points as well as range of total scores vary in different psychiatric rating scales, thus, to maintain homogeneity, the total scores of these psychiatric rating scales are scaled down to 100 and denoted as RSS. RSS is the response variable under consideration. The regressors considered suitable for the study are classified into two categories: 1) clinical information related to vital signs and laboratory test result reports consisting of 23 variables, viz. Diastolic Blood Pressure (DBP) (mmHg), Systolic Blood Pressure (SBP) (mmHg), Pulse Rate (pulse per min), Haemoglobin (hb) (g/dL), Red Blood Cell (RBC) (million/μL), Mean Corpuscular Hemoglobin (MCH) (pg)**,** Mean Corpuscular Volume (MCV) (fL)**,** Mean Corpuscular Hemoglobin Concentration (MCHC) (g/dL)**,** Total Leukocyte Count (TLC) (cells/L), Platelet (thousand/μL), Blood Urea (b.urea) (mg/dL), Serum Creatinine (sr.cr) (mg/dL), Sodium (NA) (mEq/L), Potassium (K) (mEq/L), Serum Bilirubin (S.Bil) (mg/dL), Alanine Aminotransferase (ALT) (IU/L), Aspartate Aminotransferase (AST) (IU/L), Alkaline Phosphatase (ALP) (IU/L), Total Cholesterol (TCHOL) (mg/dL), High-Density Lipoprotein (HDL) (mg/dL), Triglycerides (S.TG) (mg/dL), Haematocrit or Packed-Cell Volume (PCV) (%) and Random Blood Sugar (RBS) (mg/dL). 2) The second category consists of clinical information on 8 psychiatric variables, viz. family history (fh), number of episodes experienced by the patient (epinew), onset of illness (onset), course of illness (course), alcohol or tobacco abuse (abuse), type of disorder (discode), suicidal ideation or self-harm (sui_sharm) and insight of illness (insight). The codes used for categorical variables are defined as follows:

i. Family history (fh): '0' and '1' indicate absence and presence of family history, respectively.

ii. Onset of illness (onset): '1', '2', '3', '4' and '5' indicate abrupt/sudden, acute, chronic, insidious, and sub-acute, respectively.

iii. Course of illness (course): '1', '2', '3' and '4' indicate continuous and progressive, continuous, episodic, and fluctuating, respectively.

iv. Abuse: '1' and '2' indicate absence and presence of alcohol or tobacco abuse, respectively.

v. Type of disorder (discode): '1', '2', '3' and '4' indicate Bipolar affective disorder (BPAD), Depression/Depressive disorder, Others, and Schizophrenia, respectively.

vi. Suicidal ideation or self-harm (sui_sharm): '1' implies absence while '2' indicates presence of suicidal ideation and/or self-harm in the patient.

vii. Insight: The grades of insight are as suggested by Sadock (2009).

Two other covariates considered are: age and gender. For gender, categories '1' and '2' indicate female and male, respectively.

3. Methods

Let there be n observations of a response variable Y and p associated predictor variables $X=(X_1, X_2,..., X_p)^T$. In this study, the response variable Y indicates the severity of illness quantified in terms of the total score of the psychiatric rating scale, denoted as RSS. (Here, $p=33$ and $n=78$). Out of these 33 predictors, $(X_1, X_2,..., X_{23})$ represent 23 features of laboratory test results and vital signs, viz. $X_1{\sim}DBP$, $X_2{\sim}SBP$, $X_3{\sim}Pulse\ Rate$, $\ldots, X_{23}{\sim}PCV$, X_{24}, X_{25} and X_{26} represent covariates age, gender and number of episodes while the remaining 7 variables represent thecategorical predictors of psychiatric factors, i.e. $X_{27} \sim family\ history(fh)$, $X_{28} \sim onset\ of\ il\ln ess(onset)$,...., $X_{33} \sim Insight$.

3.1. Method of imputation of missing observations

For imputing the missing values in the predictors, the imputation method given by Stekhoven and Bühlmann (2012) is used. Under this method, the missing values are predicted using a Random Forest (RF) trained on the observed parts of the dataset. The performance of the imputation method is assessed using the normalized root mean squared error (NRMSE) (Oba et al. (2003)) for the continuous variables and the proportion of falsely classified entries (PFC) over the categorical missing values. For both continuous as well as categorical variables, a value close to 0 indicates good performance.

3.2. Multicollinearity detection

Amongst the numerous approaches to detect multicollinearity in the data, namely determinant approach, Farrar and Glauber test (Farrar & Glauber (1967)), condition index (Belsley (1991)), Leamer's method (Greene (1993)) and variance inflation factor (VIF), the VIF is the most commonly used method. Let R_i^2 denote the coefficient of multiple determination of X_i regressed on the remaining $(p-1)$ explanatory variables. For the X_i, VIF is defined as

$$VIF_i = \frac{1}{(1-R_i^2)}. \tag{1}$$

A VIF of 5 or more indicates serious or excessive multicollinearity (Akinwande, Dikko and Samson (2015); Jongh et al. (2015)).

3.3. Dummy coding

Dummy coding is a method of representing a categorical variable into a series of dichotomous variables. For the categorical/qualitative predictors with K-levels, K indicatordummy/binary variables are created. Suppose X_i is a K-level factor input, then let X_{ij} $(j=1,2,...,K)$ be such that $X_{ij}=I\left(X_i=j\right)$. Together this group of X_{ij} represents the effect of X_i (Hastie, Tibshirani & Friedman (2009)).

3.4. Regularization techniques

Regularization is the process of penalizing the coefficients of predictor variables so that the resulting model has better predictive power. In this paper, the following types of regularization techniques, viz. ridge, group LASSO and adaptive LASSO are used to identify the predictors of severity of illness (Hoerl and Kennard (1970), Hastie, Tibshirani, & Wainwright (2015); James et al. (2013); Yuan and Lin (2006); Zou (2006)).

3.4.1. Ridge regression

Ridge regression is a variant of least squares regression in which the sum of squared errors is minimized, with an upper bound on the sum of squared values of the model parameters. In particular, the ridge regression coefficient estimates are obtained by solving the l_2 optimization problem

$$\underset{\beta_0,\beta}{\text{minimize}} \sum_{j=1}^{n}\left(y_j-\beta_0-\sum_{i=1}^{p}x_{ji}\beta_i\right)^2 \text{ subject to } \sum_{i=1}^{p}\beta_i^2 \le t \tag{2}$$

This equation is equivalent to solving

$$\hat{\beta}_i(\lambda)=\underset{\beta}{\arg\min}\left\{\sum_{j=1}^{n}\left(y_j-\beta_0-\sum_{i=1}^{p}x_{ji}\beta_i\right)^2+\lambda\sum_{i=1}^{p}\beta_i^2\right\} \tag{3}$$

where λ, known as the tuning parameter, controls the strength of the penalty. The larger the value of λ, the greater the amount of shrinkage. The second term, $\lambda\sum_{i=1}^{p}\beta_i^2$ is called shrinkage penalty.

3.4.2. Least absolute shrinkage and selection operator (LASSO) regression

LASSO is a regularization and variable selection method for statistical models. Under this technique, the sum of squared errors is minimized, with an upper bound on the sum of the absolute values of the model parameters. The LASSO estimate is defined by the solution to the l_1 optimization problem

$$\underset{\beta_0,\beta}{\text{minimize}} \sum_{j=1}^{n}\left(y_j-\beta_0-\sum_{i=1}^{p}x_{ji}\beta_i\right)^2 \text{ subject to } \left\|\beta\right\|_1 \le t \tag{4}$$

where $\|\beta\|_1 = \sum_{i=1}^{p} |\beta_j|$ is the l_1 norm of β and t is a user-specified parameter. This optimization problem is equivalent to the parameter estimation that follows

$$\hat{\beta}_i(\lambda) = \arg\min_{\beta} \left\{ \sum_{j=1}^{n} \left(y_j - \beta_0 - \sum_{i=1}^{p} x_{ji}\beta_i \right)^2 + \lambda \|\beta\|_1 \right\} \tag{5}$$

where λ is as defined in section 3.4.1. When the optimization problem is minimized, some coefficients shrink to zero, i.e. $\hat{\beta}_i(\lambda) = 0$, for some values of i, resulting in exclusion of some predictors.

Zhao and Yu (2006) showed that variable selection with LASSO could be consistent if the underlying model satisfies some irrepresentable conditions. The irrepresentable condition that should be satisfied is defined as follows:

Let $\boldsymbol{X} = (\boldsymbol{X_1}, \boldsymbol{X_2})'$, where $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ is the subset of $\boldsymbol{X}$ that contains the relevant and irrelevant predictor variables, respectively. Let $\boldsymbol{\beta_1}$be the coefficients of $\boldsymbol{X_1}$. The covariance matrix of $\boldsymbol{X}$ can be computed as $\boldsymbol{\Sigma} = n^{-1}\boldsymbol{X'X}$, which is a symmetric matrix. Let $\boldsymbol{C_{11}} = n^{-1}\boldsymbol{X_1}'\boldsymbol{X_1}$ and $\boldsymbol{C_{22}} = n^{-1}\boldsymbol{X_2}'\boldsymbol{X_2}$ be the covariance matrix of relevant and irrelevant predictor variables, respectively. Let $\boldsymbol{C_{12}} = n^{-1}\boldsymbol{X_1}'\boldsymbol{X_2}$ and $\boldsymbol{C_{21}} = n^{-1}\boldsymbol{X_2}'\boldsymbol{X_1}$ be the covariances between relevant and irrelevant variables. Then, $\boldsymbol{\Sigma}$ can be expressed in block-wise form as

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{C_{11}} & \boldsymbol{C_{12}} \\ \boldsymbol{C_{21}} & \boldsymbol{C_{22}} \end{pmatrix}$$

Assuming $\boldsymbol{C_{11}}$ is invertible, the irrepresentable condition can be defined as:

$$\left|\boldsymbol{C_{21}}\boldsymbol{C_{11}}^{-1} sign(\boldsymbol{\beta_1})\right|_{\infty} < 1 \text{, and the inequality holds elementwise.} \tag{6}$$

These conditions are restrictive and may not hold for all datasets. Thus, the adaptive LASSO model, which is an improvement over LASSO, is used.

3.4.3. Adaptive LASSO

The adaptive LASSO is an extension of LASSO, in which adaptive weights are used for penalizing different coefficients in the l_1 penalty (Zou (2006)). Suppose that $\hat{\beta}$ is a root- n–consistent estimator to β. Let $\widehat{\beta_\imath}$ be the ordinary least square estimate, $\gamma > 0$, and the weight vector is defined as $\hat{w} = 1/\left|\hat{\beta}\right|^{\gamma}$, then the adaptive LASSO estimates $\widehat{\beta^{(n)}}$ are given by

$$\hat{\beta}_i^{(n)} = \arg\min_{\beta} \left\{ \sum_{j=1}^{n} \left(y_j - \beta_0 - \sum_{i=1}^{p} x_{ji}\beta_i \right)^2 + \lambda_n \sum_{i=1}^{p} \hat{w}_i |\beta_i| \right\} \tag{7}$$

3.4.4. Group LASSO

The group LASSO is a generalization of LASSO for performing group-wise variable selection (Yuan and Lin (2006)). Suppose that u predictors are divided into L groups, with u_l being the number in group l. Let X_l represent the predictors corresponding to l^{th} group, with corresponding coefficient vector β_l. The group LASSO minimizes the convex criterion

$$\arg\min_{\beta}\left(\frac{1}{2}\sum_{j=1}^{n}\left(y_j-\beta_0-\sum_{l=1}^{L}X_{jl}\beta_l\right)^2+\lambda\sum_{l=1}^{L}\sqrt{u_l}\left\|\beta_l\right\|\right), \tag{8}$$

where the $\sqrt{u_l}$ terms account for the varying group sizes. This procedure encourages sparsity at both the group and individual levels. That is, for some values of λ, an entire group of predictors may drop out of the model (Hastie, Tibshirani & Friedman (2009)).

Figure 1 presents the steps followed during the course of this study.

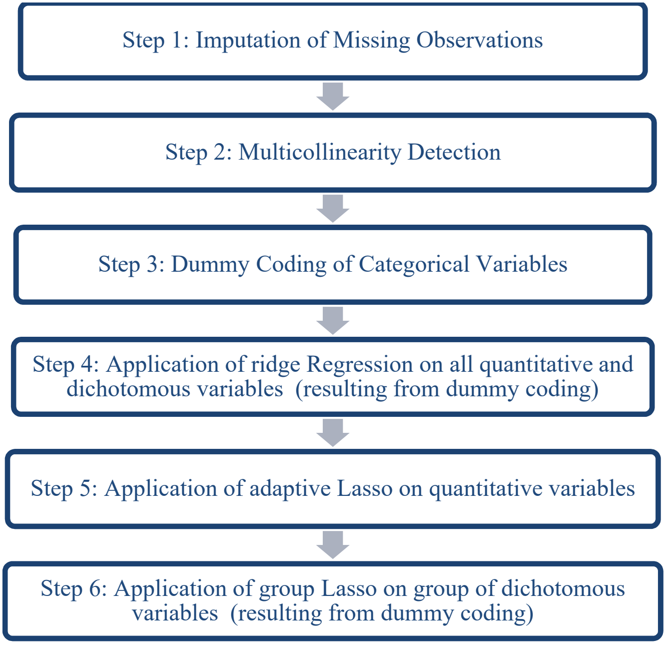

Step 1: Imputation of Missing Observations
Step 2: Multicollinearity Detection
Step 3: Dummy Coding of Categorical Variables
Step 4: Application of ridge Regression on all quantitative and dichotomous variables (resulting from dummy coding)
Step 5: Application of adaptive Lasso on quantitative variables
Step 6: Application of group Lasso on group of dichotomous variables (resulting from dummy coding)

Figure 1. Flowchart of steps followed during the course of the study

4. Results

This section displays the results obtained on stepwise application of methods (discussed in previous section) to the dataset considered.

4.1. Imputation of missing observations

The missingness in the data can be visualized graphically in Figure 2.

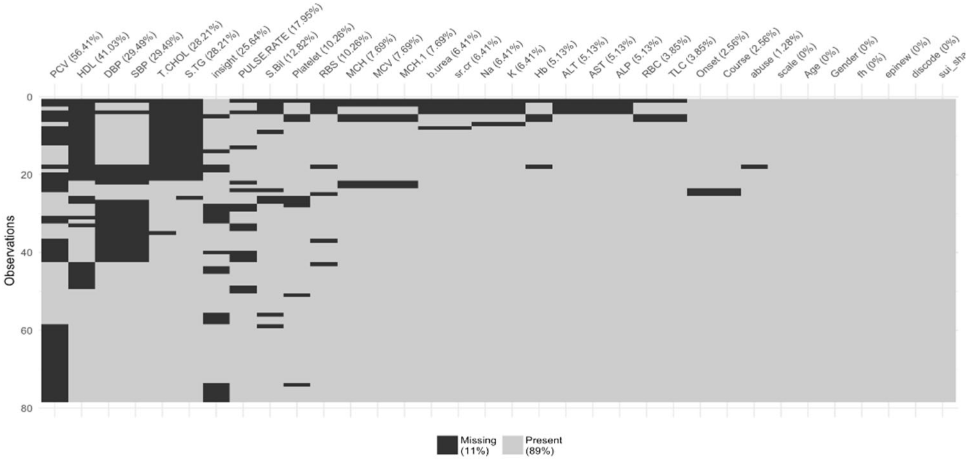


Figure 2. Visual representation of missingness in data.

In Figure 2, black colour shows the location of missing values with respect to each variable. The information on the percentage of missing values overall (in the legend), and in each variable is also provided. Missing observations are imputed using the non-parametric random forest algorithm described in section 3.1. Table 1 presents the Out-of-bag (OOB) error associated with imputation of missing observations.

Table 1. Estimated Out-of-bag (OOB) imputation error

Error type	Result
NRMSE	0.2921
PFC	0.3278

The NRMSE and PFC are not far from zero, indicating not much error is committed in imputing data. The descriptive statistics of quantitative variables and the summary of frequencies for categorical variables after imputation are presented in Tables 2 and 3.

Table 2. Descriptive statistics of quantitative variables after imputation

S. no.	Variable	Min	Max	Median	Mean	Stdev
1	DBP (mmHg)	81.60	141.00	118.21	117.78	9.63
2	SBP (mmHg)	13.60	96.50	76.16	76.08	9.16
3	Pulse rate (per min)	70.00	89.60	81.94	81.42	2.78
4	Hb (g/dL)	7.45	17.00	13.20	13.14	2.07
5	RBC (million/µL)	2.69	6.21	4.61	4.60	0.65
6	MCH (pg)	20.80	101.80	32.60	49.55	27.53
7	MCV (fL)	8.78	120.00	75.60	66.32	27.49
8	MCHC (g/dL)	12.60	40.60	32.47	32.13	3.27
9	TLC (cells/L)	2305.25	12600.00	6850.00	6981.77	2121.81
10	Platelet (thousand/µL)	1.11	11.90	1.99	2.30	1.48

Table 2. Descriptive statistics of quantitative variables after imputation (cont.)

S. no.	Variable	Min	Max	Median	Mean	Stdev
11	b.urea (mg/dL)	1.80	46.00	21.71	22.23	7.25
12	sr.cr (mg/dL)	0.60	1.70	1.00	1.01	0.18
13	Na (mEq/L)	131.50	154.00	140.63	141.03	4.11
14	K (mEq/L)	3.36	6.40	4.28	4.27	0.47
15	S.Bil (mg/dL)	0.30	2.60	0.70	0.78	0.38
16	ALT (IU/L)	12.00	170.75	27.00	35.37	26.89
17	AST (IU/L)	16.00	175.00	34.00	42.22	28.26
18	ALP (IU/L)	1.00	358.00	154.90	161.35	69.13
19	RBS (mg/dL)	59.00	273.50	106.88	114.28	38.50
20	TCL (mg/dL)	102.00	221.00	158.87	161.04	26.66
21	HDL (mg/dL)	27.00	282.00	48.44	58.82	37.87
22	S.TG (mg/dL)	32.00	426.50	123.35	126.87	59.94
23	PCV (%)	2.22	77.00	40.13	39.52	9.00
24	Age (years)*	20.00	70.00	40.50	41.94	10.48
25	Number of episodes *	1.00	5.00	2.00	2.08	0.98
26	RSS*	4.48	68.75	37.50	36.53	13.99

Note: *There were no missing values for these quantitative variables: age, RSS and episodes.

Table 3. Summary of frequencies of categorical variables after imputation

Variable	Category	Frequency	% Total
Gender*	Female	42	53.85
	Male	36	46.15
Family History (fh)	Absent	22	28.21
	Present	56	71.79
Onset	Abrupt	17	21.79
	Acute	25	32.05
	Chronic	1	1.28
	Insidious	32	41.03
	Sub-Acute	3	3.85
Course	Continuous and Progressive	30	38.46
	Continuous	12	15.38
	Episodic	22	28.21
	Fluctuating	14	17.95
Abuse	Absent	46	58.97
	Present	32	41.03
Type of Disorder (discode)*	BPAD	17	21.79
	Depression	5	6.41
	Others	17	21.79
	Schizophrenia	39	50
Suicidal ideation or self-harm (sui_sharm)*	Absent	62	79.49
	Present	16	20.51
Insight	Grade 1	29	37.18
	Grade 2	12	15.38
	Grade 3	20	25.64
	Grade 4	14	17.95
	Grade 5	3	3.85

Note: *There were no missing values for these categorical variables: gender, discode and sui_sharm.

4.2. Multicollinearity detection

The inclusion of a large number of variables, which are also observed to be interdependent and correlated, lead to the problem of multicollinearity. Thus, a check for detection of multicollinearity among regressors is performed using Variance Inflation Factor (VIF). Table 4 presents VIF for each regressor.

Table 4. Variance Inflation Factor (VIF) for regressors

S. no.	Variables	VIF	S no.	Variables	VIF
1	DBP (mmHg)	7.18	18	ALP (IU/L)	1.77
2	SBP (mmHg)	5.19	19	RBS (mg/dL)	2.89
3	Pulse rate (per min)	1.54	20	TCL (mg/dL)	2.67
4	Hb (g/dL)	7.23	21	HDL (mg/dL)	2.10
5	RBC (million/µL)	5.30	22	S.TG (mg/dL)	3.38
6	MCH (pg)	9.77	23	PCV (%)	2.53
7	MCV (fL)	10.77	24	Age (years)	2.51
8	MCHC (g/dL)	2.66	25	Number of Episodes	2.97
9	TLC (cells/L)	1.67	26	Gender	5.51
10	Platelet (thousand/µL)	1.57	27	Family History (fh)	2.03
11	b.urea (mg/dL)	1.82	28	Onset	1.93
12	sr.cr (mg/dL)	1.73	29	Course	2.52
13	Na (mEq/L)	1.67	30	Abuse	4.73
14	K (mEq/L)	1.51	31	Type of disorder (discode)	1.95
15	S.Bil (mg/dL)	2.21	32	Suicidal ideation or self-harm (sui_sharm)	1.35
16	ALT (IU/L)	3.26	33	Insight	1.88
17	AST (IU/L)	3.42			

A VIF of 5 or more indicates serious or excessive multicollinearity. Thus, the problem of multicollinearity exists in the data due to high values of VIF for regressors DBP, SBP, Hb, RBC, MCH, MCV and gender.

4.3. Dummy coding

The dummy coding is applied to the 8 categorical variables consisting of clinical information on psychiatric factors related to mental disorders. These 8 categorical variables are transformed into 26 dichotomous variables with each variable representing each category. For example, if X_{onset} represents the variable onset with 5 categories, then it is transformed into 5 binary/dichotomous variables X_{onsetj} $(j = 1,2,...,5)$ such that $X_{onsetj} = I\left(X_{onset} = j\right)$

4.4. Ridge regression

The ridge regression is applied to a total of 51 regressors including 25 quantitative and 26 binary variables with response variable being psychiatric rating scale score (RSS). The quantitative variables include 23 variables representing clinical information related to vital signs and laboratory test result reports (defined in Section 2), age and number of episodes. The binary variables represent categories of psychiatric variables obtained as a result of dummy coding. The model space is searched using 13-fold cross-validation to obtain the optimum value of the tuning/regularization parameter $\lambda = 21.2001$. Figure 3 presents the mean absolute cross validation error curve plotted as function of $\log(\lambda)$ along with the upper and lower standard deviation curves. It is evident from the figure that the mean absolute cross-validation error is minimum when $\log(\lambda)$ is approximately 3.

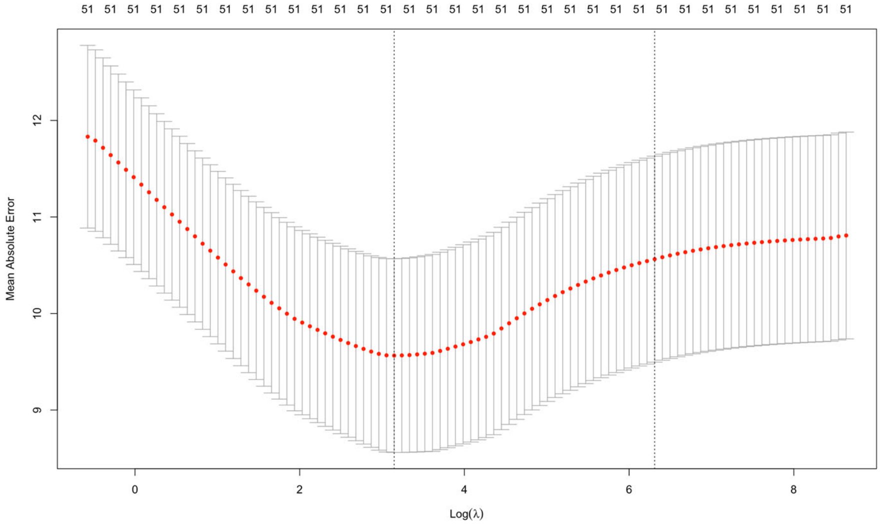


Figure 3. Mean absolute cross validation error curve plotted as function of $\log(\lambda)$for ridge regression

The coefficients derived on applying the ridge regression to the variables under consideration are presented in Table 5.

Table 5. Regression coefficients estimated from ridge regression

Regressor	Coefficient	Regressor	Coefficient
Intercept	37.8463	Gender: Female	0.3749
DBP (mmHg)	-0.0434	Gender: Male	-0.3742
SBP (mmHg)	-0.0511	Family History: Absent	-2.3078
Pulse rate (per min)	0.1400	Family History: Present	2.3071
Hb (g/dL)	-0.0600	Onset: Abrupt	-0.4918

Table 5. Regression coefficients estimated from ridge regression (cont.)

Regressor	Coefficient	Regressor	Coefficient
RBC (million/μL)	-1.3414	Onset: Acute	-0.1071
MCH (pg)	-0.0095	Onset: Chronic	2.0362
MCV (fL)	0.0125	Onset: Insidious	0.2534
MCHC (g/dL)	0.1158	Onset: Sub Acute	0.5442
TLC (cells/L)	-0.0003	Course: Continuous and Progressive	-2.5869
Platelet (thousand/μL)	-0.0334	Course: Continuous	-0.5289
b.urea (mg/dL)	-0.0209	Course: Episodic	1.1966
sr.cr (mg/dL)	6.6989	Course: Fluctuating	2.9792
Na (mEq/L)	-0.0097	Abuse: Absent	0.6838
K (mEq/L)	0.9443	Abuse: Present	-0.6838
S.Bil (mg/dL)	-1.9304	Type of Disorder: BPAD	-0.6246
ALT (IU/L)	0.0081	Type of Disorder: Depression	2.6030
AST (IU/L)	-0.0120	Type of Disorder: Others	-2.8586
ALP (IU/L)	0.0063	Type of Disorder: Schizophrenia	1.7502
RBS (mg/dL)	-0.0015	Suicidal ideation or self-harm: Absent	-0.3341
TCL (mg/dL)	-0.0239	Suicidal ideation or self-harm: Present	0.3341
HDL (mg/dL)	0.0007	Insight: Grade 1	1.6457
S.TG (mg/dL)	-0.0053	Insight: Grade 2	0.9930
PCV (%)	-0.0786	Insight: Grade 3	-1.4074
Age (years)	-0.0728	Insight: Grade 4	-1.3659
Number of Episodes	1.2645	Insight: Grade 5	-1.1939

It is evident from Table 5 that the coefficients estimated by the ridge regression for 18 regressors out of 51 have values close to 0 indicating that they do not have much effect on the severity of illness.

4.5. LASSO regression

In this study, the LASSO model is applied to the quantitative and categorical predictors separately. The group LASSO is applied to the categorical variables while the adaptive LASSO is used for quantitative regressors.

4.5.1. Adaptive LASSO

The adaptive LASSO is applied to the 25 quantitative variables including 23 variables consisting of clinical information related to vital signs and laboratory test result reports (defined in Section 2), age and number of episodes. The optimum value of the regularization parameter $\lambda = 2.4328$ is obtained using 13-fold cross-validation. Figure 6 presents the mean absolute cross validation error curve plotted as function of $\log(\lambda)$ along with the upper and lower standard deviation curves. It is evident from the figure that the mean absolute cross-validation error is minimum when $\log(\lambda)$ is approximately 0.9.

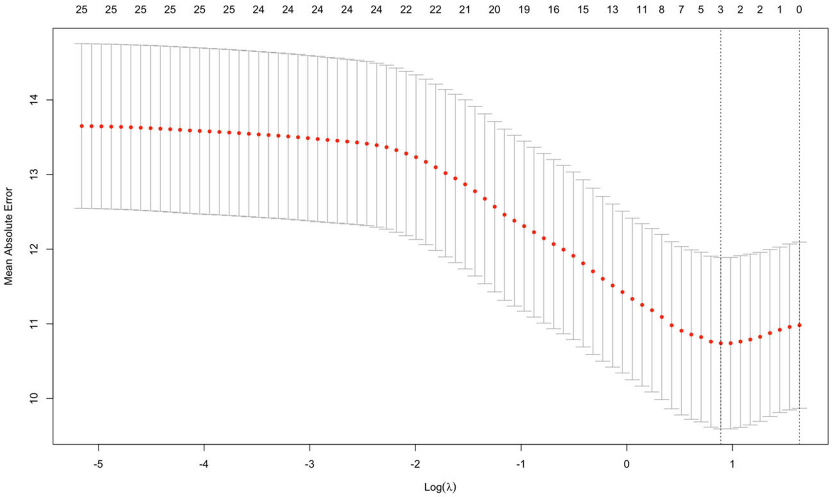


Figure 4. Mean absolute cross validation error curve plotted as function of $\log(\lambda)$for adaptive LASSO model

The predictors selected from the adaptive LASSO along with their coefficients are presented in Table 6.

Table 6. Regression coefficients estimated from adaptive LASSO regression

S. no.	Regressor (Unit)	Coefficient	S. no.	Regressor (Unit)	Coefficient
1	Intercept	40.8275	14	Na (mEq/L)	0.0000
2	DBP (mmHg)	0.0000	15	K (mEq/L)	0.0000
3	SBP (mmHg)	0.0000	16	S.Bil (mg/dL)	0.0000
4	Pulse rate (per min)	0.0000	17	ALT (IU/L)	0.0000
5	Hb (g/dL)	0.0000	18	AST (IU/L)	0.0000
6	**RBC (million/μL)**	**-2.1370**	19	ALP (IU/L)	0.0000
7	MCH (pg)	0.0000	20	RBS (mg/dL)	0.0000
8	MCV (fL)	0.0000	21	TCL (mg/dL)	0.0000
9	MCHC (g/dL)	0.0000	22	HDL (mg/dL)	0.0000
10	TLC (cells/L)	0.0000	23	S.TG (mg/dL)	0.0000
11	Platelet (thousand/μL)	0.0000	24	PCV (%)	0.0000
12	b.urea (mg/dL)	0.0000	25	Age (years)	0.0000
13	**sr.cr (mg/dL)**	**3.1596**	26	**Number of Episodes**	**2.7509**

The adaptive LASSO selected only 3 relevant predictors out of a total of 25 variables by shrinking the coefficients of less other regressors to zero. All of these predictors have coefficients far from 0. Thus, laboratory test results on Red Blood Cell (RBC), serum creatinine (sr.cr), and number of episodes are found to be the relevant predictors of severity of mental illness as measured by the psychiatric rating scales.

4.5.2. Group LASSO

The group LASSO is applied to the 26 dichotomous variables obtained from 8 psychiatric variables after applying dummy coding with each binary variable representing each category. The optimum value of $\lambda = 0.4829$is obtained using 13-fold cross-validation. Figure 5 presents the mean absolute cross validation error curve plotted as function of $\log(\lambda)$ along with the upper and lower standard deviation curves. It is evident from the figure that the mean absolute cross-validation error is minimum when $\log(\lambda)$ is approximately -0.7.

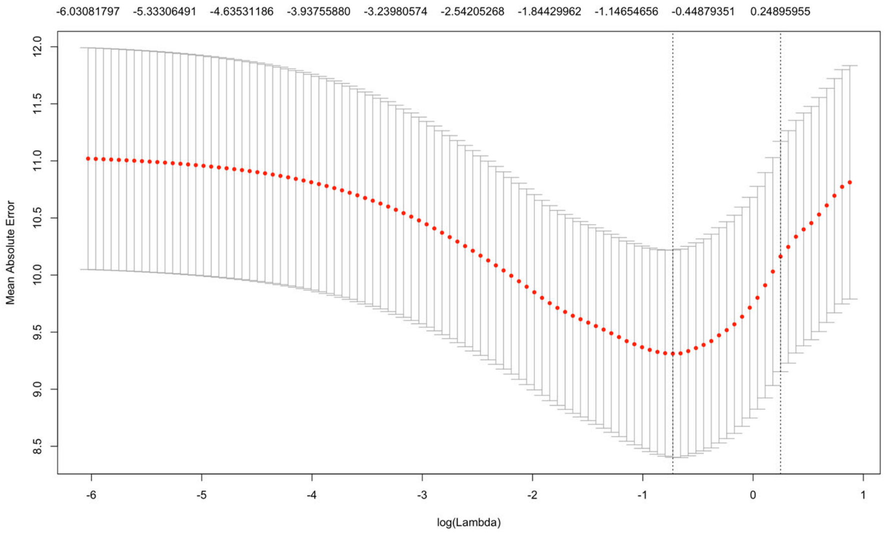


Figure 5. Mean absolute cross validation error curve plotted as function of $\log(\lambda)$ for group LASSO model.

The predictors selected by the group LASSO along with their coefficients are presented in Table 7.

Table 7. Regression coefficients estimated from group LASSO regression

S. no.	Regressor	Category	Coefficient
1	Intercept		34.6398
2	Gender	Female	0.0585
3		Male	-0.0585
4	Family History (fh)	**Absent**	**-3.6916**
5		**Present**	**3.6918**
6	Onset	Abrupt	0.0000
7		Acute	0.0000
8		Chronic	0.0000
9		Insidious	0.0000
10		Sub-Acute	0.0000

Table 7. Regression coefficients estimated from group LASSO regression (cont.)

S. no.	Regressor	Category	Coefficient
11	Course	**Continuous and Progressive**	**-3.7530**
12		**Continuous**	**-1.2243**
13		**Episodic**	**2.0239**
14		**Fluctuating**	**2.9536**
15	Abuse	Absent	0.0241
16		Present	-0.0241
17	Type of Disorder (discode)	Bipolar affective disorder (BPAD)	-0.0389
18		**Depression**	**1.1385**
19		**Others**	**-3.4340**
20		**Schizophrenia**	**2.3347**
21	Suicidal ideation or self-harm (sui_sharm)	Absent	0.0000
22		Present	0.0000
23	Insight	**Grade 1**	**2.1232**
24		Grade 2	0.4696
25		**Grade 3**	**-1.3771**
26		**Grade 4**	**-1.0185**
27		Grade 5	-0.1970

The group LASSO selected 6 groups of predictors out of a total of 8 groups of psychiatric variables. Thus, gender, family history, course, alcohol and/or tobacco abuse, type of disorder and insight of illness are found to be relevant predictors of severity of mental illness.

All the calculations were performed in R software using adalasso, coefplot, gglasso, glmnet, mctest, missForest, pastecs and summarytools packages.

5. Discussion

Recently, a large number of research studies have focused on establishing diagnostic tests for mental disorders based on reports of blood tests and psychiatric factors. Richards et al. (2016) predicted severity of depression based on gender, age, employment status, marital status, previous diagnosis of depression, recent experience of life stressors using multiple linear regression. Huang et al. (2014) predicted the diagnosis and severity of depression based on a large sample of electronic health record (EHR) data consisting of information on demographic variables, structured variables such as ICD diagnosis codes, prescription codes, and unstructured variables such as progress notes, pathology reports, radiology reports, and transcription reports. This motivated us to predict the severity of illness based on the laboratory and pathological reports and certain psychiatric aspects. Further, the information on these basic variables is generally readily available for all mental disorders.

Missingness is a commonly encountered problem in medical data. However, ignoring or removing missing data leads to an important loss of information and results in biased estimation. We have used multiple imputation to deal with missingness since in addition to restoring the natural variability of the missing values, it incorporates the uncertainty due to the missing data, which results in valid statistical inference (Kang (2013)). Multicollinearity is commonly observed in datasets with large number of regressors. Variance Inflation factor (VIF) is the most common approach for detecting multicollinearity. There is no set VIF threshold available in the literature to be used as a standard rule. In this study, we employed a VIF threshold of 5 for collinearity diagnostics since a VIF value that is near or above 5, indicates that the regressors may be highly correlated (Akinwande, Dikko and & Samson (2015); Jongh et al. (2015)).

When there are a large number of predictors, the correlation between them (multicollinearity) generally limits the usefulness of classic regression methods. Regularization techniques such as ridge, LASSO, and elastic net are particularly useful in such cases. In this study, we applied both ridge and extensions of LASSO viz. the adaptive and group LASSO models on the data and observed that adaptive and the group LASSO models did not extract any of the 18 regressors for which the coefficients were estimated to be close to 0 by the ridge regression. Further, we compared the ridge and the LASSO models using the Bayesian Information Criterion (BIC) and observed that the BIC values for the group LASSO (BIC=1057.617) and the adaptive LASSO (BIC=1131.936) were lower than the ridge regression model (BIC=1148.786). Thus, in this study, the group and adaptive LASSO models performed better than the ridge model.

The LASSO (l_1) penalty function performs variable selection and dimension reduction by shrinking coefficients, while the ridge (l_2) penalty function shrinks the coefficients of correlated variables towards their average (Kim et al. (2017)). In general, LASSO is preferred over the ridge model in terms of interpretability since it extracts the relevant predictors. However, in medical data, it is not advisable to completely ignore or remove the less relevant predictors due to their clinical implication. Even if the objective of a study is to extract relevant predictors, it is suggested to perform both LASSO and the ridge regression since the ridge regression supports the results of the LASSO regression and will help to make a decision depending upon the clinical relevance of the regressor based on a chosen level of significance.

In the adaptive LASSO, the weights are based on the ordinary least square estimates. The weights are data-dependent and adaptively chosen from the data with large coefficients receiving small weights and small coefficients receiving large weights.

In this study, it was observed that from a wide range of clinical variables consisting of information on both laboratory test results and psychiatric aspects, the following are the relevant predictors of the severity of mental illness: Red Blood Cells (RBC), Serum Creatinine (Sr.Cr), number of episodes, gender, family history, course of illness, alcohol and tobacco abuse, type of disorder and insight of an illness. Our results are in accordance with previous studies. Setoyama et al. (2016) found that serum creatinine is commonly associated with severity of depression in three independent cohort sets regardless of the presence or absence of medication and diagnostic difference. Barbato (1998), Hafner (2005) and Richards et al. (2016) have identified gender as one of the relevant predictors of severity of mental illness. Lu et al. (2018) found that positive family history is a strong predictor of schizophrenia. Marzo et al. (2006) showed that patients with multi-episode bipolar disorder would be more prone to have higher levels of cognitive impairment suggesting that patients with a higher number of episodes and recurring or episodic course result in severe outcomes. Studies in the past showed that a higher amount of alcohol and tobacco consumption is found to be associated with greater severity of illness (Goldstein, Velyvis, & Parikh (2006); Krishnadas et al. (2012); Dwivedi, Chatterjee, & Singh (2017)). Jacob (2016) showed that patients with good insight have a less severe disease.

This paper adds to the literature of medical research aimed at identifying the biomarkers for diagnosis and predictors of the severity status of mental disorders. The clinicians can use the relevant factors to build a profile of the patient and his needs. This work will help in developing valid and efficient approaches to diagnose the disorders at an early stage. It will also aid clinicians in devising effective strategies for treatment planning.

Generally, the predictive accuracy of the regularization method is tested on a test dataset after fitting the regression model on the training dataset. This procedure could not be adopted in this paper due to the small sample size. To maintain consistent selection of predictors, the tuning parameter for fitting regularization models is selected using 13-fold cross-validation. However, a limitation of using the cross validation method in the case of a small sample size could suffer from overfitting.

References

Akinwande, M. O., Dikko, H. G., and Samson, A., (2015). Variance inflation factor: as a condition for the inclusion of suppressor variable(s) in regression analysis. *Open Journal of Statistics*, pp. 754–767.

American Psychiatric Association, (2013). *Diagnostic and statistical manual of mental disorders*, 5th edition Arlington, VA: American Psychiatric Publishing.

Barbato, A., (1998). *Schizophrenia and Public Health. Nations For Mental Health*, Division of Mental Health and Prevention of Substance Abuse, Geneva: World Health Organization.

Bahn, S., Schwarz, E., Harris, L. W., Martins-De-Souza, D., Rahmoune, H., and Guest, P. C., (2013). Biomarker blood tests for diagnosis and management of mental disorders: focus on schizophrenia. *Archives of Clinical Psychiatry*, São Paulo, 40(1), pp. 02–09.

Brådvik, L., (2018). Suicide Risk and Mental Disorders. *International journal of environmental research and public health*, 15 (9), pp. 2028–2031.

Belsley, D., (1991). *Conditioning diagnostics: collinearity and weak data in regression*, New York: Wiley.

Brewer, B. R., Pradhan, S., Carvell, G., and Delitto, A., (2009). Application of modified regression techniques to a quantitative assessment for the motor signs of Parkinson's Disease." *IEEE transactions on neural systems and rehabilitation engineering: a publication of the IEEE Engineering in Medicine and Biology Society*, 17 (6), pp. 568–575.

Canan, F., Dikici, S., Kutlucan, A., and Celbek, G., Coskun, H., Gungor, A., Aydin, Y. and Kocaman, G., (2012). Association of mean trombosit volume with DSM-IV major depression in a large community-based population: the MELEN study. *Journal of psychiatric research*, 46 (3), pp. 298–302. 10.1016/j.jpsychires.2011.11.016.

Dwivedi, A. K., Chatterjee, K., and Singh, R., (2017). Lifetime alcohol consumption and severity in alcohol dependence syndrome. *Industrial Psychiatry Journal*, 26(1), pp. 34–38.

Farrar, and Glauber, R., (1967). Multicollinearity in regression analysis: The problem revisited. *The Review of Economics and Statistics*, 49 (1), 92–107.

Goldstein, B., Velyvis, V., and Parikh, S. V., (2006). The association between moderate alcohol use and illness severity in bipolar disorder: a preliminary report. *The Journal of Clinical Psychiatry*, 67 (1), pp. 102–106.

Greene, W. H., (1993). T*he econometric approach to efficiency analysis. In the measurement of productive efficiency and productivity change*, by Harold O. Fried, C. A. Knox Lovell, and Shelton S. Schmidt, pp. 68–119. United Kingdom.

Haenisch, F., Cooper, J. D., Reif, A., Kittel-Schneide, S., Steiner, J., Leweke, F. M., Rothermundt, M., Beveren, N., Crespo-Facorro, B., Niebuhr, D., Cowan, D., Weber, N., Yolken, R., Penninx, B. and Bahn, S., (2016). Towards a blood-based diagnostic panel for bipolar disorder. *Brain, Behavior, and Immunity*, 52, pp. 49–57. https://doi.org/10.1016/j.bbi.2015.10.001

Hafner, H., (2005). *Gender Differences in Schizophrenia. In Estrogen Effects in Psychiatric Disorders*, by N. Bergemann, and A. (eds.) Riecher-Rӧssler. Austria: SpringerWienNewYork.

Hastie, T., Tibshirani, R., and Friedman, J., (2009). The elements of statistical learning: data mining, inference and prediction. *Second Edition*. California: Springer.

Hastie, T., Tibshirani, R., and Wainwright, M., (2015). *Statistical learning with sparsity: The Lasso and generalizations*. New York: Chapman and Hall/CRC Press.

Hoerl, A. E., Kennard, R. W., (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12 (1), pp. 55–67. DOI: 10.1080/00401706.1970.10488634.

Huang, S. H., Lependu, P., Iyer, S. V., Ai-Seale, M., Carrell, T. D., and Shah, N. H., (2014). Toward personalizing treatment for depression: predicting diagnosis and severity. *Journal of the American Medical Informatics Association*, 21 (6), pp. 1069–1075.

Jacob, K. S., (2016). Insight in psychosis: An indicator of severity of psychosis, an explanatory model of illness, and a coping strategy. *Indian journal of psychological medicine*, 38(3), pp. 194–201.

Jain, R., (1985). Ridge regression and its application to medical data. *Computers and Biomedical Research*, 18, pp. 363–368.

James, G., Witten, D., Hastie, T., and Tibshirani, R., (2013). *An introduction to statistical learning: with applications in R*, New York: Springer.

Jongh, P. J. De, Jongh, E. De, Pienaar, M., Gordon-Grant, H., Oberholzer, M., and Santana, L., (2015). The impact of pre-selected variance inflation factor thresholds on the stability and predictive power of logistic regression models in credit scoring. *Orion*, 31(1), pp. 17–37, DOI: https://doi.org/10.5784/31-1-162.

Kang, H., (2013). The prevention and handling of the missing data. *Korean journal of anesthesiology*, 64 (5), pp. 402–406.

Kim, M. H., Banerjee, S., Park, S. M., and Pathak, J., (2017). *Improving risk prediction for depression via Elastic Net regression – Results from Korea National Health Insurance Services Data*. AMIA ... Annual Symposium proceedings. AMIA Symposium, 2016, pp. 1860–1869.

Krishnadas, R., Jauhar, S., Telfer, S., Shivashankar, S., and Mccreadie, R., (2012). Nicotine dependence and illness severity in schizophrenia. *The British journal of psychiatry*, 201 (4), pp. 306–12.

Laursen, T. M., Labouriau, R., Licht, R. W., Bertelsen, A., Munk-Olsen, T., and Mortensen, P. B., (2005). Family history of psychiatric illness as a risk factor for schizoaffective disorder: A Danish Register-Based Cohort Study. *Arch Gen Psychiatry*, 62 (8), pp. 841–848. doi:10.1001/archpsyc.62.8.841

Lu, Y., Pouget, J. G., Andreassen, O. A., Djurovic, S., Esko, T., Hultman, C. M., Metspalu, A., Milani, L., Werge, T., and Sullivan, P. F., (2018). Genetic risk scores and family history as predictors of schizophrenia in Nordic registers. *Psychological medicine*, 48(7), pp. 1201–1208.

Marzo, S. D., Giordano, A., Pacchiarotti, I., Colom, F., Sánchez-Moreno, J., and Vieta, E., (2006). The impact of the number of episodes on the outcome of bipolar disorder. *The European Journal of Psychiatry*, 20, pp. 21–28.

Mcdaniel, K., Edland, S., and Heyman, A., (1995). Relationship between level of insight and severity of dementia in Alzheimer disease. CERAD Clinical Investigators. Consortium to Establish a Registry for Alzheimer's Disease. *Alzheimer Dis Assoc Disord*, 9 (2), pp. 101–104.

Milne, B., Caspi, A., Harrington, H., Poulton, R., Rutter, M., and Moffitt, T., (2009). Predictive value of family history on severity of illness: The case for depression, anxiety, alcohol dependence, and drug dependence. *Arch Gen Psychiatry*, 66 (7), pp. 738–747.

Oba, S., Sato, M. A., Takemasa, I., Monden, M., Matsubara, K., and Ishii, S., (2003). A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19, pp. 2088–2096.

Richards, D., Richardson, T., Timulak, L., Viganò, N., Mooney, J., Doherty, G., Hayes, C., Sharry, J., (2016). Predictors of depression severity in a treatment-seeking sample. *International Journal of Clinical and Health Psychology*, 16 (3), pp. 221–314.

Sadock, B., (2009). *Psychiatric report, medical record and medical error*. In S. V. Sadock BJ, Kaplan and Sadock's Comprehensive Textbook of Psychiatry (9th ed., pp. 907–18). Philadelphia: Lippincott Williams and Wilkins.

Setoyama, D., Kato, T. A., Hashimoto, R., Kunugi, H., Hattori, K., Hayakawa, K., Sato-Kasai, M., Shimokawa, N., Kaneko, S., Yoshida, S., Goto, Y. I., Yasuda, Y., Yamamori, H., Ohgidani, M., Sagata, N., Miura, D., Kang, D., and

Kanba, S., (2016). Plasma metabolites predict severity of depression and suicidal ideation in psychiatric patients-A Multicenter Pilot Analysis. *PLoS One*, 11(12). e0165267

Stegenga, B. T., Kamphuis, M. H., King, M., Nazareth, I., and Geerlings, M. I., (2010). The natural course and outcome of major depressive disorder in primary care: the PREDICT-NL study. *Social psychiatry and psychiatric epidemiology,* 47 (1), pp. 87–95.

Stekhoven, D. J., Bühlmann, P., (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28 (1), pp. 112–118.

Tibshirani, R., (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*. Series B (Methodological), pp. 267–288.

Upadhya, S. S., Cheeran, A. N., (2018). Performance comparison of regression techniques in predicting Parkinson disease severity score using speech features. *Biomedical Engineering: Applications, Basis and Communications*, 30(4). https://doi.org/10.4015/S1016237218500254

World Health Organization, (1992). T*he Icd-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines*. Geneva: World Health Organization.

WORLD HEALTH ORGANIZATION, (2000). Cross-national comparisons of the prevalences and correlates of mental disorders. WHO International Consortium in Psychiatric Epidemiology, *Bull*, 78 (4), pp. 413–426.

World Health Organization, (2003). Investing in mental health. *World Health Organization*, pp. 1–48.

Yuan, M., Lin, Yi., (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society*. Series B 68, part 1, pp. 49–67.

Zhao, P., Yu, B., (2006). On Model Selection Consistency of Lasso. *Journal of Machine Learning Research*, 7, pp. 2541–2563.

Zou, H., and Hastie, T., (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*, Series B, 67(2), pp. 301–320.

Zou, H., (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association: Theory and Methods*, 101(476), pp. 1418–1429.

Zimmerman, M., Morgan, T. A., and Stanton, K., (2018). The severity of psychiatric disorders. *World psychiatry: official journal of the World Psychiatric Association* (WPA), 17 (3), pp. 258–275.

STATISTICS IN TRANSITION new series, June 2022
Vol. 23 No. 2, pp. 153–161, DOI 10.2478/stattrans-2022-0021
Received – 06.12.2021; accepted – 05.04.2022

Regression model of water demand for the city of Lodz as a function of atmospheric factors

Czesław Domański[1], Robert Kubacki[2]

ABSTRACT

One of the Sustainable Development Goals (Goal 6) set by the United Nations is to provide people with access to water and sanitation through sustainable water resources management. Water supply companies carrying out tasks commissioned by local authorities ensure there is an optimal amount of water in the water supply system. The aim of this study is to present the results of the work on a statistical model which determined the influence of individual atmospheric factors on the demand for water in the city of Lodz, Poland, in 2010-2019. In order to build the model, the study used data from the Water Supply and Sewage System Company (Zakład Wodociągów i Kanalizacji Sp. z o.o.) in the city of Lodz complemented with data on weather conditions in the studied period. The analysis showed that the constructed models make it possible to perform a forecast of water demand depending on the expected weather conditions.

Key words: water demand, atmospheric factors, regression model.

1. Introduction

As the global climate changes and the urban population continues to grow, water resources in many of the world's cities are likely to be under increasing stress from reduced water supply and increased demand (Bates et al., 2008).

There have been several studies investigating the role of weather and climate variables in municipal water consumption (e.g. Balling and Gober, 2006; Ghiassi et al., 2008).

Previous studies used maximum and minimum temperatures and precipitation as explanatory variables to estimate water consumption. In addition, the interactions among different weather and climate variables that influence water use are not well understood (Praskievicz and Chang, 2009).

[1] Institute of Statistics and Demography, University of Lodz, Poland. E-mail: czedoman@uni.lodz.pl. ORCID: https://orcid.org/0000-0001-6144-6231.

[2] Poland. E-mail: robertkubacki@o2.pl. ORCID: https://orcid.org/0000-0003-0591-9529.

The aim of this study is an attempt to verify the hypothesis as to whether weather factors can better describe the phenomenon of water demand for the city of Lodz, Poland. Water is needed by everyone. Correctly predicting its demand is important to achieve two opposing objectives. Firstly, its quantity should be sufficient to satisfy the city's needs. Secondly, it should not be wasted. The data obtained clearly show that daily water demand varies from day to day, week to week and month to month. This is compounded by trends related to changing behaviour of the population and other users of the water supply system. In this study it was possible to obtain the total number of cubic metres pumped per day into the system. Data on individual consumers (households, industry, education, health) are only available in an aggregated form. Nevertheless, when reading this study, we should be aware that households are responsible for the consumption of 69% of the total volume of water in the city.

2. Regression models

2.1. Multiple regression

A multiple regression model is written as:

$$y_i = \beta_0 + \sum_{j=1}^{d} x_{ij}\beta_j + \varepsilon_i, \;\; i = 1,2, \dots n, x \in \boldsymbol{R}^d \qquad \varepsilon_i \sim N(0, \sigma^2), \tag{1}$$

where β_0 corresponds to the intercept, $\beta_1, \dots, \beta_d$ correspond to the model coefficients, x_i to the observation/measurement data, and ε to the residuals.

The objective function for the residual sum of squares is written as

$$\mathcal{L} = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - f(x_i; \beta))^2, \tag{2}$$

By plugging in the regression model equation from above we get

$$\mathcal{L} = \sum_{i=1}^{n} (y_i - \beta_0 + \sum_{j=1}^{d} x_{ij}\beta_j)^2, \tag{3}$$

where n corresponds to the number of observations and d corresponds to the number of features of the data set (Walesiak and Gatnar, 2009).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(pred_i - obs_i)^2}{n}} \tag{4}$$

The RMSE is the square root of the sum of the squared difference between the observed and predicted values, normalized by the number of observations n.

The lower RMSE the better the model fits the data (Géron, 2017).

Overfitting reduces the generalization properties of a model. When there are many correlated variables in a linear regression model, their coefficients can become poorly determined and exhibit high variance; hence, the values of the coefficients become huge. A wildly large positive coefficient on one variable can be canceled by a similarly large negative coefficient on its correlated cousin. By imposing a size constraint on the coefficients, this problem is alleviated (Harrington, 2012). Regularization methods constrain the model parameters in some way and thus are suitable to prevent overfitting.

In many regularization models an additional term is added to the optimization function for the optimal parameter estimates $\hat{\beta}_{opt}$.

$$\hat{\beta}_{opt} = arg\min\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 + \lambda g(\beta) \tag{5}$$

where g is a function of the coefficients β, which encourages the desired properties of β, and λ is a regularization parameter.

2.2. Ridge regression

Ridge regression, sometimes referred to as $\mathcal{L}_2$ - regularized regression, is a method to shrink the regression coefficients by imposing a penalty on their size. The Ridge regression uses a squared penalty on the regression coefficient vector β (Patterson and Gibson, 2018).

$$\beta_{RR} = \text{arg min}\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 + \lambda\|\beta\|^2 \tag{6}$$

Here, $\lambda > 0$ is a regularization parameter that controls the amount of shrinkage: the larger the value of $\lambda > 0$, the greater the amount of shrinkage. The coefficients are shrunk toward zero but do not reach zero. If $\lambda \rightarrow 0$ the parameter estimates β_{RR} approach the parameter estimates of the least-square solution β_{LS}.

$$Case\ \lambda \rightarrow 0 : \beta_{RR} \rightarrow \beta_{LS} \tag{7}$$

$$Case\ \lambda \rightarrow \infty : \beta_{RR} \rightarrow \vec{0} \tag{8}$$

We can solve the ridge regression problem using exactly the same procedure as for least squares,

$$\mathcal{L} = \|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 + \lambda\|\beta\|^2 = (\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta) + \lambda\beta^T\beta \tag{9}$$

First, take the gradient of $\mathcal{L}$ with respect to β and set to zero,

$$\nabla\mathcal{L} = -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\beta + 2\lambda\beta = 0 \tag{10}$$

Then, solve for β to find that

$$\beta_{RR} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}, \tag{11}$$

where **I** corresponds to the identity matrix.

2.3. LASSO regression

The LASSO (least absolute shrinkage and selection operator), also referred to as $\mathcal{L}_1$-regularized regression, is a shrinkage method like the ridge regression, with subtle but important differences. The LASSO estimate is defined by

$$\beta_{LASSO} = \arg\min||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda||\beta|| \tag{12}$$

where

$$\lambda > 0, \tag{13}$$

$$||\beta|| = \sum_{j=1}^{d}|\beta_j| \tag{14}$$

The LASSO method performs both regularization and variable selection. During the LASSO model fitting process only a subset of the provided features is selected for the use in the final model. The LASSO forces certain coefficients to be set to zero, effectively choosing a simpler model that does not include those coefficients. In contrast to the ridge regression, which can be solved analytically, numerical optimization (e.g. coordinate descent) is warranted to find the solution for the LASSO regression (Grus, 2018).

The degree of regularization depends on the regularization parameter λ. Thus, it is useful to evaluate the regression function for a sequence of λ.

3. Reference data

For the water demand analysis, data obtained from the Water Supply and Sewerage Works in the city of Lodz were used. The data from the period 2010-2019 included the amount of water injected into the water supply system each day. The set contains 3652 observations. Weather data obtained from www.ogimet.com were used as explanatory variables. The data contain a summary of the weather condition for all weather stations available on the website. Data from the station closest to the place of water intake for the city of Lodz were used for the study. The features comprising the weather description included: temperature (maximum, minimum, average), dew point temperature, humidity, wind direction, intensity and gust, atmospheric pressure, precipitation, cloud cover, sunshine, horizontal visibility and snow cover. Weather data can be obtained free of charge, but obtaining a complete set of data required writing a program in VBA to retrieve data cyclically after 50 observations.

4. The model estimation

Raw data on pumping volumes and weather factors were combined and subjected to preliminary analysis. Missing data were filled in. Filling in data gaps to preserve as many observations as possible for modelling concerned only weather data.

In addition, this concerned, e.g. the amount of snowfall during holiday periods, which were marked with a "-" and were replaced with a value of 0. From the preliminary observation of the data it can be concluded that the amount of water pumped to the water supply systems in the city of Lodz is decreasing every year. The variable describing YEAR takes the values 1, 2, 3, ..., 10 for successive years of observation 2010, 2011, 2012, ..., 2019. Moreover, it was possible to observe that the amount of pumped water changes depending on the month. The lowest average value of pumped water per month is observed in August. For this purpose, a set of zero-one variables was created for each month of the year with August omitted to prevent collinearity between the variables. Also for the days of the week it was observed that the average amounts of pumped water differ. On Sundays, on average, the least water is pumped into the system. This resulted in dedicated zero-one variables describing the days of the week except Sundays. When observing the outlier variables, it was possible to observe that the lowest amounts of pumped water fall on public holidays. For this purpose, the variable SWIETO was created, which takes the value 1 if the following holidays were celebrated on that day: 1st of January, Easter and Easter Monday (movable holidays), 1st and 3rd of May, 15th of August, 1st and 11th of November, 25th and 26th of December.

Other variables used to build the models are presented in Table 1.

Table 1. Other variables used to estimate the models

Variable name	Description
T_MAX	maximum temp. obs. over a 24h period for a given weather station
T_MIN	minimum temp. obs. over a 24h period for a given weather station
T_AVG	average temp. obs. over a 24h period for a given weather station
DEW_POINT	dew point – temp. below which water vapour starts condensing. Expressed in degrees Celsius
HUMIDITY	humidity of the air; it takes values from 0 to 100
WIND_SPEED	wind speed (km/h)
WIND_GUST	wind gusts (km/h)
ATM_PRESSURE	atmospheric pressure, at sea level (hPa)
PRECIPITATION	total precipitation in the last 24 hours (mm)
CLOUD_COVER	total cloud cover
CLOUD_LOW	low cloud cover
SUNSHINE	number of hours of sunshine in the last 24 hours (hours)
VISIBILITY	visibility expressed in km
SNOW	total snowfall in centimetres in the last 24 hours
T_MAX4	zero-one variable taking value 1 for a max. temperature greater than 29 degrees Celsius
HOLIDAY_M1	zero-one variable with value 1 if the day before was a holiday
HOLIDAY_M2	a zero-one variable with value of 1 if there was a holiday two days before

Source: own calculations.

The dataset was split into two subsets. The 2019 data were left as a test set (it was not used in any of the model building stages). Data from 2010-2018 were used for model estimation.

Five competing predictive models were built. The first containing only intercept. The second model with explanatory variables produced only from calendar and holiday variables. The third model was enriched with weather variables. The fourth model used type-one regularization (ridge regression) and the fifth model with type-two regularization (lasso regression). The use of regularisation methods still ensures an easy interpretation of the results while reducing the variance of the random component.

All estimated models were compared with a common measure of RMSE.

The results for the first model with intercept are presented in Table 2.

Table 2. Estimated parameter of model (1) with intercept

Parameter	Estimate	Std. error	P(>|t|)
(Intercept)	111 150	159.3	<2E-16

Source: own calculations.

All estimated parameters in other models are statistically significant and the sign of the estimate is as expected.

RSME measure was used to compare the forecasting performance of the models. The calculated RMSE values for the learning set and the test set for all models are shown in Table 3.

Table 3. The calculated RMSE values for the learning set and the test set in the constructed models

Model	Train RMSE	Test RMSE
Model 1 (Intercept)	9127	8318
Model 2 (Calendar & Holiday)	6083	8899
Model 3 (Weather)	5608	7892
Model 4 (Ridge regression)	5326	7258
Model 5 (Lasso regression)	5329	7294

Source: own calculations.

Comparing the data presented in Table 4, we can conclude that model 1 predicts water demand better than model 2. The inclusion of weather variables in model 3 improves RSME on both the learning set and the test set. Even better results are obtained when using regularization methods (lasso and ridge). Finally, model 4 (ridge regression) was selected as the best model.

The results for the fourth model (ridge regression) are presented in Table 4.

Table 4. Estimated parameters of model (4) – ridge regression

Parameter	Description	Estimate
(Intercept)	(Intercept)	80767.520942
YEAR	Year	-2147.757928
JANUARY	January	9010.219430
FEBRUARY	February	10145.548916
MARCH	March	10397.937947
APRIL	April	8995.271739
MAY	May	7495.161123
JUNE	June	8566.800711
JULY	July	1604.157143
SEPTEMBER	September	5924.705100
OCTOBER	October	8775.879775
NOVEMBER	November	9511.791982
DECEMBER	December	10101.703019
HOLIDAY	Holiday	-11239.859751
MO	Monday	4687.873426
TU	Tuesday	5558.797385
WE	Wednesday	5992.229451
TH	Thursday	6114.855399
FR	Friday	5087.399054
SA	Saturday	4227.928486
T_MAX	Max. temperature	-209.509809
T_MIN	Min. temperature	-215.081234
T_AVG	Avg. temperature	1098.767247
DEW_POINT	Dew point	-554.396231
HUMIDITY	Humidity	70.395127
WIND_SPEED	Wind speed	-26.091345
WIND_GUST	Wind gust	-7.264865
ATM_PRESSURE	Atmospheric pressure	20.991359
PRECIPITATION	Precipitation	-31.693193
CLOUD_COVER	Cloud cover	-474.172728
CLOUD_LOW	Low Cloud cover	46.187129
SUNSHINE	Sunshine	-14.184110
VISIBILITY	Visibility	121.005270
SNOW	Snow	259.730182
T_MAX4	1 if temp. exceeds 29 Celsius degrees	4266.155406
HOLIDAY_M1	Holiday (day before)	-6052.783226
HOLIDAY_M2	Holliday (2 days before)	-3259.093499

Source: own calculations.

The best performance of the objective function in the ridge regression model was obtained for parameter $\lambda = 0.1$. R^2 coefficient in this model is 0.6594.

Best λ estimation, which minimizes the residuals (difference between observations and predicions), was achieved by recalculating 100 models with different values of λ.

In Figure 1 we observe calculated mean square error values for selected $\log(\lambda)$ values.

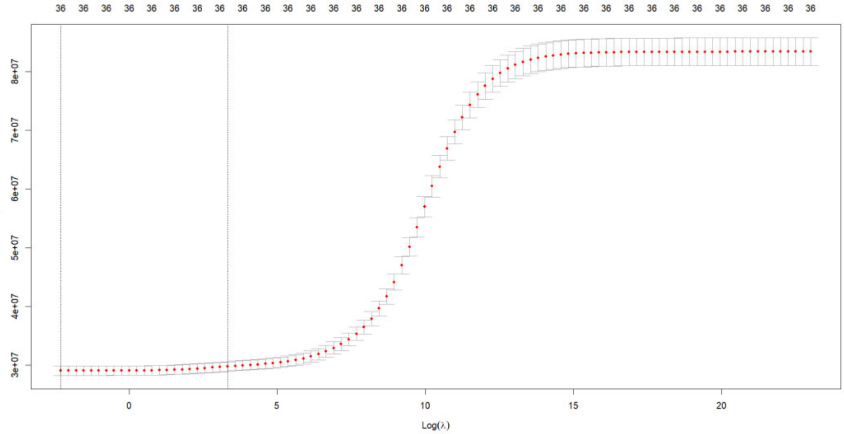


Figure 1. Mean square error values for $\log(\lambda)$ values used in the ridge regression model

5. Conclusions

This study examined the relation between daily weather variables and water use in the city of Lodz, Poland. Similar to previous studies, we found that maximum daily temperature is a good predictor of water demand. We also found that holidays are significant in decreasing the water demand. Moreover, like wind speed is a good predictor of water demand. It is likely that higher wind speed increases evaporation of water, which induces a cooling effect and thus decreases daily water consumption. Together, all these variables explain between 65% of the variations in the city of Lodz. Relatively similar results (up to 61% of the variations explained) were achieved by other authors using ARIMA model (Praskievicz and Chang, 2009).

Further models will also incorporate non-climatic variables such as sociodemographic, prices or structural variables (Zhang and Brown, 2005), which provide our models with greater explanatory power.

Acknowledgement

This paper was presented on the 39th International Scientific Conference “Multivariate Statistical Analysis” (MSA 2021), which financed its publication.

References

Balling, R. C., Jr. Gober, P., (2006). Climate variability and residential water use in the city of Phoenix, Arizona. *Journal of Applied Meteorology and Climatology*, Vol. 46, pp. 1130–1137.

Bates, B. C., Kundzewicz, Z. W., Wu, S. and Palutikof, J. P., (2008). Climate Change and Water: *Technical Paper of the Intergovernmental Panel on Climate Change*. Geneva, Switzerland: IPCC Secretariat.

Géron, A., (2017). Hands-On Machine Learning with Scikit-Learn and TensorFlow, Boston: O'Reilly.

Ghiassi, M., Zimbra, D. K. and Saidane, H., (2008). Urban water demand forecasting with a dynamic artificial neural network model. *Journal of Water Resources Planning and Management*, Vol. 134, pp. 138–146.

Grus, J., (2018). Data science od podstaw, Katowice: Helion.

Harrington, P., (2012). Machine Learning in Action, Shelter Island: Manning.

Patterson, J., Gibson, A., (2018). Deep Learning. *Praktyczne wprowadzenie*, Katowice: Helion.

Praskievicz, S., Chang, H., (2009). Identifying the Relationships Between Urban Water Consumption and Weather Variables in Seoul, Korea, *Physical Geography,* Vol. 30, pp. 324-337.

Walesiak, M., Gatnar, E., (2009). Statystyczna analiza danych z wykorzystaniem programu R, Warszawa: PWN.

Zhang, H. H. Brown, D. F., (2005). Understanding urban residential water use in Beijing and Tianjin, China. *Habitat International*, Vol. 3, pp. 469–491.

STATISTICS IN TRANSITION new series, June 2022
Vol. 23 No. 2, pp. 163–183, DOI 10.2478/stattrans-2022-0022
Received – 01.08.2021; accepted – 06.04.2022

Advances on Permutation Multivariate Analysis of Variance for big data

Stefano Bonnini[1], **Getnet Melak Assegie**[2]

ABSTRACT

In many applications of the multivariate analyses of variance, the classic parametric solutions for testing hypotheses of equality in population means or multisample and multivariate location problems might not be suitable for various reasons. Multivariate multisample location problems lack a comparative study of the power behaviour of the most important combined permutation tests as the number of variables diverges. In particular, it is useful to know under which conditions each of the different tests is preferable in terms of power, how the power of each test increases when the number of variables under the alternative hypothesis diverges, and the power behaviour of each test as the function of the proportion of true alternative hypotheses. The purpose of this paper is to fill the gap in the literature about combined permutation tests, in particular for big data with a large number of variables. A Monte Carlo simulation study was carried out to investigate the power behaviour of the tests, and the application to a real case study was performed to show the utility of the method.

Key words: big data, MANOVA, permutation test, multivariate analysis.

1. Introduction

In many applications of the multivariate analyses of variance (MANOVA), the classic parametric solutions for testing hypotheses of equality in population means or multisample and multivariate location problems might not be suitable for various reasons. For instance, the strong and implausible assumptions of iid observations and multivariate normality are the main reasons for considering parametric methods neither flexible nor robust and consequently often unsuitable. Moreover, in the presence of big data with a high number of response variables, great attention should be paid when the number of response variables is larger than the sample sizes, because of the loss of degrees of freedom.

[1] Department of Economics and Management, University of Ferrara, Italy. E-mail: bnnsfn@unife.it. ORCID: https://orcid.org/0000-0002-7972-3046.

[2] University of Parma, Italy. E-mail: mlkgnt@unife.it. ORCID: https://orcid.org/0000-0001-7288-9636.

Even if there is not a unique definition, in statistics, a dataset is usually classified as "big data" if it represents a collection of informative data, extensive in terms of volume, velocity and variety, such that specific analytical technologies and methods are required for the extraction of value or knowledge (Baro et al., 2015). Big data are typical of many empirical disciplines such as biomedicine, economics, biology, ICT, education and research, financial services, social media, automotive industries, etc. (Özköse et al., 2015). Frequently, the high volume of big data depends on the multivariate nature of the dataset, due to the large number of variables. In addition, the variety of big data, due to the presence of different types of variables (quantitative and qualitative) and to the variability and heterogeneity of data, makes inferential problems more complex and requires robust and valid techniques to make inferences. For instance, in studies focused on social media, text, video, audio, and image data are jointly analysed. Hence, tests of hypotheses for big data must be addressed with appropriate methods that lead to reliable decisions, in short times and taking into account the variability and heterogeneity of the information.

A typical approach to variable oriented multivariate problems consists in the application of exploratory methods based on the dimensionality reduction such as principal component analysis (PCA) or factor analysis (FA) (Johnson and Wichern, 2007; Farcomeni and Greco, 2016). For two-sample multivariate testing problems, in the presence of numeric data, a typical solution is the Hotelling T-square test. These methods are based on strong assumptions such as the linearity of the relationships between variables or normality.

Linearity is a very strong and often unrealistic assumption. Normality is a reasonable assumption only with large sample sizes due to asymptotic properties of the statistics. Nevertheless, even in cases where linearity and normality are reasonable assumptions, especially in inferential problems, in the presence of many variables the estimation of a large number of unknown parameters, such as covariances or correlations, is required. Moreover, when the sample size is less than the number of variables, a problem related to the degrees of freedom arises and some typical parametric methods, such as the Hotelling T-square test, are not applicable.

In such problems, nonparametric methods are preferable because they do not require that the underlying probability law belongs to a given family of distributions and no parameters need to be estimated. In particular, permutation tests follow a distribution-free approach and are almost as powerful as parametric methods based on normality when this assumption is true but much more powerful when the true underlying distribution deviates from the Gaussian (Pesarin, 2001; Anderson, 2001).

Solutions for multivariate tests within the family of permutation methods consider the dependence between response variables without modelling it explicitly, and consequently without the need of estimating parameters or assuming linearity

(Pesarin and Salmaso, 2010a; Bonnini et al., 2014; Arboretti et al., 2018). Permutation solutions for multivariate location problems have been proposed and studied mainly in terms of power and robustness with respect to the underlying distribution, especially comparing their performance with that of the classic parametric tests (Pillar, 2013; Anderson, 2001; Pesarin, 2001). An interesting proposal is based on the combination of the univariate permutation tests of the marginal variables (Pesarin, 2001). Pesarin and Salmaso (2010a,b) proved that the power of the most commonly used combined permutation tests, with fixed sample size and divergent number of variables under the alternative hypothesis, tends to one in the two-sample problem.

According to the type of the combining function used, a different combined test is obtained. Hence a deep study with the goal of comparing different combined tests, especially for big data with a large number of variables, is important and suitable, in order to find the most powerful test under different scenarios. To the best of our knowledge, for the multivariate multisample location problem, a comparative study of the power behaviour of the most important combined permutation tests as the number of variables diverges is missing. In particular, it is useful to know under which conditions each of the different tests is preferable in terms of power, how the power of each test increases when the number of variables under the alternative hypothesis diverges and the power behaviour of each test as a function of the proportion of true alternative hypotheses.

The purpose of this paper is to fill this gap in the literature about combined permutation tests. The paper is organized as follows. Section 2 is dedicated to a review of the literature on the MANOVA problem. The method of combined permutation tests is described in Section 3. In Section 4 the results of a comparative simulation study are reported and discussed. In Section 5, the application of the method to a real case study is presented. Finally, the conclusions are in Section 6.

2. Literature review

The goal of several empirical studies is the comparison of two or more populations in the presence of multivariate response variables. Often, regardless of the number of factors, the problem consists in testing the significance of treatment effects or the presence of a shift in some location parameters. In what follows, the variation of population means is investigated using multivariate analysis of variance (MANOVA). To test whether there is a significant difference between group means, various parametric multivariate tests based on strong assumptions have been proposed. The most commonly used are the Hotelling T-square test (Hotelling, 1992), the test of Wilks (1932) and the proposal of Pillai (1955). The main assumptions of these tests are normality, constant variances and continuous responses. Moreover, these methods cannot be applied for big datasets when the number of response variables is greater than the sample size.

Nonparametric solutions have been proposed to overcome the limits of the tests mentioned above due to the lack of robustness with respect to the assumptions (Pesarin and Salmaso, 2010a; Bonnini et al., 2014; Pillar, 2013; Bonnini, 2016). For instance, Anderson (2001) introduced a nonparametric solution based on the permutation test for an ecological problem. The permutation test statistic was the Fisher F ratio obtained from a distance matrix, and the simulation results proved the appropriateness of the permutation test for both one-way and two-way MANOVA. Pillar studied the accuracy and power of permutation tests for MANOVA based on different test statistics. According to his study, the sum of squares between groups with the Euclidean distance was preferable to the Chord distance and the sum of Fs of univariate ANOVA. Moreover, the simulation study revealed that the permutation test was powerful also under heteroscedastic and with unbalanced samples.

In the literature, several works concerning applications of permutation tests for one-way and two-way MANOVA have been published. A non-exhaustive list includes the following papers: Mantel and Valand (1970), Mielke et al. (1976), Clarke (1993), Pillar and Orlóci (1996), Legendre and Anderson (1999), Mielke and Berry (1999), McArdle and Anderson (2001), Arboretti et al. (2018), Finch (2016). However, the extension of the permutation test for two-way MANOVA requires great attention in permuting the statistical units between groups. This is because the exchangeability condition is guaranteed only within the levels of one factor by considering the second factor as a block. Thus, constrained permutations are essential (Anderson, 2001). The two-sample multivariate problem has been frequently considered. See for instance Pesarin and Salmaso (2010), Polko-Zajac (2020), Bonnini and Melak Assegie (2019). Instead, the multi-sample case has been addressed by fewer authors (see Bonnini, 2016). In some cases permutation solutions for complex problems such as multiaspect tests (Polko-Zajac, 2019), directional alternatives (Bonnini et al., 2014; Arboretti and Bonnini, 2009), tests for categorical data (Arboretti and Bonnini, 2008; Bonnini, 2014) have been developed. In this paper, we focus on multi-sample location problems for numeric variables and nondirectional alternative hypotheses.

3. Methods

3.1. Multivariate permutation test

The permutation test is a distribution-free test based on the assumption of exchangeability under the null hypothesis (Pesarin, 2001). To apply the permutation principle, the sample data are partitioned into groups based on the treatment levels in an experimental study and pseudogroups in an observational study. To this end, the structure of the dataset for $S \geq 2$ independent samples and V-dimensional response is represented by:

$$\boldsymbol{Y} = \{Y_{igq} | i = 1,2, \dots, n_g, g = 1,2, \dots, S, q = 1,2, \dots, V\} \tag{1}$$

The dataset $\boldsymbol{Y}$ takes values on the V-dimensional sample space Ω for which a σ-algebra $\mathcal{A}$ and a nonparametric family $\mathcal{P}$ of non-degenerate unknown distributions are defined, and supposed to be exchangeable.

Hypothesis testing based on the permutation approach requires a clear formulation of the null hypothesis. The null hypothesis in the MANOVA problem is defined as the equality of S multivariate (unknown) distributions:

$$H_o : \{P_1 = P_2 = \cdots = P_S\} = \{\boldsymbol{Y_1} \ =^d \boldsymbol{Y_2} \ =^d \dots =^d \boldsymbol{Y_S}\}. \tag{2}$$

Under homoscedasticity, the difference between the groups is due to a shift in location. Thus, the null hypothesis could be formulated as equality of group means for each response variable. Let $\boldsymbol{Y_g}$ be a V-variate numeric random variable such that $\boldsymbol{Y_g} = \boldsymbol{\mu} + \boldsymbol{\delta_g} + \boldsymbol{\varepsilon_g}$, with $\boldsymbol{\mu}$ vector of V unknown location parameters, $\boldsymbol{\delta_g}$, $g = 1, \dots, S$, vectors of V treatment effects and $\boldsymbol{\varepsilon_g}, g = 1, \dots, S$, exchangeable V-dimensional random vectors that follow an unknown probability distribution with equal variance-covariance matrix $\boldsymbol{\Sigma}$ and such that $E(\boldsymbol{\varepsilon_g}) = \boldsymbol{0}$.

The null hypothesis is:

$$H_o : \{\boldsymbol{\delta_1} \ = \boldsymbol{\delta_2} =, \dots, = \boldsymbol{\delta_S} = \boldsymbol{0}\} \tag{3}$$

A further decomposition of the null hypothesis with respect to the marginal distributions of the multivariate response can be considered. The multivariate hypothesis can be broken down into V partial null hypotheses:

$$H_o : \bigcap_{q=1}^{V} (\delta_{1q} =, \dots, = \delta_{Sq} = 0) \equiv \bigcap_{q=1}^{V} H_{oq} \tag{4}$$

where the intersection symbol means that the null hypothesis of the overall problem is true if all the V partial null hypotheses are true. Accordingly, with a similar approach, the alternative multivariate hypothesis H_1 of inequality in distribution may be represented as follows:

$$H_1 : \bigcup_{q=1}^{V} \overline{H}_{oq} \tag{5}$$

where the union symbol indicates that the alternative hypothesis is true if at least one partial null hypothesis is false and $\overline{H}_{oq}$ denote the negation of the q-th partial null hypothesis. It is worth noting that directional alternatives are also possible but the purpose of this paper is to focus on two-tailed multi-sample multivariate problems.

When the overall null hypothesis is true and the equality in distribution holds, the vector of V observations concerning a generic statistical unit comes from any of the S populations with equal probability. In other words, the exchangeability of units with respect to the populations/samples is satisfied. In order to determine the null distribution of the test statistic, all the possible assignments of the n units to the S samples can be considered. Without loss of generality, let us assume that the n_1 units of the first sample correspond to the first n_1 rows of the observed dataset $\boldsymbol{Y}$, the n_2 units of the second sample correspond to the next n_2 rows of the dataset, and so on, until the n_S units of the S-th sample that correspond to the last n_S rows of the dataset. Each possible assignment is equivalent to a permutation of the rows of the dataset or to resampling without replacement the n units with $n = n_1 + n_2 + \cdots + n_S$.

For computational convenience, instead of considering the exact test, based on all the $\frac{n!}{\prod_{g=1}^{S} n_g!}$ possible assignments of the n units to the S groups, a random sample of permutations is used according to the Conditional Monte Carlo method.

3.2. Partial tests

The application of the method of Combined Permutation Test to the permutation MANOVA presented above consists in carrying out one univariate permutation test for each partial hypothesis and in combining the p-values of the univariate tests. The dependence between the univariate partial test statistics, according to the permutation distribution, is taken into account in the resampling strategy by permuting the rows of the observed dataset instead of permuting the elements of each column independently of the other columns.

A suitable test statistic for each partial permutation test is the so-called Treatment Sum of Squares (SS_{Treat}), which depends on the deviations of the within-group sample means from the total sample mean. Hence, the q^{th} partial test statistic or equivalently the test statistic of the q^{th} partial test, with $q = 1,2,\ldots,V$, is

$$T_q = \sum_{g=1}^{S} n_g \left(\bar{Y}_{gq} - \bar{Y}_{\cdot q}\right)^2 \tag{6}$$

with $\bar{Y}_{\cdot q} = \frac{\sum_g n_g \bar{Y}_{gq}}{\sum_g n_g} = \frac{\sum_g n_g \bar{Y}_{gq}}{n}$, where $\bar{Y}_{gq}$ represents the mean of the values of the q-th variable observed in the g-th sample.

The multivariate permutation distribution of the test statistic $\boldsymbol{T} = (T_1, T_2, \dots, T_V)$ under the null hypothesis is obtained through the following procedure:

1) compute the vector of observed values of $\boldsymbol{T}$ from the dataset $\boldsymbol{Y}$:

$$\boldsymbol{T_{obs}} = \boldsymbol{T}(\boldsymbol{Y}) = (T_{1,obs}, T_{2,obs}, \dots, T_{V,obs})$$

2) randomly permute the rows of the dataset (or reassign statistical units to groups) and compute the values of the test statistics as a function of the permuted dataset: $\boldsymbol{T}^p = \boldsymbol{T}(\boldsymbol{Y}^p)$
3) repeat step (2) R times independently and compute the permutation test statistics. Let $T_{q,r}^p$ be the value of the q-th partial test statistic related to the r-th permutation of the dataset $\boldsymbol{Y_r^p}$. Hence

$$\boldsymbol{T_r^p} = \boldsymbol{T}(\boldsymbol{Y_r^p}) = (T_{1,r}^p, T_{2,r}^p, \dots, T_{V,r}^p)$$

4) estimate the significance level function of the partial tests

$$\hat{\lambda}_{q,r}^p = \lambda(T_{q,r}^p) = \frac{\sum_{j=1}^{R} \mathrm{I}(T_{q,j}^p \geq T_{q,r}^p) + 0.5}{R+1} \tag{7}$$

with $r = 1, 2, \dots, R$, $q = 1, 2, \dots, V$, and $I(E)$ indicator function of E, which takes value 1 if E is true and 0 otherwise. The p-value of the q-th partial test is $\hat{\lambda}_{q,obs}^p = \lambda(T_{q,obs}^p)$.

3.3. Combination

According to the method based on the combination of dependent permutation tests, the test statistic for the overall problem is obtained by combining the p-values of the partial tests. The synthesis of the information provided by the partial tests regarding the marginal variables is provided by the application of a suitable combining function φ. Hence, the test statistic useful for the overall test, the multivariate analysis of variance, is

$$T_{comb} = \varphi(\lambda_1, \lambda_2, \dots, \lambda_V).$$

The proposal of combining p-values of partial tests in order to solve multivariate, multi-aspect, multi-strata tests, or other complex testing problems that can be broken down into partial univariate tests, appeared for the first time in the literature twenty years ago in Pesarin (2001) and was later studied and developed by several authors. For extended but not exhaustive reviews, see Pesarin and Salmaso (2010a) and

Bonnini et al. (2014). Since, for the combination of the partial tests, $\varphi(\cdot)$ must satisfy some simple, mild and easily attainable conditions, several different functions can be used and each of them corresponds to a different solution with specific properties within the family of combined permutation tests.

A suitable combining function $\varphi: (0,1)^V \rightarrow \mathbb{R}$ must satisfy the following properties:

1) $\forall(\lambda'_q, \lambda''_q)$ in (0,1), $\lambda'_q < \lambda''_q \Leftrightarrow \varphi(\dots, \lambda'_q, \dots) \geq \varphi(\dots, \lambda''_q, \dots)$ ceteris paribus (non-increasing monotony)
2) $\exists \lambda_q \epsilon \{\lambda_1, \lambda_2, \dots, \lambda_V\}$ s.t. $\lambda_q \rightarrow 0 \Leftrightarrow \varphi(\lambda_1, \lambda_2, \dots, \lambda_V) \rightarrow \bar{\varphi} < \infty$ (finite supremum)
3) $\forall \alpha \epsilon (0,1)$, $\exists T_{comb,\alpha} < \bar{\varphi}$ where $T_{comb,\alpha}$ is the test critical value (finite critical value)

The most popular combining functions in the literature of combined permutation tests are Fisher, Liptak and Tippett functions. The Fisher omnibus combining function is

$$T_F = -2 \sum_q log(\lambda_q) \tag{8}$$

where $log(x)$ denotes the natural logarythm of x. Liptak`s combining function is based on the transformation of the complement to one of the p-values through the inverse of the cumulative distribution function (or the quantile function) of the standard normal distribution:

$$T_L = \sum_q \Phi^{-1}(1 - \lambda_q) \tag{9}$$

where $\Phi(x) = P(X \leq x)$ with $X \sim \mathcal{N}(0,1)$. Tippett combination is based on an order statistic and considers, as observed value of the combined test statistic, the complement to one of the most significant p-value:

$$T_T = max_q\{1 - \lambda_q\} \tag{10}$$

Under the null distribution, if the V partial tests are independent and continuous, the Tippett function follows the uniform distribution in (0,1).

Without loss of generality, let us assume that the null hypotheses of the overall and partial problems are rejected for large values of the respective test statistics. It is trivial to show that all three combination rules defined above satisfy this condition. Given that the observed value of the combined test statistic is

$$T_{comb,obs} = \varphi(\hat{\lambda}^p_{1,obs}, \hat{\lambda}^p_{2,obs}, \dots, \hat{\lambda}^p_{V,obs}).$$

the p-value of the permutation MANOVA with the combined permutation test is given by

$$\hat{\lambda}_{comb,obs} = \lambda(T_{comb,obs}) \tag{11}$$

The three presented tests can have much different power behaviours under different conditions, hence a comparative analysis to deepen their properties, advantages and limits is important to support the analyst in the decision about which test to use based on the power.

4. Simulation study

The power behaviour of the three combined permutation tests defined in the previous section for the MANOVA problem was investigated through a Monte Carlo simulation study. Different scenarios, under the null and the alternative hypothesis, were considered in order to compare the power of the three proposals as a function of the sample sizes, of the number of samples, of the number of components of the multivariate response and of the proportion of true partial alternative hypotheses when H_0 is false.

Data were simulated according to the one-way MANOVA model. We considered multivariate datasets with two different sizes from the point of view of the number of responses: $V = 50$ and $V = 100$. With regard to the number of compared samples, $S = 3$ and $S = 5$ are the cases taken into account. Simulation study has been carried out generating data from V-variate normal random variables hence under the "probabilistic condition most favorable to the classic parametric tests" and under homoscedasticity. For all the S populations, the variance of each of the V components of the multivariate response and the correlation between any pair of variables was set equal to 1 and to 0.3 respectively. Hence, the $V \times V$ covariance matrix of each population is $\Sigma = [\sigma_{kq}]$ with $\sigma_{qq} = 1$, $q = 1,2,\dots,V$, and $\sigma_{kq} = 0.3$, $k \neq q \epsilon \{1,2,\dots,V\}$.

The number of simulated datasets and the number R of permutations were both equal to 1000. In the simulations, we considered the balanced design with size n_1=$n_2 = \cdots = n_S = n$. The two sample sizes taken into account are $n = 10$ and $n = 30$. In the simulations, $\boldsymbol{\mu} = \mathbf{0}$. Let p be the proportion of true partial alternative hypothesis. Then, the V-variate normal distribution of the random variable that simulates data for the g-th sample $(g = 1,2,\dots,S)$ has a vector of means with $(1 - p)\ V$ zeros and pV values equal to $\tau(g - 1)$. Formally

$$\boldsymbol{\delta_g} = \tau(g - 1)\begin{pmatrix}\mathbf{1}\\ \mathbf{0}\end{pmatrix}$$

where $\mathbf{1}$ is a vector of pV elements equal to 1 and $\mathbf{0}$ is a vector of $(1 - p)\ V$ elements equal to 0. To consider different shifts in the population locations, the simulations were carried out with $\tau = 0.5$ and $\tau = 1.0$. Moreover, the different proportions p of true alternative hypotheses used in the scenarios are $0.00, 0.05/0.06, 0.10, 0.20, 0.30, 0.40, 0.50, 0.70, 0.90, 1$. The first positive proportion in the list is 0.05 if $q = 100$ (5 true partial alternative hypotheses) and 0.06 if $q = 50$ (3 true partial alternative hypotheses). The significance level chosen in all the scenarios is $\alpha = 0.05$. All simulations were carried out with the R programming software version 4.1.0. Specific scripts were created by the authors for this purpose.

Table 1 shows the rejection rates of the tests under all different cases when the number of variables V is equal to 100. The performance of the tests under H_0 can be evaluated from the column corresponding to $p = 0.00$ (no true partial alternative hypotheses). It is evident that, in most cases, the rejection rates are either less than or very close to the nominal α level 0.05. The test based on the Tippet combination exceeds α more frequently than the others but the probability of wrong rejection of H_0 seems to be not far from 0.05, hence we can say that all the tests are well approximated.

When $p > 0$, the power behaviour of the tests can be assessed under H_1. Unbiasedness of all the tests is demonstrated because the rejection rates are greater under the alternative hypothesis than under the null hypothesis. Moreover, the greater the sample size the higher the power, thanks to the consistency of the tests. As expected, the power is increasing function of the shift of the population locations that depends on τ. Finally, the greater the number of samples the higher the rejection rates of the tests. Focusing on the effect of p on the estimated probability of rejecting H_0 when it is false, the increasing monotonic relationship is evident for all the tests. The growth rate of the power with respect to p is high and, when 100% of the partial alternative hypotheses is true, the rejection of H_0 is sure or almost sure.

From the comparative analysis, it emerges that the Liptak test is always the worst, except in the case in which all the partial alternative hypotheses are true. As said, in this scenarios, the power of all the combined tests tends to one and the tests are equivalent. In general the lower performance of the test based on the Liptak combination is evident and it is uniformly less powerful than the other permutation MANOVAs. This is consistent with Pesarin's (2001) statement about the preferability of other tests than Liptak, except for p=1. When the proportion of true partial alternative hypotheses is low, the combined test based on Tippett's rule is by far the best. Also this conclusion is not surprising, according to Pesarin (2001) but, in our simulation study, the extent of the difference in performance of the test based on Tippet's function can be evaluated. Moreover, according to this results, Tippet's combination is never less performant than the others, except in the first setting, when $S = 3$, $n = 10$ and $\tau = 0.5$ when $p \geq 0.90$, where the differences in the rejection rates of the various tests are negligible.

Table 1. Rejection rates of combined permutation tests for $V = 100$ and $\alpha = 0.05$.

S	**n**	**τ**	**φ**	**Proportion of true partial alternative hypotheses (p)**									
				0.00	**0.05**	**0.10**	**0.20**	**0.30**	**0.40**	**0.50**	**0.70**	**0.90**	**1.00**
3	10	0.5	F	0.047	0.082	0.108	0.250	0.462	0.642	0.776	0.870	0.918	0.924
			L	0.045	0.078	0.074	0.156	0.308	0.466	0.596	0.792	0.882	0.914
			T	0.050	0.426	0.546	0.640	0.752	0.798	0.846	0.858	0.898	0.928
		1.0	F	0.036	0.106	0.310	0.918	1	1	1	1	1	1
			L	0.034	0.078	0.138	0.418	0.806	0.882	0.916	0.930	0.986	1
			T	0.054	0.988	1	1	1	1	1	1	1	1
	30	0.5	F	0.056	0.104	0.240	0.890	1	1	1	1	1	1
			L	0.056	0.080	0.132	0.340	0.822	0.884	0.902	0.938	0.988	1
			T	0.058	0.940	0.990	1	1	1	1	1	1	1
		1.0	F	0.046	0.124	0.342	0.996	1	1	1	1	1	1
			L	0.046	0.086	0.160	0.432	0.872	0.870	0.878	0.956	0.984	1
			T	0.052	1	1	1	1	1	1	1	1	1
5	10	0.5	F	0.046	0.136	0.374	0.968	1	1	1	1	1	1
			L	0.042	0.100	0.180	0.486	0.812	0.904	0.956	0.965	0.986	1
			T	0.052	0.984	1	1	1	1	1	1	1	1
		1.0	F	0.052	0.144	0.359	0.988	1	1	1	1	1	1
			L	0.054	0.104	0.168	0.502	0.838	0.862	0.924	0.934	0.984	1
			T	0.056	1	1	1	1	1	1	1	1	1
	30	0.5	F	0.050	0.130	0.370	0.994	1	1	1	1	1	1
			L	0.052	0.076	0.178	0.442	0.802	0.89	0.892	0.922	0.980	1
			T	0.054	1	1	1	1	1	1	1	1	1
		1.0	F	0.044	0.124	0.352	0.996	1	1	1	1	1	1
			L	0.034	0.072	0.156	0.500	0.840	0.898	0.904	0.944	0.980	1
			T	0.050	1	1	1	1	1	1	1	1	1

Source: author computations; F: Fisher, L: Liptak, T: Tippett, τ: location shift, φ:combining function

In Table 2, the rejection rates of the tests when the number of variables is $V = 50$ are reported. Again the good performance of the tests under the null hypothesis ($p = 0.00$) is proved by the values of the estimated power. These values are usually not greater than $\alpha = 0.05$ even if sometimes they exceed the significance level, especially in the case of Tippet's combination. Nevertheless, when greater than 0.05, the rejection rates under H_0 are not far from α and then the tests are well approximated. Hence, this conclusion is valid regardless of the number of variables V.

Table 2 confirms also that the probability of right rejection of the null hypothesis of MANOVA by the combined permutation tests increases with the sample size n, with the number of samples S, with the shift parameter τ and with the proportion of true partial alternative hypotheses p. Another empirical evidence of the simulation study is that in general the power is greater with 100 variables than with 50 variables. This statement seems obvious thinking to the tendency of the power to one when the number of variables diverges in the two-sample problem proved by Pesarin and Salmaso (2010b). They focus on the relationship between power of the overall test and non-centrality parameter in the case 100% of the variables are under the alternative hypothesis. According to our results, the power of the multi-sample tests in the case $V = 100$ is much greater than in the case $V = 50$ only when the percentage of true partial alternative hypotheses is low, otherwise the difference seems not evident and always in the same direction. Hence, in our opinion, for the power behaviour, the proportion of true partial alternative hypothesis matters and it is more important than the absolute number of true partial alternatives. For instance, when $V = 50$ and $p = 0.40$, the number of true partial alternative hypothesis is 20, exactly as when $V = 100$ and $p = 0.20$. But in the former case, when $S = 3$, $n = 10$ and $\tau = 0.5$, the rejection rates of the tests based on the Fisher, Liptak and Tippett combination are 0.626, 0.490 and 0.636 respectively; instead in the latter case, under the same scenario, 0.250, 0.156 and 0.640 respectively. Hence, even if the number of true alternative hypotheses is the same, the power of the tests based on the Fisher and Liptak combinations is much lower when the proportion of true partial alternative hypotheses is smaller. Tippett represents an exception. Consider, under the same scenario, the case $V = 50$ and $p = 0.20$ (rejection rate 0.466) and $V = 100$ and $p = 0.10$ (rejection rate 0.546). Hence, with the same proportion p, the power increases with V only in the case of Tippett's combination.

In general, the case $V = 50$, confirms that the Liptak combination is the best choice only when $p = 1$ but in this situation the power of the other tests is very similar. In most of the considered settings, the Tippett combination is preferable because the power quickly tends to 1 as the proportion of true alternative hypotheses diverges. When $S = 3$, $n = 10$ and $\tau = 0.5$ this is the most powerful test up to $p = 0.40$. For larger values of p it becomes the less powerful test.

Table 2. Rejection rates of combined permutation tests for $V = 50$ and $\alpha = 0.05$.

S	n	τ	φ	Proportion of true partial alternative hypotheses (p)									
				0.00	**0.06**	**0.10**	**0.20**	**0.30**	**0.40**	**0.50**	**0.70**	**0.90**	**1.00**
3	10	0.5	F	0.050	0.082	0.096	0.230	0.414	0.626	0.766	0.870	0.924	0.886
			L	0.054	0.066	0.074	0.142	0.258	0.490	0.668	0.828	0.912	0.888
			T	0.057	0.268	0.316	0.466	0.556	0.636	0.726	0.768	0.818	0.810
		1.0	F	0.042	0.054	0.260	0.892	0.996	1	1	1	1	1
			L	0.038	0.068	0.120	0.446	0.812	0.938	0.950	0.986	0.988	1
			T	0.050	0.094	0.966	1	1	1	1	1	1	1
	30	0.5	F	0.052	0.100	0.230	0.824	1	1	1	1	1	1
			L	0.054	0.070	0.138	0.348	0.756	0.936	0.954	0.974	0.992	1
			T	0.056	0.876	0.960	0.994	0.998	0.998	1	1	0.998	1
		1.0	F	0.038	0.128	0.278	0.996	1	1	1	1	1	1
			L	0.040	0.092	0.140	0.424	0.846	0.938	0.948	0.974	0.980	1
			T	0.050	1	1		1	1	1	1	1	1
5	10	0.5	F	0.038	0.164	0.310	0.904	0.998	1	1	1	1	1
			L	0.044	0.132	0.162	0.408	0.798	0.944	0.952	0.970	0.988	1
			T	0.051	0.946	0.994	0.996	1	1	1	1	1	1
		1.0	F	0.048	0.182	0.174	0.978	1	1	1	1	1	1
			L	0.052	0.114	0.356	0.452	0.826	0.950	0.948	0.970	0.994	1
			T	0.054	1	1	1	1	1	1	1	1	1
	30	0.5	F	0.052	0.126	0.316	0.990	1	1	1	1	1	1
			L	0.048	0.076	0.156	0.458	0.836	0.940	0.956	0.976	0.996	1
			T	0.054	1	1	1	1	1	1	1	1	1
		1.0	F	0.051	0.136	0.348	0.986	1	1	1	1	1	1
			L	0.049	0.072	0.156	0.468	0.852	0.938	0.954	0.976	0.990	1
			T	0.053	1	1	1	1	1	1	1	1	1

Source: author's computations; F: Fisher, L: Liptak, T: Tippett, τ: location shift, φ:combining function

5. Case Study about organizational well-being of University Workers

Organizational well-being is the first element that influences effectiveness, efficiency, productivity and development of a public organization. As part of objective 3 of the 2014-2016 Positive Action Plan proposed by the Equality Opportunities Office of the University of Ferrara (UNIFE), the Rector's Delegate for Equal Opportunities presented a project in order to promote the improvement of the working well-being of the administrative-technical staff. This project consists in the definition of interventions aimed at improving quality of working life based on findings deriving from empirical surveys.

A questionnaire was administered to a sample of 120 employees of UNIFE in order to assess the degree of work-related stress, to detect the opinions of employees with respect to the organization and the working environment and identify possible actions for the improvement of the general conditions of the public employees at UNIFE. One goal of the survey was also to test the existence of possible differences in organizational well-being among sub-groups of employees defined by gender and age.

The 120 respondents represent a random sample of the population of the technical-administrative staff. In order to test for the joint effect of gender and age on the organizational well-being at UNIFE, a simple random sample of 30 employees was selected from each of the following four groups:

- FU50: 50 years old or younger females,
- FO50: over 50 years old females,
- MU50: 50 years old or younger males,
- MO50: over 50 years old males.

The questionnaire, consisting of 79 questions, was administered to the respondents from the 4th to the 11th of December 2014. The questionnaire was designed by the Italian National Anti-Corruption Authority (ANAC) and the National Institute for Occupational Accident Insurance (INAIL) that decided to adopt a Likert scale, based on the first 6 integer values representing the level of agreement with respect to the 79 statements (1= not at all, …, 6=completely). It is worth noting that the permutation analysis of variance can be applied to numeric variables. The assumption of normality but not even that of continuity is required. Hence, it is a valid approach in the case of both interval and discrete scales. The results of the simulation study can be extended to testing problems for interval variables, and consequently applied to the case study. Even if, strictly speaking, the response variables in the considered application on organizational well-being are ordinal, it is common practice to treat them as interval data. In general, interval and discrete variables can be considered as the result of the discretization of continuous variables.

Furthermore, unlike the parametric approach, the permutation test does not require that a specific underlying family of distributions is known or assumed. The null permutation distribution of the test statistics can be determined regardless of whether the underlying distribution of the data is continuous or not. The 79 statements are reported in Appendix 1.

Let Y_{vg} be the random variable that represents the response concerning the v-th statement of an employee belonging to group g, with $v = 1,2, \dots ,79$ and $g \in G = \{FU50, FO50, MU50, FO50\}$. The testing problem can be represented by the following hypotheses:

$$H_0 : \bigcap_{v=1}^{79} \left[Y_{v,FU50} =^d Y_{v,FO50} =^d Y_{v,MU50} =^d Y_{v,MO50} \right]$$

vs

$$H_1 : \bigcup_{v=1}^{79} \left[\exists g', g'' \in G \text{ s.t. } Y_{v,g'} \neq^d Y_{v,g''} \right]$$

The significance level is $\alpha = 0.05$. According to the simulation study, the most suitable testing method seems to be the combined permutation test based on the Tippett combining function. The application of this test provides a p-value of 0.755, much greater than α. Hence the null hypothesis cannot be rejected. At the significance level 0.05, there is no empirical evidence to reject the null hypothesis of no difference of the organizational well-being between groups in favor of the hypothesis that the organizational well-being of the groups is not the same. In other words, we cannot conclude that there is a significance effect of gender and age on the employees' well-being. The analysis was carried out by the authors by creating specific R scripts for the implementation of the methodology.

It is worth noting that the final p-value of the combined test is invariant with respect to the combination strategy. In other words, if we perform a two-level combination, i.e. the first within-domain combination of partial tests and the second combination with respect to the domains, the final result is the same as obtained by permuting the partial tests all together at the same time (see Pesarin, 2001). If we had significance in the overall test, it would be useful to identify the partial tests that contribute to the overall significance. This can be done with a suitable adjustment of the p-values of the partial tests for controlling the Family Wise Error rate and avoiding the inflation of the type I error of the final combined test.

In this case, an interesting two-stage combination strategy could be of interest, because the questionnaire is divided into sections corresponding to partial aspects of organizational well-being. Each aspect corresponds to a set of questions and consequently to a domain of variables (construct). In the case of significance of the overall combined test, the analysis of the adjusted p-values of the partial combined tests related to the constructs would make sense. Unfortunately, the overall null hypothesis is not rejected.

This result proves that, in the University of Ferrara, the organizational well-being of the employees in terms of risks, working environment, respect, relationship with colleagues and office manager, transparency, motivation, etc. is not affected by age and gender. It could be considered as evidence of gender-age equality within the organization.

6. Conclusions

The purpose of the work is to deepen the study of the power behaviour of combined permutation tests for MANOVA problems with big data. The assessment of the convergence rate of the power to one as the proportion of variables under the alternative hypothesis increases and a comparison between the three most commonly used members within this family of tests represent the main scientific added value of the paper.

These nonparametric multi-sample location tests are well approximated, consistent, unbiased and powerful also for small sample sizes. The power is also an increasing function of the number of samples and of the number of variables of the dataset. The asymptotic behaviour of the tests when the number of variables diverges was studied and the simulations proved that the proportion of true partial alternative hypotheses is more important than the absolute number of variables of the dataset in explaining the increase of power. The test based on the Tippett combination represents an exception to this general rule.

This test seems to be much more powerful than the others when the proportion of true partial alternative hypotheses is not large but competitive also when the proportions are close to one. This is the only condition in which the test based on Liptak combination is competitive but, for small proportions of true alternatives, this test is by far the least powerful.

Definitely, it seems that, among the distribution free solutions to the multivariate analysis of variance in the family of combined permutation tests, the method based on the Tippet combination is in general preferable, especially if there are no preventive information about the possible percentage of variables (or marginal distributions) under the alternative hypothesis. Instead of the Tippett combination, the Fisher rule can be applied when the percentage is close to 100%. The Liptak combination seems to be non-convenient in general.

This methodological tool is an important and useful solution of testing problems for big data, especially when the number of variables is very large and the sample sizes are small. The usefulness and the effectiveness of the method is confirmed by the application to the case study concerning the survey on the organizational well-being at the University of Ferrara discussed in the paper.

Acknowledgments

The authors wish to thank the anonymous reviewers who have contributed significantly to the quality of the paper with their comments and suggestions.

The work was partially supported by the University of Ferrara, which funded the FIR project "Measuring and assessing social sustainability, inclusion and accessibility at university: methods and applications".

References

Anderson, M. J., (2001). A new method for non-parametric multivariate analysis of variance. *Austral ecology*, 26(1), pp. 32–46.

Arboretti, R., Bonnini, S., (2008). Moment-based multivariate permutation tests for ordinal categorical data. *Journal of Nonparametric Statistics*, 20(5), pp. 383–393.

Arboretti, R., Bonnini, S., (2009). Some new results on univariate and multivariate permutation tests for ordinal categorical variables under restricted alternatives. *Statistical Methods and Applications: Journal of the Italian Statistical Society*, 18(2), pp. 221–236.

Arboretti, R., Ceccato, R., Corain, L., Ronchi, F. and Salmaso, L., (2018). Multivariate small sample tests for two-way designs with applications to industrial statistics. *Statistical Papers*, 59(4), pp. 1483–1503.

Baro, E., Degoul, S., Beuscart, R. and Chazard, E., (2015). *Toward a literature-driven definition of big data in healthcare.* BioMed research international (https://doi.org/10.1155/2015/639021).

Bonnini, S., And Melak Assegie, G., (2019). *Permutation multivariate tests for treatment effect: theory and recent developments.* In SUSAN SSACAB 2019, pp. 30–30. The Biostatistics Research Unit of the South African Medical Research Council.

Bonnini, S., (2014). Testing for heterogeneity with categorical data: permutation solution versus bootstrap method. *Communications in Statistics: Theory and Methods*, 43(4), pp. 906–917.

Bonnini, S., (2016). Multivariate approach for comparative evaluations of customer satisfaction with application to transport services. *Communications in Statistics: Simulation and Computation*, 45(5), pp. 1554–1568.

Bonnini, S., Corain, L., Marozzi, M. and Salmaso, L., (2014). *Nonparametric hypothesis testing: rank and permutation methods with applications in R. John Wiley & Sons*.

Bonnini, S., Prodi, N., Salmaso, L., Visentin, C., (2014). Permutation approaches for stochastic ordering. *Communications in Statistics: Theory and Methods*, 43(10-12), pp. 2227–2235.

Clarke, K.R., (1993). Non-parametric multivariate analyses of changes in community structure. *Australian journal of ecology*, 18(1), pp.117–143.

Farcomeni, A. and Greco, L., (2016). Robust methods for data reduction. CRC press.

Finch, W.H., (2016). Comparison of multivariate means across groups with ordinal dependent variables: a Monte Carlo simulation study. *Frontiers in Applied Mathematics and Statistics*, 2, p. 2.

Hotelling, H., (1992). The generalization of Student's ratio. I*n Breakthroughs in statistics*, (pp. 54-65). Springer, New York, NY.

Johnson, R., (1997). Wichern. D., (2007). Applied multivariate statistical analysis. Prentice-Hall: London.

Legendre, P. and Anderson, M. J., (1999). Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. *Ecological monographs*, 69(1), pp.1–24.

Mantel, N., Valand, R. S., (1970). A technique of nonparametric multivariate analysis. *Biometrics*, pp. 547-558.

McArdle, B. H., Anderson, M. J., 2001. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology*, 82(1), pp. 290–297.

Mielke Jr, P. W., Berry, K. J., (1999). Multivariate tests for correlated data in completely randomized designs. *Journal of Educational and Behavioral Statistics*, 24(2), pp. 109–131.

Mielke Jr, P. W., Berry, K. J., Johnson, E. S., (1976). Multi-response permutation procedures for a priori classifications. *Communications in Statistics: Theory and Methods*, 5(14), pp. 1409–1424.

Özköse, H., Arı, E. S. and Gencer, C., (2015). Yesterday, today and tomorrow of big data. *Procedia-Social and Behavioral Sciences*, 195, pp. 1042–1050.

Pesarin, F., (2001). *Multivariate permutation tests: with applications in biostatistics*, Vol. 240. Wiley: Chichester.

Pesarin, F., Salmaso, L., (2010a). *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons: Chichester.

Pesarin, F., Salmaso, L., (2010b). Finite-sample consistency of combination-based permutation tests with application to repeated measures designs. *Journal of Nonparaetric Statistics*, 22(5), pp. 669–684.

Pillai, K. S., (1955). Some new test criteria in multivariate analysis. *The Annals of Mathematical Statistics*, pp. 117–121.

Pillar, V., (2013). How accurate and powerful are randomization tests in multivariate analysis of variance?. *Community Ecology*, 14(2), pp. 153–163.

Pillar, V.D.P., Orlóci, L., (1996). On randomization testing in vegetation science: multifactor comparisons of relevé groups. Journal of Vegetation Science, 7(4), pp. 585–592.

Polko-Zajac, D., (2019). On permutation location-scale tests. *Statistics in Transition*, 20(4), pp. 153-166.

Polko-Zajac, D., (2020). A comparative study on the power of parametric and permutation tests for a multidimensional and two-sample location problem. *Argumenta Oeconomica Cracoviensia*, 2(23), pp. 69–79

Wilks, S. S., (1932). Certain generalizations in the analysis of variance. *Biometrika*, pp. 471–494.

Appendix 1

Code	Statement
A.01	My working place is safe
A.02	I have been informed about the risks connected to my job
A.03	I am satisfied about the environment of my working place
A.04	I have suffered harassment
A.05	My dignity has been harmed at work
A.06	At work the smoking ban is respected
A.07	I usually take enough breaks
A.08	I can work hard
A.09	I am not comfortable when I am working
A.10	The colleagues are not polite with me
A.11	I am allowed to take a break when I wish
A.12	I don't have the chance to take enough breaks
B.10	At work I have suffered bullying
B.01	In the workplace I am respected in my trade union membership
B.02	In the workplace I am respected in my political orientation
B.03	In the workplace I am respected in my religious faith
B.04	My gender identity is an obstacle to my enhancement at work
B.05	In the workplace I am respected in my ethnicity and race
B.06	In the workplace I am respected in relation to my mother tongue
B.07	My age is an obstacle to my enhancement at work
B.08	In the workplace I am respected in relation to my mother tongue
C.01	The workload is assigned with equity
C.02	The responsibilities are assigned with equity
C.03	My salary is proportional to the commitment
C.04	The pay is differentiated according to quantity and quality of work
C.05	My manager makes work decisions impartially
D.01	At UNIFE the path of professional development of each employee is well defined and clear
D.02	At UNIFE the career opportunities depend on merit
D.03	UNIFE gives the possibility to develop skills and aptitudes of individuals in relation to the requirements of the different roles
D.04	My current role is appropriate to my professional profile
D.05	I am satisfied with my professional path within UNIFE
E.01	I know what is expected of my work
E.02	I have the skills to do my job
E.03	I have the resources and tools to do my job
E.04	I have an adequate level of autonomy in my work
E.05	My work gives me a sense of personal fulfilment
E.06	I know how to do my job
E.07	I understand what is expected of me at work
E.08	I have freedom of choice in deciding how to do my job
E.09	I have unattainable deadlines

Code	Statement
E.10	I have to work very hard
E.11	I have a say in deciding how fast I can do my job
E.12	I'm getting pressure to work overtime
E.13	I have freedom of choice in deciding what to do at work
E.14	I have to do my job very quickly
E.15	I have deadlines impossible to meet
E.16	I have a say in how to do my job
E.17	My working hours can be flexible
E.18	Job requests made to me by various people/offices are difficult to combine
F.01	I feel part of a team
F.02	I help colleagues even if it's not my job
F.03	I am esteemed and treated with respect by colleagues
F.04	In my group, those who have information make it available to everyone
F.05	The organization pushes to work in a group and to collaborate
F.06	If the job becomes difficult, I can count on the help of my colleagues
F.07	At work my colleagues show me the respect I deserve
F.08	I receive support information that helps me in my work
F.09	There are frictions or conflicts between colleagues
F.10	My colleagues give me the help and support I need
F.11	Colleagues are willing to listen to my work problems
G.01	My organization invests in people, including through adequate training
G.02	The rules of conduct are clearly defined
G.03	Organisational tasks and roles are well defined
G.04	The circulation of information within the organisation is appropriate
G.05	My organisation promotes measures to reconcile working time and life time
G.06	I have clear duties and responsibilities
G.07	I must neglect some tasks because I have too much to do
G.08	I know the goals of my department/office
G.09	Staff are always consulted on changes in work
G.10	I'm supported in emotionally challenging jobs
G.11	Workplace relations are strained
H.01	I am proud when I tell someone that I work at UNIFE
H.02	I am proud when UNIFE achieves good results
H.03	I am sorry if someone has a bad opinion of UNIFE
H.04	Values and behaviours at UNIFE are similar to mine
H.05	If possible, I would change company
I.01	Relative and friends think that UNIFE is important for the collectivity
I.02	Students think that UNIFE is important for the collectivity
I.03	People think that UNIFE is important for the collectivity

STATISTICS IN TRANSITION new series, June 2022
Vol. 23, No. 2, pp. 185–196, DOI 10.2478/stattrans-2022-0023
Received – 29.11.2021; accepted – 16.03.2022

Scaled Fisher consistency for the partial likelihood estimation in various extensions of the Cox model

Tadeusz Bednarski, **Piotr B. Nowak,**[1]
Magdalena Skolimowska-Kulig[2]

ABSTRACT

The Cox proportional hazards model has become the most widely used procedure in survival analysis. The theoretical basis of the original model has been developed in various extensions. In the recent years, vital research has been undertaken involving the incorporation of random effects to survival models. In this setting, the random effect is a variable (*frailty*) which embraces a variation among individuals or groups of individuals which cannot be explained by observable covariates. The right choice of the frailty distribution is essential for an accurate description of the dependence structure present in the data. In this paper, we aim to investigate the accuracy of inference based on the primer Cox model in the existence of unobserved heterogeneity, that is, when the data generating mechanism is more complex than presumed and described by the kind of an extension of the Cox model with undefined frailty. We show that the conventional partial likelihood estimator under the considered extension is Fisher-consistent up to a scaling factor, provided symmetry-type distributional assumptions on covariates. We also present the results of simulation experiments that reveal an exemplary behaviour of the estimators.

Key words: frailty models, Cox model, Fisher consistency.

1. Introduction

Statistical analysis of time-to-event data through survival regression models has become common practise in a variety of disciplines including mainly demography and medicine but also economics, actuarial science, reliability research and others. The regression framework allows for the inclusion of relevant factors, like gender, socio-economic status, or received treatment, which explain variation among the individuals or items being studied. However, such an analysis is nearly always susceptible to the omission of influential covariates and leaves unexplained variation. In some cases, the unobserved heterogeneity may cause inferential perturbations that are beyond the control of the researcher. One way of accounting for this estimation problem is to extend the model by including an unobserved random effect - a frailty variable, which allowed heterogeneity in longevity endowment. The notion of frailty was introduced and applied to the population data by Vaupel et al. (1979). The term *frailty* indicates that some individuals are frailer than others, that is, the event under consideration is more likely to happen for them. In its classical and mostly applied form, a

[1] Institute of Economic Sciences, University of Wrocław, Poland. E-mail: piotr.nowak2@uwr.edu.pl. ORCID:https://orcid.org/0000-0002-7404-2946.

[2] Institute of Economic Sciences, University of Wrocław, Poland. E-mail: magdalena.skolimowska-kulig@uwr.edu.pl. ORCID:https://orcid.org/0000-0002-4748-7624.

frailty model assumes proportional hazards and includes an unobservable random variable acting multiplicatively on the baseline hazard function. In recent years a number of papers and textbooks have appeared discussing extensions of common survival models to a wide variety of frailty models that are suitable to handle more complex survival data. A comprehensive review of frailty modelling in survival data analysis can be found in Hougaard (2000) and Wienke (2011). Kalbfleisch and Prentice (2002) give detailed theoretical treatments using the counting process theory. More applied presentations are given by Klein and Moeschberger (2003), and Therneau and Grambsch (2000). Aalen et al. (2008) provide an insight into the theoretical and applied structure of frailty models used in survival and event history analysis on the counting processes basis. Henderson and Oman (1999) investigate the consequences of ignoring frailty in analysis and fitting misspecified Cox proportional hazards models to the marginal distributions. The usual approach to statistical inference with unobserved frailty assumes a parametric family of distributions for frailty, usually gamma but also inverse Gaussian, positive stable, compound Poisson, or more general the power variance function family. For particular types of parametric frailty models the maximisation of the marginal likelihood leads to estimates of the parameters in the model, but for semiparametric frailty models more complex estimation techniques are needed (see Duchateau and Janssen, 2008). Certainly, modelling the frailty distribution is a remedy for biased estimation of regression parameters, but its limited choice relies mainly on their mathematical tractability.

Our objective is to investigate whether the traditionally used partial likelihood estimation method can be worthwhile when the model is misspecified, more precisely, the existing heterogeneity is neglected. In our considerations we apply the approach taken by Bednarski and Skolimowska-Kulig (2018, 2019) and Bednarski and Nowak (2021). They focused on the fundamental requirement needed in sound statistical inference about parameters, the Fisher consistency of estimators. They studied the behaviour of the standard estimators, like maximum likelihood for the exponential model or partial likelihood for the Cox model under extended models, with no assumption about distributional structure of frailty. Then, of course, the Fisher consistency condition need not be true, but it is shown that the commonly used procedures for the estimation of regression parameters in certain hazard-based survival models generate consistent up-to-scale estimators for extensions of these models. In the article we demonstrate that the partial likelihood estimator for the Cox regression model is Fisher consistent up to a scaling factor under an extended model with unobserved generalised frailty. The limitation we make is the distribution of covariates, which is assumed to be elliptically symmetric. In our approach to the scaled Fisher consistent estimation, we adapt the general ideas of Ruud (1983) and Stoker (1986) who studied regression coefficient estimators in particular regression models with assumed misspecification, and showed their up-to-scale consistency.

2. Maximisation criterion for the Cox estimator

In this part we remind the criterion that yields the regression parameter estimator in the Cox proportional hazards model (Cox, 1972). We assume that the survival time T, given the covariate vector X, has the conditional distribution function

$$F(t|x) = 1 - \exp\left(-\Lambda(t)e^{\beta' x}\right),$$

where $\Lambda(t) = \int_0^t \lambda(u)du$, λ is the baseline hazard function and β is the parameter vector. Suppose that we observe a sample $(T_i \wedge C_i, X_i)$, $i = 1, 2, \ldots, n$, where the censoring variable C is independent of the survival time T given the covariate vector X. To formulate the partial likelihood let us denote by $t_{(1)} < t_{(2)} < \ldots < t_{(m)}$ the unique failure times. The partial likelihood for the Cox model can be written as

$$L(\beta) = \prod_{i=1}^{m} \frac{e^{\beta' X_{(i)}}}{\sum_{j \in R_i} e^{\beta' X_j}},$$

where the risk set $R_i = \{k : T_k \wedge C_k \geq t_{(i)}\}$ consists of subjects that have not failed or been censored by time $t_{(i)}$ and $X_{(i)}$ is the covariate vector for the subject that fails at $t_{(i)}$.

The Cox estimator maximises $L(\beta)$, or equivalently

$$\left(\prod_{i=1}^{m} \frac{e^{\beta' X_{(i)}}}{\frac{1}{n}\sum_{j \in R_i} e^{\beta' X_{(j)}}}\right)^{1/n}.$$

Thus, it is given as

$$\arg\max_{\beta} \frac{1}{n} \sum_{i=1}^{n} \left(\beta' X_i - \ln \frac{1}{n} \sum_{T_j \geq T_i} e^{\beta' X_j}\right) \mathbb{I}_{\{T_i \leq C_i\}}.$$

If $F_n(t, c, x)$ denotes the empirical distribution function of the sample (T_i, C_i, X_i), $i = 1, 2, \ldots, n$, and sums are replaced by empirical integrals, then the above expression can be stated as

$$\arg\max_{\beta} \int \left(\beta' y - \ln \int_{t \wedge b \geq w} e^{\beta' x} dF_n(t, b, x)\right) \mathbb{I}_{\{w \leq c\}} dF_n(w, c, y).$$

Since F_n converge uniformly in probability to a true distribution F we can expect that under sufficiently stringent conditions, the maximising $\hat{\beta}_n$ converge in probability to

$$\arg\max_{\beta} \int \left(\beta' y - \ln \int_{t \wedge b \geq w} e^{\beta' x} dF(t, b, x)\right) \mathbb{I}_{\{w \leq c\}} dF(w, c, y),$$

if the latter solution exists.

3. Fisher consistency and scaled Fisher consistency

Since the right time censoring present in the Cox regression model plays no essential role in the argumentation presented in the paper we skip it in order to make the notation more concise. If the underlying cumulative distribution F comes from the Cox model, that is there is a parameter vector β_0 and a nonnegative baseline hazard λ yielding $\Lambda(t) = \int_0^t \lambda(s)ds$ such that

$$F(t|x) = 1 - \exp\left(-\Lambda(t)e^{\beta_0' x}\right),$$

then

$$\beta_0 = \arg\max_{\beta} \int \left(\beta' y - \ln \int_{t \geq w} e^{\beta' x} dF(t,x)\right) dF(w,y) \tag{1}$$

for every parameter vector β_0. The last property means that the Cox estimator is Fisher consistent at the model. The consistency is independent of the baseline hazard λ and it holds under censoring independent of survival time T given the covariate values X.

In general, all statistical estimators based on random samples are defined by explicit or implicit functionals of the corresponding empirical distribution functions. If values of such a functional coincide with true parameters when the empirical distribution is replaced by the true model distribution then we say that the functional is Fisher consistent. Without Fisher consistency the estimator cannot even be consistent asymptotically. Therefore, when studying estimation proposals we would put its Fisher consistency property in the first place. In practice, Fisher-consistent functionals (estimators) associated with a given parametric family are used even if we think the family imperfectly describes a real distribution and the discrepancy is deeper than the one resulting from occasional influential errors. It may therefore be justified in some instances to study what happens when the estimator-functional is used under reasonable nonparametric extensions of the original model. We make it precise in the case of the partial likelihood estimator. Define then a Cox model with generalised frailty and regression parameter β_0 via cumulative distribution function of time, conditional on covariates $X = x$ and frailty $A = a$ as

$$F(t|x,a) = 1 - \exp\left(-\Lambda(t,a)e^{\beta_0' x}\right), \tag{2}$$

where the cumulated hazard $\Lambda(t,a) = \int_0^t \lambda(t,a)dt$ takes now into account possibly complex individual changes in time to event distribution. To simplify forthcoming notations we will use the same letter F for model distributions from the Cox model with extended frailty as in the case of the strict Cox model. The scaled Fisher consistency of the Cox estimator means here that for each parameter value β_0 there exists $c > 0$, possibly depending on β_0, such that

$$c\beta_0 = \arg\max_{\beta} \int \left(\beta' y - \ln \int_{t \geq w} e^{\beta' x} dF(t,x,a)\right) dF(w,y,b) \tag{3}$$

for $F(t,x,a) = F(t|x,a)G(x)H(a)$, where G and H are the distributions of covariates and frailty respectively, and the random variables X and A are assumed independent.

4. Results

Consider the extended Cox model with generalised frailty independent on the covariates, given by (2) and the problem of maximisation of

$$\int \left(\beta' y - \ln \int_{t \geq w} e^{\beta' x} dF(t|x,a)dG(x)dH(a)\right) dF(w|y,b)dG(y)dH(b) \tag{4}$$

with respect to β. Since the above expression can be written as

$$\int \left(-\ln \int_{t \geq w} e^{\beta' x} dF_0(t|x,a) dG_0(x) dH(a) \right) dF_0(w|y,b) dG_0(y) dH(b),$$

where $F_0(t|x,a) = 1 - \exp\left(-\Lambda_0(t,a)e^{\beta_0' x}\right)$, $\Lambda_0(t,a) = \Lambda(t,a)e^{\beta_0' EX}$ and G_0 is the distribution of centred covariates $X - EX$, the maximisation can be reduced to covariates with expectation zero.

Lemma 1. *Let β_0 be the true parameter value and the covariate vector with zero mean be such that for every vector β the conditional expectation $E((\beta - proj_{\beta_0}\beta)X|\beta_0'X)$ is almost surely zero. Then β maximising (4) equals $c\beta_0$ for some real c.*

Proof. For F with centred covariates the maximisation of (4) is equivalent to the minimisation of

$$\int \left(\ln \int_{t \geq w} e^{\beta' x} dF(t|x,a) dG(x) dH(a) \right) dF(w|y,b) dG(y) dH(b),$$

which in turn can be stated as

$$\int \left(\ln \int_{t \geq w} e^{\beta' x} dF(t|x,a) dG_1(x|\beta_0' x) dG_2(\beta_0' x) dH(a) \right) dF(w|y,b) dG(y) dH(b), \quad (5)$$

where β_0 denotes the true parameter value, G_1 is the conditional distribution of X given $\beta_0'X$ while G_2 is the distribution of $\beta_0'X$.

Notice that for any nonzero parameter vector β we can write $\beta = c\beta_0 + \beta_1$, where $c\beta_0 = proj_{\beta_0}\beta$ is the orthogonal projection of β on β_0. Then (5) becomes

$$\begin{aligned}
&\int \left(\ln \int_{t \geq w} e^{c\beta_0' x} dF(t|\beta_0' x, a) \int e^{\beta_1' x} dG_1(x|\beta_0' x) dG_2(\beta_0' x) dH(a) \right) \\
&\qquad dF(w|y,b) dG(y) dH(b) = \\
&\int \left(\ln \int_{t \geq w} e^{c\beta_0' x} dF(t|\beta_0' x, a) E\left(e^{\beta_1' X} | \beta_0' X = \beta_0' x \right) dG_2(\beta_0' x) dH(a) \right) \\
&\qquad dF(w|y,b) dG(y) dH(b) \geq \\
&\int \left(\ln \int_{t \geq w} e^{c\beta_0' x} dF(t|\beta_0' x, a) e^{E(\beta_1' X | \beta_0' X = \beta_0' x)} dG_2(\beta_0' x) dH(a) \right) \\
&\qquad dF(w|y,b) dG(y) dH(b). \quad (6)
\end{aligned}$$

If for every vector β the conditional expectation $E(\beta_1'X|\beta_0'X)$ is almost surely zero then the right side of the inequality (6) equals

$$\int \left(\ln \int_{t \geq w} e^{c\beta_0' x} dF(t|x,a) dG(x) dH(a) \right) dF(w|y,b) dG(y) dH(b)$$

and we can conclude that the minimising value of β, if it exists, must be equal to $c\beta_0$. □

The following theorem is an immediate consequence of the above lemma and the scaled

Fisher consistency condition (3).

Theorem 1. *If β_0 is the true parameter value and $E(X|\beta_0'X) \in lin(\beta_0)$ almost surely then the partial likelihood estimator is scaled Fisher consistent under the Cox model with generalised frailty.*

Proof. Notice that for X such that $E(X|\beta_0'X) \in lin(\beta_0)$ almost surely and for any $\beta = c\beta_0 + \beta_1$, where $c\beta_0$ is the orthogonal projection of β on $lin(\beta_0)$, we have $E(\beta_1'X|\beta_0'X) = 0$ almost surely. □

There are important cases when the condition assumed in the above theorem holds. One of them is when X is spherically symmetric, that is if for every orthogonal matrix Γ the random vector ΓX is distributed as X.

Corollary 1. *If X has a spherically symmetric distribution then the partial likelihood estimator is scaled Fisher consistent under the Cox model with generalised frailty.*

The following conclusion results directly from the preceding considerations.

Corollary 2. *If β_0 is the true parameter value, M a nonsingular matrix, $X = MZ$ and $E(Z|\gamma_0'Z) \in lin(\gamma_0)$ almost surely for $\gamma_0 = \beta_0'M$ then the partial likelihood estimator is scaled Fisher consistent under the Cox model with generalised frailty.*

It is also quite straightforward to show that if $X = MZ$, where M is a nonsingular matrix and Z is spherically symmetric, then again we have the scaled Fisher consistency of the partial likelihood estimator.

Corollary 3. *If $X = MZ$ and Z has a spherically symmetric distribution then the partial likelihood estimator is scaled Fisher consistent under the Cox model with generalised frailty.*

Another special case covered by Theorem 1 is considered below.

Corollary 4. *Let the random vector $X = (X_1, \dots, X_k)'$ be exchangeable, i.e. $(X_{\pi(1)}, \dots, X_{\pi(p)})'$ and X have the same distribution for any permutation π of the set $\{1, 2, \dots, n\}$. If $\beta_0' = (b, b, \dots, b)$, $b \neq 0$, then the partial likelihood estimator is scaled Fisher consistent under the Cox model with generalised frailty.*

Proof. The exchangeability of X implies that $E(X_1|\beta_0'X) = \cdots = E(X_k|\beta_0'X)$. On the other hand $E(X_1 + \cdots + X_k|\beta_0'X) = X_1 + \cdots + X_k$ and finally $E(X_1|\beta_0'X) = (X_1 + \cdots + X_k)/k$. Therefore $E(X|\beta_0'X) \in lin(\beta_0)$ and the Fisher consistency holds. □

5. Simulation studies

In this section we present the results of simulation studies conducted to investigate how the violation of the symmetry assumption on the regressors distribution or the omission of the covariates may affect properties of the partial likelihood estimation of the regression parameters. The experiments also show exemplary up-to-scale estimation under the Cox model with various choices of generalised frailty. All simulations were run with the R programming language.

Example 1. *Estimation under non-symmetric covariates*

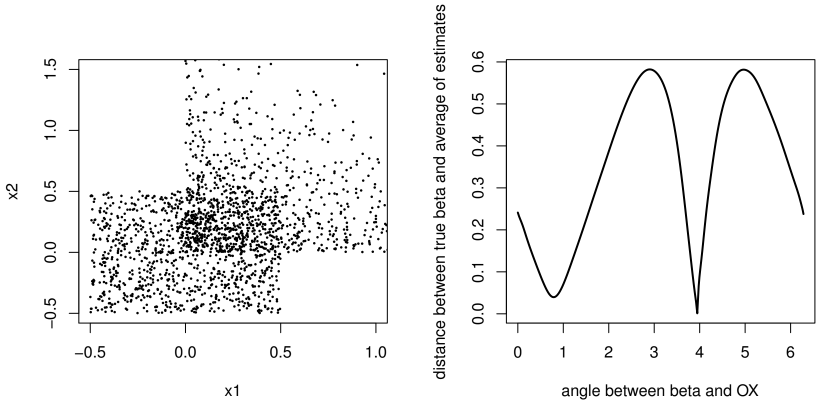
The sample of size 500 was generated from model (2) with $\Lambda(t,a) = ta$ and frailty given as a mixture of two gamma distributions. The covariates are from a two-dimensional uniform distribution mixed with a two-dimensional independent chi-square $\chi^2(2)$ (see the left panel of Figure 1). The curve shown in the right-hand panel of Figure 1 represents the distance $d(\alpha)$ between the true beta $\beta_0' = (\cos(\alpha), \sin(\alpha))$ and the normalized averaged of estimates for $\alpha \in [0,2\pi]$. The values on x axis are angles between the horizontal axis and true betas.

From the above description it follows that the density of the covariate vector $X = (X_1,X_2)'$ has the form

$$f(x_1,x_2) = \frac{1}{2} \cdot \mathbb{I}_{(-0.5,0.5)^2}(x_1,x_2) + \frac{1}{2}\exp(-(x_1+x_2)/2) \cdot \mathbb{I}_{(0,\infty)^2}(x_1,x_2).$$

Obviously X is not elliptically symmetric, however, since $f(x_1,x_2) = f(x_2,x_1)$, it is exchangeable. Thus, for $\beta_0' = (\sqrt{2}/2, \sqrt{2}/2)$ and $\beta_0' = (-\sqrt{2}/2, -\sqrt{2}/2)$ by Corollary 4 the partial likelihood estimation is scaled Fisher consistent. Hence, we can see that function $d(\alpha)$ attains minimum for $\alpha = \pi/4$ and $\alpha = 5/4\pi$.

Figure 1: Typical covariates and distance between true parameter and average estimates.



Example 2. *Up-to-scale estimation for various types of generalised frailty*

This example provides Monte Carlo simulations for different choices of generalised frailty $\Lambda(t,a)$. Five forms of generalised frailty for A distributed as shifted binomial distribution $binom(1,0.5)+1$ are considered (see Table 1). Observe that all functions $\Lambda(t,a)$ have the following properties: for $a \in \{1,2\}$ $\Lambda(0,a) = 0$ and $\Lambda(t,a) > 0$ for every positive t, $\lim_{t\to\infty}\Lambda(t,a) = \infty$ and they are continuously differentiable and strictly increasing on $t \in (0,\infty)$. The conditional distribution of the survival time T given $[X = x, A = a]$ was generated using the formula $\Lambda^{-1}(-\ln(U)\exp(-\beta_0'x),a)$, where U follows the uniform distribution on the interval $[0,1]$. The true parameter value was taken as $\beta_0' = (1,0.5,0.2)$

and the regressors $X = (X_1, X_2, X_3)'$ were used: either with standard normal or exponential distributions and $Cor(X_i, X_j) = 0.7$ for $i \neq j$. Estimations were repeated 5000 times for a sample size of 500. For each combination of the covariates and the generalised frailty two vectors are given as a result. The first one refers to scales - the means of ratios of estimates and the true parameters. Under normally distributed regressors the scaled Fisher consistency holds, so we expect the scales to be the same. The second vector in each cell consists of standard deviations of estimates. Simulations indicate good asymptotic behaviour of the estimators in the model with normal regressors as the differences in scales are very slight. Other choices of elliptically symmetric regressors, not presented in this example, lead to similar results. In the case of non-symmetric covariates the estimation brings worse results.

Table 1: Results of simulation experiment for different choices of generalised frailty. The first vector in each cell refers to the means of ratios of components of estimates and the true parameters. The second one refers to the standard deviations of the vector estimates of true parameter values.

Generalised frailty	Normal regressors	Nonormal regressors
$\Lambda(t,a) = a\sqrt{t}$	(0.9361, 0.9398, 0.9324) (0.0791, 0.0756, 0.0729)	(0.9260, 0.9357, 0.9434) (0.0865, 0.0739, 0.0728)
$\Lambda(t,a) = a\sqrt{t} + t$	(0.6228, 0.6228, 0.6122) (0.0770, 0.0700, 0.0676)	(0.2747, 0.3125, 0.3886) (0.0864, 0.0784, 0.0761)
$\Lambda(t,a) = t^2 + at - t$	(0.8855, 0.8817, 0.8839) (0.0781, 0.0726, 0.0704)	(0.6628, 0.6889, 0.7401) (0.0835, 0.0784, 0.0731)
$\Lambda(t,a) = t^{3a/2-1}$	(0.7588, 0.7589, 0.7572) (0.0775, 0.0733, 0.0708)	(0.4612, 0.5104, 0.5937) (0.0924, 0.0844, 0.0792)
$\Lambda(t,a) = t^{3a/2-1} + t^2$	(0.8560, 0.8520, 0.8498) (0.0804, 0.0726, 0.0695)	(0.4799, 0.5271, 0.6110) (0.0949, 0.0844, 0.0817)

Example 3. *The effect of variable's omission in the Cox model*

The above considerations show a wide range of distributional possibilities for the explanatory variables for which the estimation of regression parameters in the Cox model is scale Fisher consistent under the extended model with generalised frailty. As a particular case of $\Lambda(t,a)$ assume that $\Lambda(t,a) = a\Lambda(t)$. The frailty variable A has a special interpretation in survival analysis for the Cox model, where it is supposed to describe proportional changes of cumulated hazard $\Lambda(t)$ for individual units within the population. Since

$$P(T > t|x,a) = \exp\left(-\Lambda(t)a\exp(\beta_0'x)\right) = \exp\left(-\Lambda(t)\exp(\beta_0'x + \ln(a))\right)$$

it can as well be interpreted as a missing (independent) covariate. In practical data analysis it would be difficult to specify in any reasonable way the distributional form of the missing covariate.

A Monte Carlo experiment was conducted to investigate properties of the partial likelihood estimation when variables are omitted and data are generated from the true Cox model. In order to demonstrate the effect of variable's omission the following Cox model was taken: $\beta_0' = (-1,-1,0.5,1)$, $\Lambda(t) = t^2$ and $X = (X_1,X_2,X_3,X_4)'$ where: (X_1,X_2) has the distribution described in Example 1, (X_3,X_4) has the two dimensional normal distribution with correlation of $1/\sqrt{2}$ and vectors (X_1,X_2) and (X_3,X_4) are independent.

The sample size of $n = 500$ was taken and the estimation was repeated 5000 times. Simulation results, given by the means of ratios of estimates and parameters, and by standard deviations of estimates, are presented in Table 2. The up-to-scale consistent estimation of the corresponding regression coefficients is revealed for covariates vector (X_1,X_2) and (X_3,X_4). For the estimation based on the entire set of regressors the estimation is consistent with the scale of one. For other regressor vectors it can be observed that departure from the elliptically symmetric distribution implies the lack of scaled consistency in estimation.

Table 2: Results of simulation experiment. The first and the second vector in each cell refers to mean scales and standard deviations of estimates of true parameter values corresponding to the subset of regressors.

Subsets of X	Scales \Sd	Subsets of X	Scales\Sd
(X_1,X_2)	(0.4299, 0.4228) (0.0774, 0.0816)	(X_1,X_2,X_3)	(0.6637, 0.6611, 1.9944) (0.0856, 0.0844, 0.0328)
(X_1,X_3)	(0.7613, 1.9430) (0.0895, 0.0330)	(X_1,X_2,X_4)	(0.9226, 0.9330, 1.1621) (0.0824, 0.0840, 0.0257)
(X_1,X_4)	(1.0396, 1.1299) (0.0841, 0.0247)	(X_1,X_3,X_4)	(1.1106, 0.9455, 0.9476) (0.0795, 0.0336, 0.0303)
(X_3,X_4)	(0.8820, 0.8854) (0.0345, 0.0301)	(X_1,X_2,X_3,X_4)	(1.0060, 1.0001, 1.0007, 1.0020) (0.0775, 0.0765, 0.0327, 0.0295)

Example 4. *Mayo Clinic Primary Biliary Cirrhosis Data*

The example is based on the data from the Mayo Clinic trial in PBC, available in the package `survival` of `R` program. In this example we consider four explanatory variables:

- age
- edema (0 for no edema, 0.5 for moderate and 1 for severe edema)
- bili (serum bilirunbin mg/dl)
- protime (standardised blood clotting time)

Before fitting the Cox model we logarithmically transformed variables bili and protime. The assumption of elliptical symmetry was checked by three tests implemented in the package `ellipticalsymmetry` in `R` program. We chose test MPQ by Manzottii et al. (2002) (Test 1), test by Schott (2002) (Test 2) and test by Huffer and Park (2007) (Test 3). The results are summarized in Table 3. Let $\hat{\beta}_0$ denote the estimate of the coefficients in the Cox model for

Table 3: Fitting the Cox model for PBC data with dropped regressors.

Subsets of regressors	P-values of tests for elliptical symmetry			Scales
	Test 1	Test 2	Test 3	
age, edema	<2.2e-16	<2.2e-16	<2.2e-16	(0.7585, 1.7932)
age, ln(bili)	0.1412	0.3712	0.0024	(1.0506, 1.1204)
age, ln(protime)	0.0078	0.0954	0.0008	(0.9026, 1.5432)
edema, ln(bili)	<2.2e-16	<2.2e-16	<2.2e-16	(1.1993, 0.9885)
edema, ln(protime)	<2.2e-16	0.0118	<2.2e-16	(1.6157, 1.2628)
ln(bili), ln(protime)	0.6853	0.0600	0.6400	(1.0273, 1.1492)
age, edema, ln(bili)	<2.2e-16	<2.2e-16	<2.2e-16	(1.0126, 1.1380, 1.0403)
age, edema, ln(protime)	<2.2e-16	0.0002	<2.2e-16	(0.7483, 1.5422, 1.2809)
edema, ln(bili), ln(protime)	<2.2e-16	0.0004	<2.2e-16	(1.0657, 0.9487, 0.9526)
age, ln(bili), ln(protime)	0.0015	0.0933	4.8e-05	(1.0335, 1.0550, 1.2253)

regressors: age, edema, ln(bili), ln(protime) and let $\hat{\beta}$ denote the estimate of the coefficients in the Cox model based on the subset of this regressors. For each subset of regressors we computed scales as ratios of the coefficients for corresponding variables in $\hat{\beta}_0$ and $\hat{\beta}$. All tests detect correctly the lack of elliptical symmetry when regressors include discrete variable edema. The lack of elliptical symmetry of explanatory variables may imply the scales not being the same after omitting regressors. For the regressors (age, ln(bili)) and (ln(bili), ln(protime)) the differences in the scales are slight. They seem to be elliptical symmetric (see p-values in Table 3).

6. Concluding discussion

An important property of the Cox model is that the baseline hazard is an unspecified function and makes the Cox model of a semiparametric type. A key reason for the popularity of the Cox model is that even though the baseline hazard is not specified, reasonably good estimates of regression coefficients, hazard ratios of interest, and adjusted survival curves can be obtained for a wide variety of data situations. Frailty models arise naturally from the Cox model with unobserved covariates, which form the frailty parameter and handle right-censoring and left-truncation, which is crucial in time-to-event analysis. Frailty gives way to explain additional time variability that could not be grasped by the original Cox model. The usual investigation of the partial likelihood estimator for the Cox regression model involved an interest in the consistency of the partial likelihood estimator under the Cox model with frailty, which presumes the time distribution dependent on a single baseline hazard λ, multiplied by a positive random variable A called frailty. Fisher consistency of the Cox estimator was studied under the independence of frailty A from the covariates X and under analytically convenient frailty distributions. Nevertheless, attributing to population individuals the same baseline up to a proportionality factor (frailty) and making consistent estimation dependent on purely analytic properties of frailty distribution seemed far from satisfactory. In fact studies show that the Fisher consistency does not hold under arbitrary frailty. What we could naturally hope for then, would be the so-called scaled Fisher consistency - regression parameters could be estimated consistently up to an unknown scaling factor. In the paper we demonstrate that this is attainable, the classical partial likelihood procedure leads to the estimator satisfying this condition up to a scaling factor under the extended Cox model with generalised frailty and an elliptically symmetric distribution of the covariates. The simulation studies indicate the lack of this property in the case of violating the assumption. The Cox model with generalised frailty is of great importance in various analyses of time-to-event data. However, it should be noted that the omission of an influential variable or misspecification of the frailty distribution may lead to severe estimation errors. In this light, considering estimation consistent up to scale may result in meaningful comparisons of impact of covariates on hazards.

References

Aalen, O. O., Borgan, O., Gjessing, H. K., (2008). *Survival and Event History Analysis. A Process Point of View*, Springer.

Bednarski, T., (1993). Robust estimation in Cox regression model. *Scandinavian Journal of Statistics*, 20, pp. 213–225.

Bednarski, T., Nowak, P. B., (2021). Scaled Fisher consistency of the partial likelihood estimator in the Cox model with arbitrary frailty. *Probability and Mathematical Statistics*, 41(1), pp. 77–87.

Bednarski, T., Skolimowska-Kulig, M., (2018). Scaled consistent estimation of regression parameters in frailty models. *Acta Universitatis Lodziensis Folia Oeconomica*, 5(338), pp. 133–142.

Bednarski, T., Skolimowska-Kulig, M., (2019). On scale Fisher consistency of maximum likelihood estimator for the exponential regression model under arbitrary frailty. *Statistics and Probability Letters*, 150, pp. 9–12.

Cox, D. R., (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society B*, 34, pp. 187–220.

Duchateau, L., Janssen, P., (2008). *The Frailty Model*, Springer-Verlag.

Henderson, R., Oman, P., (1999). Effect of frailty on marginal regression estimates in survival analysis. *Journal of the Royal Statistical Society B*, 61, pp. 367–379.

Hougaard, P., (2000). *Analysis of Multivariate Survival Data*, Springer Verlag.

Huffer, F. W., Park, C., (2007). A test for elliptical symmetry. *Journal of Multivariate Analysis* 98 (2), 256–281.

Kalbfleisch, J. D., Prentice, R. L., (2002). *The Statistical Analysis of Failure Time Data*, Hoboken, N.J.: Wiley.

Klein, J. P., Moeschberger, M. L., (2003). *Survival Analysis. Techniques for Censored and Truncated Data*, New York: Springer-Verlag.

Manzotti, A., Pérez, F. J., Quiroz, A. J., (2002). A Statistic for Testing the Null Hypothesis of Elliptical Symmetry. *Journal of Multivariate Analysis*, 81 (2), 274–285.

Ruud, P., (1983). Sufficient conditions for the consistency of maximum likelihood estimation despite misspecification of distribution in multinomial discrete choice models. *Econometrica*, 51(1), pp. 225–228.

Schott, J. R., (2002). Testing for elliptical symmetry in covariance-matrix-based analyses. *Statistics & Probability Letters*, 60 (4), pp. 395–404.

Stoker, T., (1986). Consistent estimation of scaled coefficients. *Econometrica*, 54(6), pp. 1461–1481.

Therneau, T. M., Grambsch, P. M., (2000). *Modeling Survival Data: Extending the Cox Model*, New York: Springer-Verlag.

Vaupel, J. W., Manton, K. G., Stallard, E., (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16, pp. 439–454.

Wienke, A., (2010). *Frailty Models in Survival Analysis*, Chapman & Hall/CRC.

STATISTICS IN TRANSITION new series, June 2022
Vol. 23 No. 2, pp. 197–208, DOI 10.2478/stattrans-2022-0024
Received – 20.06.2020; accepted – 28.07.2021

On the quick estimation of probability of recovery from COVID-19 during first wave of epidemic in India: a logistic regression approach

Hemlata Joshi[1], S. Azarudheen[2], M. S. Nagaraja[3], Singh Chandraketu[4]

ABSTRACT

The COVID-19 pandemic has recently become a threat all across the globe with the rising cases every day and many countries experiencing its outbreak. According to the WHO, the virus is capable of spreading at an exponential rate across countries, and India is now one of the worst-affected country in the world. Researchers all around the world are racing to come up with a cure or treatment for COVID-19, and this is creating extreme pressure on the policy makers and epidemiologists. However, in India the recovery rate has been far better than in other countries, and is steadily improving. Still in such a difficult situation with no effective medicine, it is essential to know if a patient with the COVID-19 is going to recover or die. To meet this end, a model has been developed in this article to estimate the probability of a recovery of a patient based on the demographic characteristics. The study used data published by the Ministry of Health and Family Welfare of India for the empirical analysis.

Key words: COVID-19, epidemic, coronavirus disease, recovery estimation, logistic regression, logit analysis.

1. Introduction

Coronaviruses are the group of related RNA viruses which has ribonucleic acid as its genetic material. These viruses cause diseases in humans, other mammals and birds and sickness may range from common cold to severe respiratory diseases. COVID-19

[1] CHRIST (Deemed to be University), Bangalore, India. E-mail: hemlata.joshi28@gmail.com. ORCID: https://orcid.org/0000-0002-4051-6330

[2] CHRIST (Deemed to be University), Bangalore, India. E-mail: azarudheen.s@christuniversity.in. ORCID: https://orcid.org/0000-0001-7568-4273.

[3] CHRIST (Deemed to be University), Bangalore, India. E-mail: nagaraja.ms@christuniversity.in. ORCID: https://orcid.org/0000-0002-6900-8436.

[4] CHRIST (Deemed to be University), Bangalore, India. E-mail: chandraketu.lko@gmail.com. ORCID: https://orcid.org/0000-0003-2367-5396.

is the most recent disease that has jumped off to humans. Initially the eruption of the novel coronavirus was documented in China's Wuhan at the beginning of December 2019 and then circularized all across the world. Often during coughing or sneezing, the infection of coronavirus disease disseminates from one human to others via droplets raised from the respiratory system of the infected humans (WHO, 2020). The COVID-19 symptoms generally include fever, dry cough, tiredness, and in severe cases, infection can lead pneumonia, shortness of breath, chest pain, loss of speech or movement, kidney failure, and even death (WHO, 2020), but approximately 20 percent of the cases have been deemed to be severe (Singh et al., 2020). The World Health Organization (WHO) announced this COVID-19 a pandemic on 11 March 2020 and ingeminated the call for countries to take quick actions and scale up response to treat, detect and reduce transmission to save people's lives. The developed countries such as the United States of America, Italy, Spain, France, UK, etc. are struggling to overcome the disease spreading by novel coronavirus. According to WHO, by the end of May 2020 it has spread in around 188 countries, the total number of cases have exceeded 6 million and approximately 3.7 lakh deaths worldwide. In India, the first case of coronavirus infection was observed in Kerela on 30 January 2020 and for the two months, the spread of the coronavirus disease was extremely slow may be due to the strict nationwide lockdown. After that, the Government of India gave the conditional relaxation in the nationwide lockdown and during this period of lockdown, the coronavirus cases started increasing with the exponential rate. Although the incubation period for the coronavirus disease has not been confirmed yet, from the pooled analysis it is seen that the symptoms may appear in 2 days to 14 days (Singhal, 2020) and the Government of India has declared minimum 14 days quarantine period for the suspected cases. In the absence of any efficacious medicine or vaccination, the social distancing has been consented as a most efficient scheme for cutting the severity of this coronavirus disease all across the globe (Ferguson et al., 2020; Singh et al., 2020).

As India is the second largest most populated country and majority of the population live under the inadequate hygiene and with insufficient medical facilities such as lack of testing kits, labs and health personnel, etc., and with the relaxation in lockdown, the coronavirus disease may start spreading at community level. In the middle of June, the total confirmed COVID-19 cases crossed 3.43 lakh with an increase of more than ten thousand cases in a single day and the new cases was rising at the record pace while the deaths have come up to 9900 with 380 fatalities. If the same rate continues, India will reach the sixth position in the most affected countries by COVID-19, and presently India is the 7th worst affected country after the USA, Brazil, Russia, UK, Spain and Italy (WHO), and in terms of the fatality rate, India is at the twelfth position while it is ranked 8th in terms of recovery rate from coronavirus disease currently.

The Prime Minister of India Mr. Narendra Modi stated that currently India is being listed amongst the countries with the least number of deaths due to coronavirus and also said that the death rate can still be reduced if we all follow all the guidelines suggested by WHO. PM Modi also said that the decision of nationwide lockdown on time served better in controlling the speed of spreading of coronavirus disease in India. According to the ICMR's serological survey, about 0.73% of the population was exposed to the virus by the mid-June and India could have 200 million COVID infected people by September. The Indian Council of Medical Research (ICMR) said that India was not in community transmission yet but a large chunk of the population is at risk and physical distancing and other similar measures need to continue. The return of millions of migrants to villages in Bengal, UP, Bihar, Orissa, Chhattisgarh, Jharkhand, etc. will lead to a surge of infections in these rural hinterlands.

As COVID-19 is a new pandemic, it has become a challenging task in front of the scientists and researchers to fight with this coronavirus disease in the absence of vaccine. Thus, to know its behaviour and nature a lot of research is being done all across the globe, so that it could help the scientists or epidemiologists to possibly cure humans from its infections. The published data on COVID-19 pandemic are analysed by many researchers by using various mathematical modelling approaches (Rao et al., 2020; Chen et al., 2020). Huang et al. (2020) worked on the clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Modelling and forecasting of the COVID-19 pandemic is done by Anastassopoulou et al. (2020), Corman et al. (2020), Rothe et al. (2020) and Gamero, J. et al. (2020) and many interesting results have been obtained using the principles of mathematical modelling. Nikolay et al. (2020) used the coronavirus data and compared the Verhult model with the half-logistic curve of growth with polynomail variable transfer model. Further, they have compared the Verhulst growth model with Verhulst curve of growth with polynomail variable transfer model on the Covid-19 data and also have studied the intrinsic properties of some models of growth with polynomial variable transfer that give a very good approximation of the specific data on the pandemics in Cuba. Zaliskyi et al. (2020) built a mathematical model for COVID-19 data of European countries. In this article, an effort has been made to estimate the probability of recovery from the coronavirus disease using the indirect method of estimation. For this a logistic regression techniques has been used and for the empirical analysis, the available information about the demographic variables such as age and gender of the patients, which was published by the Ministry of Health and Family Welfare, Government of India, is utilized.

2. The Model and Methodology

Here, the variable of the interest is the status of the patient whether the patient recovered or deceased after the infection of COVID-19. The status of the patient can take only two values – either 0 if the patient deceased due to COVID-19 or 1 if recovered, and we want to estimate the probability of dying or survived after getting the infection of COVID-19 as a function of the indicator variables such as gender (male or female) and various age groups (0–20, 21–40, 41–60 and 60 and over). Since the response variable is of a dichotomous type, the logistic regression modelling technique is applied for the estimation of the probability whether the patient will die or recover by using various age groups and gender of the patients.

Let π denote the probability of recovery from the corona disease of a patient for the given values of p predictor variables and the relationship between the probability π and p predictors can be represented by the logistic model (see Chatterjee, S. and Hadi, Ali S. (2006)), i.e.

$$\begin{aligned}\pi &= Pr(Y = 1 \mid X_1 = x_1,..., X_p = x_p) \\ &= \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2 ......, \beta_p X_p+\varepsilon}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2 ..., \beta_p X_p+\varepsilon}}\end{aligned} \tag{1}$$

The function given in equation (1) is the logistic regression function. It is non-linear in the regression coefficients $\beta_0, \beta_1 ... \beta_p$ and it is linearised by the logit transformation, i.e. if the probability of an event that the patient recovers from the corona disease is π then the ratio $\frac{\pi}{1-\pi}$ obtained is the odds for the recovery from the coronavirus disease.

Since

$$\begin{aligned}1-\pi &= Pr(Y=0 \mid X_1 = x_1,...,X_p = x_p) \\ &= \frac{1}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+.....+\beta_p x_p+\varepsilon}}\end{aligned} \tag{2}$$

Then,

$$\frac{\pi}{1-\pi} = e^{\beta_0+\beta_1 x_1+\beta_2 x_{2+}+......+\beta_p x_p+\varepsilon} \tag{3}$$

Taking natural log both sides, we get

$$\begin{aligned}logit(\pi) &= log(\frac{\pi}{1-\pi}) \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \varepsilon\end{aligned} \tag{4}$$

Here, the function $logit\ (\pi)$ in equation (4) is a linear function of explanatory variables x_i (i=1,…,p) in terms and it is called the logit function and the range of π in equation (1) is between 0 and 1 while the range of the values of $log(\frac{\pi}{1-\pi})$ is between $-\infty$ and ∞, which makes the logits more appropriate linear regression fitting, and the disturbance term ε satisfies all the basic assumptions of ordinary least squares.

Now, our predictor variables are categorical type so the dummy variables are created for each of the categorical predictors. If the regression model contains an intercept term, the number of dummies defined should always be one less than the number of categories of that variable. Let G be the dummy variable for the gender of the patient which have only two categories (male and female), i.e. $G=1$ if the patient is male, 0 otherwise. Similarly, the dummy variables for the age having four age groups is $A_t; t=1,2,3$ and it can be defined as

$$\begin{aligned} A_1 &= 1; \quad \text{If the patient lies in the age group } 21-40 \\ &= 0; \quad \text{Otherwise} \\ A_2 &= 1; \quad \text{If the patient lies in the age group } 41-60 \\ &= 0; \quad \text{Otherwiseand} \\ A_3 &= 1; \quad \text{If the patient lies in the age group 60 and above} \\ &= 0; \quad \text{Otherwise} \end{aligned} \tag{5}$$

Here, the female category in the dummy variable G and the age group 0–20 in the A_t dummy variable are taken in the reference category and the logit model can be written as

$$\begin{aligned} logit(\pi) &= log(\frac{\pi}{1-\pi}) \\ &= \beta_0 + \beta_1 A_1 + \beta_2 A_2 + \beta_3 A_3 + \beta_4 G + \varepsilon \end{aligned} \tag{6}$$

3. Empirical Analysis

For the estimation of the probability of recovery of a patient infected by coronavirus disease in India, the data issued by the Ministry of Health and Family Welfare (MoHFW, India) are utilized. In the analysis, 427 patients have been included due to the lack of availability of data on all the patients and the data on the patients' status from all over India are taken from between 30 January 2020 to 30 May 2020, which is shown in Table 1. From the available data, an effort has been made to estimate the probability of recovery from coronavirus disease in India. For this, the logistic regression technique is used and the developed model is shown in equation (6), where age group and gender of the patients are the indicator variables and π is the probability of recovery of a patient from coronavirus disease. The analysis is done using *RStudio* (R Core Team (2020)) and the results obtained are shown in Table 2. The estimated model is given as

$$\begin{aligned} logit\ (\pi) &= log\ \left(\frac{\pi}{1-\pi}\right) \\ &= 0.0401 + 0.9346\ A_1 - 1.5913\ A_2 - 2.0101\ A_3 - 0.1071\ G \end{aligned} \tag{7}$$

Now, from Table 2, it can be seen that the age groups 41–60 and 60 and over are significant at 0.05 level of significance as their p-values are smaller than the 0.05and the log odds of recovery from the corona disease are –1.5913 and –2.0101 for the age group 41–60 and 60 and over respectively. For a better understanding of the results, the exponentiated terms of the regression coefficients has also been computed, which is shown in Table 3. If we look at the exponentiated terms of these log odds of significant variables, i.e. $exp(-1.5913) = 0.20365$ and $exp(-2.0101) = 0.13397$, these exponentiated terms show the odds of recovery from the coronavirus disease means that recovery odds for the patients in the age group 41–60 years is equal to 0.2036 times the recovery odds for the patients in the age group 0–20 years. Similarly, the patients aged 60 and over have 0.13397 times the odds of being recovered from Covid-19 disease compared to the patients in the age group 0–20 years on average, holding all else constant. From these two odds ratios, it can also be discovered that the odds of recovery from the corona disease is higher in the patients aged between 41–60 than the patients whose age is 60 and over. From Table 2, it can be assured that for the patients in the age group 0–20 and who are male, the probability of recovery from coronavirus disease is 0.6597 and the probability of recovery for the male patients aged between 41–60 is 0.6818. Also, the predicted recovery probability from coronavirus disease of patients aged 60 and over is 0.6746, which is slightly lower than the patients aged between 41–60 and higher than the patients of aged between 0–20. But on average, it can also be seen that the probability of recovery from coronavirus disease during the first wave of pandemic is almost same in all the patients and lies between the probability 0.6597–0.6818. If we look at the coefficient of gender (male) in Table 2, which is also statistically insignificant, it means there is no strong evidence for a gender difference in risk of dying due to coronavirus disease. This implies that the probability of recovery from coronavirus disease is same in males and females, keeping all else constant.

To test the goodness of fit of the model to the data, the log likelihood ratio R^2, sometimes called McFadden R-squared, the C-Statistic (Concordance Statistic) and Chi-Square goodness of fit test, has been used. The McFadden R-square is defined as:

$$R^2_{MF} = 1 - \frac{LL_{full}}{LL_0} \tag{8}$$

where LL_{full} is the full log likelihood model and LL_0 is the log likelihood function of the model with the intercept only. Backhaus et al. (2000) suggested that a McFadden R^2 value is in the range 0.2–0.4 indicates a good fit of the model and the obtained value of the R^2_{MF} is 1-384.12/482.96= 0.20465463 and shows the model is sufficiently well fitted to the data and the C-statistics can be computed by considering all possible pairs consisting of one patients who recovered from the coronavirus disease and one patients who deceased. The obtained C-statistics is the proportion of such pairs in which the patients who experienced a recovery from coronavirus disease had a higher estimated probability of experiencing the recovery than the patients who did not experience the recovery from the coronavirus disease. The value of C-statistics can lie between 0.50 to 1.00 The closer the C-statistic is to 1, the better a model is able to classify outcomes correctly. The value of C-statistics between 0.70 and 0.80 signals the model is good fitted to the data and the value between 0.50 to 0.70 indicates poor models (Hosmer & Lemeshow, 2000). Here, the obtained C-statistic is 0.7599994, which also indicates that the model is good enough and is able to classify outcomes correctly.

The Chi-square goodness of fit test is also used to test the goodness of fit of the model. For this, the standardized residuals are calculated as

$$r_i = \frac{y_i - \hat{y_i}}{\sqrt{\hat{y}_i(1 - \hat{y}_i)}}$$

And then the Chi-squared statistics is obtained as

$$\chi^2 = \sum_{i=1}^{n} r_i^2$$

The χ^2 statistics follows a χ^2 distribution with n-(p+1) degree of freedom, where p are the number of covariates. The obtained χ^2 value is 427.228 with 422 degree of freedom and the corresponding p-values is 0.4199021. This indicates that we cannot reject the null hypothesis that the model is exactly correct and it shows that the model fits the data well. From Figures (1 and 2), it can also be seen that the observed and expected number of cases of recovered and deceased is almost same, which also indicates that the model fits the data well.

4. Conclusion

The coronavirus has wreaked havoc all across the world with the rising cases of COVID-19 every day and with the absence of any effective treatment. In these gravedigger circumstances, the Government of India adopted many preventive steps such as lockdown, social distancing and urging people to live with extra cleanliness and India benefited somewhat from the strict lockdown but this nationwide lockdown cannot be continued for so long as it is not the solution for this pandemic, and it also not good for the country's economy. Hence, it is necessary to estimate the probability of recovery from the coronavirus disease as most of the Indian population is living in poor hygienic conditions. In this article, a probability model is developed using the indirect method of estimation based on some demographic factors, and it is found that the probability of recovery from coronavirus disease is statistically same in both males and females. Also, the coronavirus patients in the age group 0–40 years have almost equal probability of being recovered from this disease. In the patients aged between 41–60, the odds of recovery from the coronavirus disease is equal to 0.2036 times the recovery odds of the patients of the age group 0–20 years, while the patients aged 60 and over have 0.13397 times the odds of recovery from coronavirus compared to the patients of the age group 0-20 years on average. Also, the odds of recovery from coronavirus is higher in the patients of the age group 41–60 years than in the patients aged 60 and over.

References

Anastassopoulou, C., Russo, L., Tsakris, A. and Siettos, C., (2020). Data-based analysis, modelling and forecasting of the COVID-19 outbreak, *PLOS ONE*, 15(3), e0230405. https://doi.org/10.1371/journal.pone.0230405.

Backhaus, K., Erichson, B., Plinke, W. and Weiber, R., (2000). *Multivariate analysenmethoden*, Berlin: Springer.

Chen, Yi C., Lu, P. E., Chang, C. S. and Liu, T. H., (2020). *A Time-dependent SIR model for COVID-19 with Undetectable Infected Persons*, http://gibbs1.ee.nthu.edu.tw/A Time Dependent SIR Model For Covid 19.pdf.

Chatterjee, S. and Hadi, Ali S., (2006) Regression analysis by example. *John Wiley & Sons*, Inc., Hoboken, New Jersey.

Corman, V. M., Landt, O., Kaiser, M., Molenkamp, R., Meijer, A., Chu, D. K., Bleicker, T., Brunink, S., Schneider, J. and Schmidt, M. L., (2020). *Detection of 2019 novel coronavirus (2019-ncov) by realtime rt-pcr, Euro surveillance*, 25(3), 2000045.

Ferguson, N. M., Laydon, D., Nedjati-Gilani, G., Imai, N., Ainslie, K., Baguelin, M., Bhatia, S., Boonyasiri, A., Cucunubã, Z., Cuomo-Dannenburg, G., Dighe, A. Dorigatti, I., Fu, H., Gaythorpe, K., Green, W., Hamlet, A., Hinsley, W., Okell, L. C., Elsland, S. V., Thompson, H., Verity, R., Volz, E., Wang H., Wang, Y., Walker, P. Gt., Walters, C., Winskill, P., Whittaker, C., Donnelly, C. A., Riley, S. and Ghani, A. C., (2020). *Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand*, Imperial College COVID-19 Response Team.

Gamero, J., Tamayo, J. A. and Martinez-Roman J. A., (2020). *Forecast of the evolution of the contagious disease caused by novel corona virus (2019-ncov) in China*, arXiv preprint ar Xiv: 2002, 04739.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J. and Gu, X., (2020). *Clinical features of patients infected with 2019 novel coronavirus in Wuhan*, China, The Lancet, 395(10223), pp. 497–506.

Hosmer Dw, Lemeshow S., (2000). *Applied Logistic Regression (2nd Edition)*, New York, NY: John Wiley & Sons;.

Nikolay K., Anton I. and Asen R., (2020). On the half–logistic model with ”polynomial variable transfer”. Application to approximate the specific ”data corona virus”. *International Journal of Differential Equations and Applications*, 19(1), pp. 45–61.

Nikolay K., Anton I. and Asen R. (2020). On the Verhulst growth model with polynomial variable transfer. Some applications. *International Journal of Differential Equations and Applications*, 19(1), pp. 15-32.

Maksym Z., Roman O. B., Yuliia P., Maksim I. and Irakli P., (2020). Mathematical model building for COVID-19 diseases data in European Countries. *IDDM'2020: 3rd International Conference on Informatics & Data-Driven Medicine*, November 19–21, 2020, Växjö, Sweden, Session 1: Artificial intelligence, CEUR Workshop Proceedings.

Rao Srinivasa A., S. R., Krantz S., G., Kurien T. and Bhat R., (2020). Model based retrospective estimates for COVID-19 or coronavirus in India: continued efforts required to contain the virus spread. *Current Science*, 118(7), pp. 1023-1025.

R Core Team, (2020). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria, URL https://www.R-project.org/.

Rothe, C., Schunk, M., Sothmann, P., Bretzel, G., Froeschl, G., Wallrauch, C., Zimmer, T., Thiel, V., Janke C. and Guggemos, W., (2020). Transmission of 2019-ncov infection from an asymptomatic contact in Germany. *New England Journal of Medicine*, 382(10), pp. 970-971.

Singh, B. P., Singh, G., (2020). *Modeling Tempo of COVID-19 Pandemic in India and Significance of Lockdown*, https://doi.org/10.1002/pa.2257.

Singh, B. P., (2020). *Forecasting Novel Corona Positive Cases in India using Truncated Information: A Mathematical Approach*, medRxiv preprint, doi: https://doi.org/10.1101/2020.04.29.20085175.

Singh, R., Adhikari, R., (2020). *Age-structured impact of social distancing on the COVID-19 epidemic in India*, arXiv: 2003, 12055.

Singhal, T., (2020). A review of coronavirus disease-2019 (COVID-19). *The Indian Journal of Pediatrics*, pp. 1-6.

World Health Organization, (2020). *Updated WHO advice for international traffic in relation to the outbreak of the novel coronavirus 2019-nCoV*, Available at: https://www.who.int/ith/2020-24-01-outbreak-of-Pneumonia-caused-by-new-coronavirus/en/ (accessed January 2020).

World Health Organization, (2020). *Coronavirus disease (COVID-19) Weekly Epidemiological Update and Weekly Operational Update*, Available at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports (accessed March 2020).

Appendix

Table 1. Number of patients deceased or recovered from the corona disease in India during 30 January 2020 to 30 May 2020

Age Group	**Patient Status**				**Total**
	Deceased		**Recovered**		
	Female	**Male**	**Female**	**Male**	
0-20	6	2	4	4	16
21-40	6	15	21	31	73
41-60	48	104	11	19	182
60 and over	51	87	6	12	156
Total	111	208	42	66	427

Table 2. β Coefficients showing the log odds ratios of recovery from the coronavirus disease

Deviance Residuals:				
Min	1Q	Median	3Q	Max
-1.61	-0.59	-0.51	0.8	2.1
Coefficients:				
Group	Estimate	Standard Error	z value	Pr(>|z|)
Intercept	0.0401	0.5103	0.0790	0.9373
21-40	0.9346	0.5676	1.6470	0.0996
41-60	-1.5913	0.5441	-2.9250	0.0034*
60 and over	-2.0101	0.5632	-3.5690	0.0003*
Gender(Male)	-0.1071	0.2695	-0.3970	0.6913
Null Deviance:	482.96 on 426 degree of freedom			
Residual Deviance:	384.14 on 422 degree of freedom			
AIC:	394.14			
Number of Fisher scoring iterations: 4				

The p-values denoted by * are significant at 0.05 level of significance

Table 3. Exponentiated estimated coefficients showing the odds ratios and their respective confidence intervals

Group	**Estimates**	**95% Confidence Interval**	
		Lower limit	**Upper limit**
Intercept	1.04	0.38	2.89
21-40	2.55	0.83	7.89
41-60	0.20	0.07	0.60
60 and over	0.13	0.04	0.41
Gender (Male)	0.90	0.53	1.53

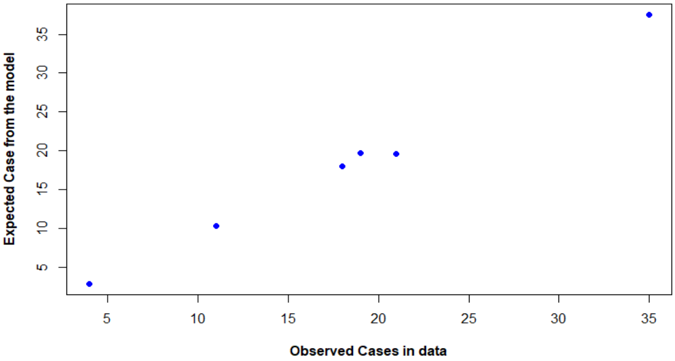


Figure 1. Observed and expected number of cases recovered from the corona disease in groups

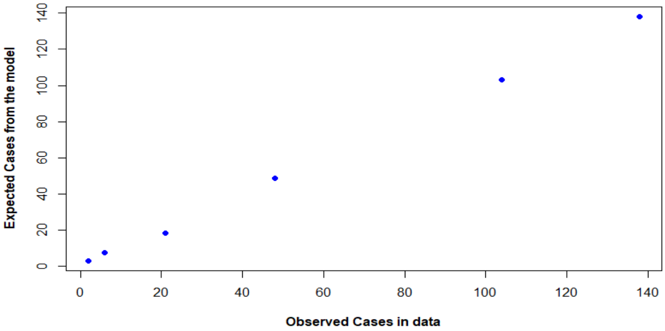


Figure 2. Observed and expected number of cases deceased from the corona disease in groups

STATISTICS IN TRANSITION new series, June 2022
Vol. 23, No. 2 pp. 209–210

Laureates of the Jerzy Spława-Neyman Medal

The ***3rd Congress of Polish Statistics***, organized by Statistics Poland and the Polish Statistical Association, took place from 26 to 28 April 2022 in Cracow. This year's edition of the Congress commemorated the 110th anniversary of the Polish Statistical Association.

The Congress of Polish Statistics was one of the most important scientific events and was attended by international academics, scientists, stakeholders of official statistics as well as by representatives of public administration and media.

Its goal was to strengthen the multi-directional cooperation between the Polish and international community of statisticians – representing scientific and decision-making centres involved in the advancement of theoretical and application areas of broadly understood statistical sciences and related disciplines, with the intention of also improving the functioning of national systems of public statistics.

During the Congress the Jerzy Spława-Neyman Medal, awarded by the Chapter of the Polish Statistical Society, was presented to honour the following eminent Professors:

1. Danny Pfeffermann (University of Southampton, UK),
2. Partha Lahiri (University of Maryland, USA),
3. Włodzimierz Okrasa (Cardinal Stefan Wyszynski University (UKSW).

Pfeffermann Danny is a Professor of Statistics at the University of Southampton, UK, and Professor Emeritus at the Hebrew University of Jerusalem, Israel. As of 2013, he has been the Government Statistician and Director General of the Central Bureau of Statistics in Israel. He is a past President of the Israel Statistical Society and a past President of the International Association of Survey Statisticians (IASS). For the last 20 years, he has also served as a consultant for the US Bureau of Labor Statistics. Danny Pfeffermann is a Fellow of the American Statistical Association and an elected member of the International Statistical Institute (ISI). He has received a BA degree in Mathematics and Statistics and MA and PhD degrees in Statistics from the Hebrew University of Jerusalem. He is a recipient of numerous prestigious awards, including Waksberg award in 2011, the West Medal by the Royal Statistical Society in 2017, the Julius Shiskin Memorial Award for Economic Statistics in 2018, and the SAE 2018 Award for his distinguished contribution to the SAE methodology and the advancement of Official Statistics in the Central Bureau of Statistics in Israel.

Lahiri Partha is a Professor of Survey Methodology and Mathematics at the University of Maryland College Park and an adjunct research professor at the Institute of Social Research, University of Michigan, Ann Arbor. His areas of research interest include data linkages, Bayesian statistics, survey sampling and small-area estimation. Dr. Lahiri has served on the editorial board of a number of international journals and on many advisory committees, including the U.S. Census Advisory committee and U.S. National Academy panel. He has also served as an advisor or consultant for various international organizations such as the United Nations and the World Bank. He is a Fellow of the American Statistical Association and the Institute of Mathematical Statistics and an elected member of the International Statistical Institute. Dr. Lahiri is the recipient of the 2020 SAE award for his outstanding contribution to the research, application, and education of small area estimation.

Okrasa Włodzimierz is a Professor and Head of the Research Methods and Evaluation Department at the Institute of Sociological Sciences in Cardinal Stefan Wyszynski University in Warsaw, and also serves as an Advisor to the President of Statistics Poland, and as an Editor-in-Chief of the *Statistics in Transition new series*. He was teaching and researching in Polish and American universities and was an ASA Senior Research Fellow at the US Bureau of Labor Statistics (1990–1991), a Program Director for statistics and economics in the Social Science Research Council, New York (1991–1993). He then worked for the World Bank in Washington, D. C. (1994–2000), and was a Head of the Unit at the European Science Foundation (2000–2003, Strasbourg). Elected Member of the International Statistical Institute (ISI) actively participating in international scientific events; he is a Vice-President of the Polish Statistical Association; member of the State Scientific Council for Statistics, and of the Statistics Poland's Methodological Committee. He is the author or co-author of over one hundred scientific publications. He is a laureate of research grants - international (e.g. US National Science Foundation; The British Academy, Ford Foundation, World Bank, UNDP, IRIS) and national, as well as numerous scientific awards (e.g. Medal of National Education Committee and Gold Medal for long-term service).

STATISTICS IN TRANSITION new series, June 2022
Vol. 23 No. 2, pp. 211–214

39th Conference Multivariate Statistical Analysis MSA 2021. Conference Review

Marta Małecka[1], **Artur Mikulec**[2]

The 39th international scientific conference Multivariate Statistical Analysis MSA 2021 was held on November 8–10, 2021 in Łódź, at the Faculty of Economics and Sociology of the University of Łódź (3/5 P.O.W. Street). The conference was organized by the Department of Statistical Methods of the University of Łódź and its co-organizers: the Institute of Statistics and Demography of the University of Łódź, the Committee of Statistics and Econometrics of the Polish Academy of Sciences and the Polish Statistical Association, branch in Łódź. The honorary patronage over the conference was taken by Elżbieta Żądzińska – Rector of the University of Łódź and Dominik Rozkrut – President of Statistics Poland.

The conference was organized in cooperation with the MASEP conference (Measurement and Assessment of Social and Economic Phenomena), organized by the Department of Economic and Social Statistics of the University of Łódź. The conference received financial support from the Minister of Education and Science (MEiN) as part of the "Excellent Science" program (DNK/SP/515427/2021). Educational activities related to the conference were supported by the National Bank of Poland as part of an educational project (NBP-DEW-WPE-AB-0662-0226-2021), which corresponds to NBP's priority areas of economic education – "New horizons of economic thought". The company StatSoft Polska Sp. z o.o. was also the content partner of the conference. Prof. Czesław Domański was the chairman of the Scientific Committee and Alina Jędrzejczak, Assoc. Prof. of the University of Łódź was the chairman of the Organizing Committee. The scientific secretaries of the conference included Marta Małecka, Asst. Prof. of the University of Łódź, Artur Mikulec, Asst. Prof. of the University of Łódź and Aleksandra Baszczyńska, Assoc. Prof. of the University of Łódź.

[1] University of Lodz, Department of Statistical Methods, Poland. E-mail: marta.malecka@uni.lodz.pl.

[2] University of Lodz, Department of Statistical Methods, Poland. E-mail: artur.mikulec@uni.lodz.pl.

The main goal of the MSA 2021 conference was to organize an international forum for discussion and exchange of ideas and views on the development of statistics as a science. The specific objectives were:

- presentation of the latest achievements in the field of multidimensional statistical analysis,
- disseminating knowledge in the field of data analysis and the application of statistical methods in other scientific disciplines, especially in economics, sociology and finance, exchange of experiences,
- creating a bridge between science (statistics) and research practice (individual users, business and administration).

The MSA 2021 conference was held in the hybrid formula: in-person and online. 84 people (62 in-person and 22 online) from various academic centres in Poland participated in the conference, including representatives from: Gdańsk, Katowice, Kraków, Lublin, Łódź, Poznań, Radom, Szczecin, Warsaw and Wrocław, representatives of Statistics Poland and the Statistical Office in Łódź. The conference hosted also representatives of academic society from abroad: Czech Republic, India, Iran and Italy. During the conference 15 sessions (plenary and parallel) were held, with 63 papers presented (41 stationary and 22 on-line).

The conference was opened by the Chairperson of the Organizing Committee, Alina Jędrzejczak. On behalf of Elżbieta Żądzińska – Rector of the University of Łódź – the conference participants were welcomed by Agnieszka Kurczewska, Vice-Rector for External Relations. Then the short welcome speeches were given by Ewa Kusideł, Vice-Dean for Science (Faculty of Economics and Sociology, UŁ) and the Chairman of the Scientific Committee – Czesław Domański (University of Łódź).

According to the tradition of the MSA conference, the first plenary session (PLENARY I, November 8th) was devoted to prominent representatives of the historical statistical thought and to the memories of recently deceased statisticians. This session was chaired by Bronisław Ceranka (Poznań University of Life Sciences). The first lecture devoted to the life and scientific work of Antoni Łomnicki (1881–1941) – a probabilist and statistician – was given by Mirosław Krzyśko (Adam Mickiewicz University in Poznań). The next speaker was Czesław Domański (University of Łódź), who presented the memories of Józef Kleczyński (1841–1900) – a precursor of population estimation between censuses – and the memories of Kazimierz Władysław Kumaniecki (1880–1941) – the initiator of the Polish Statistical Society and of the first Statistical Yearbook – Polish Statistics. In this session, the profiles of three famous Polish statisticians who passed away last year were recalled:

- "Dominik Szynal – creator of the probabilistic environment in Lublin" – Mariusz Bieniek (Maria Curie-Skłodowska University),

- "Daniel Kosiorowski – an outstanding Krakow statistician" – Józef Pociecha (University of Economics in Krakow),
- "Ryszard Walkowiak – statistician and naturalist" – Małgorzata Graczyk (University of Life Sciences in Poznań).

During the conference, there were four open lectures given by invited speakers. Two of them were presented during the plenary session on the first day of the conference (PLENARY II, November 8th). This session was chaired by Czesław Domański (University of Łódź). The invited lectures included:

- "Harnessing the power of Earth Observation for Official Statistics" – Dominik Rozkrut (Statistics Poland),
- "About the sampling plans depending on the position statistics of the auxiliary variable" – Janusz Wywiał (University of Economics in Katowice).

The other two invited lectures were given at the closing session on the last day of the conference (PLENARY II, November 10th). This session was chaired by Alina Jędrzejczak (University of Łódź). The topics of these lectures were as follows:

- "The Appearance of the Rawlsian Paradox when Neglecting Income Dependence of the Random Equivalence Scales" – Stanisław Maciej Kot (Gdańsk University of Technology),
- "Graphical and Computational Tools to Guide Parameter Choice for Robust Clusterwise Regression" – Francesca Greselin (University of Milan).

Parallel sessions of the conference covered a broad area of topics related to the theory and applications of mathematical statistics. The scope of the topics included in particular the following groups of issues:

1. Theory of statistical methods. The papers presented at the conference covered both the topics related to the estimation and statistical inference. There were papers from the area of taxonomic issues. Topics related to dealing with outliers, fuzzy numbers, Big Data, bootstrapping techniques and text recognition were also discussed.
2. Macroeconomic applications. In this thematic group, there were issues related to macroeconomic interventions, inflation and the use of modern data collection methods such as scanner data on web-scraped data.
3. Demographic and social issues. An extensive group of papers concerned issues related to the labour market. There were debates relating to many social issues such as people with disabilities or elderly people, retirement benefits and the quantification of poverty. In the area of demography, topics such as birth dynamics and demography of cities were discussed. A number of papers dealt with social issues related to the COVID-19 pandemic.

4. Sustainable development. Among the social topics, a special place was taken by discussions related to sustainable development, in particular: the impact of economic activity on the environment, air quality, water demand depending on weather conditions or the transformation of cities.
5. Business applications. In the field of business applications, the conference discussions covered the following areas: energy consumption forecasts, logistics, duration of business entities, micro-enterprise statistics, investment attractiveness of voivodships, industrial transformation, identification of bid rigging, organizational security culture.
6. Financial market. The use of statistical methods in financial market analysis was a separate group of topics within statistical applications. Presented papers dealt with issues such as: the relationship between the COVID-19 pandemic and exchange rates, the cryptocurrency market, banking scoring models, financial efficiency of insurance companies and modelling of systemic risk in the insurance sector.

The parallel sessions were chaired by:

November 8th	
SESSION IIIA	Jerzy Korzeniewski (University of Łódź)
SESSION IIIB	Sławomir Śmiech (Cracow University of Economics)
SESSION IVA	Andrzej Sokołowski (Cracow University of Economics)
SESSION IVB	Małgorzata Graczyk (Poznań University of Life Sciences)
November 9th	
SESSION IA	Jacek Białek (University of Łódź)
SESSION IB	Maria Grzelak (University of Łódź)
SESSION IIA	Sławomir Bukowski (Kazimierz Pulaski University of Technology and Humanities in Radom)
SESSION IIB	Wojciech Zieliński (Warsaw University of Life Sciences)
SESSION IIIA	Beata Bieszk-Stolorz (University of Szczecin)
SESSION IIIB	Mirosław Krzyśko (Adam Mickiewicz University, Poznań)
November 10th	
SESSION IA	Grażyna Trzpiot (University of Economics in Katowice)
SESSION IB	Stanisław Maciej Kot (Gdańsk University of Technology)

A detailed list of presenting authors and topics is available at: https://sites.google.com/view/msa2021pl/archiwum/msa-2021.

The debates were summed up and the conference was closed by the Chairman of the Scientific Committee, Prof. Czesław Domański. He thanked all participants for their active participation in this year's edition of the conference. He also gave thanks to the co-organizers and partners, and all institutions cooperating in the organization of the conference. The chairman announced that the Multivariate Statistical Analysis MSA 2022 conference will be held at the Training and Conference Center of the University of Łódź in Łódź on November 7–9, 2022.

About the Authors

Arora Sangeeta is working as a Professor in the Department of Statistics, Panjab University, Chandigarh-India. Her main research areas are applied statistics, income inequality & Lorenz dominance, statistical inference and Bayesian statistics, statistical quality control and environmental statistics. She has published over 50 publications in different national/ international journals of repute and has authored one book on Bayesian inference and is a life member of the Indian Society of Probability and Statistics, and the Indian Society for Medical Statistics. She is actively involved in teaching, research supervision and reviewing of various national and international journals.

Assegie Getnet Melak is a PhD candidate (expected to be a graduate in May 26 2022) at Parma University, Italy in the department of economics and business sciences. His main areas of interest include: multivariate analysis, multivariate time series, longitudinal data analysis, survival analysis, nonparametric tests, permutation tests, big data, multivariate regression, panel regression, simulation, R programming, sustainable development goals, industrial policy, firm performance. Getnet has published 7 research papers in international/national journals and conferences.

Azarudheen S. is an Assistant Professor at the Department of Statistics, CHRIST (Deemed to be University), India. His research interests are statistical modelling, acceptance sampling, reliability analysis, statistical inference and data analysis in particular. Dr. Azarudheen S has published over 10 research papers in international/national journals and conferences. He is also an expert in R and SPSS. He has published few book chapters in reputed book series. Professor Azarudheen S. is an active member of many scientific professional bodies.

Bonnini Stefano is an Associate Professor of Statistics at the Department of Economics and Management, University of Ferrara. His main areas of interest include: multivariate analysis, nonparametric statistics, permutation tests, categorical data analysis, composite indicators. Currently he is a member of the Italian Statistical Society, the Research Data Alliance (RDA) and CMStatistics (ERCIM Working Group on Computational and Methodological Statistics). Professor Bonnini is an Associate Editor of the Journal of Applied Statistics, an editorial board member of the Baltic Journal of Modern Computing and Guest Editor for a special issue of the journal Mathematics.

Borkowski Maeusz, MA in economics, currently he is a PhD attendee at Doctoral School in the Social Sciences (economics and finance discipline) at the University of Bialystok. His main areas of interest include: economic development, institutional economics and quantitative methods in economics (especially partial least squares structural equation modelling – PLS-SEM).

Brudz Magdalena is an Assistant at the Department of Operational Research, Faculty of Economics and Sociology, University of Lodz. Simultaneously, at the same faculty, at the Department of Spatial Econometrics, she is finishing her doctoral dissertation entitled "The impact of IT solutions on the transformational processes of the labour market in Poland". Her main area of current research interests is regional research related to the labour market and quality of life.

Domański Czesław is a Full Professor at the Department of Statistical Methods, Faculty of Economics and Sociology, University of Lodz. His research interests are: tests based on runs theory and order statistics, (multivariate) normality tests and non-classical methods of statistical inference. Currently, he is a member of the Scientific Statistical Council of the President of Statistics Poland, the main council of the Polish Statistical Association and Committee on Statistics and Econometrics at the Polish Academy of Sciences.

Gagui Abdelmalek is an Associate Professor at the Department of Mathematics, Faculty of Sciences, University of Amar Telidji Laghouat Algeria. His research interests are nonparametric statistics estimation, functional data analysis, locally linear models.

Gupta Sat is a Full Professor of Mathematics and Statistics. He is working in the field of sampling and a prominent figure in providing novel data collection and handling procedures.

Grover Gurprit is a Professor of Statistics and Head of Department of Statistics, University of Delhi. Her area of interest is in the field of biostatistics, demography, statistical quality control and reliability. She has over 35 years of research experience. She has supervised over 30 students for their PhD (Doctoral) theses and MPhil dissertations. She has published over 75 research papers in international and national journals and authored 3 books. She can be contacted at: gurpritgrover@yahoo.com

Jangra Vikas is a full time Research Scholar at the Department of Statistics, Panjab University Chandigarh. His main areas of interest include: income distribution, income inequality, poverty measurement, Bayesian inference.

Jewczak Maciej is an Assistant Professor at the Department of Operational Research, Faculty of Economics and Sociology, University of Lodz. He is an author/co-author of nearly 45 scientific studies and an active participant in many scientific conferences, seminars, workshops and trainings. His research interests focus on the use of quantitative tools and techniques in analyses of various aspects of health, spatial econometrics and statistics, quality of life, social policy, welfare and logistics.

Jibrin Sanusi A. is a Senior Lecturer of Statistics at Kano University of Science and Technology (KUST), Wudil, Kano-Nigeria. He is specialized in time series (interminable long memory), econometrics, R programming, statistical and mathematical software. He has published over 30 research papers in international/national journals and conferences. He has also published one book. He is a member of SLU Journal of Science and Technology editorial boards, the Proprietor of Elhaljibrin Consultancy Services and a member of Statistics and Mathematics Professional Associations in Nigeria.

Joshi Hemlata is an Assistant Professor at the Department of Statistics, CHRIST (Deemed to be University), India. Her research areas are regression modelling and mathematical demography. She has published over 12 articles in international/national journals and conferences. Hemlata Joshi has achieved a young researcher award from the Institute of Scholars, India and she is also an active member of many scientific professional bodies.

Kaushik Sakshi is a Team Lead, Biometrics, Biostatistics department in a leading Clinical Research Organisation (CRO) Veranex Solutions. She has over 6 years of clinical research experience. Her key areas of interest include biostatistics, statistical inference including both frequentist and Bayesian inference, adaptive clinical trial designs and missing data imputation methods. She has published 4 research papers and one accepted in international journals.

Khan M. I. is an Assistant Professor at the Department of Mathematics, Faculty of Science, Islamic University of Madinah, Saudi Arabia. His research has been focused in the area of mathematical statistics and ordered random variables. Dr. Khan is a life member of the Indian Society for Probability and Statistics (ISPS), Indian Bayesian Society (IBS) and Kerala Statistical Association (KSA). Dr. Khan is also a reviewer of several statistics journals. He has attended over 20 conferences and delivered invited talks presentations at various universities and institutions across the globe.

Kubacki Robert is working as a Head of Analytical CRM in a Polish commercial bank. He completed his PhD in Economics from University of Lodz, in 2018. His research interests include machine learning and data mining.

Mahajan Kalpana K. is a former Professor, Department of Statistics, Panjab University, Chandigarh. Her Research interests are: nonparametric inference, Bayesian inference, income inequality measurement and environmental statistics. She is a life member of a number of various societies including Indian Society of Probability and Statistics and Indian Medical Council of Statistics. She is currently engaged in research supervision and review of articles for various statistical journals.

Mustafa Abdelfattah is an Associate Professor at the Department of Mathematics, Faculty of Science, Mansoura University, Egypt. Currently, he is a member of Mathematics Department, Faculty of Science, Islamic University of Madinah, KSA. His main areas of interest include: mathematical statistics, reliability engineering, lifetime distributions and estimation. He has published over 50 research papers in international/national journals and conferences. He has also published one Arabic book in statistics to Princess Sattam Bin Abdul Aziz, KSA.

Nagaraja M. S. is an Assistant Professor at the Department of Statistics, CHRIST (Deemed To Be University), Bangalore, India. His main areas of interest include: agricultural statistics, regression analysis and multivariate analysis. He has published 14 research papers in international/national journals and conferences.

Nowak Piotr Bolesław is an Assistant Professor in the Institute of Economic Sciences, University of Wroclaw. His research interests include mathematical statistics and its applications, in particular survival analysis, applied mathematics in economy and medicine.

Rahman Rosmanjawati Abdul is a Senior Lecturer of Statistics at Universiti Sains Malaysia, USM, in Penang, Malaysia. Her research interests are Applied Statistics and Time Series Analysis, focusing on analyzing financial, economic, and environmental data. She has published various papers in international and national journals.

Sabharwal Alka is a Professor of Statistics at Kirori Mal College, University of Delhi. Her areas of interest are in the field of stochastic processes and biostatistics. Previously she has published many research papers on problems related to diabetes and its complications, chronic kidney problems, mental disorders, and adolescent behavioural problems through statistical modelling. Currently, she is working on psychological and behavioural problems related to young adults. She can be contacted at: alkasabh@gmail.com

Skolimowska-Kulig Magdalena is an Assistant Professor at the Institute of Economic Sciences, Faculty of Law, Administration and Economics, University of Wrocław. Her main research interests focus on mathematical statistics and probability models with their applications in the reliability theory and economics.

Singh Chandraketu is an Assistant Professor at the Department of Statistics, CHRIST (Deemed to be University), Bengaluru, India. He obtained his PhD degree in 2021 from IIT (ISM), Dhanbad, Dhanbad, India. His research interests are survey sampling and statistical inference. He has published over 20 research papers in Indian and foreign journals of repute. He presented his research problems in international and national conferences.

Sohail Umair is working as an Assistant Professor of Statistics in University of Narowal, Narowal, Pakistan. His research interests are missing values, data imputation, and randomized response.

Sohil Fariha is an Assistant Professor of Education. His research interests are statistical inference and data analysis in particular. Dr. Fariha has published several research papers in international/national journals and conferences.

Shabbir Javid is a Full Professor of Statistics. His research interests are survey sampling and randomized responses. Professor Javid has published over 230 research papers in international/national journals and conferences.

GUIDELINES FOR AUTHORS

We will consider only original work for publication in the Journal, i.e. a submitted paper must not have been published before or be under consideration for publication elsewhere. Authors should consistently follow all specifications below when preparing their manuscripts.

Manuscript preparation and formatting

The Authors are asked to use *A Simple Manuscript Template (Word or LaTeX) for the Statistics in Transition Journal (published on our web page:* http://stat.gov.pl/en/sit-en/editorial-sit/).

- ***Title and Author(s)***. The title should appear at the beginning of the paper, followed by each author's name, institutional affiliation and email address. Centre the title in **BOLD CAPITALS**. Centre the author(s)'s name(s). The authors' affiliation(s) and email address(es) should be given in a footnote.
- ***Abstract.*** After the authors' details, leave a blank line and centre the word **Abstract** (in bold), leave a blank line and include an abstract (i.e. a summary of the paper) of no more than 1,600 characters (including spaces). It is advisable to make the abstract informative, accurate, non-evaluative, and coherent, as most researchers read the abstract either in their search for the main result or as a basis for deciding whether or not to read the paper itself. The abstract should be self-contained, i.e. bibliographic citations and mathematical expressions should be avoided.
- ***Key words***. After the abstract, Key words (in bold) should be followed by three to four key words or brief phrases, preferably other than used in the title of the paper**.**
- ***Sectioning***. The paper should be divided into sections, and into subsections and smaller divisions as needed. Section titles should be in bold and left-justified, and numbered with **1.**, **2.**, **3.**, etc.
- ***Figures and tables***. In general, use only tables or figures (charts, graphs) that are essential. Tables and figures should be included within the body of the paper, not at the end. Among other things, this style dictates that the title for a table is placed above the table, while the title for a figure is placed below the graph or chart. If you do use tables, charts or graphs, choose a format that is economical in space. If needed, modify charts and graphs so that they use colours and patterns that are contrasting or distinct enough to be discernible in shades of grey when printed without colour.
- ***References.*** Each listed reference item should be cited in the text, and each text citation should be listed in the References**.** Referencing should be formatted after the Harvard Chicago System – see http://www.libweb.anglia.ac.uk/referencing/harvard.htm. When creating the list of bibliographic items, list all items in alphabetical order. References in the text should be cited with authors' name and the year of publication. If part of a reference is cited, indicate this after the reference, e.g. (Novak, 2003, p.125).