SORT

Statistics and Operations Research Transactions

Volume
45

Number 2, July-December 2021

Generalitat de Catalunya
Institut d'Estadística de Catalunya

SORT

Statistics and Operations Research Transactions

Volume 45, Number 2, July-December 2021

ISSN: 1696-2281
eISSN: 2013-8830

Articles

9 771696 228009 45002

www.idescat.cat/sort/

Aims

SORT (*Statistics and Operations Research Transactions*) -formerly *Qüestiió*- is an international journal launched in 2003, published twice-yearly by the Institut d'Estadística de Catalunya (Idescat), co-edited by the Universitat Politècnica de Catalunya, Universitat de Barcelona, Universitat Autònoma de Barcelona, Universitat de Girona, Universitat Pompeu Fabra, Universitat de Lleida i Universitat Rovira i Virgili and the cooperation of the Spanish Section of the International Biometric Society and the Catalan Statistical Society. SORT promotes the publication of original articles of a methodological or applied nature or motivated by an applied problem in statistics, operations research, official statistics or biometrics as well as book reviews. We encourage authors to include an example of a real data set in their manuscripts. SORT is an Open Access journal which does not charge publication fees.

SORT is indexed and abstracted in the *Science Citation Index Expanded* and in the *Journal Citation Reports* (Clarivate Analytics) from January 2008. The journal is also described in the *Encyclopedia of Statistical Sciences* and indexed as well by: *Current Index to Statistics*, Índice *Español de Ciencia y Tecnología*, *MathSci*, *Current Mathematical Publications and Mathematical Reviews*, and *Scopus*.

SORT represents the third series of the *Quaderns d'Estadística i Investigació Operativa* (*Qüestiió*), published by Idescat since 1992 until 2002, which in turn followed from the first series *Quaderns d'Estadística, Sistemes, Informàtica i Investigació Operativa* (1977-1992). The three series of *Qüestiió* have their origin in the *Cuadernos de Estadística Aplicada e Investigación Operativa*, published by the UPC till 1977.

Secretary and subscriptions to SORT

Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona (Spain)
Tel. +34 - 93 557.30.76 - 93 557.30.00
Fax. +34 - 93 557.30.01
E-mail: sort@idescat.cat

Publisher: Institut d'Estadística de Catalunya (Idescat)

ISSN 1696-2281
eISSN: 2013-8830
DL B-46.085-1977
Key title: SORT
Numbering: 1 (december 1977)

www.idescat.cat/sort/

ISSN: 1696-2281
eISSN: 2013-8830
SORT 45 (2) July-December (2021)

SORT

Statistics and Operations Research Transactions

Coediting institutions

Universitat Politècnica de Catalunya
Universitat de Barcelona
Universitat de Girona
Universitat Autònoma de Barcelona
Universitat Pompeu Fabra
Universitat de Lleida
Universitat Rovira i Virgili
Institut d'Estadística de Catalunya

Supporting institutions

Spanish Region of the International Biometric Society
Societat Catalana d'Estadística

Generalitat
de Catalunya
Institut d'Estadística
de Catalunya

SORT

Volume 45
Number 2
July-December 2021
ISSN: 1696-2281
eISSN: 2013-8830

Articles

SORT 45 (2) July-December 2021, 93-120 DOI: 10.2436/20.8080.02.111

Nonparametric estimation of the probability of default with double smoothing

Rebeca Peláez[1], Ricardo Cao[2] and Juan M. Vilar[2]

Abstract

In this paper, a general nonparametric estimator of the probability of default is proposed and studied. It is derived from an estimator of the conditional survival function for censored data obtained with a double smoothing, on the covariate and on the variable of interest. An empirical study, based on modified real data, illustrates its practical application and a simulation study shows the performance of the proposed estimator and compares its behaviour with smoothed estimators only in the covariate. Asymptotic expressions for the bias and the variance of the probability of default estimator are found and asymptotic normality is proved.

MSC:62G05, 62G07, 62G08, 62G20, 62N02, 62P20.

Keywords: Censored data; kernel method, probability of default, risk analysis, survival analysis.

1. Introduction

Credit risk is an important research area. It is useful for financial companies to assess the risk of insolvency caused by unpaid loans. Estimating the probability of default on consumer credits, loans and credit cards is one of the main problems that banks, savings banks, savings cooperatives and other credit companies must address. For a fixed time, t, and a horizon time, b, the probability of default (PD) can be defined as the probability that a credit that has been paid until time t becomes unpaid not later than time $t+b$. To estimate the PD, banks and financial institutions typically use features of the credit and the clients. They usually build some linear combination (credit scoring) based on these informative variables and the probability of default is allowed to depend on this scoring

[1] Research Group MODES, Department of Mathematics, CITIC, University of A Coruña, A Coruña, Spain.

[2] Research Group MODES, Department of Mathematics, CITIC, University of A Coruña and ITMATI, A Coruña, Spain.

Received: February 2021

Accepted: June 2021

x, $PD(t|x)$. A common aproach in credit scoring is using logistic regression to build the index. The logistic model of credit scoring has been studied by Wiginton (1980), Srinivasan and Kim (1987), Steenackers and Goovaerts (1989), Thomas, Crook and Edelman (1992) and Samreen and Zaidi (2012), among others.

It can be deduced from the definition of the PD that it is a relevant measure in other fields apart from the financial one. For example, companies that provide energy services (electricity, gas), water, streaming services (TV, cinema, music), telephone or internet are interested in estimating the probability that a customer who receives their services at time t will leave the company before time $t+b$.

There is an extensive literature in which survival analysis methods are used for solving credit risk problems. Among others, we mention the work of Naraim (1992), Stepanova and Thomas (2002), Hanson and Schuermann (2004), Glennon and Nigro (2005), Allen and Rose (2006), Baba and Goko (2006), Beran and Djaïdja (2007) and Cao, Vilar and Devia (2009). A common feature of all these papers is the use of parametric or semiparametric regression techniques for modelling the time to default, including exponential models, Weibull models and Cox's proportional hazards models, which are typical in this literature. Nonparametric curve estimation is a flexible approach that only uses the information that the data provides without making assumptions about the shape of the curve. Therefore, it is very convenient in this context. Following this idea, Cao et al. (2009) proposed a PD estimator using Beran's estimator for the conditional survival function, Beran (1981). This work was expanded in the paper of Peláez, Cao and Vilar (2021b) who studied four nonparametric estimators of the probability of default in credit risk derived from estimators of the conditional survival function for censored data.

In the recent work, Peláez, Cao and Vilar (2021a), a general nonparametric estimator of the conditional survival function with double smoothing is proposed and studied. This survival estimator is not only smoothed in the covariate but also in the time variable. A large simulation study shows there that the estimator with double smoothing improves on the corresponding nonparametric estimator of the survival function which is smoothed only in the covariate. Here, a general nonparametric estimator of the PD with double smoothing is proposed and studied. It is derived from the survival estimator with double smoothing studied in Peláez et al. (2021a).

The remainder of this paper is organized as follows. In Section 2, the nonparametric estimator of the probability of default with double smoothing is defined, the doubly smoothed PD estimator based on Beran's estimator is applied to a set of modified real data and its asymptotic properties are presented. In Section 3, a simulation study shows the improvement obtained by using the double smoothing in several nonparametric estimators of the probability default. Finally, Section 4 contains some concluding remarks. Appendix A includes terminology, assumptions and detailed theoretical results. Appendix B includes a sketch of proof of the theoretical results.

2. Nonparametric PD estimator with double smoothing

Let $\{(X_i, Z_i, \delta_i)\}_{i=1}^n$ be a simple random sample of (X, Z, δ) where X is the covariate, $Z = \min\{T, C\}$ is the follow-up time variable, T is the time to occurrence of the event, C is the censoring time and $\delta = I_{\{T \leq C\}}$ is the uncensoring indicator. It is assumed that an unknown relationship between T and X exists. In credit risk, usually, X is the credit scoring, Z is the observed maturity, T is the time to default and C is the time until the end of the study or the anticipated cancellation of the credit. The distribution function of T is denoted by $F(t)$ and the survival function by $S(t)$. The functions $F(t|x)$ and $S(t|x)$ are the conditional distribution and survival functions of T given $X = x$ evaluated at t. In this context, let x be a fixed value of the covariate X and b any fixed value (typically, $b = 12$ in months). Then the probability of default in a time horizon $t + b$ from a maturity time, t, is defined as follows:

$$\begin{aligned} PD(t|x) &= P(T \leq t + b | T > t, X = x) \\ &= \frac{F(t+b|x) - F(t|x)}{1 - F(t|x)} = 1 - \frac{S(t+b|x)}{S(t|x)}. \end{aligned} \tag{1}$$

Therefore, an estimator of $PD(t|x)$ could be obtained by replacing $S(t+b|x)$ and $S(t|x)$ in (1) with appropriate estimators. Following this idea, Cao et al. (2009) and Peláez et al. (2021b) used nonparametric estimators of the conditional survival function, $\widehat{S}_h(t|x)$ with $h = h_n$ being the smoothing parameter for the covariate, to obtain the corresponding nonparametric estimator of $PD(t|x)$ denoted by $\widehat{PD}_h(t|x)$.

In Peláez et al. (2021a) the following nonparametric estimator of the conditional survival function with double smoothing is proposed and studied:

$$\widetilde{S}_{h,g}(t|x) = 1 - \sum_{i=1}^{n} s_{(i)} \mathbb{K}\left(\frac{t - Z_{(i)}}{g}\right), \tag{2}$$

where $s_{(i)} = \widehat{S}_h(Z_{(i-1)}|x) - \widehat{S}_h(Z_{(i)}|x)$ with $i = 2, ..., n$ and $s_{(1)} = 1 - \widehat{S}_h(Z_{(1)}|x)$, $Z_{(i)}$ is the i-th element of the sorted sample of Z, $\mathbb{K}(t)$ is the distribution function of a kernel K, $\mathbb{K}(t) = \int_{-\infty}^{t} K(u)du$, and $g = g_n$ is the smoothing parameter for the time variable. This survival estimator, defined in (2), is not only smoothed in the covariate but also in the time variable. It is based on the idea of estimating the survival function in a point t conditional on x by means of a weighted mean of the values that the estimator $\widehat{S}_h(u|x)$ takes in points near t.

Estimating the probability of default, $PD(t|x)$, by means of the nonparametric estimator of the conditional survival function with double smoothing is the aim of this paper. For this purpose, $S(t|x)$ in (1) is replaced by the doubly smoothed nonparametric estimator, $\widetilde{S}_{h,g}(t|x)$, obtaining the following nonparametric estimator of $PD(t|x)$:

$$\widetilde{PD}_{h,g}(t|x) = 1 - \frac{\widetilde{S}_{h,g}(t+b|x)}{\widetilde{S}_{h,g}(t|x)}. \tag{3}$$

Since $\widehat{S}_h(t|x)$ is any arbitrary conditional survival estimator, the probability of default estimator $\widetilde{PD}_{h,g}(t|x)$ is very general. From now on, this paper focuses on the doubly smoothed Beran's estimator $\widetilde{S}^B_{h,g}(t|x)$ associated through (2) with the classic Beran's estimator of the survival function given by

$$\widehat{S}^B_h(t|x) = \prod_{i=1}^{n}\left(1-\frac{I_{\{Z_i\leq t,\,\delta_i=1\}}w_{i,n}(x)}{1-\sum_{j=1}^{n}I_{\{Z_j<Z_i\}}w_{j,n}(x)}\right), \tag{4}$$

where

$$w_{i,n}(x)=\frac{K\big((x-X_i)/h\big)}{\sum_{j=1}^{n}K\big((x-X_j)/h\big)}$$

with $i=1,...,n$ and $h=h_n$ is the smoothing parameter for the covariable.

Using $\widetilde{S}^B_{h,g}(t|x)$ in (3), the smoothed probability of default estimator based on Beran's survival estimator is obtained. It is denoted by $\widetilde{PD}^B_{h,g}(t|x)$.

However, any other estimator of the conditional survival function could be considered to obtain the corresponding smoothed estimator defined in (2) and then, to estimate the probability of default through the expression given in (3). In particular, two other survival estimators are considered in this work: the Weighted Nadaraya-Watson estimator (WNW) and the Van Keilegom-Akritas estimator (VKA). The WNW estimator was built following the idea of Cai (2003), where the survival estimator is based on local linear regression. Since the weighted local linear estimator presents problems when estimating probabilities, a constant fit is proposed in Peláez et al. (2021b). The VKA estimator was defined in Van Keilegom and Akritas (1999) and Van Keilegom, Akritas and Veraverbeke (2001). The expressions for both estimators are shown in Section 2 of Peláez et al. (2021a) and they are denoted by $\widehat{S}^{WNW}_h(t|x)$ and $\widehat{S}^{VKA}_h(t|x)$. Their smoothed versions are built according to Equation (2), obtaining the following smoothed survival estimators: $\widetilde{S}^{WNW}_{h,g}(t|x)$ and $\widetilde{S}^{VKA}_{h,g}(t|x)$. Replacing $\widetilde{S}_{h,g}(t|x)$ with $\widetilde{S}^{WNW}_{h,g}(t|x)$ and $\widetilde{S}^{VKA}_{h,g}(t|x)$ in Equation (3) gives the nonparametric smoothed estimators of $PD(t|x)$ denoted by $\widetilde{PD}^{WNW}_{h,g}(t|x)$ and $\widetilde{PD}^{VKA}_{h,g}(t|x)$.

2.1. Application to real data

In order to illustrate the use of these smoothed estimators in the context of credit risk, a real data set is analysed using the doubly smoothed Beran's estimator. The data consists of a sample of 10000 consumer credits from a Spanish bank registered between July 2004 and November 2006. They are also considered by Peláez et al. (2021b), where the PD is estimated using parametric and non-parametric methods for $S(t|x)$ which are not smoothed in t, as is the case in this paper. The data set provides the credit scoring computed for each borrower, the observed lifetime of the credit in months and the uncensoring indicator. To obtain each customer's credit scoring, the financial institution adjusted a scoring model on several informative variables collected in the dataset: gender, marital status, profession, place of residence, type of housing, age, employment

history and bank account balance. See Devia (2016) for more details. Due to confidentiality, the estimated coefficients of the original explanatory variables are not reported here. The resulting credit scoring is used as a covariate in this analysis. The sample censoring percentage is 92.8%; equivalently, the proportion os credits for which the default is observed is 7.2%. An intentionally biased subsample was obtained from the original sample, so as so as not to show the true solvency situation of the bank and thus preserve confidentiality.

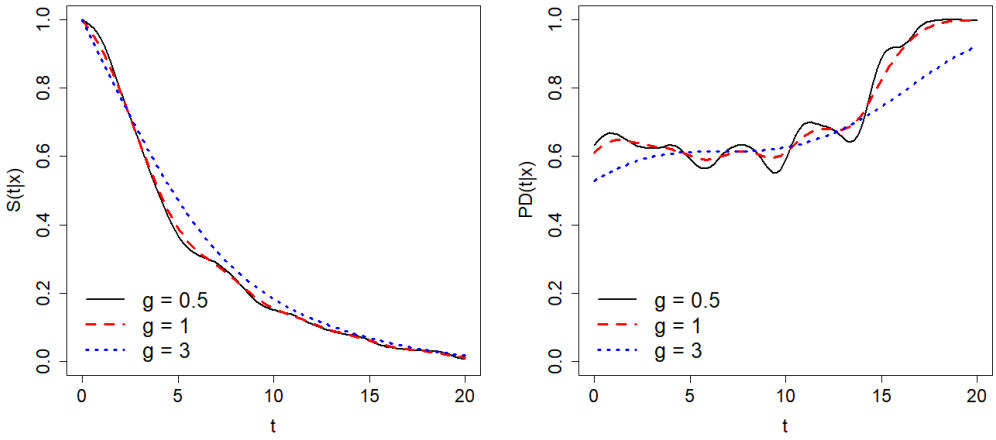


Figure 1. *Estimation of $S(t|x)$ (left) and estimation of $PD(t|x)$ (right) at horizon $b = 5$ for $x = 0.5$ by means of the smoothed Beran estimator on the consumer credits dataset for $h = 0.05$ and three different values of g.*

The probability of default for $x = 0.5$ at horizon $b = 5$ months is estimated in a time grid along the interval $[0,25]$ using the smoothed Beran's estimator. The estimation is obtained with some different possible values of the time variable smoothing parameter, while the covariate bandwidth is fixed to a reasonable value ($h = 0.05$), since it has a very slight influence on the estimation. Figure 1 shows the results.

Beran's estimation and the smoothed Beran's estimation of the conditional survival function and the *PD* for $h = 0.05$ and $g = 3$ are shown in Figure 2. Although the survival estimations are very similar with both estimators, it can be seen how the roughness of the curve estimation is reduced and the jumps are removed when using the smoothed Beran's estimator. This is even more remarkable when estimating the probability of default.

According to the smoothed Beran's estimation, the probability of default has an increasing tendency. It follows from it that the higher the debt maturity, the higher the probability of falling into default for an individual with this credit scoring.

Finally, sample quartiles of the credit scoring are considered for the group of clients with observed default (uncensored group) and the group with unobserved default (censored group). Figure 3 shows the PD estimation by means of the smoothed Beran's

estimator for these values of the credit scoring at horizon $b = 5$ months with $h = 0.05$ and $g = 3$.

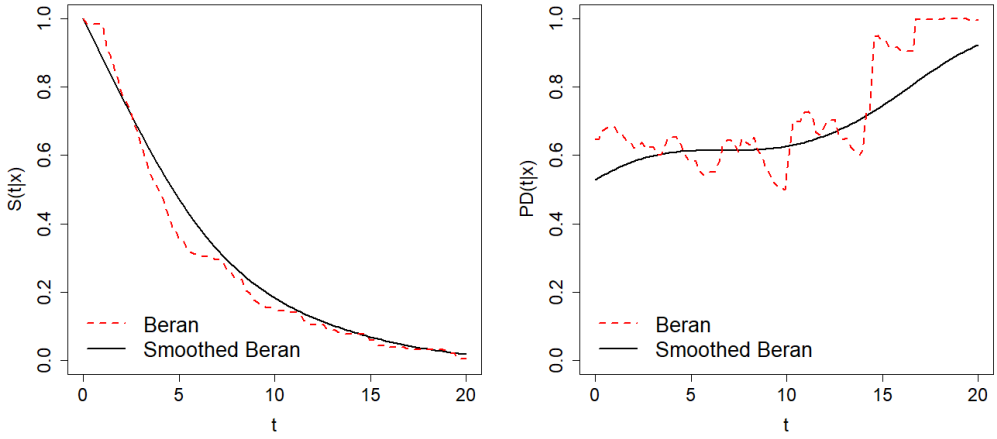


Figure 2. *Estimation of $S(t|x)$ (left) and estimation of $PD(t|x)$ (right) at horizon $b = 5$ for $x = 0.5$ by means of Beran's estimator (dashed line) and smoothed Beran's estimator (solid line) using the bandwidths $h = 0.05$ and $g = 3$ on the consumer credits dataset.*

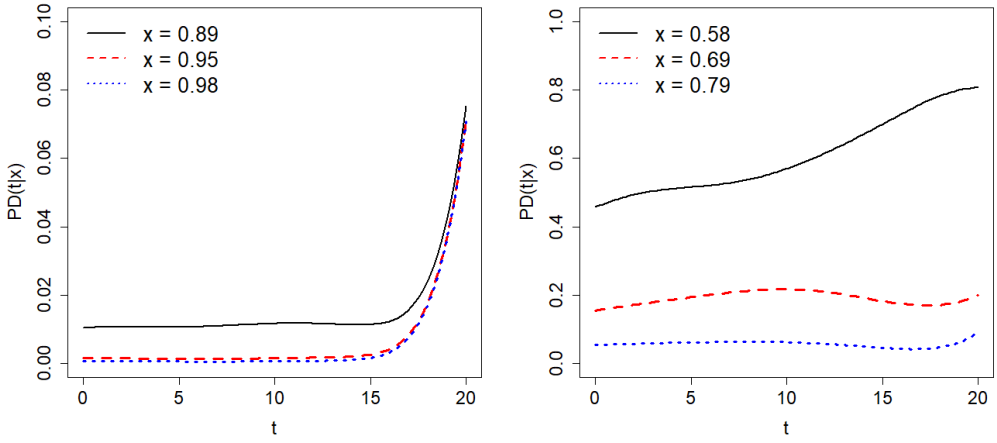


Figure 3. *Smoothed Beran's estimation of $PD(t|x)$ at horizon $b = 5$, for large (left) and small (right) values of the score x, using bandwidths $h = 0.05$ and $g = 3$. The large values chosen are the three sample quartiles of the score for nondefaulted credits, while the small values are the three sample quartiles of the score for the defaulted credits.*

2.2. Asymptotic results of the doubly smoothed Beran's PD estimator

Asymptotic properties of the smoothed Beran's estimator of the PD, $\widetilde{PD}^{B}_{h,g}(t|x)$, are obtained using the results for the smoothed Beran's survival estimator presented in Peláez

et al. (2021a). An intuitive idea of these results is shown here. The simplified expression of the asymptotic bias of $\widetilde{PD}^B_{h,g}(t|x)$ is as follows:

$$\mathrm{ABias}\big(\widetilde{PD}^B_{h,g}(t|x)\big) \quad = \quad c_1 h^2 + c_2 g^2,$$

and the asymptotic variance of $\widetilde{PD}^B_{h,g}(t|x)$ is given by

$$\mathrm{AVar}\big(\widetilde{PD}^B_{h,g}(t|x)\big) \quad = \quad c_3 \frac{1}{nh} + c_4 \frac{g}{nh} + c_5 \frac{h}{n},$$

for some real constants c_1, c_2, c_3, c_4 and c_5. For detailed expressions of these constants and the asymptotic normality of the estimator, see Appendix A. For proofs of these results see Appendix B.

It is difficult to use the theoretical bias and variance in an applied context in order to compare estimators or to obtain optimal smoothing parameters, since the constants c_1, c_2, c_3, c_4 and c_5 involved are complex and depend on too many population functions.

3. Simulation study

Intuitively, the improvement coming from smoothing in the time variable in the conditional survival function estimator will lead to a similar gain for nonparametric PD estimators. The aim of this section is to explore this by simulation.

Two models are considered and three different censoring scenarios are distinguished for each model. Model 1 is close to a proportional hazards model, while Model 2 moves away from this Cox's model. The covariate X follows a $U(0,1)$ distribution in both models.

For Model 1, the time to occurrence of the event conditional on the covariate, $T|X=x$, follows a Weibull distribution with parameters $d=2$ and $A(x)^{-1/d}$ where $A(x)=1+5x$, and the censoring time conditional on the covariate, $C|X=x$, follows a Weibull distribution with parameters $d=2$ and $B(x)^{-1/d}$ where $B(x)=10+b_1x+20x^2$, for some suitable values of b_1. The conditional survival function, the probability of default and the censoring conditional probability of this model are the following:

$$S(t|x) = e^{-A(x)t^d},$$

$$PD(t|x) = 1 - \frac{e^{-A(x)(t+b)^d}}{e^{-A(x)t^d}},$$

$$P(\delta = 0|X=x) = \frac{B(x)}{A(x)+B(x)}.$$

Setting $x=0.6$, the chosen values are $b_1=-27$, $b_1=-22$ and $b_1=-2$, so that the censoring probability is 0.2, 0.5 and 0.8, respectively. The time horizon is $b=0.1$ (20% of the time range) and the estimation is obtained in a time grid $0<t_1<\cdots<t_{n_t}$ of size n_t where $t_{n_t}+b=F^{-1}(0.95|x=0.6)$.

Model 2 considers an exponential distribution with parameter $\Gamma(x) = 2 + 58x - 160x^2 + 107x^3$ for the time to occurrence of the event conditional on the covariate, $T|X = x$ and an exponential distribution with parameter $\Delta(x) = 10 + d_1 x + 20x^2$, for some suitable values of d_1, for the censoring time conditional on the covariate, $C|X = x$. In this case, the conditional survival function, the probability of default and the censoring conditional probability are given by:

$$S(t|x) = e^{-\Gamma(x)t},$$

$$PD(t|x) = 1 - e^{-\Gamma(x)b},$$

$$P(\delta = 0|X = x) = \frac{\Delta(x)}{\Gamma(x) + \Delta(x)}.$$

Setting $x = 0.8$, the chosen values are $d_1 = -113/4$, $d_1 = -55/2$ and $d_1 = -123/5$, so that the censoring conditional probability is 0.2, 0.5 and 0.8, respectively. The time horizon is $b = 0.7$ (20% of the time range) and the PD is estimated in a time grid $0 < t_1 < \cdots < t_{n_t}$ of size n_t where $t_{n_t} + b = F^{-1}(0.95|x = 0.8)$.

The standard Gaussian kernel truncated in the range $[-50, 50]$ is used for both covariate and time variable smoothing. The sample size is $n = 400$, and the size of the lifetime grid is $n_t = 100$. The boundary effect is corrected using the reflexion principle proposed in Silverman (1986).

These models were previously used in the simulation study of Peláez et al. (2021a). This makes it possible to compare the results obtained in both studies.

First, the performance of Beran's PD estimator, $\widehat{PD}_h^B(t|x)$, and the smoothed Beran's PD estimator, $\widetilde{PD}_{h,g}^B(t|x)$, are compared.

The optimal bandwidth for $\widehat{PD}_h^B(t|x)$, h_1, is taken as the value which minimises a Monte Carlo approximation of the mean integrated squared error (MISE) given by

$$MISE_x(h) = E\left(\int \left(\widehat{PD}_h^B(t|x) - PD(t|x)\right)^2 dt\right)$$

based on $N = 100$ simulated samples. The value of $MISE_x(h)$ using this smoothing parameter is approximated from $N = 1000$ simulated samples and used, along with its square root (*RMISE*), as a measure of the estimation error which is committed by $\widehat{PD}_h^B(t|x)$.

The smoothed PD estimator $\widetilde{PD}_{h,g}^B(t|x)$ depends on two bandwidths: h that measures the smoothing degree introduced in the covariate and g that measures the smoothing in the time variable. Two strategies are used in order to obtain these smoothing parameters.

Strategy 1

It consists in fixing the covariate smoothing parameter to the the optimal h_1 for Beran's estimator and approximating the optimal smoothing parameter g. The error to

minimise is

$$MISE_x(h_1,g) = E\left(\int \left(\widetilde{PD}^B_{h_1,g}(t|x) - PD(t|x)\right)^2 dt\right)$$

considered as a function of the bandwidth g. It is approximated from $N = 100$ simulated samples in a grid of 50 g values and the bandwidth which provides the smaller error is chosen as g_1. Then, $N = 1000$ samples are simulated to approximate $MISE_x(h_1,g_1)$ which is the measure of the estimation error of $\widetilde{PD}^B_{h,g}(t|x)$.

Strategy 2

The optimal bandwidth (h_2,g_2) is chosen (from a meshgrid of 50 values of h and 50 values of g) as the pair which minimises some Monte Carlo approximation of

$$MISE_x(h,g) = E\left(\int \left(\widetilde{PD}^B_{h,g}(t|x) - PD(t|x)\right)^2 dt\right)$$

based on $N = 100$ simulated samples. Then, the value of the *MISE* committed by $\widetilde{PD}^B_{h_2,g_2}(t|x)$ is approximated from $N = 1000$ simulated samples.

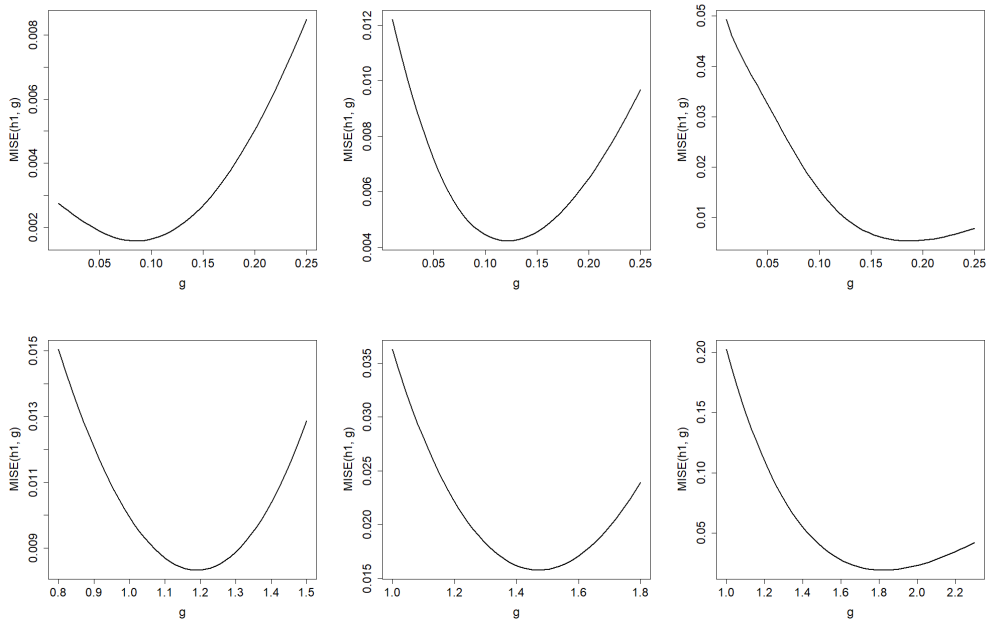


Figure 4. *$MISE_x(h_1,g)$ function approximated via Monte Carlo for the smoothed Beran's estimator using $N = 100$ simulated samples from Model 1 (top) and Model 2 (bottom) with $P(\delta = 0|x) = 0.2$ (left), $P(\delta = 0|x) = 0.5$ (center) and $P(\delta = 0|x) = 0.8$ (right).*

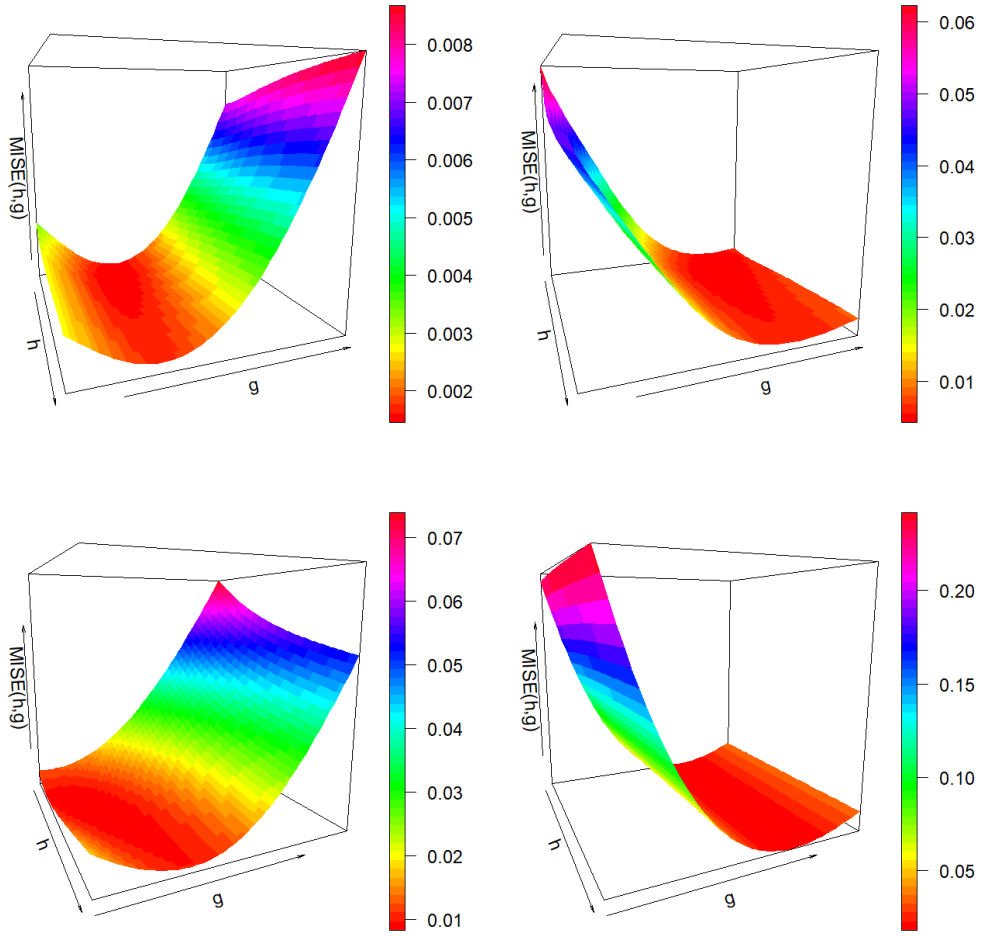


Figure 5. *$MISE_x(h,g)$ function approximated via Monte Carlo for the smoothed Beran's estimator using $N = 100$ simulated samples from Model 1 (top) and Model 2 (bottom) with $P(\delta = 0|x) = 0.2$ (left) and $P(\delta = 0|x) = 0.8$ (right).*

The main advantage of using Strategy 1 is its lower computational cost, but it provides rather worse results than Strategy 2. It should be noted that neither the bandwidth obtained with Strategy 1 nor Strategy 2 can be used in practice but they produce a fair comparison since the estimators are built using the best possible smoothing parameters.

The error curve $MISE_x(h_1,g)$, which is minimised to obtain the optimal time smoothing parameter according to Strategy 1, is shown in Figure 4 for each level of censoring conditional probability and each model. It follows from these graphs that the optimal bandwidth g is easily approximated by Strategy 1.

The function $MISE_x(h,g)$ for Models 1 and 2 for the lowest and highest censoring levels is shown in Figure 5. These plots show the two-dimensional functions to be min-

imised using Strategy 2. The red area is the region where this minimum is reached and its coordinates provide the optimal smoothing bandwidths. It is clear that the choice of the time bandwidth (g) notably affects the estimation the estimation error, whereas h seems not to affect much the quality of the estimator.

On the contrary, the value of g for which the smallest error is committed does not seem to depend too much on the value of the covariate bandwidth (h). Figure 6 shows $MISE_x(h,g)$ as a function of g for some fixed values of h within the interval where the optimum is reached. The curves are similar and close for all values of h, mainly at the highest level of censoring conditional probability.

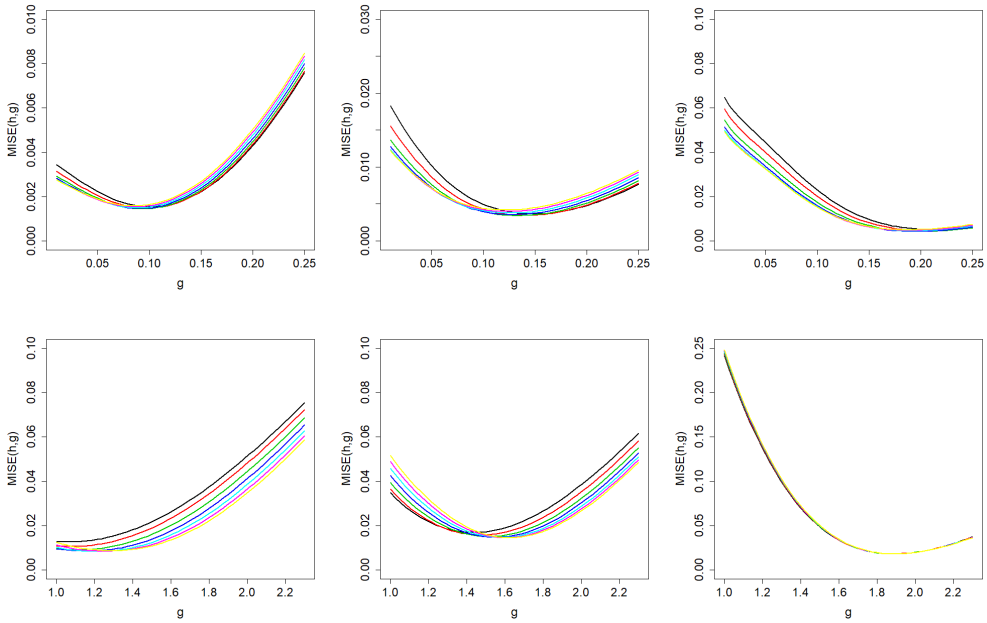


Figure 6. *$MISE_x(h,g)$ as a function of g, approximated via Monte Carlo for the smoothed Beran's estimator using $N = 100$ simulated samples from Model 1 (top) and Model 2 (bottom) for some fixed equispaced values of $h \in [0.1, 0.5]$ with $P(\delta = 0|x) = 0.2$ (left), $P(\delta = 0|x) = 0.5$ (center) and $P(\delta = 0|x) = 0.8$ (right).*

The optimal bandwidths and the estimation errors that are committed by Beran's estimator and the smoothed Beran's estimator with both Strategies 1 and 2 for each model are shown in Table 1. The value of R_i is defined as follows:

$$R_i(x) = \frac{RMISE\left(\widetilde{PD}^B_{h_i,g_i}(\cdot|x)\right)}{RMISE\left(\widehat{PD}^B_{h_1}(\cdot|x)\right)},$$

with $i = 1,2$ depending on whether Strategy 1 or 2 is used. They help to compare the behaviour of the estimators and quantify the improvement of the double smoothing over

the smoothed estimator only in the covariate. The closer to 0 the value of R_1 or R_2, the greater the improvement with respect to Beran's estimator. The relation between R_1 and R_2 (R_1 greater than R_2 or viceversa) also informs which of the two strategies reduces the error most.

Table 1. *Optimal bandwidths, RMISE, R_1 and R_2 of the PD estimation for Beran's estimator, the smoothed Beran's estimator with Strategy 1 and the smoothed Beran's estimator with Strategy 2 in each level of censoring conditional probability for Model 1 and Model 2.*

		Model 1			**Model 2**		
$\mathbf{P(\delta = 0 \mid x)}$		**0.2**	**0.5**	**0.8**	**0.2**	**0.5**	**0.8**
$\widehat{\mathbf{PD}}^{\mathbf{B}}_{\mathbf{h_1}}$	$\mathbf{h_1}$	0.35714	0.34694	0.39796	0.10306	0.12265	0.14224
	RMISE	0.05437	0.11195	0.25738	0.27128	0.49813	0.67999
$\widetilde{\mathbf{PD}}^{\mathbf{B}}_{\mathbf{h_1,g_1}}$	$\mathbf{h_1}$	0.35714	0.34694	0.39796	0.10306	0.12265	0.14224
	$\mathbf{g_1}$	0.08347	0.12265	0.18633	1.18571	1.47755	1.82245
	RMISE	0.04065	0.06574	0.07246	0.25222	0.24154	0.20558
	$\mathbf{R_1}$	0.74765	0.58723	0.28153	0.92974	0.48489	0.30233
$\widetilde{\mathbf{PD}}^{\mathbf{B}}_{\mathbf{h_2,g_2}}$	$\mathbf{h_2}$	0.21429	0.15714	0.18980	0.10816	0.25918	1.00000
	$\mathbf{g_2}$	0.09327	0.13735	0.19612	1.21122	1.61020	1.90204
	RMISE	0.03845	0.05941	0.06208	0.09210	0.12350	0.13434
	$\mathbf{R_2}$	0.70719	0.53068	0.24120	0.33950	0.24793	0.19756

In all cases, *RMISE* values are lower for the smoothed Beran's estimator with both Strategies 1 and 2 than for Beran's estimator and this difference becomes bigger when increasing the censoring conditional probability. The estimator $\widetilde{PD}^{B}_{h,g}(t|x)$ with optimal bandwidth (h_2, g_2) (Strategy 2) provides more accurate estimations than the others, since the relation $0 < R_2 < R_1 < 1$ is satisfied for all cases.

When the censoring conditional probability is 0.2 or 0.5 in Model 1, the time smoothing with Strategy 1 reduces the error by about 35% and this improvement is about 60% when the conditional probability of censoring is 0.8. This improvement increases by an additional $5-10\%$ when using Strategy 2. The error reduction in Model 2 with respect to the nonsmoothed PD estimator is more significant, reaching 50% and 70% when using Strategy 1 and censoring is moderate or heavy, respectively. This reduction is bigger when using Strategy 2, reaching $75-80\%$.

A brief study not included here shows that the results of these simulations hold when the distribution of X is not uniform but a more realistic asymmetric distribution if X denotes the credit scoring.

The computation time of both estimators should be considered in the comparison. Table 2 shows the CPU times (in seconds) that Beran's estimator and the smoothed Beran's estimator spend on estimating the probability of default curve in a 100-point time grid and a fixed value of x, for different values of the sample size. The smoothing parameters are fixed to the optimal ones for estimating estimating the curve.

Table 2. *CPU time (in seconds) for estimating $PD(t|x)$ in a time grid of size 100 for each estimator and different sample sizes.*

n	50	100	200	400	1200
Beran	0.01	0.01	0.01	0.02	0.03
SBeran	0.03	0.03	0.03	0.05	0.20

Table 2 shows that the second smoothing increases the CPU time and the Beran's PD estimator with double smoothing is more affected by the increase in sample size than Beran's estimator.

It is expected that the two strategies used to find the optimal bandwidths will also have different computational efficiency. Table 3 shows the CPU time (in minutes) for each strategy and several number of trials to check this. For both strategies the size of each sample is $n = 400$ and the PD is estimated in a time grid of size 100. The number of simulated samples to approximate the *MISE* by Monte Carlo is the parameter that varies in order to compare the CPU time of each strategy. Strategy 1 has a clear computational advantage over Strategy 2, since Strategy 2 is significantly slower.

Table 3. *CPU time (in minutes) for approximating the optimal bandwidth (h,g) for $\widetilde{PD}^{B}_{h,g}(t|x)$ with Strategies 1 and 2.*

N	50	100	150	200
Strategy 1	3.01	4.28	5.40	7.32
Strategy 2	80.61	204.51	228.01	296.95

Since the improvement in statistical efficiency that the time variable smoothing provides to Beran's PD estimator has been verified, it is interesting to check if other PD estimators based on other estimators for the survival function are equally improved by applying this type of smoothing. With this aim, in a second simulation study, the behaviours of the smoothed Weighted Nadaraya-Watson estimator (SWNW), $\widetilde{PD}^{WNW}_{h,g}(t|x)$, and the smoothed Van Keilegom-Akritas estimator (SVKA), $\widetilde{PD}^{VKA}_{h,g}(t|x)$, are compared to each other as well as to the smoothed Beran's estimator.

Since the computational times of these estimators are pretty high, only Strategy 1 is used to look for the optimal smoothing parameters, since Strategy 2 would further increase the computation time of the simulations.

In order to quantify the improvement that the smoothing provides to the PD estimators and compare the performance of the three estimators, the ratios $R^{\bullet}_{S}$ and $R^{\bullet}_{c}$ are defined as follows:

$$R^{\bullet}_{S}(x) = \frac{RMISE\left(\widetilde{PD}^{\bullet}_{h_1,g_1}(\cdot|x)\right)}{RMISE\left(\widehat{PD}^{\bullet}_{h_1}(\cdot|x)\right)},$$

$$R^{\bullet}_{c}(x) = \frac{RMISE\left(\widetilde{PD}^{\bullet}_{h_1,g_1}(\cdot|x)\right)}{RMISE\left(\widetilde{PD}^{B}_{h_2,g_2}(\cdot|x)\right)},$$

being $\bullet = B, WNW, VKA$ and they are included in Table 4 along with the approximation of the optimal smoothing parameters and the error committed by each estimator.

Table 4. *Optimal bandwidths, RMISE, $R_S^\bullet$ and $R_c^\bullet$ of the PD estimation for the smoothed Beran's estimator, the smoothed WNW estimator and the smoothed VKA estimator with Strategy 1 for each level of censoring conditional probability for Models 1 and 2.*

	$\mathbf{P(\delta=0\|x)=0.2}$			$\mathbf{P(\delta=0\|x)=0.5}$			$\mathbf{P(\delta=0\|x)=0.8}$		
	SBeran	**SWNW**	**SVKA**	**SBeran**	**SWNW**	**SVKA**	**SBeran**	**SWNW**	**SVKA**
Model 1									
$\mathbf{h_1}$	0.35714	0.38776	0.25918	0.34694	0.90102	0.22857	0.39796	1.00000	0.23469
$\mathbf{g_1}$	0.08347	0.14020	0.06327	0.12265	0.20531	0.11653	0.18633	0.28367	0.19347
RMISE	0.04065	0.03513	0.06418	0.06574	0.03260	0.09957	0.07246	0.04705	0.09816
$\mathbf{R_S^\bullet}$	0.74765	0.50036	0.88744	0.58723	0.19457	0.76112	0.28153	0.14115	0.38976
$\mathbf{R_c^\bullet}$	1.05722	0.91365	1.66918	1.10655	0.54873	1.67598	1.16720	0.75789	1.58119
Model 2									
$\mathbf{h_1}$	0.10306	0.09143	0.04567	0.12265	0.10694	0.05380	0.14224	0.11857	0.12837
$\mathbf{g_1}$	1.18571	1.55102	1.44286	1.47755	1.77551	1.45714	1.82245	1.92857	1.52857
RMISE	0.25222	0.12628	0.49730	0.24154	0.13406	0.37621	0.20558	0.13375	0.11410
$\mathbf{R_S^\bullet}$	0.92974	0.33177	1.63226	0.48489	0.19828	0.88273	0.30233	0.16480	0.16868
$\mathbf{R_c^\bullet}$	2.73855	1.37112	5.39957	1.95580	1.08551	3.04623	1.53030	0.99561	0.84934

The values of $R_S^\bullet$ report the influence of the smoothing. The smaller the value, the better the estimation obtained with the smoothed estimator compared to the corresponding nonsmoothed estimator. Since its value is less than 1 in almost all cases of Models 1 and 2, it is confirmed that the smoothing in the time variable is an improvement of any of the estimators, mainly when censoring is heavy. In addition, the smaller the value of $R_S^\bullet$, the greater the improvement that smoothing provides to the estimator. In this line, the doubly smoothed WNW estimator is the estimator whose error is reduced the most.

The value of $R_c^\bullet$ is useful to compare the behaviour of the three estimators with the behaviour of $\widetilde{PD}^B_{h_2,g_2}(t|x)$ (the smoothed Beran's estimator with Strategy 2). Since almost all the $R_c^\bullet$ values obtained are greater than 1, it can be concluded that the smoothed Beran's estimator with Strategy 2 provides more accurate estimations of the probability of default. Moreover, the closer to 1 the value of $R_c^\bullet$, the better the estimators. Thus, in general terms, the smoothed Beran's estimator with Strategy 1 is the second best option for estimating the probability of default.

In some cases the smoothed WNW estimator presents an $R_c^\bullet$ less than 1, which indicates that the error it makes is occasionally smaller than the errror committed by the smoothed Beran's estimator with Strategy 2. Therefore, the smoothed WNW estimator appears to be competitive with Beran's in some contexts.

It is also appropriate to analyse the differences between the computacional times os these techniques. Table 5 shows the CPU time (in seconds) that is needed by each

estimator to obtain the probability of default curve in a time grid of size 100 and a fixed value of x for different values of the sample size.

Table 5. *CPU time (in seconds) for estimating $PD(t|x)$ in a time grid of size 100 for every estimator and different sample sizes.*

n	Beran	SBeran	SWNW	SVKA
50	0.01	0.03	2.30	0.42
100	0.01	0.03	6.33	1.80
200	0.01	0.03	25.97	7.34
400	0.02	0.05	140.62	53.99
1200	0.03	0.20	1459.35	507.36

The time variable smoothing clearly implies an increase of the CPU time. The three doubly smoothed PD estimators which were considered have higher CPU times than Beran's estimator. It should be noted that the smoothed Beran's estimator is least affected by the increase of the sample size and it is the fastest of the three doubly smoothed estimators. The CPU time of the smoothed VKA increases very fast with the sample size but the slowest method and most affected by the sample size is the smoothed WNW estimator.

4. Conclusions

A general doubly smoothed estimator of the probability of default is proposed in this paper. Asymptotic properties of the smoothed PD estimator based on the smoothed Beran's estimator for the survival function are proved and its asymptotic distribution is found. This doubly smoothed Beran's estimator of the PD showed a remarkably good behavior in the scenarios analysed in the simulation study. The time variable smoothing results in a significant improvement of the PD estimator, since the estimation error (*MISE*) is reduced, mainly when using Strategy 2 for approximating the optimal bandwidth. However, the computational time is increased. These same evidences were observed in any of the smoothed PD estimators studied by simulation. Nevertheless, the smoothed Beran's estimator of the PD turned out to have the most stable behaviour and to be the fastest of all. The selection of the smoothing parameters for the smoothed Beran's estimator is still an outstanding problem. The study of automatic methods probably based on the bootstrap is an appealing idea to be considered for future work.

Acknowledgements

This research has been supported by MINECO Grant MTM2017-82724-R, and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-015, ED431C-2020-

14 and Centro Singular de Investigación de Galicia ED431G 2019/01), all of them through the ERDF.

References

Allen, L. N. and Rose, L. C. (2006). Financial survival analysis of defaulted debtors. *Journal of the Operational Research Society 57*(6), 630–636.

Baba, N. and Goko, H. (2006). Survival analysis of hedge funds. *Bank of Japan, Working Papers Series* (06-E-05).

Beran, J. and Djaïdja, A. K. (2007). Credit risk modeling based on survival analysis with inmunes. *Statistical Methodology 4*, 251–276.

Beran, R. (1981). Nonparametric regression with randomly censored survival data. *Technical report, University of California*.

Billingsley, P. (1968). *Convergence of Probability Measure. Wiley Series in probability and Mathematical Statistics: Tracts on probability and statistics*, Volume 9. Wiley.

Cai, Z. (2003). Weighted local linear approach to censored nonparametric regression. In M. G. Akritas and D. N. Politis (Eds.), *Recent Advances and Trends in Nonparametric Statistics*, pp. 217–231.

Cao, R., Vilar, J. M. and Devia, A. (2009). Modelling consumer credit risk via survival analysis (with discussion). *Statistics and Operations Research Transactions 33*(1), 3–30.

Dabrowska, D. M. (1989). Uniform consistency of the kernel conditional Kaplan-Meier estimate. *The Annals of Statistics 17*(3), 1157–1167.

Devia, A. (2016). *Contribuciones al análisis estadístico del riesgo de crédito*. PhD Thesis, Universidade da Coruña.

Glennon, D. and Nigro, P. (2005). Measuring the default risk of small business loans: a survival analysis approach. *Journal of Money, Credit and Banking 37*, 923–947.

Hanson, S. G. and Schuermann, T. (2004). Estimating probabilities of default. *Staff Report Federal Reserve Bank of New York 190*, 923–947.

Iglesias-Pérez, M. C. and González-Manteiga, W. (1999). Strong representation of a generalized product-limit estimator for truncated and censored data with some applications. *Journal of Nonparametric Statistics 10*(3), 213–244.

Naraim, B. (1992). Survival analysis and the credit granting decision. In L. C. Thomas, J. N. Crook, and D. B. Edelman (Eds.), *Credit Scoring and Credit Control, Oxford University Press*, pp. 109–121.

Peláez, R., Cao, R. and Vilar, J. M. (2021a). Nonparametric estimation of the conditional survival function with double smoothing. Unpublished paper, URL: http://dm.udc.es/preprint/Pelaez_Cao_Vilar_Survival_estimator_double_smoothing_Nov2020.pdf.

Peláez, R., Cao, R. and Vilar, J. M. (2021b). Probability of default estimation in credit risk using a nonparametric approach. *TEST 30*, 383–405.

Samreen, A. and Zaidi, F. (2012). Design and development of credit scoring model for the commercial banks of pakistan: forecasting creditworthiness of individual borrowers. *International Journal of Business and Social Science 17*, 155–166.

Silverman, B. W. (1986). *Density Estimation for Statistics & Data Analysis*. Monographs on Statistics and Applied Probability. Chapman and Hall.

Srinivasan, V. and Kim, Y. H. (1987). Credit granting: a comparative analysis of clasification procedures. *Journal of Finance 42*, 665–681.

Steenackers, A. and Goovaerts, M. J. (1989). A credit scoring model for personal loans. *Insurance: Mathematics and Economics 8*, 31–34.

Stepanova, M. and Thomas, L. (2002). Survival analysis methods for personal loan data. *Operations Research 50*, 277–289.

Thomas, L. C., Crook, J. N. and Edelman, D. B. (1992). *Credit Scoring and Credit Control*. Oxford University Press.

Van Keilegom, I. and Akritas, M. (1999). Transfer of tail information in censored regression models. *The Annals of Statistics 27*(5), 1745–1784.

Van Keilegom, I., Akritas, M. and Veraverbeke, N. (2001). Estimation of the conditional distribution in regression with censored data: a comparative study. *Computational Statistics & Data Analysis 53*, 457–481.

Wiginton, J. C. (1980). A note on the comparison of logit and discriminant models of consumer credit behaviour. *Journal of Financial and Quantitative Analysis 15*, 757–770.

A. Asymptotic results of the doubly smoothed Beran's estimator of the PD

Asymptotic properties of the smoothed Beran's estimator of the PD, $\widetilde{PD}^{B}_{h,g}(t|x)$, are shown in this section. The following notation is used.

Let $R:\mathbb{R}\longrightarrow\mathbb{R}$ be any function and define the constants

$$c_R=\int R(t)^2dt,\quad d_R=\int t^2R(t)dt,$$

and the functions

$$R_l(u)=u^lR(u),\quad \mathbb{R}_l(u)=\int_{-\infty}^{u}R_l(t)dt. \tag{5}$$

Given any function $f:\mathbb{R}^k\longrightarrow\mathbb{R}$, its first derivatives with respect to the first and second variables are denoted as follows:

$$f'(x_1,...,x_k)=\frac{\partial f(x_1,...,x_k)}{\partial x_1},\quad \dot{f}(x_1,...,x_k)=\frac{\partial f(x_1,...,x_k)}{\partial x_2}$$

Correspondingly, the second derivatives with respect to the first or second variable are denoted by $f''(x_1,...,x_k)$ and $\ddot{f}(x_1,...,x_k)$, respectively. Finally, let $f*g$ be the convolution of any two functions f and g.

The required assumptions are listed below. They are standard in the literature and not too restrictive in this context. They were previously assumed in Peláez et al. (2021a), Dabrowska (1989) and Iglesias-Pérez and González-Manteiga (1999) in the nonparametric conditional survival function estimation setup.

A.1. X, T, C are absolutely continuous random variables.

A.2. The density function of X, m, has support $[0,1]$.

A.3. Let $H(t)=P(Z\leq t)$ be the distribution function of Z and $H(t|x)$ be the conditional distribution function of $Z|X=x$,

(a) Let $I=[x_1,x_2]$ be an interval contained in the support of m such that,

$$0<\gamma=\inf\{m(x):x\in I_c\}<\sup\{m(x):x\in I_c\}=\Gamma<\infty$$

for some $I_c=[x_1-c,x_2+c]$ with $c>0$ and $0<c\Gamma<1$.

(b) For any $x\in I$, the random variables $T|X=x$ and $C|X=x$ are independent.

(c) Denoting $l_{H(\cdot|x)}=\inf\{t/H(t|x)>0\}$ and $u_{H(\cdot|x)}=\inf\{t/H(t|x)=1\}$, for any $x\in I_c$, $0\leq l_{H(\cdot|x)}$, $0\leq u_{H(\cdot|x)}$

(d) There exist $l,u,\theta\in\mathbb{R}$ with $l<u$, satisfying $\inf\{1-H(u|x):x\in I_c\}\geq\theta>0$. Therefore $1-H(t|x)\geq\theta>0$ for every $(t,x)\in[l,u]\times I_c$.

A.4. The first and second derivatives of m, $m'(x)$ and $m''(x)$, respectivaly, exist and are continuous in I_c.

A.5. Let $H_1(t) = P(Z \leq t, \delta = 1)$ be the subdistribution function of Z when $\delta = 1$. The corresponding density functions of $H(t)$ and $H_1(t)$ are bounded away from 0 in $[l,u]$.

A.6. Let $H_1(t|x)$ the conditional subdistribution function of $Z|X = x$ when $\delta = 1$. The first and second derivatives with respect to t of the functions $H(t|x)$ and $H_1(t|x)$, i.e. $H'(t|x)$, $H_1'(t|x)$, $H''(t|x)$ and $H_1''(t|x)$, exist and are continuous in $[l,u] \times I_c$.

A.7. The second partial derivatives first with respect to x and second with respect to t of the functions $H(t|x)$ and $H_1(t|x)$, i.e. $\dot{H}'(t|x)$ and $\dot{H}_1'(t|x)$ respectively, exist and are continuous in $[l,u] \times I_c$.

A.8. The kernel, K, is a symmetric, continuous and differentiable density function with compact support $[-1,1]$.

A.9. The smoothing parameters $h = h_n$ and $g = g_n$ satisfy $h \to 0$, $g \to 0$, and $nh^3 \to \infty$ and $nhg^2 \to \infty$ when $n \to \infty$.

Using the asymptotic results for the smoothed Beran's estimator of the conditional survival function given in Peláez et al. (2021a), the asymptotic properties of the estimator $\widetilde{PD}^B_{h,g}(t|x)$ are obtained. The following are the functions required to state these results:

$$
\begin{aligned}
&\xi(Z,\delta,t,x) = \frac{1_{\{Z\leq t,\delta=1\}}}{1-H(Z|x)} - \int_0^t \frac{1_{\{u\leq Z\}}}{\big(1-H(u|x)\big)^2} dH_1(u|x),\\
&\eta(Z,\delta,t,x) = \int K(u)\big(1-F(t-gu|x)\big)\xi(Z,\delta,t-gu,x)du,\\
&\Phi_\xi(u,t,x) = E\big[\xi(Z_1,\delta_1,t,x)|X_1 = u\big],\\
&J(t|x) = \big(1-F(t|x)\big)L(t|x),\\
&L(t|x) = \int_o^t \frac{dH_1(z|x)}{\big(1-H(z|x)\big)^2},\\
&D_g(u,t_1,t_2,x) = \mathrm{Cov}\big[\eta(Z_1,\delta_1,t_1,x),\eta(Z_1,\delta_1,t_2,x)|X_1 = u\big]m(u),\\
&N(u,t_1,t_2,x) = E\big[\xi(Z_1,\delta_1,t_1,x)\xi(Z_1,\delta_1,t_2,x)\ X_1 = u\big],\\
&D(t,x) = \big(1-F(t|x)\big)^2\Big(m''(x)N(x,t,t,x) + m(x)N''(x,t,t,x) + 2m'(x)N'(x,t,t,x)\\
&\qquad\qquad -2c_K m(x)\Phi'_\xi(x,t,x)\Phi'_\xi(x,t,x)\Big),\\
&B_1(t,x) = \frac{d_K\big(1-F(t|x)\big)}{2m(x)}\big(2\Phi'_\xi(x,t,x)m'(x) + \Phi''_\xi(x,t,x)\big),\\
&B_2(t,x) = -\frac{1}{2}d_K F''(t|x),\\
&C_1(t_1,t_2,x) = \frac{2c_K}{m(x)}J(t_1|x)\big(1-F(t_2|x)\big)\mathbb{K} * K\left(\frac{t_2-t_1}{g}\right),
\end{aligned}
$$

$$
\begin{aligned}
C_2(t_1,t_2,x) &= \frac{c_K}{m(x)}\left(2J(t_1|x)f(t_2|x)\mathbb{K}*K_1\left(\frac{t_1-t_2}{g}\right)\right.\\
&\qquad \left.+2J'(t_1|x)\big(1-F(t_2|x)\big)\mathbb{K}*K_1\left(\frac{t_2-t_1}{g}\right)\right),\\
C_3(t_1,t_2,x) &= \frac{d_{K^2}}{m^2(x)}\Big(m(x)\big(1-F(t_1|x)\big)\big(1-F(t_2|x)\big)\Phi'_\xi(x,t_1,x)\Phi'_\xi(x,t_2,x)\\
&\qquad +\frac{1}{2}D''_g(x,t_1,t_2,x)\Big),\\
V_1(t,x) &= \frac{c_K}{m(x)}\big(1-F(t|x)\big)^2L(t|x),\\
V_2(t,x) &= \frac{c_K(c_{\mathbb{K}}-1)}{m(x)}\big(1-F(t|x)\big)^2L'(t|x),\\
V_3(t,x) &= \frac{d_{K^2}}{m^2(x)}\Big(m(x)\big(1-F(t|x)\big)^2\big(\Phi'_\xi(x,t,x)\big)^2+\frac{1}{2}D(t,x)\Big).
\end{aligned}
$$

Another assumption related to the differentiability of the above functions is required:

A.10 Let $(t,x)\in[l,u]\times I_c$. The first derivative of $L(u|x)$ with respect to u exists at (t,x). The second derivative of $m(u)$ exists et $u=x$. The second derivative of $S(u|x)$ exists at (t,x) and $(t+b,x)$. The second derivative of $\Phi_\xi(u,t,x)$ exists at (x,t,x). The second derivative of $J(u|x)$ exists at (t,x). The second derivative of $D_g(u,t_1,t_2,x)$ exists at $(x,t,t+b,x)$. The second derivatie of $N(u,t_1,t_2,x)$ exists at (x,t,t,x).

Theorem A.1. *Let $(t,x)\in[l,u]\times I_c$ be such that $S(t|x)>0$. Under assumptions A.1-A.10, expressions for the asymptotic bias of $\widetilde{PD}^B_{h,g}(t|x)$, $ABias\big(\widetilde{PD}^B_{h,g}(t|x)\big)$, and the asymptotic variance of $\widetilde{PD}^B_{h,g}(t|x)$, $AVar\big(\widetilde{PD}^B_{h,g}(t|x)\big)$, are the following:*

$$
\begin{aligned}
ABias\big(\widetilde{PD}^B_{h,g}(t|x)\big) &= \frac{\big(1-PD(t|x)\big)B_1(t,x)-B_1(t+b,x)}{S(t|x)}h^2\\
&\quad+\frac{\big(1-PD(t|x)\big)B_2(t,x)-B_2(t+b,x)}{S(t|x)}g^2,
\end{aligned}
$$

$$
\begin{aligned}
AVar\big(\widetilde{PD}^B_{h,g}(t|x)\big) &= \left(\frac{V_1(t+b,x)}{S(t|x)^2}-2\frac{S(t+b|x)C_1(t,t+b,x)}{S(t|x)^3}+\frac{S(t+b|x)^2V_1(t,x)}{S(t|x)^4}\right)\frac{1}{nh}\\
&\quad+\left(\frac{V_2(t+b,x)}{S(t|x)^2}-2\frac{S(t+b|x)C_2(t,t+b,x)}{S(t|x)^3}+\frac{S(t+b|x)^2V_2(t,x)}{S(t|x)^4}\right)\frac{g}{nh}\\
&\quad+\left(\frac{V_3(t+b,x)}{S(t|x)^2}-2\frac{S(t+b|x)C_3(t,t+b,x)}{S(t|x)^3}+\frac{S(t+b|x)^2V_3(t,x)}{S(t|x)^4}\right)\frac{h}{n}.
\end{aligned}
$$

Theorem A.2. *Under the assumptions of Theorem A.1 and assuming* $C_h := \lim_{n\to\infty} n^{1/5}h > 0$ *and* $C_g := \lim_{n\to\infty} n^{1/5}g > 0$*, the limit distribution of* $\widetilde{PD}^B_{h,g}(t|x)$ *is given by*

$$\sqrt{nh}\big(\widetilde{PD}^B_{h,g}(t|x) - PD(t|x)\big) \xrightarrow{d} N(\mu, s_0),$$

where

$$\begin{aligned}\mu &= C_h^{5/2}\frac{\big(1-PD(t|x)\big)B_1(t,x) - B_1(t+b,x)}{S(t|x)} \\ &\quad + C_h^{1/2}C_g^{4/2}\frac{\big(1-PD(t|x)\big)B_2(t,x) - B_2(t+b,x)}{S(t|x)}\end{aligned}$$

and

$$\begin{aligned}s_0^2 &= \frac{V_1(t+b,x)}{S(t|x)^2} - 4\frac{S(t+b|x)}{S(t|x)^3}\frac{c_K\big(1-F(t|x)\big)\big(1-F(t+b|x)\big)L(t|x)}{m(x)} \\ &\quad + \frac{S(t+b|x)^2 V_1(t,x)}{S(t|x)^4}.\end{aligned}$$

Remark 1. *Assuming* $C_h := \lim_{n\to\infty} n^{1/5}h > 0$*, but* $n^{1/5}g \to 0$*, the asymptotic distribution of the smoothed Beran's PD estimator is* $\sqrt{nh}\big(\widetilde{PD}^B_{h,g}(t|x) - PD(t|x)\big) \xrightarrow{d} N(\widetilde{\mu}, s_0)$*. with*

$$\widetilde{\mu} = C_h^{5/2}\frac{\big(1-PD(t|x)\big)B_1(t,x) - B_1(t+b,x)}{S(t|x)}.$$

Assuming $n^{1/5}h \to 0$*,* $n^{1/5}g \to 0$ *and* $\dfrac{nh}{(\ln n)^3} \to \infty$*, the asymptotic distribution of the smoothed Beran's PD estimator is* $\sqrt{nh}\big(\widetilde{PD}^B_{h,g}(t|x) - PD(t|x)\big) \xrightarrow{d} N(0, s_0)$.

B. Proofs

Proofs of the results shown in Appendix A are done using results from papers Peláez et al. (2021b) and Peláez et al. (2021a).

Proof of Theorem A.1.

Denote $P = S(t+b|x)$, $Q = S(t|x)$ and $PD(t|x) = 1 - \dfrac{P}{Q}$. Similarly, $\widetilde{P} = \widetilde{S}^B_{h,g}(t+b|x)$, $\widetilde{Q} = \widetilde{S}^B_{h,g}(t|x)$ and $\widetilde{PD}^B_{h,g}(t|x) = 1 - \dfrac{\widetilde{P}}{\widetilde{Q}}$. As a consequence of the proof of Theorem 1 in Peláez et al. (2021b):

$$ABias\big(\widetilde{PD}^B_{h,g}(t|x)\big) = \alpha_1 + \alpha_2 + \alpha_3, \tag{6}$$

$$AVar\big(\widetilde{PD}^{B}_{h,g}(t|x)\big) = \beta_1 + \beta_2 + \beta_3, \tag{7}$$

where

$$\alpha_1 = \frac{P}{Q} - \frac{E(\widetilde{P})}{E(\widetilde{Q})}, \quad \alpha_2 = \frac{Cov(\widetilde{P},\widetilde{Q})}{E(\widetilde{Q})^2}, \quad \alpha_3 = -\frac{E\left[\frac{\widetilde{P}}{\widetilde{Q}}\big(\widetilde{Q}-E(\widetilde{Q})\big)^2\right]}{E(\widetilde{Q})^2} \tag{8}$$

and

$$\beta_1 = \frac{Var(\widetilde{P})}{E(\widetilde{Q})^2}, \quad \beta_2 = -2\frac{E(\widetilde{P})Cov(\widetilde{P},\widetilde{Q})}{E(\widetilde{Q})^3}, \quad \beta_3 = \frac{E(\widetilde{P})^2 Var(\widetilde{Q})}{E(\widetilde{Q})^4}. \tag{9}$$

The asymptotic expressions for the bias, the covariance and the variance of the survival estimator $\widetilde{S}^{B}_{h,g}(t|x)$ are obtained from Theorems 3 and 4 of Peláez et al. (2021a):

$$\text{Bias}\big(\widetilde{S}^{B}_{h,g}(t|x)\big) = B_1(t,x)h^2 + B_2(t,x)g^2 + o(h^2), \tag{10}$$

$$\begin{aligned}\text{Cov}\big(\widetilde{S}^{B}_{h,g}(t_1|x),\widetilde{S}^{B}_{h,g}(t_2|x)\big) &= C_1(t_1,t_2,x)\frac{1}{nh} + C_2(t_1,t_2,x)\frac{g}{nh} \\ &\quad + C_3(t_1,t_2,x)\frac{h}{n} + R_n(t,x),\end{aligned} \tag{11}$$

$$\text{Var}\big(\widetilde{S}^{B}_{h,g}(t|x)\big) = V_1(t,x)\frac{1}{nh} + V_2(t,x)\frac{g}{nh} + V_3(t,x)\frac{h}{n} + R_n(t,x), \tag{12}$$

where $R_n(t,x) = o\left(\frac{g^2}{nh} + \frac{h}{n}\right)$.

Considering Equations (8)-(12), detailed expressions for α_1, α_2 and α_3 are obtained as follows:

$$\begin{aligned}\alpha_1 &= \frac{P}{Q} - \frac{P + B_1(t+b,x)h^2 + B_2(t+b,x)g^2 + o(h^2) + o(g^2)}{Q + B_1(t,x)h^2 + B_2(t,x)g^2 + o(h^2) + o(g^2)} \\ &= \frac{PQ + PB_1(t,x)h^2 + PB_2(t,x)g^2 + o(h^2) + o(g^2)}{Q\Big(Q + B_1(t,x)h^2 + B_2(t,x)g^2 + o(h^2) + o(g^2)\Big)} + \\ &\quad \frac{-PQ - QB_1(t+b,x)h^2 - QB_2(t+b,x)g^2 + o(h^2) + o(g^2)}{Q\Big(Q + B_1(t,x)h^2 + B_2(t,x)g^2 + o(h^2) + o(g^2)\Big)} \\ &= \frac{PB_1(t,x)h^2 - QB_1(t+b,x)h^2 + o(h^2) + o(g^2)}{Q\Big(Q + B_1(t,x)h^2 + B_2(t,x)g^2 + o(h^2) + o(g^2)\Big)} + \\ &\quad \frac{PB_2(t,x)g^2 - QB_2(t+b,x)g^2 + o(h^2) + o(g^2)}{Q\Big(Q + B_1(t,x)h^2 + B_2(t,x)g^2 + o(h^2) + o(g^2)\Big)} \\ &= \frac{\big(1 - PD(t|x)\big)B_1(t,x) - B_1(t+b,x)}{S(t|x)}h^2 \\ &\quad + \frac{\big(1 - PD(t|x)\big)B_2(t,x) - B_2(t+b,x)}{S(t|x)}g^2 + o(h^2) + o(g^2),\end{aligned} \tag{13}$$

$$\alpha_2 = \frac{C_1(t,t+b,x)}{S(t|x)^2}\frac{1}{nh} + \frac{C_2(t,t+b,x)}{S(t|x)^2}\frac{g}{nh} + \frac{C_3(t,t+b,x)}{S(t|x)^2}\frac{h}{n} + R_n(t,x), \tag{14}$$

$$\begin{aligned} \alpha_3 &= \frac{E\left[\frac{\widetilde{P}}{\widetilde{Q}}\left(\widetilde{Q}-E(\widetilde{Q})\right)^2\right]}{E(\widetilde{Q})^2} \leq \frac{\mathrm{Var}(\widetilde{Q})}{E(\widetilde{Q})^2} \\ &= \frac{V_1(t,x)}{S(t|x)^2}\frac{1}{nh} + \frac{V_2(t,x)}{S(t|x)^2}\frac{g}{nh} + \frac{V_3(t,x)}{S(t|x)^2}\frac{h}{n} + R_n(t,x). \end{aligned} \tag{15}$$

Plugging (13), (14) and (15) into (6) and using Assumption A.9,

$$\begin{aligned} \mathrm{ABias}\left(\widetilde{PD}^B_{h,g}(t|x)\right) &= \frac{\left(1-PD(t|x)\right)B_1(t,x) - B_1(t+b,x)}{S(t|x)}h^2 \\ &\quad + \frac{\left(1-PD(t|x)\right)B_2(t,x) - B_2(t+b,x)}{S(t|x)}g^2, \end{aligned}$$

and the bias part in Theorem A.1 is proved.

Now, expressions (9), (10), (11) and (12) lead to

$$\beta_1 = \frac{V_1(t+b,x)}{S(t|x)^2}\frac{1}{nh} + \frac{V_2(t+b,x)}{S(t|x)^2}\frac{g}{nh} + \frac{V_3(t+b,x)}{S(t|x)^2}\frac{h}{n} + R_n(t,x), \tag{16}$$

$$\begin{aligned} \beta_2 &= -2\frac{S(t+b,x)C_1(t,t+b,x)}{S(t|x)^3}\frac{1}{nh} - 2\frac{S(t+b,x)C_2(t,t+b,x)}{S(t|x)^3}\frac{g}{nh} \\ &\quad -2\frac{S(t+b,x)C_3(t,t+b,x)}{S(t|x)^3}\frac{h}{n} + R_n(t,x), \end{aligned} \tag{17}$$

$$\begin{aligned} \beta_3 &= \frac{S(t+b,x)^2V_1(t,x)}{S(t|x)^4}\frac{1}{nh} + \frac{S(t+b,x)^2V_2(t,x)}{S(t|x)^4}\frac{g}{nh} \\ &\quad + \frac{S(t+b,x)^2V_3(t,x)}{S(t|x)^4}\frac{h}{n} + R_n(t,x), \end{aligned} \tag{18}$$

and plugging Equations (16), (17) and (18) in (7) the variance part in Theorem A.1 is proved. ■

Proof of Theorem A.2

From Equations (1) and (3) follows:

$$\frac{\widetilde{S}^B_{h,g}(t+b|x)}{\widetilde{S}^B_{h,g}(t|x)} - \frac{S(t+b|x)}{S(t|x)} = -\left(\widetilde{PD}^B_{h,g}(t|x) - PD(t|x)\right). \tag{19}$$

On the other hand, denoting $a_1 = \dfrac{1}{S(t|x)}$, $a_2 = -\dfrac{S(t+b|x)}{S(t|x)^2}$ and

$$C\left(\widetilde{S}^B_{h,g}(t|x)\right) = \frac{S(t|x)\left(\widetilde{S}^B_{h,g}(t+b|x) - S(t+b|x)\right) - S(t+b|x)\left(\widetilde{S}^B_{h,g}(t|x) - S(t|x)\right)}{\widetilde{S}^B_{h,g}(t|x)S(t|x)},$$

it holds

$$\begin{aligned}\frac{\widetilde{S}^{B}_{h,g}(t+b|x)}{\widetilde{S}^{B}_{h,g}(t|x)}-\frac{S(t+b|x)}{S(t|x)} &= a_1\big(\widetilde{S}^{B}_{h,g}(t+b|x)-S(t+b|x)\big)+a_2\big(\widetilde{S}^{B}_{h,g}(t|x)-S(t|x)\big)\\ &\quad +C\big(\widetilde{S}^{B}_{h,g}(t|x)\big)\left(1-\frac{\widetilde{S}^{B}_{h,g}(t|x)}{S(t|x)}\right),\end{aligned}$$

and considering (19):

$$\begin{aligned}PD(t|x) &- \widetilde{PD}^{B}_{h,g}(t|x)\\ &= a_1\big(\widetilde{S}^{B}_{h,g}(t+b|x)-S(t+b|x)\big)+a_2\big(\widetilde{S}^{B}_{h,g}(t|x)-S(t|x)\big)\\ &\quad +C\big(\widetilde{S}^{B}_{h,g}(t|x)\big)\left(1-\frac{\widetilde{S}^{B}_{h,g}(t|x)}{S(t|x)}\right).\end{aligned} \tag{20}$$

Since $\widetilde{S}^{B}_{h,g}(t|x)$ is a consistent estimator of $S(t|x)$, $\widetilde{S}^{B}_{h,g}(t|x)\xrightarrow{p}S(t|x)$. Thus,

$$1-\frac{\widetilde{S}^{B}_{h,g}(t|x)}{S(t|x)}\xrightarrow{p}0.$$

Therefore, the asymptotic distribution of $\sqrt{nh}\big(\widetilde{PD}^{B}_{h,g}(t|x)-PD(t|x)\big)$ is the same as the asymptotic distribution of the linear combination

$$a_1\sqrt{nh}\big(\widetilde{S}^{B}_{h,g}(t+b|x)-S(t+b|x)\big)+a_2\sqrt{nh}\big(\widetilde{S}^{B}_{h,g}(t|x)-S(t|x)\big).$$

From Lemma 1 in Peláez et al. (2021a), $\widetilde{S}^{B}_{h,g}(t|x)$ is split up into the following terms

$$\widetilde{S}^{B}_{h,g}(t|x) = S(t|x)+\sum_{i=1}^{n}\varphi_{n,i}(t,x)+B_2(t,x)g^2+R_n(t|x), \tag{21}$$

where $\varphi_{n,i}(t,x)=\dfrac{1}{nh}\dfrac{K\big((x-X_i)/h\big)}{m(x)}\eta(Z_i,\delta_i,t,x)$ are independent and identically distributed random variables for all $i=1,...,n$ and $R_n(t|x)$ is negligible with respect to the other terms:

$$R_n(t|x)=O_p\left(\frac{\ln n}{nh}\right)^{3/4}+o(g^2)+O_p\left(h^2+\frac{1}{\sqrt{nh}}\right)\sum_{i=1}^{n}\varphi_{n,i}(t,x).$$

Using (21),

$$\begin{aligned}&a_1\sqrt{nh}\big(\widetilde{S}^{B}_{h,g}(t+b|x)-S(t+b|x)\big)+a_2\sqrt{nh}\big(\widetilde{S}^{B}_{h,g}(t|x)-S(t|x)\big)\\ &\quad=\sum_{i=1}^{n}\widetilde{\varphi}_{n,i}(t,x)+a_1B_2(t+b,x)g^2\sqrt{nh}+a_2B_2(t,x)g^2\sqrt{nh}+\widetilde{R}_n(t,x),\end{aligned} \tag{22}$$

where

$$\widetilde{\varphi}_{n,i}(t,x) = \sqrt{nh}\big(a_1\varphi_{n,i}(t+b,x)+a_2\varphi_{n,i}(t,x)\big) \tag{23}$$

and

$$\begin{aligned}\widetilde{R}_n(t,x) &= \sqrt{nh}\big(a_1R_n(t+b,x)+a_2R_n(t,x)\big)\\ &= \sqrt{nh}(a_1+a_2)O_p\left(\frac{\ln n}{nh}\right)^{3/4}+\sqrt{nh}(a_1+a_2)o(g^2)\\ &\quad +O_p\left(h^2+\frac{1}{\sqrt{nh}}\right)\sum_{i=1}^{n}\widetilde{\varphi}_{n,i}(t,x).\end{aligned} \tag{24}$$

Since $h\to 0$ and $nh\to\infty$, the term $O_p\left(h^2+\frac{1}{\sqrt{nh}}\right)\sum_{i=1}^n\widetilde{\varphi}_{n,i}(t,x)$ in (24) is negligible with respect to $\sum_{i=1}^n\widetilde{\varphi}_{n,i}(t,x)$ in (22). Given that $g\to 0$, the term $\sqrt{nh}(a_1+a_2)o(g^2)$ in (24) is negligible with respect to $a_1B_2(t+b,x)g^2\sqrt{nh}+a_2B_2(t,x)g^2\sqrt{nh}$ in (22). Finally, the term $\sqrt{nh}(a_1+a_2)O_p\left(\frac{\ln n}{nh}\right)^{3/4}$ in (24) is negligible with respect to $\sum_{i=1}^n\widetilde{\varphi}_{n,i}(t,x)$ in (22) because $\frac{nh}{(\ln n)^3}=\frac{C_hn^{4/5}}{(\ln n)^3}\to\infty$. The variance of the dominant term in (22) is $O(1)$:

$$\begin{aligned}\mathrm{Var}\big(\textstyle\sum_{i=1}^n\widetilde{\varphi}_{n,i}(t,x)\big) &= n\mathrm{Var}\big(\widetilde{\varphi}_{n,1}(t,x)\big)\\ &= n^2h\Big(a_1^2\mathrm{Var}\big(\varphi_{n,1}(t+b,x)\big)+a_2^2\mathrm{Var}\big(\varphi_{n,1}(t,x)\big)\\ &\quad +2a_1a_2\mathrm{Cov}\big(\varphi_{n,1}(t+b,x),\varphi_{n,1}(t,x)\big)\Big).\end{aligned} \tag{25}$$

From the proof of Theorem 3 in Peláez et al. (2021a),

$$\begin{aligned}&\mathrm{Cov}\big(\varphi_{n,1}(t_1,x),\varphi_{n,1}(t_2,x)\big)\\ &= \frac{2c_K}{m(x)n^2}\big(1-F(t_1|x)\big)\big(1-F(t_2|x)\big)L(t_1|x)\mathbb{K}*K\left(\frac{t_2-t_1}{g}\right)\frac{1}{h}+O\left(\frac{g}{n^3h}\right).\end{aligned}$$

In particular, for $t_1=t$, $t_2=t+b$, $\mathbb{K}*K\left(\frac{t_2-t_1}{g}\right)=\mathbb{K}*K\left(\frac{b}{g}\right)$ and

$$\begin{aligned}\lim_{n\to\infty}\mathbb{K}*K\left(\frac{b}{g}\right) &= \lim_{u\to\infty}\int_{-\infty}^{+\infty}K(y)\mathbb{K}(u-y)dy\\ &= \int_{-\infty}^{+\infty}\lim_{u\to\infty}\mathbb{K}(u-y)K(y)dy=\int_{-\infty}^{+\infty}K(y)dy=1.\end{aligned}$$

Consequently,

$$\begin{aligned}&\mathrm{Cov}\big(\varphi_{n,1}(t+b,x),\varphi_{n,1}(t,x)\big)\\ &= \frac{2c_K}{m(x)n^2}\big(1-F(t|x)\big)\big(1-F(t+b|x)\big)L(t|x)\frac{1}{h}+O\left(\frac{g}{n^3h}\right)+o\left(\frac{1}{n^2h}\right).\end{aligned} \tag{26}$$

For $t_1=t_2$,

$$\begin{aligned}
\mathbb{K}*K\left(\frac{t_2-t_1}{g}\right) &= \mathbb{K}*K(0) = \int \mathbb{K}(u)K(-u)du \\
&= \int \mathbb{K}(u)K(u)du = \int K(u)\left(\int_{-\infty}^{u} K(v)dv\right)du = \int\int_{\{v\le u\}} K(u)K(v)dudv \\
&= \frac{1}{2}\left(\int\int_{\{v\le u\}} K(u)K(v)dudv + \int\int_{\{u\le v\}} K(v)K(u)dvdu\right) \\
&= \frac{1}{2}\int\int_{\mathbb{R}^2} K(u)K(v)dudv = \frac{1}{2}.
\end{aligned}$$

So,

$$\mathrm{Var}\big(\varphi_{n,1}(t,x)\big) = \frac{c_K}{m(x)n^2}\big(1-F(t|x)\big)^2\big)L(t|x)\frac{1}{h} + O\left(\frac{g}{n^3h}\right). \tag{27}$$

Replacing (26) and (27) in (25),

$$\begin{aligned}
&\mathrm{Var}\big(\textstyle\sum_{i=1}^n \widetilde{\varphi}_{n,i}(t,x)\big) \\
&\quad = a_1^2\frac{c_K}{m(x)}\big(1-F(t+b|x)\big)^2L(t+b|x) + a_2^2\frac{c_K}{m(x)}\big(1-F(t|x)\big)^2L(t|x) \\
&\quad +4a_1a_2\frac{c_K}{m(x)}\big(1-F(t|x)\big)\big(1-F(t+b|x)\big)L(t|x) + O\left(\frac{g}{n}\right) + o(1).
\end{aligned}$$

Thus, $\mathrm{Var}\big(\sum_{i=1}^n \widetilde{\varphi}_{n,i}(t,x)\big) = O(1)$ and the linear combination can be expressed as (22) with $\widetilde{R}_n(t,x)$ negligible with respect to the term $\sum_{i=1}^n \widetilde{\varphi}_{n,i}(t,x)$. Therefore, we proceed to analyse the asymptotic distribution of $\sum_{i=1}^n \widetilde{\varphi}_{n,i}(t,x)$.

As the variables $\varphi_{n,i}(t,x)$ are independent and identically distributed for all $i = 1,...,n$, the variables $\widetilde{\varphi}_{n,i}(t,x)$ are also so. In addition, $\mathrm{Var}\big(\widetilde{\varphi}_{n,i}(t,x)\big)$ exists and it is finite for all $i = 1,...,n$. In this scenario, if Lindeberg's condition for triangular arrays (see Theorem 7.2 in Billingsley (1968)) is satisfied, then

$$\sum_{i=1}^{n}\left(\widetilde{\varphi}_{n,i}(t,x) - E\big[\widetilde{\varphi}_{n,i}(t,x)\big]\right) \xrightarrow{d} N(0,s_0), \tag{28}$$

where

$$\begin{aligned}
s_0^2 &= a_1^2\frac{c_K}{m(x)}\big(1-F(t+b|x)\big)^2L(t+b|x) + a_2^2\frac{c_K}{m(x)}\big(1-F(t|x)\big)^2L(t|x) \\
&\quad +4a_1a_2\frac{c_K}{m(x)}\big(1-F(t|x)\big)\big(1-F(t+b|x)\big)L(t|x).
\end{aligned} \tag{29}$$

We will now check Lindeberg's condition:

$$\lim_{n\to\infty}\frac{1}{s_0^2}E\left[\sum_{i=1}^{n}\left(\widetilde{\varphi}_{n,i}(t,x) - E\big[\widetilde{\varphi}_{n,i}(t,x)\big]\right)^2\mathbb{1}_{n,i}\right] = 0 \tag{30}$$

for every $\varepsilon > 0$, where $\mathbb{1}_{n,i}$ denotes the indicator function given by

$$\mathbb{1}_{n,i} = \mathbb{1}\Big(\ \widetilde{\varphi}_{n,i}(t,x) - E[\widetilde{\varphi}_{n,i}(t,x)]\ > \varepsilon s_0\Big).$$

Using assumption A.3d, $\xi(Z,\delta,t,x)$ is found out to be bounded:

$$\begin{aligned}|\xi(Z,\delta,t,x)| &= \frac{1_{\{Z\le t,\delta=1\}}}{1-H(Z|x)} - \int_0^t \frac{dH_1(u|x)}{\big(1-H(u|x)\big)^2} \\ &\le \frac{1_{\{Z\le t,\delta=1\}}}{1-H(Z|x)} + \int_0^t \frac{dH_1(u|x)}{\big(1-H(u|x)\big)^2} \le \frac{1}{\theta} + \int_0^t \frac{dH_1(u|x)}{\theta^2} \\ &\le \frac{1}{\theta} + \frac{H(t|x)}{\theta^2} \le \frac{1}{\theta} + \frac{1}{\theta^2}\end{aligned}$$

and, consequently, η is also bounded:

$$\begin{aligned}|\eta(Z,\delta,t,x)| &\le \int K(u)\big(1-F(t-gu|x)\big)\left(\frac{1}{\theta}+\frac{1}{\theta^2}\right)du \\ &= \left(\frac{1}{\theta}+\frac{1}{\theta^2}\right)\left(\big(1-F(t|x)\big)+\frac{g^2}{2}d_K\big(1-F''(t|x)\big)\right)+O(g^2).\end{aligned}$$

Since η is bounded, K and $m(x)$ have compact support and $nh\to\infty$, $\big\{\widetilde{\varphi}_{n,i}(t,x), i=1,...,n, n\in\mathbb{N}$ is a sequence of random variables which is bounded by a convergent to zero sequence. Hence, there exists $n_0\in\mathbb{N}$ such that for all $i=1,...,n$, $\mathbb{1}_{n,i}=0$ for all $n\ge n_0$ and accordingly,

$$\lim_{n\to\infty}\frac{1}{s_0^2}E\left[\sum_{i=1}^n\Big(\widetilde{\varphi}_{n,i}(t,x)-E\big[\widetilde{\varphi}_{n,i}(t,x)\big]\Big)^2\mathbb{1}_{n,i}\right]=0,$$

which proves Lindeberg's condition given in (30).

Furthermore, from Theorem 3 in Peláez et al. (2021a),

$$E\big(\varphi_{n,1}(t,x)\big)=B_1(t,x)\frac{h^2}{n}+o\left(\frac{h^2}{n}\right),$$

so,

$$\begin{aligned}E\Big(\textstyle\sum_{i=1}^n\widetilde{\varphi}_{n,i}(t,x)\Big) &= nE\big(\widetilde{\varphi}_{n,1}(t,x)\big) \\ &= a_1n\sqrt{nh}E\big(\varphi_{n,1}(t+b,x)\big)+a_2n\sqrt{nh}E\big(\varphi_{n,1}(t,x)\big) \\ &= \sqrt{nh^5}\big(a_1B_1(t+b,x)+a_2B_1(t,x)+o(h^2)\big).\end{aligned}$$

Therefore, taking into account that $h=C_hn^{-1/5}$, we have

$$\sum_{i=1}^n\widetilde{\varphi}_{n,i}(t,x)\xrightarrow{d}N(\mu_0,s_0),$$

where

$$\mu_0=C_h^{5/2}\big(a_1B_1(t+b,x)+a_2B_1(t,x)\big).$$

Consequently, recalling (22) and assuming $g=C_gn^{-1/5}$,

$$a_1\sqrt{nh}\big(\widetilde{S}^B_{h,g}(t+b|x)-S(t+b|x)\big)+a_2\sqrt{nh}\big(\widetilde{S}^B_{h,g}(t|x)-S(t|x)\big)\xrightarrow{d}N(\mu_1,s_0),$$

where

$$\mu_1 = \mu_0 + C_h^{1/2} C_g^{4/2} \big(a_1 B_2(t+b,x) + a_2 B_2(t,x) \big).$$

Finally, using equation (20) with $a_1 = \dfrac{1}{S(t|x)}$ and $a_2 = -\dfrac{S(t+b|x)}{S(t|x)^2}$, the asymptotic distribution of the PD estimator holds:

$$\sqrt{nh}\big(\widetilde{PD}^B_{h,g}(t|x) - PD(t|x)\big) \xrightarrow{d} N(\mu, s_0),$$

where $\mu = -\mu_1$. Then,

$$\begin{aligned}
\mu &= C_h^{5/2} \left(\frac{S(t+b|x)}{S(t|x)^2} B_1(t,x) - \frac{B_1(t+b,x)}{S(t|x)} \right) \\
&\quad + C_h^{1/2} C_g^{4/2} \left(\frac{S(t+b|x)}{S(t|x)^2} B_2(t,x) - \frac{B_2(t+b,x)}{S(t|x)} \right) \\
&= C_h^{5/2} \frac{\big(1 - PD(t|x)\big) B_1(t,x) - B_1(t+b,x)}{S(t|x)} \\
&\quad + C_h^{1/2} C_g^{4/2} \frac{\big(1 - PD(t|x)\big) B_2(t,x) - B_2(t+b,x)}{S(t|x)}
\end{aligned}$$

and

$$\begin{aligned}
s_0^2 &= \frac{1}{S(t|x)^2} \frac{c_K \big(1 - F(t+b|x)\big)^2 L(t+b|x)}{m(x)} + \frac{S(t+b|x)^2}{S(t|x)^4} \frac{c_K \big(1 - F(t|x)\big)^2 L(t|x)}{m(x)} \\
&\quad - 4 \frac{S(t+b|x)}{S(t|x)^3} \frac{c_K \big(1 - F(t|x)\big)\big(1 - F(t+b|x)\big) L(t|x)}{m(x)} \\
&= \frac{V_1(t+b,x)}{S(t|x)^2} - 4 \frac{S(t+b|x)}{S(t|x)^3} \frac{c_K \big(1 - F(t|x)\big)\big(1 - F(t+b|x)\big) L(t|x)}{m(x)} \\
&\quad + \frac{S(t+b|x)^2 V_1(t,x)}{S(t|x)^4}.
\end{aligned}$$

■

SORT 45 (2) July-December 2021, 121-142 DOI: 10.2436/20.8080.02.112

Modified almost unbiased two-parameter estimator for the Poisson regression model with an application to accident data

Mustafa I. Alheety[1], Muhammad Qasim[2], Kristofer Månsson[2]
and B.M. Golam Kibria[3]

Abstract

Due to the large amount of accidents negatively affecting the wellbeing of the survivors and their families, a substantial amount of research is conducted to determine the causes of road accidents. This type of data come in the form of non-negative integers and may be modelled using the Poisson regression model. Unfortunately, the commonly used maximum likelihood estimator is unstable when the explanatory variables of the Poisson regression model are highly correlated. Therefore, this paper proposes a new almost unbiased estimator which reduces the instability of the maximum likelihood estimator and at the same time produce smaller mean squared error. We study the statistical properties of the proposed estimator and a simulation study has been conducted to compare the performance of the estimators in the smaller mean squared error sense. Finally, Swedish traffic fatality data are analyzed to show the benefit of the proposed method.

MSC: 62J05; 62J07.

Keywords: Applied traffic modeling, Maximum likelihood estimator, Mean squared error matrix, Poisson regression, Simulation study, Traffic fatality.

1. Introduction

According to the World Health Organization (2015), fatalities caused by motor vehicle collisions leads to more than 1.2 million deaths worldwide. This large amount of acci-

[1] Department of Mathematics,College of Education for Pure Sciences, University of Anbar, Iraq.

[2] Department of Economics, Finance and Statistics, Jönköping University, Sweden.

[3] Department of Mathematics and Statistics, Florida International University, Miami, FL 33199, USA, Email: kibriag@fiu.edu

Received: August 2020
Accepted: July 2021

dents negatively affects the wellbeing of the survivors and their families (Donaldson, Brooke and Faux, 2009). Therefore a great interest exists in developing new models and methods to estimate the causes of accidents. Examples of previous research where new methods are suggested have appeared in Ivan, Wang and Bernardo (2000), Lyon et al. (2003), Lord, Manar and Vizioli (2005b), Chiou and Fu (2013) and Shi, Abdel-Aty and Lee (2016) among others. This paper is motivated by the work of Shi et al. (2016) and focuses on the issue of multicollinearity which is defined as the situation when two or more explanatory variables are highly correlated.

The problem of multicollinearity has significant impact on the performance of ordinary least squares (OLS) estimation of unknown regression coefficients. Furthermore, it leads to instability and a high variance of the parameters estimated by OLS and eventually provides the wrong sign of the regression coefficients. Another consequence of multicollinearity is the wider confidence interval, decreased statistical power which result in increased probability of type II error in hypothesis testing in terms of the parameters. As a solution to this problem for linear regression models, Hoerl and Kennard (1970a, 1970b) proposed the ridge regression (RR) method, which is a biased or shrinkage estimator, as an alternative to ordinary least squares. They showed that one may reduce the variance of the estimated coefficients substantially by introducing a small amount of bias. This method was generalized in order to be used for models estimated by maximum likelihood estimator (MLE) such as the logit and Poisson models by Schaefer, Roi and Wolfe (1984) and Månsson and Shukur (2011), among others. Kibria, Månsson and Shukur (2015) proposed several estimators for estimating the ridge parameter k based on Poisson ridge regression (PRR) model. Liu (1993) by taking the advantage of ridge regression and Stein estimator (1956), proposed a new biased estimator and showed its merit for the linear regression model. The ridge (Hoerl and Kennard, 1970a), Liu (1993) and Liu-type estimators have been developed for other generalized linear models such as negative binomial regression, Poisson regression, zero inflated Poisson regression, gamma regression and beta regression models, for instances, see Månsson (2011), Månsson (2013), Asar and Genç (2018), Cetinkaya and Kaciranlar (2019), Toker, Ustundağ and Qasim (2019), Qasim et al. (2020a, 2020b), Kibria, Månsson and Shukur (2013), Huang and Yang (2014), Kurtoglu and Ozkale (2016), Qasim, Amin and Amanullah (2018), Lukman et al. (2020), Amin, Qasim and Amanullah (2019), Amin et al. (2020a, 2020b), Karlsson, Månsson and Kibria (2020), Qasim, Månsson and Kibria (2021) among others.

In this paper, we propose a new general biased estimator for Poisson regression model, which will be called the modified almost unbiased two-parameter Poisson estimator (MAUTPPE). The previous methods suggested by Månsson and Shukur (2011) and Shi et al. (2016) have disadvantages of inducing much bias. This is an unattractive property to applied researchers of these estimators and therefore, in this paper, we suggest a bias correction that substantially reduces the bias and still solves the problem of multicollinearity. As an illustration of this new method, we model traffic fatality data of Sweden. We show a substantial increase of predictive power of this new method as compared to MLE and the standard ridge regression method.

The organization of the paper is as follows. The proposed estimator and its superiority are given in Section 2. The estimation of the shrinkage parameters are outlined in Section 3. To compare the performance of the estimators, a simulation study has been conducted in Section 4. An application about the traffic fatalities in Sweden is given in Section 5. Finally some concluding remarks are given in Section 6.

2. Statistical methodology

2.1. Maximum likelihood estimator for the Poisson regression model

The Poisson regression model is used when the dependent variable (y_i) comes in the form of count data and distributed as P(μ_i) , where μ_i is a parameter of the Poisson distribution and it can be written as $\mu_i = \exp(x_i\beta)$ as mean response function for the Poisson regression model, where x_i is the i-th row of X which is a $n \times (p+1)$ data matrix with p explanatory variables and β is a $(p+1) \times 1$ vector of coefficients. The traditional MLE is used to estimate β. The log likelihood of this model corresponds to:

$$L(\beta; y) = \sum_{i=1}^{n} \exp(x_i\beta) + \sum_{i=1}^{n} y_i \log(\exp(x_i\beta)) + \log(\prod_{i=1}^{n} y_i!) \tag{1}$$

Solving $L(\beta; y)$ with respect to β results in:

$$\frac{\partial L}{\partial \beta} = \sum_{i=1}^{n} (y_i - \exp(x_i\beta))x_i = 0$$

Now, we use the iteratively re-weighted least squares (IRLS) algorithm to get the MLE which can be written as follows:

$$\hat{\beta} = (X^{\intercal}\hat{W}X)^{-1}X^{\intercal}\hat{W}Z = (S)^{-1}X^{\intercal}\hat{W}Z, \tag{2}$$

where $S = X^{\intercal}\hat{W}X$, $\hat{W} = \text{diag}(\hat{\mu}_i)$ and Z is the column vector with

$$Z_i = \log(\hat{\mu}_i)\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}$$

The MLE of $\hat{\beta}$ is asymptotically unbiased estimator of β. When the explanatory variables are suffering for high correlation, the matrix S is ill-conditioned and the MLE becomes unstable with high variance. To solve this problem, Månsson and Shukur (2011) introduced the Poisson ridge estimator (PRE) as follows:

$$\hat{\beta}_{\text{PRR}} = (S + kI_p)^{-1}S\hat{\beta}, k > 0 \tag{3}$$

Also, Månsson et al. (2012) and Qasim et al. (2019) proposed the Poisson Liu regression estimator (PLRE) as:

$$\hat{\beta}_{\text{PLE}} = (S + I_p)^{-1}(S + dI_p)\hat{\beta}$$

$$= [I_p - (1-d)(S+I_p)^{-1}]\hat{\beta}, \qquad 0 < d < 1 \tag{4}$$

In order to get an estimator that performs better than the PRE and PLRE, Asar and Genç (2018) proposed the following two-parameter Poisson estimator (TPPE) as:

$$\hat{\beta}_{\text{TPE}} = T_{k,d}\hat{\beta}, \qquad k > 0, \quad 0 < d < 1 \tag{5}$$

where $T_{k,d} = (S+kI_p)^{-1}(S+kdI_p)$.

2.2. The proposed estimator

The TPPE (Asar and Genç, 2018) is the biased estimator and it has disadvantage of inducing considerable bias. This is an unattractive property to applied researchers. Therefore, in this section, we propose a bias correction that substantially reduces the bias and is more efficient than TPPE as well as improved estimators. The new estimator, which we called the modified almost unbiased two-parameters Poisson estimator, denoted by $\hat{\beta}_{\text{MAUTPPE}}$ and defined as follows:

$$\hat{\beta}_{\text{MAUTPPE}} \quad = \quad F_{k,d}\hat{\beta}, \qquad k > 0, \quad 0 < d < 1 \tag{6}$$

where $F_{k,d} = [I_p - (1-d)^2(S+I_p)^{-2}](I_p + kS^{-1})^{-1}, \quad 0 < d < 1, k > 0$.

The estimator in (6) is motivated from the following fact: The bias of $\hat{\beta}_{\text{PLE}}$ in Eq. (4) is given as

$$\text{Bias}(\hat{\beta}_{\text{PLE}}) = -(1-d)(S+I_p)^{-1}\beta.$$

Hence, by following Kadiyala (1984), the biased corrected of $\hat{\beta}_{\text{PLE}}$ can be defined as

$$\tilde{\beta}_{\text{PLE}} = \hat{\beta}_{\text{PLE}} + (1-d)(S+I_p)^{-1}\hat{\beta}.$$

Therefore, by following Ohtani (1986), we replace the $\hat{\beta}$ by $\hat{\beta}_{\text{PLE}}$ to get the almost unbiased PLRE, $\tilde{\beta}_{\text{PLE}}$:

$$\begin{aligned} \tilde{\beta}_{\text{PLE}} &= [I_p - (1-d)(S+I_p)^{-1}]\hat{\beta}_{\text{PLE}} \\ &= [I_p - (1-d)^2(S+I_p)^{-2}](I_p + kS^{-1})^{-1}\hat{\beta} \end{aligned} \tag{7}$$

Now, if we replace $\hat{\beta}$ in Eq. (7) by $\hat{\beta}_{\text{PRE}}$ from Eq. (3), we get the proposed estimator in Eq. (6).

The properties of the MAUTPPE are obtained as follows:

$$E(\hat{\beta}_{\text{MAUTPPE}}) = F_{k,d}\beta$$

The bias of the MAUTPPE:

$$\begin{aligned} \text{Bias}(\hat{\beta}_{\text{MAUTPPE}}) &= (F_{k,d} - I_p)\beta \\ &= [(I_p - (1-d)^2(S+I_p)^{-2})(I_p + kS^{-1})^{-1} - I_p]\beta \\ &= = S^{-1}\{-k(S+I_p)^2 - S(1-d)^2\}(S+I_p)^{-2}(S+kI_p)^{-1}S\beta \\ &= B_1^*. \end{aligned} \tag{8}$$

The variance covariance matrix of the MAUTPPE is given as:

$$\text{Cov}\left(\hat{\beta}_{\text{MAUTPPE}}\right) = F_{k,d}S^{-1}F_{k,d}. \tag{9}$$

2.3. Properties of the estimators

We use the spectral decomposition in order to find the matrix mean square error (MMSE) and scalar mean squared error (SMSE). So, we can rewrite the matrix S as $S = P\Lambda P^{\intercal}$, where P and Λ are the eigenvectors and eigenvalues of the matrix S, respectively, such that $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$. Since MAUTPPE is the biased estimator, we have to use the MMSE as a criterion for goodness of fit where it is containing all relevant information regarding the estimators (such as, variance and biased). The MMSE of an estimator $\tilde{\beta}$ of β can be written as:

$$\begin{aligned} \text{MMSE}(\tilde{\beta}) &= E(\tilde{\beta}-\beta)(\tilde{\beta}-\beta)^{\intercal} \\ &= \text{Var}(\tilde{\beta}) + (\text{Bias}(\tilde{\beta}))(\text{Bias}(\tilde{\beta}))^{\intercal} \end{aligned}$$

$$\begin{aligned} \text{MMSE}(\hat{\beta}_{\text{MAUTPPE}}) &= P(I_p + k\Lambda^{-1})^{-1}(I_p - (1-d)^2(\Lambda + I_p)^{-2}) \\ & \quad \Lambda^{-1}(I_p - (1-d)^2(\Lambda + I_p)^{-2})(I_p + k\Lambda^{-1})^{-1}P^{\intercal} + B_1 B_1^{\intercal}, \end{aligned} \tag{10}$$

where k and d are the biasing parameters and $B_1 = \text{Bias}(\hat{\beta}_{\text{MAUTPPE}}) = (F_{k,d} - I_p)\alpha$, where $\alpha = P^{\intercal}\beta$.

If we take the trace of MMSE, then we get SMSE as follows:

$$\text{SMSE}(\tilde{\beta}) = \text{tr}(\text{MMSE}(\tilde{\beta})) \tag{11}$$

So,

$$\text{MMSE}(\hat{\beta}_{\text{MAUTPPE}}) = \sum_{j=1}^{p} \frac{\lambda_j \left\{(\lambda_j+1)^2 - (1-d)^2 \ \ ^2 + \alpha^2 \left\{k(\lambda_j+1)^2 + \lambda_j(1-d)^2 \ \ ^2\right.\right.}{(\lambda_j+k)^2(\lambda_j+1)^2} \tag{12}$$

Asar and Genç (2018) computed the MMSE and SME of the TPPE as:

$$\text{MMSE}(\hat{\beta}_{\text{TPE}}) = P(\Lambda+k)^{-1}(\Lambda+kdI_p)\Lambda^{-1}(\Lambda+kdI_p)(\Lambda+k)^{-1}P^{\intercal} + B_2 B_2^{\intercal},$$

$$\text{SMSE}(\hat{\beta}_{\text{TPE}}) = \sum_{j=1}^{p}\Big(\frac{(\lambda_j+kd)^2}{\lambda_j(\lambda_j+k)^2}\Big) + \sum_{j=1}^{p}\Big(\frac{\alpha_j^2(d-1)^2k^2}{(\lambda_j+k)^2}\Big)$$

where $B_2 = \text{Bias}(\hat{\beta}_{\text{TPE}}) = P(\Lambda+kI_p)^{-1}\alpha(d-1)k$.

The MMSE and SMSE of the MLE are defined respectively as follows:

$$\text{MMSE}(\hat{\beta}) = S^{-1} = P\Lambda^{-1}P^{\intercal}.$$

$$\text{SMSE}(\hat{\beta}) = \sum_{j=1}^{p}\frac{1}{\lambda_j}$$

2.4. The performance of the proposed estimator

2.4.1. The comparison between the MLE and MAUTPPE

The comparison between MLE and MAUTPPE are illustrated using matrix mean squared error (MMSE):

$$\mathrm{MMSE}(\hat{\beta}) = S^{-1}.$$
$$\mathrm{MMSE}(\hat{\beta}_{\mathrm{MAUTPPE}}) = F_{k,d}S^{-1}F_{k,d} + B_1B_1^{\tau},$$

We state the following theorem to demonstrate the comparison between MLE and MAUTPPE.

Theorem 2.1. *Under MMSE criterion, the MAUTPPE ($\hat{\beta}_{MAUTPPE}$) is superior to the MLE ($\hat{\beta}$), namely, MMSE($\hat{\beta}$) − MMSE($\hat{\beta}_{MAUTPPE}$) ≥ 0 if and only if:*

$$B_1^{\tau}[S^{-1} - F_{k,d}S^{-1}F_{k,d}]^{-1}B_1 \leq 1$$

Proof. The difference of MMSE values between MLE and MAUTPPE can be found as

$$\begin{aligned}\Delta_1 = \mathrm{MMSE}(\hat{\beta}) - \mathrm{MMSE}(\hat{\beta}_{\mathrm{MAUTPPE}}) &= S^{-1} - (F_{k,d}S^{-1}F_{k,d} + B_1B_1^{\tau}) \\ &= D_1 - B_1B_1^{\tau},\end{aligned}$$

where $D_1 = S^{-1} - F_{k,d}S^{-1}F_{k,d}$.

Let $D_1 = P\Upsilon P^{\tau} = P\mathrm{diag}\{\gamma_1,\ldots,\gamma_p\}P^{\tau}$ by using the spectral decomposition, where

$$\Upsilon = \Lambda^{-1} - (I+k\Lambda^{-1})^{-1}(I-(1-d)^2(\Lambda+I)^{-2})\Lambda^{(-1)}(I-(1-d)^2(\Lambda+I)^{-2})(I+k\Lambda^{-1})^{-1}.$$

Therefore,

$$\gamma_j = \frac{1-\left[\left(1+\frac{k}{\lambda_j}\right)^{-2}\left(1-\frac{(1-d)^2}{(\lambda_j+1)^2}\right)\left(1-\frac{(1-d)^2}{(\lambda_j+1)^2}\right)\right]}{\lambda_j}, \quad j=1,\ldots,p$$

Since $\frac{(1-d)^2}{(\lambda_j+1)^2} < 1$ and $\left(1+\frac{k}{\lambda_j}\right)^{-2} < 1$ for $k>0, 0<d<1$ and $\lambda_j > 0$. Then

$$\left(1+\frac{k}{\lambda_j}\right)^{-2}\left(1-\frac{(1-d)^2}{(\lambda_j+1)^2}\right)\left(1-\frac{(1-d)^2}{(\lambda_j+1)^2}\right) > 1;$$

and that means $\gamma_j > 0, \forall j$.

This implies that D_1 is positive definite Now, in order to find the conditions that make Δ_1 is positive definite, we have to introduce the Lemma 2.1:

Lemma 2.1 (See Farebrother, 1976)**.** *Let M be a positive definite matrix and α be a vector, then $M - \alpha\alpha^{\tau} \geq 0$ if and only if $\alpha^{\tau}M^{-1}\alpha \leq 1$.*

Therefore, by applying Lemma 2.1, the proof is completed.

2.4.2. The comparison between the TPPE and MAUTPPE estimators

The properties of TPPE are obtained as follows:

$$\begin{aligned} \text{Bias}(\hat{\beta}_{\text{TPPE}}) &= k(d-1)(S+kI_p)^{-1}\beta \\ &= B_2 \end{aligned}$$

and

$$\text{Cov}(\hat{\beta}_{\text{TPPE}}) = T_{k,d}S^{-1}T_{k,d}.$$

The MMSE of TPPE is given as follows:

$$\text{MMSE}(\hat{\beta}_{\text{TPPE}}) = T_{k,d}S^{-1}T_{k,d} + B_2B_2^{\top}. \tag{13}$$

The following theorem is demonstrated the comparison between TPPE and MAUTPPE.

Theorem 2.2. *For $0 < d < 1$ for fixed k, under Poisson regression model, the MAUTPPE $\hat{\beta}_{MAUTPPE}$ is superior to TPPE $\hat{\beta}_{TPPE}$ in the sense of MMSE if and only if*

$$B_1^{\top}D_2^{-1}B_1 \leq 1.$$

Proof. The difference of MMSE values between them can be given by:

$$\begin{aligned} \Delta_2 &= \text{MMSE}(\hat{\beta}_{\text{TPPE}}) - \text{MMSE}(\hat{\beta}_{\text{MAUTPPE}}) \\ &= PD_2P^{\top} + B_2B_2^{\top} - B_1B_1^{\top} \\ &= P\,\text{diag}\left\{ \frac{(\lambda_j+kd)^2}{\lambda_j(\lambda_j+k)^2} - \frac{\left(1+\frac{k}{\lambda_j}\right)^{-2}\left(1-\frac{(1-d)^2}{(\lambda_j+1)^2}\right)\left(1-\frac{(1-d)^2}{(\lambda_j+1)^2}\right)}{\lambda_j} \right\}_{j=1}^{p} P^{\top} + B_2B_2^{\top} - B_1B_1^{\top}, \end{aligned}$$

where

$$\begin{aligned} D_2 &= (\Lambda+kI_p)^{-1}(\Lambda+kd)\Lambda_{-1}(\Lambda+kd)(\Lambda+kI_p)^{-1} \\ &- (I_p-(1-d)^2(\Lambda+I_p)^{-2})(I+k\Lambda^{-1})^{-1}\Lambda^{-1}(I_p+k\Lambda^{-1})^{-1} \\ &\quad (I_p-(1-d)^2(\Lambda+I_p)^{-2}) \end{aligned}$$

Since $B_2B_2^{\top}$ is nonnegative definite, we focus upon the quantity

$$\frac{(\lambda_j+kd)^2}{\lambda_j(\lambda_j+k)^2} - \frac{\left(1+\frac{k}{\lambda_j}\right)^{-2}\left(1-\frac{(1-d)^2}{(\lambda_j+1)^2}\right)\left(1-\frac{(1-d)^2}{(\lambda_j+1)^2}\right)}{\lambda_j}$$

for searching on the condition or conditions that make Δ_2 is positive definite.

Therefore, Δ_2 is positive definite if

$$\frac{(\lambda_j+kd)^2}{(\lambda_j+k)^2} \geq \left(1+\frac{k}{\lambda_j}\right)^{-2}\left(1-\frac{(1-d)^2}{(\lambda_j+1)^2}\right)\left(1-\frac{(1-d)^2}{(\lambda_j+1)^2}\right)$$

Let k be fixed, then after some simplifications for the above expression, we get:

$$(1-d)^2+(\lambda+1)^{-2}\frac{k}{\lambda_j}d \geq 0$$

Since $0<d<1$, $k>0$ and $\lambda_j>0$, the above inequality is hold and after applying Lemma 2.1, the proof is completed.

Also, we can state the following theorem:

Theorem 2.3. *For $k>0$ and let d be fixed, under Poisson regression model, the MAUTPPE is superior to TPPE in the MMSE if and only if*

$$B_1^{\top}D_2^{-1}B_1 \leq 1.$$

Proof. Same proof of Theorem 2.2.

Since the proposed estimator depends on the unknown parameters, d and k, we discuss their estimation techniques in the section follow.

3. New estimating methods for selection of k and d

It is a complicated challenge for practitioners to choose an appropriate value of k and d. Based on the work of Hoerl and Kennard (1970a), Alkhamisi et al. (2006), Kibria (2003), we propose some estimation methods for the selection of k and d.

Asar and Genç (2018) provided optimal values of k and d. Now, we derive the optimal value of k by taking derivative of SMSE($\hat{\beta}_{\text{MAUTPPE}}$) with respect to k and equating the resulting function to zero and solve for k. The procedure of estimating the optimal value is stated as:

$$\frac{\partial\left\{\text{SMSE}(\hat{\beta}_{\text{MAUTPPE}})\right\}}{\partial k} = \sum_{j=1}^{p}\left(\frac{2\alpha_j^2\left\{k(\lambda_j+1)^2+(1-d)^2\lambda_j\right.}{(\lambda_j+1)^2(\lambda_j+k)^2}\right)$$

$$-\sum_{j=1}^{p}\left(\frac{2\left\{\alpha_j^2((\lambda_j+1)^2k+(1-d)^2\lambda_j)^2+\lambda_j((\lambda_j+1)^2-(1-d)^2)^2\right\}}{(\lambda_j+1)^4(\lambda_j+k)^3}\right)$$

Equating the above equation to zero and solve for k:

$$k_j = \frac{\lambda_j^2+\left\{2-\alpha_j^2(1-d)^2\right\}\lambda_j-d^2+2d}{\alpha_j^2(\lambda_j+1)^2}, \forall \;\; j=1,2,\ldots,p.$$

Since the parameter k is positive, therefore, we suggest to apply absolute $|\ .\ |$ as

$$\hat{k_j} = |\ k_j\ |$$

. We propose the following new estimating methods for choosing the value of k based on the work of Hoerl and Kennard (1970a), Alkhamisi et al., (2006) and Kibria (2003).

$$\hat{k}_1 = \min(|\ k_j\ |).$$

$$\hat{k}_2 = \max(|\ k_j\ |).$$

$$\hat{k}_3 = \text{mean}(|\ k_j\ |).$$

$$\hat{k}_1 = \text{median}(|\ k_j\ |).$$

In addition, we derive the optimal value of d by taking derivative of $\hat{\beta}_{\text{MAUTPPE}}$ with respect to d and equating the resulting function to zero and solve for d:

$$d_j = \frac{(k\lambda_j)^{1/2}\left(\alpha_j^2\lambda_j + \alpha_j^2\right)\left(\alpha_j^2\lambda_j^2 + 1\right)^{1/2} + \alpha_j^2\lambda_j^2 + 1}{\alpha_j^2\lambda_j^2 + 1}$$

Since the value of d_j is limited between 0 and 1, therefore, we should use following estimating methods with min operator to get the value of d_j as follows:

$$\hat{d}_j = \frac{\hat{\alpha}_j^2\lambda_j^2 + 1}{(\hat{k}\lambda_j)^{1/2}\left(\hat{\alpha}_j^2\lambda_j + \hat{\alpha}_j^2\right)\left(\hat{\alpha}_j^2\lambda_j^2 + 1\right)^{1/2} + \hat{\alpha}_j^2\lambda_j^2 + 1}, \tag{14}$$

where $\lambda_j > 0$, $\alpha_j^2 > 0$ and $\hat{k} > 0$ which implies that the value of estimator $\hat{d}$ is between 0 and 1.

Now, we use the following algorithm to estimate parameters k and d.

1. Since $\hat{k}_1 - \hat{k}_4$ needs an initial value of d, we start by setting d equals some number between 0 and 1 and obtain $\hat{k}$.

2. By using Eq. (14), we estimate parameter d by plugging-in the value of k found in the first step.

3. In order to get a suitable value of $\hat{k}$, we use one of the $\hat{k}_1 - \hat{k}_4$ estimators by plugging-in the value of $\hat{d}$ found in the second step.

4. Finally, to choose the best estimate of the parameter d using one of the $\hat{k}_1 - \hat{k}_4$ from step 3 in Eq. (14) and then compute the $\hat{d}$ estimator.

4. A Simulation Study

In this section, we study the performance of the estimators using Monte Carlo simulation under different factors such as degrees of multicollinearity, different values of the shrinkage parameter d and number of explanatory variables. Different parameters are used with some specified value, illustrated in Table 1.

4.1. The design of an experiment

Following is the design of an experiment for the Poisson regression model:

1. The correlated explanatory variables are generated by considering the work of McDonald and Galarneau (1975).

$$x_{ij} = (1-\rho^2)^{0.50} w_{ij} + \rho w_{ip+1}\ ;\ \ j = 1,...,p;\ \ i = 1,\ldots,n, \tag{15}$$

   where w_{ij} are the independent standard normal pseudo-random numbers, ρ is quantified correlation between any two explanatory variables is stated as ρ^2 and x_{ij} is the number of explanatory variables. After generated correlated explanatory variables, we standardized these variables using length scaling.

2. The response variable, $Y_i (i = 1,\ldots,n)$ are generated from the Poisson distribution $P_o(\mu_i)$:

$$Y_i \sim P_o(\mu_i),$$

   where

$$\mu_i = E(Y_i) = \exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip});\ j = 1,2,...,p+1.$$

3. The parameter vectors corresponding to $p = 3,\ p = 6$ and $p = 9$ are selected by imposing the restriction on the coefficients $\beta_1, \beta_2, ..., \beta_p$ as normalized eigenvectors corresponding to the largest eigenvalues of the matrix $X^{\top}X$ so that $\sum_{j=1}^{p} \beta_j^2 = 1$ (see for more details; Kibria, 2003).

4. We use different estimators that given in Eq. (3) , (5) and (6) in this experiment. The $\widehat{\beta_{\mathrm{TPE}}}$ is estimated with the best shrinkage parameter

$$k_{\max} = \max\left[\frac{\lambda_j}{\lambda_j(1-d)\alpha_j^2 - d}\right],$$

and it was suggested by Asar and Genç (2018). For $\hat{\beta}_{\text{MAUTPPE}}$, we propose an algorithm for choosing values of the shrinkage parameters k and d. In addition, we consider initial value of d which are 0.10, 0.50 and 0.99. These values are chosen due to $0 < d < 1$ (e.g. see, Asar, Erişoğlu and Arashi, 2017).

5. In order to investigate the performance of the proposed estimators, we use MSE and bias.

$$\text{MSE}(\hat{\beta}) = \frac{\sum_{r=1}^{5000}\left[(\hat{\beta}_r - \beta)^{\top}(\hat{\beta}_r - \beta)\right]}{5000}. \tag{16}$$

$$\text{Bias}(\hat{\beta}) = \frac{\sum_{r=1}^{5000} \mid E(\hat{\beta}_r) - \beta \mid}{5000}, \tag{17}$$

where $\hat{\beta}_r$ is the estimated value of any estimator.

Table 1. *Factors, notations and values are used in the simulation.*

Factors	Notations	Values
Multicollinearity	ρ^2	0.85, 0.90, 0.95, 0.99
Number of explanatory variables	p	3, 6, 9
Initial value of shrinkage parameter	d	0.10, 0.50, 0.90
Sample size	n	25, 50, 100, 150, 200, 500
Number of Replications	R	5000

4.2. Results and Discussion

The estimated MSE and bias of the estimators are computed under different effective parameters such as sample size (n), degrees of correlation (ρ^2), initial value of the shrinkage parameter (d) and number of explanatory variables (p) and summarized them in Tables 2 to 5. All together, we created six simulation tables where we analyze the performance of MLE, TPPE and MAUTPPE by assuming different initial value of d which are 0.10, 0.50 and 0.99 (e.g. see for more details, Asar et al., 2017). To summarize the results and reduce the length of the paper, four representative tables (2-5) are included in the study. From the simulation results, it is perceived that proposed estimator MAUTPPE has the best performance as compared to the MLE and TPPE in sense of smaller MSE and bias. The MSE and bias of the MAUTPPE with $(\hat{d}, \hat{k}_2)$is minimized as compared to other shrinkage parameters ($\hat{k}_1, \hat{k}_3$ and $\hat{k}_4$).

Table 2. *Estimated MSE and bias of the estimators when* $p = 3$.

		Estimated MSE						Estimated Bias				
		MLE	TPPE	MAUTPPE				TPPE	MAUTPPE			
ρ^2	n			$(\hat{d}, \hat{k}_1)$	$(\hat{d}, \hat{k}_2)$	$(\hat{d}, \hat{k}_3)$	$(\hat{d}, \hat{k}_4)$		$(\hat{d}, \hat{k}_1)$	$(\hat{d}, \hat{k}_2)$	$(\hat{d}, \hat{k}_3)$	$(\hat{d}, \hat{k}_4)$
0.85	25	4.3700	3.6225	1.6484	1.5851	1.8466	3.0288	0.9101	0.6309	0.6387	0.6803	0.8034
	50	4.0496	3.5229	1.3957	1.4171	1.6462	2.0619	0.8912	0.6065	0.5756	0.6081	0.7142
	100	4.0184	3.3228	1.0442	1.2850	1.5154	2.0482	0.8686	0.4958	0.5541	0.5874	0.6465
	150	3.8890	3.1757	1.0151	1.2524	1.4106	2.0450	0.8444	0.4898	0.5253	0.5469	0.6337
	200	3.0153	2.0234	0.9514	0.8246	0.6644	0.1358	0.4607	0.4180	0.0182	0.3577	0.0750
	500	0.0114	0.0094	0.0071	0.0117	0.0116	0.0109	0.0071	0.0045	0.0047	0.0042	0.0038
0.90	25	4.4092	3.9826	1.7537	1.6586	2.2018	2.8577	0.9440	0.6548	0.6347	0.6954	0.7851
	50	4.0614	3.9435	1.5955	1.6100	1.9665	2.4220	0.9195	0.6114	0.6201	0.6882	0.7501
	100	4.0314	3.6039	1.4783	1.5610	1.9043	2.3502	0.9039	0.6066	0.6027	0.6506	0.7369
	150	4.0044	3.4187	1.2395	1.4477	1.8472	2.2683	0.8891	0.5597	0.5921	0.6483	0.6963
	200	3.8303	3.4145	1.2090	1.2460	1.4398	1.5617	0.8760	0.5489	0.5762	0.6040	0.6217
	500	0.0179	0.0112	0.0081	0.0170	0.0176	0.0171	0.0061	0.0044	0.0042	0.0038	0.0036
0.95	25	4.5039	3.9561	3.0520	1.8006	2.3236	3.0005	0.9703	0.7548	0.6258	0.6971	0.7840
	50	4.1655	3.8989	2.4713	1.6244	2.0000	2.7034	0.9695	0.7485	0.6182	0.6758	0.7560
	100	4.0666	3.8809	1.4129	1.5621	1.9589	2.6450	0.9651	0.5776	0.6161	0.6701	0.7466
	150	4.0547	3.7768	1.3792	1.5208	1.9571	2.3203	0.9406	0.5661	0.5757	0.6301	0.7066
	200	4.0317	3.5604	1.3585	1.5114	1.9229	2.2465	0.9067	0.5525	0.5592	0.6199	0.6848
	500	0.0734	0.0335	0.0169	0.0375	0.0541	0.0643	0.0061	0.0036	0.0060	0.0041	0.0034
0.99	25	5.5687	3.8433	3.7017	1.6010	1.9174	3.0569	1.0205	0.9152	0.6104	0.6565	0.8338
	50	4.9333	3.7302	3.5376	1.4350	1.7870	2.6267	1.0127	0.9133	0.5784	0.6214	0.7447
	100	4.2879	3.5995	3.3014	1.4221	1.7789	2.4104	0.9755	0.8361	0.5725	0.6148	0.7334
	150	4.1000	3.5868	3.2060	1.4197	1.6664	2.4072	0.9561	0.8241	0.5710	0.6019	0.6612
	200	3.9870	3.3639	3.0154	1.2841	1.5114	1.9330	0.9219	0.7374	0.5346	0.5821	0.6558
	500	0.7039	0.2528	0.0122	0.0538	0.1486	0.4296	0.0043	0.0045	0.0075	0.0045	0.0033

Table 3. *Estimated MSE of the estimators when* $p = 6$ *under consider different values of* d.

		$d = 0.10$						$d = 0.50$					$d = 0.99$				
		MLE	TPPE	MAUTPPE				TPPE	MAUTPPE				TPPE	MAUTPPE			
ρ^2	n			$(\hat{d}, \hat{k}_1)$	$(\hat{d}, \hat{k}_2)$	$(\hat{d}, \hat{k}_3)$	$(\hat{d}, \hat{k}_4)$		$(\hat{d}, \hat{k}_1)$	$(\hat{d}, \hat{k}_2)$	$(\hat{d}, \hat{k}_3)$	$(\hat{d}, \hat{k}_4)$		$(\hat{d}, \hat{k}_1)$	$(\hat{d}, \hat{k}_2)$	$(\hat{d}, \hat{k}_3)$	$(\hat{d}, \hat{k}_4)$
0.85	25	4.600	3.771	2.193	1.313	1.731	2.529	3.880	2.193	1.313	1.731	2.529	4.559	2.193	1.313	1.731	2.529
	50	4.205	3.610	1.507	1.262	1.654	2.437	3.790	1.507	1.262	1.654	2.437	4.069	1.507	1.262	1.654	2.437
	100	4.084	3.602	1.419	1.201	1.522	2.393	3.782	1.419	1.201	1.522	2.393	4.029	1.419	1.201	1.522	2.393
	150	4.040	2.993	1.107	1.143	1.318	2.175	3.518	1.107	1.143	1.318	2.175	4.001	1.107	1.143	1.318	2.175
	200	4.006	2.518	1.033	1.121	1.299	1.893	3.510	1.033	1.121	1.299	1.893	3.978	1.033	1.121	1.299	1.893
	500	0.0071	0.0910	0.0071	0.0095	0.0077	0.0071	0.0357	0.0061	0.0069	0.0066	0.0065	0.0357	0.0061	0.0069	0.0066	0.0065
0.90	25	5.135	3.997	2.628	1.339	1.739	3.127	3.882	2.628	1.339	1.739	3.127	4.894	2.628	1.339	1.739	3.127
	50	4.213	3.677	1.936	1.254	1.596	2.671	3.868	1.936	1.254	1.596	2.671	4.805	1.936	1.254	1.596	2.671
	100	4.096	3.451	1.756	1.206	1.466	2.566	3.733	1.756	1.206	1.466	2.566	4.797	1.756	1.206	1.466	2.566
	150	4.044	3.326	1.254	1.193	1.462	2.484	3.713	1.254	1.193	1.462	2.484	4.208	1.254	1.193	1.462	2.484
	200	4.003	2.744	1.065	1.126	1.408	2.165	3.656	1.065	1.126	1.408	2.165	3.997	1.065	1.126	1.408	2.165
	500	0.0141	0.0181	0.0135	0.0133	0.0134	0.0136	0.0139	0.0069	0.0078	0.0077	0.0077	0.0139	0.0069	0.0078	0.0077	0.0077
0.95	25	5.536	3.899	3.401	1.424	2.119	3.292	3.959	3.365	1.420	2.110	3.290	5.246	3.385	1.422	2.115	2.530
	50	4.322	3.834	2.800	1.371	1.836	3.280	3.942	2.784	1.381	1.844	3.279	4.289	2.797	1.376	1.832	2.915
	100	4.137	3.694	2.366	1.316	1.832	2.896	3.932	2.353	1.315	1.829	2.899	4.141	2.378	1.322	1.828	2.904
	150	4.054	3.508	1.577	1.283	1.649	2.893	3.901	1.589	1.278	1.644	2.898	4.063	1.595	1.279	1.650	3.284
	200	4.008	2.985	1.156	1.220	1.600	2.554	3.865	1.157	1.216	1.594	2.521	4.010	1.156	1.216	1.593	3.286
	500	0.0606	0.0491	0.0598	0.0474	0.0541	0.0602	0.0234	0.0071	0.0105	0.0103	0.0103	0.0234	0.0071	0.0105	0.0103	0.0103
0.99	25	9.054	3.917	5.368	1.606	2.109	3.675	5.244	5.331	1.603	2.101	3.669	8.290	5.416	1.607	2.107	3.667
	50	4.993	3.900	3.873	1.525	2.105	3.628	4.195	3.879	1.510	2.083	3.641	4.931	3.882	1.520	2.099	3.633
	100	4.541	3.845	3.836	1.374	2.039	3.606	4.084	3.836	1.370	2.025	3.601	4.498	3.839	1.369	2.029	3.611
	150	4.228	3.767	2.765	1.364	1.904	3.594	4.037	2.743	1.367	1.916	3.558	4.217	2.766	1.364	1.906	3.596
	200	4.082	3.619	2.573	1.241	1.762	3.560	3.976	2.565	1.237	1.754	3.539	4.074	2.572	1.241	1.763	3.556
	500	0.0492	0.0353	0.0186	0.0315	0.0355	0.0391	0.0353	0.0186	0.0315	0.0355	0.0391	0.0353	0.0186	0.0315	0.0355	0.0391

Table 4. *Estimated bias of the estimators when* $p = 6$ *under consider different values of* d.

		$d = 0.10$					$d = 0.50$					$d = 0.99$				
			MAUTPPE					MAUTPPE					MAUTPPE			
ρ^2	n	TPPE	$(\hat{d}, \hat{k}_1)$	$(\hat{d}, \hat{k}_2)$	$(\hat{d}, \hat{k}_3)$	$(\hat{d}, \hat{k}_4)$	TPPE	$(\hat{d}, \hat{k}_1)$	$(\hat{d}, \hat{k}_2)$	$(\hat{d}, \hat{k}_3)$	$(\hat{d}, \hat{k}_4)$	TPPE	$(\hat{d}, \hat{k}_1)$	$(\hat{d}, \hat{k}_2)$	$(\hat{d}, \hat{k}_3)$	$(\hat{d}, \hat{k}_4)$
0.85	25	0.759	0.592	0.459	0.515	0.607	0.782	0.592	0.459	0.515	0.607	0.880	0.592	0.459	0.515	0.607
	50	0.751	0.483	0.454	0.497	0.600	0.778	0.483	0.454	0.497	0.600	0.817	0.483	0.454	0.497	0.600
	100	0.742	0.480	0.446	0.485	0.585	0.774	0.480	0.446	0.485	0.585	0.813	0.480	0.446	0.485	0.585
	150	0.687	0.428	0.442	0.479	0.578	0.772	0.428	0.442	0.479	0.578	0.809	0.428	0.442	0.479	0.578
	200	0.646	0.415	0.440	0.466	0.557	0.765	0.415	0.440	0.466	0.557	0.799	0.415	0.440	0.466	0.557
	500	0.0694	0.0070	0.0095	0.0077	0.0070	0.0310	0.0081	0.0077	0.0073	0.0071	0.0310	0.0081	0.0077	0.0073	0.0071
0.90	25	0.798	0.642	0.463	0.516	0.674	0.789	0.642	0.463	0.516	0.674	0.890	0.642	0.463	0.516	0.674
	50	0.757	0.530	0.462	0.503	0.628	0.782	0.530	0.462	0.503	0.628	0.838	0.530	0.462	0.503	0.628
	100	0.723	0.525	0.457	0.497	0.621	0.780	0.525	0.457	0.497	0.621	0.820	0.525	0.457	0.497	0.621
	150	0.713	0.449	0.431	0.470	0.592	0.778	0.449	0.431	0.470	0.592	0.806	0.449	0.431	0.470	0.592
	200	0.670	0.422	0.429	0.467	0.590	0.758	0.422	0.429	0.467	0.590	0.798	0.422	0.391	0.467	0.590
	500	0.0163	0.0065	0.0068	0.0066	0.0065	0.0055	0.0023	0.0024	0.0023	0.0023	0.0055	0.0023	0.0024	0.0023	0.0023
0.95	25	0.779	0.718	0.479	0.568	0.696	0.794	0.715	0.478	0.567	0.695	0.890	0.717	0.479	0.567	0.696
	50	0.777	0.628	0.475	0.530	0.691	0.794	0.627	0.475	0.531	0.692	0.835	0.628	0.476	0.530	0.692
	100	0.741	0.593	0.471	0.526	0.656	0.789	0.592	0.472	0.526	0.656	0.813	0.595	0.471	0.525	0.658
	150	0.738	0.492	0.452	0.519	0.639	0.787	0.493	0.451	0.518	0.639	0.803	0.494	0.451	0.519	0.640
	200	0.692	0.439	0.436	0.488	0.635	0.773	0.439	0.436	0.487	0.631	0.799	0.439	0.436	0.487	0.632
	500	0.0360	0.0068	0.0105	0.0083	0.0068	0.0183	0.0073	0.0066	0.0061	0.0059	0.0183	0.0073	0.0066	0.0061	0.0059
0.99	25	0.783	0.844	0.510	0.573	0.753	0.811	0.841	0.509	0.572	0.749	0.880	0.846	0.510	0.573	0.755
	50	0.778	0.773	0.487	0.559	0.741	0.797	0.774	0.485	0.556	0.740	0.827	0.775	0.486	0.558	0.742
	100	0.763	0.761	0.463	0.550	0.741	0.794	0.761	0.462	0.548	0.740	0.809	0.762	0.463	0.548	0.740
	150	0.763	0.634	0.461	0.530	0.734	0.788	0.631	0.462	0.531	0.735	0.802	0.634	0.461	0.530	0.735
	200	0.745	0.619	0.453	0.525	0.729	0.782	0.618	0.453	0.524	0.729	0.798	0.619	0.453	0.525	0.729
	500	0.0279	0.0055	0.0075	0.0063	0.0057	0.0279	0.0055	0.0075	0.0063	0.0057	0.0279	0.0055	0.0075	0.0063	0.0057

Table 5. *Estimated MSE and bias of the estimators when* $p = 9$.

		Estimated MSE						Estimated Bias				
		MLE	TPPE	MAUTPPE				TPPE	MAUTPPE			
ρ^2	n			$(\hat{d}, \hat{k}_1)$	$(\hat{d}, \hat{k}_2)$	$(\hat{d}, \hat{k}_3)$	$(\hat{d}, \hat{k}_4)$		$(\hat{d}, \hat{k}_1)$	$(\hat{d}, \hat{k}_2)$	$(\hat{d}, \hat{k}_3)$	$(\hat{d}, \hat{k}_4)$
0.85	25	4.9867	3.9254	3.3671	1.2180	1.6399	3.7194	0.6837	0.6095	0.3820	0.4310	0.6537
	50	4.4573	3.8356	1.6960	1.1597	1.5815	3.5295	0.6735	0.4555	0.3808	0.4290	0.6320
	100	4.1008	3.5133	1.5471	1.1421	1.5178	3.1586	0.6383	0.4285	0.3775	0.4221	0.5877
	150	4.0432	3.3801	1.3994	1.1229	1.5028	2.7216	0.6138	0.4080	0.3724	0.4204	0.5519
	200	4.0187	2.7509	1.3574	1.0992	1.3052	2.1693	0.5777	0.4034	0.3688	0.4087	0.5087
	500	0.0157	0.0130	0.0121	0.0159	0.0160	0.0145	0.0189	0.0151	0.0154	0.0153	0.0147
0.90	25	5.1892	3.8377	3.6209	1.2633	1.7033	3.4648	0.6238	0.6320	0.3792	0.4333	0.6107
	50	4.2100	3.7621	2.6786	1.2155	1.6710	3.4088	0.6521	0.5576	0.3870	0.4411	0.6248
	100	4.1743	3.6540	1.7841	1.2143	1.6565	3.3917	0.6500	0.4694	0.4043	0.4586	0.5801
	150	4.0396	3.5072	1.6078	1.1832	1.6184	3.0263	0.6710	0.4380	0.3754	0.4338	0.6278
	200	4.0243	3.4173	1.1155	1.1415	1.5880	2.8179	0.6613	0.3694	0.3824	0.4362	0.5802
	500	0.0240	0.0163	0.0087	0.0234	0.0238	0.0208	0.0051	0.0052	0.0054	0.0054	0.0053
0.95	25	5.9292	3.9792	4.1619	1.3290	2.2126	3.8283	0.6920	0.6726	0.4023	0.5082	0.6664
	50	4.1767	3.8952	2.3592	1.3283	1.9382	3.6988	0.6786	0.5121	0.3945	0.4637	0.6503
	100	4.1462	3.7097	2.1484	1.2887	1.8386	3.6626	0.6576	0.5014	0.3942	0.4592	0.6375
	150	4.0412	3.7050	2.1469	1.1858	1.6857	3.4699	0.6423	0.4952	0.3781	0.4410	0.6182
	200	4.0278	3.6487	2.1459	1.1145	1.4582	3.2969	0.6413	0.4951	0.3671	0.4106	0.6094
	500	0.1204	0.0599	0.0185	0.0828	0.0960	0.0808	0.0203	0.0164	0.0252	0.0200	0.0143
0.99	25	78.7139	78.7125	9.3212	1.9733	2.7194	78.4760	0.9057	0.6977	0.4703	0.5470	0.9037
	50	11.8039	4.0722	5.4770	1.3071	1.9804	3.9385	0.6820	0.6891	0.4045	0.4825	0.6748
	100	5.6412	3.9143	4.3348	1.2536	1.9150	3.8780	0.6812	0.6647	0.3895	0.4693	0.6699
	150	4.3063	3.9019	3.8198	1.1253	1.5446	3.8354	0.6689	0.6526	0.3716	0.4268	0.6673
	200	4.0859	3.8468	3.6952	1.0334	1.5223	3.7788	0.6641	0.0491	0.3449	0.2861	0.6609
	500	0.9683	0.4018	0.0077	0.1194	0.2088	0.1181	0.0067	0.0058	0.0080	0.0063	0.0052

Table 6. *Estimated Coefficients and SMSE of the MLE, TPPE and MAUTPPE.*

Estimators	MLE	TPPE	MAUTPPE				λ_j[1]
			$(\hat{d},\hat{k}_1)$	$(\hat{d},\hat{k}_2)$	$(\hat{d},\hat{k}_3)$	$(\hat{d},\hat{k}_4)$	
Intercept	2.240	2.254	2.291	2.295	2.290	2.288	2010
Unemployment	0.075	0.067	0.051	0.046	0.045	0.045	223.2
Cars	-9.559	-6.799	-0.062	-0.282	-0.647	-0.875	186.54
Trucks	4.018	3.123	0.669	1.072	1.446	1.566	14.157
15-24 years	1.971	1.460	0.042	0.209	0.464	0.560	0.581
25-64 years	2.004	1.424	-0.030	0.101	0.202	0.242	0.224
> 64 years	1.979	1.185	-0.280	-0.755	-1.113	-1.138	0.047
MSE	27.749	14.969	8.699	7.520	6.390	6.059	

[1] $\lambda_j(\lambda_1 > \lambda_2 > \lambda_3 > \lambda_4 > \lambda_5 > \lambda_6 > \lambda_7)$ are the eigenvalues and Condition Index = $\sqrt{\frac{\lambda_{max}}{\lambda_{min}}} = 207.77$

Increasing the degree of correlation has an adverse effect on the estimators in terms of MSE. However, the estimated bias of the estimators are decreasing when the degrees of correlation is increased particular especially for MAUTPPE with $\hat{k}_2$ and $\hat{k}_3$. When the sample size increases the estimated MSE and bias are decreased. The sample size makes a good effect on the estimators in sense of large sample size. An increase in the number of explanatory variables has a negative effect in sense of estimated MSE and positive effect in sense of estimated bias for some cases. The estimated bias of the TPPE is reduced when the number of explanatory variables are increased. However, the proposed MAUTPPE has lowest bias in all cases than the TPPE. It is also noted that the estimated bias of all the estimators are reduced when the $p = 6$ and then slight increase in the estimated bias when $p = 9$ and $\rho^2 = 0.99$ for only TPPE and MAUTPPE $(\hat{d},\hat{k}_4)$. The performance of MAUTPPE $(\hat{d},\hat{k}_2)$ is significant in terms of estimated MSE and MAUTPPE $(\hat{d},\hat{k}_1)$is almost unbiased when the $\rho^2 = 0.99, n = 200$ and $p = 9$.

In addition, we analyzed the performance of TPPE and MAUTPPE by assuming different initial value of d which are 0.10, 0.50 and 0.99 (e.g. see for more details, Asar et al., 2017). These results are illustrated in Tables 3-4. The performance of TPPE and MAUTPPE do not change substantially when we consider the different initial values of d and one can see this findings in Table 3 and Table 4. The estimated MSE and bias values of the MAUTPPE are approximately same when the $\rho^2 = 0.85$ and $\rho^2 = 0.90$. One can see the insignificant change in the estimated MSE and bias of the MAUTPPE when the $\rho^2 = 0.95$ and $\rho^2 = 0.99$. Meanwhile, the estimated MSE and bias values of the TPPE are increased when the d rises. The performance of TPPE is near to MLE when the $d = 0.99$, $\rho^2 = 0.99$ and $n = 200$. For large sample size ($n = 500$), the bias of

MAUTPPE is close to zero which indicate the benefit of the proposed estimator in the sense of bias correction. Simulation results demonstrate that a bias correction estimator (MAUTPPE) substantially reduces the bias and more efficient than TPPE as well as improved estimators under certain conditions.

We can conclude that the performance of MLE is worsted in almost all condition. The MLE is not good choice in the presence of multicollinearity. The proposed MAUTPPE has quite good performance as compared to the TPPE and MLE under different conditions. However, the MAUTPPE with $(\hat{d}, \hat{k}_2)$ has better performance than the other estimators in almost all conditions.

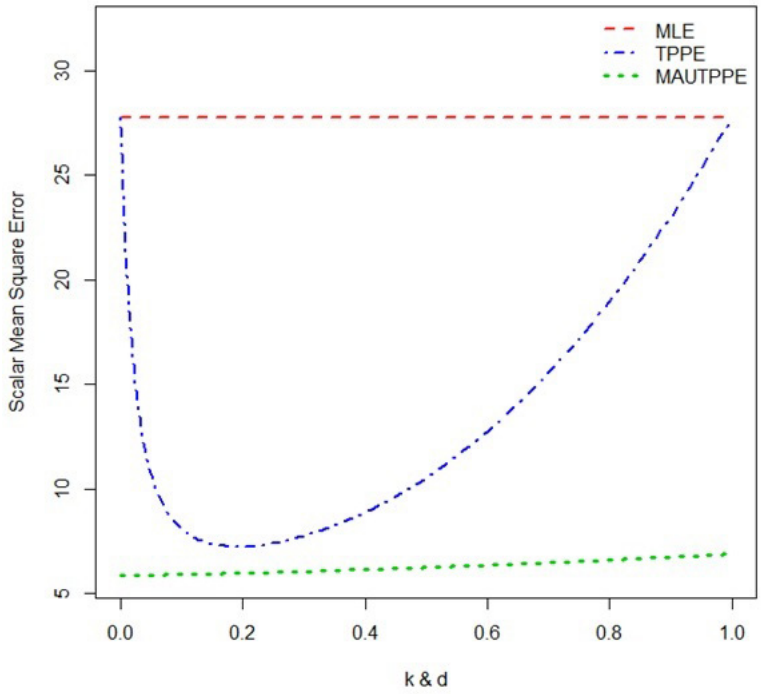


Figure 1. *Emperical estimated SMSE of MLE, TPPE and MAUTPPE.*

5. Application

To illustrate the findings of the paper, Swedish traffic fatality data for the year 2019 are analyzed in this section. The data are taken from the Statistics Sweden and Swedish transport administration. The aim of this case study is to see the impact of external factors on the traffic fatalities in Sweden, where the number of traffic fatalities is considered as dependent variable. As discussed by Wiklund, Simonsson and Forsman (2012), the

main factors are economic conditions defined as unemployment rate, traffic exposure that we measure as number of vehicles (cars and trucks), and demographic variables, all of which are considered as explanatory variables. By following the study of Stipdonk et al., (2013), we divide all individuals into three different age groups (age 15-24 years, age 25-64 years and more than 64 years). The estimated results of the model are presented in Table 6. The eigenvalues of $X^{\top}X$ matrix are 2010, 223.2, 186.54, 14.157, 0.581, 0.224 and 0.047. The condition index, $CI = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} = 207.77$, which confirmed that there are serious problems of multicollinearity. Therefore, we used TPPE and MAUTPPE to combat the multicollinearity problem. The unemployment rate coefficient is positive which shows that the number of fatalities increases and this impact is considerable low for MAUTPPE $(\hat{d},\hat{k}_4)$. The traffic exposure variables (cars and trucks) have negative and positive coefficients. This shows that more accidents occur when trucks are used and less accidents occur when cars are used. Age group 15-24 and 25-64 year's parameters are positive except MAUTPPE $(\hat{d},\hat{k}_1)$. Age group more than 64 years is positive when we use MLE and TPPE but it is negative for MAUTPPE which shows the robust results. The number of fatalities decreases when the drivers have more experience and this result can be seen only by using proposed estimator (MAUTPPE). The SMSE of MLE is inflated due to multicollinearity problem and biased estimation methods (TPPE and MAUTPPE) have lower SMSE than the MLE. One can see that a substantial decrease of the SMSE when applying MAUTPPE than the MLE and TPPE. Figure 1 illustrates the empirical SMSE of MLE, TPPE and MAUTPPE. The SMSE of MAUTPPE is smaller than the MLE and TPPE. In summary, the application shows the benefits of the proposed estimator. Program code in R for analyzing this application data set is given in Supplementary Material.

6. Some concluding Remarks

This paper proposes a new almost unbiased estimator for the parameters of the Poisson regression model. The MSE properties of the proposed estimator is investigated and a comparison is made with some existing estimators. Furthermore, a simulation study has been conducted to compare the performance of the estimators under several parametric conditions. Finally, an example illustrates the benefit of the new MAUTPPE. The overall results of the paper show the benefit of the new estimator as compared to previously suggested estimators such as TPPE and MLE. Based on both the simulation study and empirical application, we may recommend MAUTPPE with parameter combinations $(\hat{d},\hat{k}_2)$ and $(\hat{d},\hat{k}_4)$ to researchers.

References

Amin, M., Qasim, M. and Amanullah, M. (2019). Performance of Asar and Genç and Huang and Yang's Two-Parameter Estimation Methods for the Gamma Regression Model. *Iranian Journal of Science and Technology, Transactions A: Science*, 43, 2951-2963.

Amin, M., Qasim, M., Amanullah, M. and Afzal, S. (2020a). Performance of some ridge estimators for the gamma regression model. *Statistical papers*, 61, 997-1026.

Amin, M., Qasim, M., Yasin, A. and Amanullah, M. (2020b). Almost unbiased ridge estimator in the gamma regression model. *Communications in Statistics-Simulation and Computation*, 1-21.

Asar, Y. and Genç, A. (2018). A new two-parameter estimator for the Poisson regression model. *Iranian Journal of Science and Technology, Transactions A: Science*, 42, 793-803.

Asar, Y., Erişoğlu, M. and Arashi, M. (2017). Developing a restricted two-parameter Liu-type estimator: A comparison of restricted estimators in the binary logistic regression model. *Communications in Statistics-Theory and Methods*, 46, 6864-6873.

Chiou, Y. C. and Fu, C. (2013). Modeling crash frequency and severity using multinomial-generalized Poisson model with error components. *Accident Analysis and Prevention*, 50, 73-82.

Donaldson, L. H., Brooke, K. and Faux, S. G. (2009). Orthopaedic trauma from road crashes: is enough being done?. *Australian Health Review*, 33, 72-83.

Farebrother, R. W. (1976). Further results on the mean square error of ridge regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38, 248-250.

Hoerl, A. E. and Kennard, R. W. (1970a). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12, 55-67.

Hoerl, A. E. and Kennard, R. W. (1970b). Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12, 69-82.

Huang, J. and Yang, H. (2014). A two-parameter estimator in the negative binomial regression model. *Journal of Statistical Computation and Simulation*, 84, 124-134.

Ivan, J. N., Wang, C. and Bernardo, N. R. (2000). Explaining two-lane highway crash rates using land use and hourly exposure. *Accident Analysis and Prevention*, 32, 787-795.

Kandemir Çetinkaya, M. and Kaçıranlar, S. (2019). Improved two-parameter estimators for the negative binomial and Poisson regression models. *Journal of Statistical Computation and Simulation*, 89, 2645-2660.

Karlsson, P., Månsson, K. and Kibria, B. M. G. (2020). A Liu estimator for the beta regression model and its application to chemical data. *Journal of Chemometrics*, 34, e3300.

Kibria, B. M G., Månsson, K. and Shukur, G. (2013). Some ridge regression estimators for the zero-inflated Poisson model. *Journal of Applied Statistics*, 40, 721-735.

Kibria, B. M. G. (2003). Performance of some new ridge regression estimators. *Communications in Statistics-Simulation and Computation*, 32, 419-435.

Kibria, B. M. G., Månsson, K. and Shukur, G. (2015). A simulation study of some biasing parameters for the ridge type estimation of Poisson regression. *Communications in Statistics-Simulation and Computation*, 44, 943-957.

Kurtoğlu, F. and Ozkale, M. R. (2016). Liu estimation in generalized linear models: application on gamma distributed response variable. *Statistical Papers*, 57, 911-928.

Lord, D., Manar, A. and Vizioli, A. (2005b). Modeling crash-flow-density and crash-flow-V/C ratio relationships for rural and urban freeway segments. *Accident Analysis and Prevention*, 37, 185-199.

Lukman, A. F., Ayinde, K., Kibria, B. M. G. and Adewuyi, E. T. (2020). Modified ridge-type estimator for the gamma regression model. *Communications in Statistics-Simulation and Computation*, 1-15.

Lyon, C., Oh, J., Persaud, B., Washington, S. and Bared, J. (2003). Empirical investigation of interactive highway safety design model accident prediction algorithm: Rural intersections. *Transportation research record*, 1840, 78-86.

Månsson, K. (2012). On ridge estimators for the negative binomial regression model. *Economic Modelling*, 29, 178-184.

Månsson, K. (2013). Developing a Liu estimator for the negative binomial regression model: method and application. *Journal of Statistical Computation and Simulation*, 83, 1773-1780.

Månsson, K., and Shukur, G. (2011). A Poisson ridge regression estimator. *Economic Modelling*, 28, 1475-1481.

Månsson, K., Kibria, BMG., Sjolander, P. and Shukur, G. (2012). Improved Liu estimators for the Poisson regression model. *International Journal of Statistics and Probability*, 1, 2.

McDonald, G. C. and Galarneau, D. I. (1975). A Monte Carlo evaluation of some ridge-type estimators. *Journal of the American Statistical Association*, 70, 407-416.

Qasim, M., Amin, M. and Amanullah, M. (2018). On the performance of some new Liu parameters for the gamma regression model. *Journal of Statistical Computation and Simulation*, 88, 3065-3080.

Qasim, M., Kibria, B. M. G., Månsson, K. and Sjolander, P. (2020a). A new Poisson Liu regression estimator: method and application. *Journal of Applied Statistics*, 47, 2258-2271.

Qasim, M., Månsson, K., and Kibria, B. M. G. (2021). On some beta ridge regression estimators: method, simulation and application. *Journal of Statistical Computation and Simulation*, 1-14.

Qasim, M., Månsson, K., Amin, M., Kibria, B. M. G., and Sjolander, P. (2020b). Biased adjusted Poisson ridge estimators-method and application. *Iranian Journal of Science and Technology, Transactions A: Science*, 44, 1775-1789.

Schaefer, R. L., Roi, L. D. and Wolfe, R. A. (1984). A ridge logistic estimator. *Communications in Statistics-Theory and Methods*, 13, 99-113.

Shi, Q., Abdel-Aty, M. and Lee, J. (2016). A Bayesian ridge regression analysis of congestion's impact on urban expressway safety. *Accident Analysis and Prevention*, 88, 124-137.

Stein, C. (1956). Inadmissibility of usual estimator for the mean of a multivariate Normal distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, University of California Press*. 197-206.

Stipdonk, H., Bijleveld, F., Van Norden, Y. and Commandeur, J. (2013). Analysing the development of road safety using demographic data. *Accident Analysis and Prevention*, 60, 435-444.

Toker S., Ustundağ Şiray G. and Qasim M. (2019). Developing a first order two parameter estimator for generalized linear model. *11th International Statistics Congress, (ISC 2019)*; Muğla, Turkey.

Wiklund, M., Simonsson, L. and Forsman, A. (2012). *Traffic safety and economic fluctuation: long-term and short-term analyses and a literature survey*.

World Health Organization. (2015). *World report on ageing and health. World Health Organization*.

SORT 45 (2) July-December 2021, 143-162 DOI: 10.2436/20.8080.02.113

Bayesian hierarchical nonlinear modelling of intra-abdominal volume during pneumoperitoneum for laparoscopic surgery

Gabriel Calvo[1], Carmen Armero[1], Virgilio Gómez-Rubio[2] and Guido Mazzinari[3]

Abstract

Laparoscopy is an operation carried out in the abdomen through small incisions with visual control by a camera. This technique needs the abdomen to be insufflated with carbon dioxide to obtain a working space for surgical instruments' manipulation. Identifying the critical point at which insufflation should be limited is crucial to maximizing surgical working space and minimizing injurious effects. A Bayesian nonlinear growth mixed-effects model for the relationship between the insufflation pressure and the intra–abdominal volume generated is discussed as well as its plausibility to represent the data.

MSC: 62P10, 62F25.

Keywords: Intra-abdominal pressure, logistic growth function, Markov chain, Monte Carlo methods, random effects.

1. Introduction

Laparoscopy is an operation carried out in the abdomen or pelvis through small incisions with the help of a camera. It is performed by insufflating CO_2 into the abdomen

[1] Department of Statistics and Operations Research, Universitat de València, Carrer Doctor Moliner 50, 46100, Burjassot, Spain.

[2] Departamento de Matemáticas, Escuela Técnica Superior de Ingenieros Industriales, Universidad de Castilla-La Mancha, Avda, de España s/n, 02071, Albacete, Spain.

[3] Research Group in Perioperative Medicine and Department of Anaesthesiology, Hospital Universitari i Politècnic la Fe, Avinguda de Fernando Abril Martorell 106, 46026, València, Spain.

Received: June 2021.
Accepted: November 2021.

that yields a working space, i.e., pneumoperitoneum, and passing surgical instruments through small incisions using a camera to have external visual control of the procedure (Neugebauer et. al., 2010). Laparoscopy has been gaining ground since its inception because it is associated with less morbidity than the traditional method performed through a single, larger skin incision (Pache et al., 2017).

The introduction of CO_2 into the abdomen is operated by medical devices, i.e., laparoscopic insufflators, through small plastic tubes, i.e. trocars, inserted in the patient's abdominal wall. Laparoscopy technological development has been limited to improvements in camera image quality, whereas little innovation has been made in insufflation devices (Colon Cancer Laparoscopic or Open Resection Study Group, 2009).

The CO_2 insufflation pressure, i.e., intra–abdominal pressure (*IAP*), is set manually on the insufflator by the surgical team. *IAP* is measured in millimeters of mercury (mmHg), and the usual figures during laparoscopic surgery range between 12 and 15 mmHg. Although international guidelines recommend working with the lowest *IAP* value that ensures an adequate working space, the standard practice is still to initially set the *IAP* value without further adjustments regardless of the amount of generated intra–abdominal volume (*IAV*) (Neudecker et al., 2002), measured in litres (L). Operating at such high *IAP* increases perioperative morbidity since it leads to decrease abdominal blood perfusion, greater postoperative pain, peritoneal injury, and increased risk of pulmonary complications.

The abdominal compartment shows an anisotropic behaviour during pneumoperitoneum which is explained by its combination of rigid borders, e.g., spine, rib cage, and pelvis, and semirigid borders, e.g., abdominal wall muscles and the diaphragm (Becker et al., 2017). Initially, marginal gains in volume in response to pressure increments are proportional. In other words, the abdominal compliance (C_{abd}), which defines the change in volume determined by a change in pressure, follows an approximately linear relationship (Mulier et al., 2009). According to biomechanics laws, the yield stress point is eventually reached, after which applying additional pressure leads to diminishing gains in volume (Forstemann et al., 2011). Identifying this critical point at which insufflation should be limited is crucial to maximizing surgical working space while minimizing injurious *IAP* effects.

The abdomen pressure–volume dynamics during pneumoperitoneum has been discussed in previous papers (Diaz-Cambronero et al., 2019, 2020; Mazzinari et al., 2020, 2021). These studies suggest the adequacy of an increasing sigmoidal model for describing the relationship between both variables. Our aim in this work is twofold. On the one hand, we want to estimate such a model to gain knowledge about the relationship between *IAP* and *IAV*, especially about the parameters that determine the different growth stages of the process in accordance with the specific characteristics of the individuals in the target population. On the other hand, the second goal of the paper is to discuss the quality of the fit of the model to the data. This is a relevant question since the logistic growth curve is a previously used model for the topic. The hypothesis is that, in a personalised medicine environment, patient responses to insufflation can be estimated and

predicted so that an ideal *IAP* value could be determined to optimise *IAV* with the lowest risks of potential negative effects.

The statistical framework of this study is that of nonlinear growth mixed-effects models, also known as hierarchical nonlinear growth models. They have a long and important scientific tradition for describing biological, medical, and environmental growth phenomena such as pharmacokinetics (Giltinan, 2006), epidemiology (Lindsey, 2001), physiological-response processes (Peek et al., 2002), or forestry (Fang and Bailey, 2001) among others. One of the major appeals of these models is that their parameters contain direct and intuitive information on the process under study. This fact generates a multifaceted knowledge about the phenomena in question of great scientific value (Davidian, 2008).

Data for the study come from a repeated measures design (Lindstrom and Bates, 1990). In our case, the variable of interest *IAV* is measured for each individual with regard to different *IAP* values. This design generates two types of data: data from the same individual and data from several individuals. Random effects in these models are essential elements to glue together the different observations of the same individual as they could be considered as a within-individual variation (Laird and Ware, 1982).

The statistical analysis of the problem has been carried out using Bayesian inference. This statistical methodology accounts for uncertainty in terms of probability distributions (Loredo, 1989, 1992) and uses Bayes' theorem to update all relevant information. The Bayesian approach simplifies the implementation and interpretation of mixed effects models. The conditional formulation of this type of model, which explicitly includes random effects in the conditional mean, allows individual and population inferences to be made. This is due to the simplicity process of integrating out the random effects of the model, that is, moving from the conditional formulation of the model to its marginal formulation (Lee and Nelder, 2004). This feature of Bayesian models is of utmost importance in the case of growth models because it expresses in a natural probabilistic way all information about the parameters and other relevant features of the growth process through the respective posterior distribution. Furthermore, model checking can be conducted in a straightforward way to detect possible systematic bias in the model. This is particularly important for medical applications to avoid patients from receiving a suboptimal medical treatment.

The paper is organised as follows. Section 2 presents the data with a brief description that emphasises the particular features of the repeated-means design through the number of observations per individual and their *IAV* trajectories according to the *IAP* values. Section 3 introduces and formulates the statistical modelling. Section 4 accounts for posterior inferences and prediction. Section 4.1 discusses the posterior distribution of the estimation process. Sections 4.2 and 4.3 contain, respectively, some relevant results of clinical interest at specific individual levels and in general terms for different population groups. Section 5 deals with model checking by means of the cross-validated predictive density. The paper ends with an overview of the results and some conclusions.

2. Intra-abdominal volume and intra-abdominal pressure data

The data for the current modelling come from a previously published individual patient meta–analysis (Mazzinari et al., 2021) that included experimental information from three previous homogeneous clinical studies (Mazzinari et al., 2020; Diaz-Cambronero et al., 2019, 2020). All patients in these studies underwent a standardized pneumoperitoneum insufflation at a constant low flow, i.e., 3 $\mathrm{L\,min^{-1}}$, under deep neuromuscular block with a posttetanic count (PTC) between one and five assessed by quantitative monitoring. The insufflation was carried out through a leakproof trocar up to an IAP of 15mmHg for abdominal wall prestretching and then stepwise changes in IAP in the 8 to 15 mmHg pressure range were recorded. In all studies, patients' legs were placed in padded leg-holder supports with hips flexed before the initial insufflation.

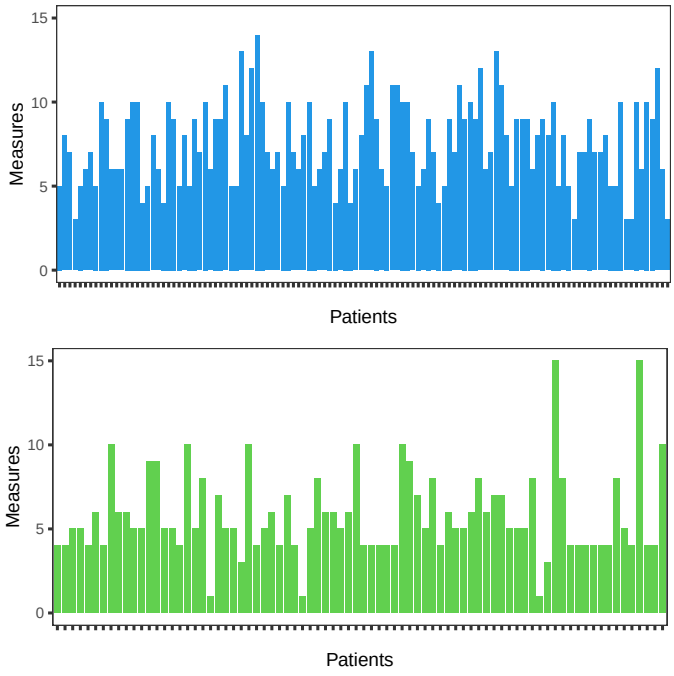


Figure 1. *Number of repeated measures in the men's group (top panel) and in the women's group (bottom panel). Each bar corresponds to a person and its ordinate is the number of measurements of that person during the study. Patients are ordered according to their age from youngest to oldest.*

The original databank had information on 204 patients, but 6 patients presented missing information on IAP, IAV, and/or age values. There are very few individuals whose missing observations do not appear to have been generated by non-ignorable mechanisms. For this reason, we decided to eliminate them directly and not engage in a very unhelpful imputation process. The final databank has 198 patients, 118 men and 80

women, and a total of 1 361 observations. We have a repeated measures design with a very different number of observations per individual: from individuals with only one observation to individuals with 15. Figure 1 shows the number of repeated measures for the group of men and women in order of age. It is interesting to note that women have in general less measurements than men in all ages.

The data have a very wide age range. The youngest patient is 23 years old and the oldest is 92, with a mean age of 64.65 years. In the men's group, the minimum and maximum also are 23 and 92, respectively, and their average is 64.49 years. Women have a minimum age of 34 and a maximum of 85, and their mean is 64.87 years.

IAP values range between 0 and 16 mmHg, and *IAV* values between 0.5 and 13 L. Figure 2 shows a spaghetti plot of *IAV* for men and women. They all show a fairly similar pattern of the *IAV* with *IAP*, although a greater range of values is observed in men, especially in large values of *IAP*. In both groups there are individuals with different behaviour but men behave more homogeneously than women.

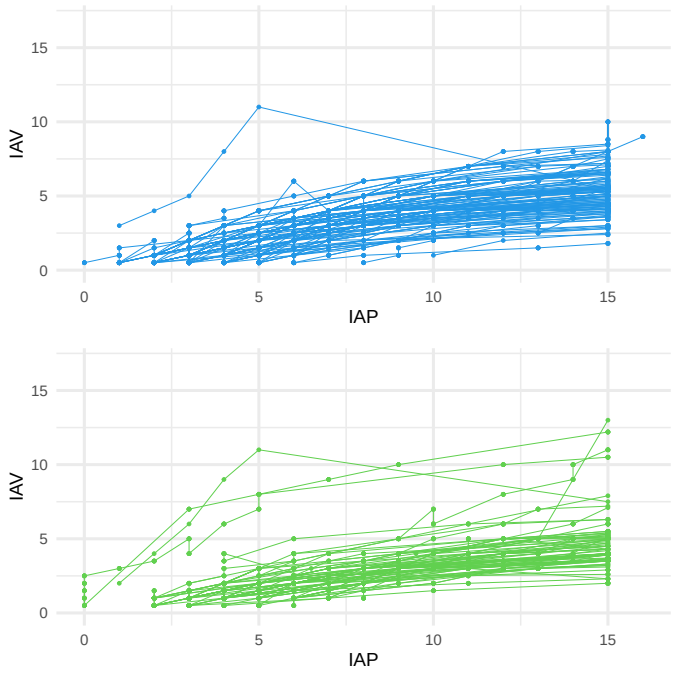


Figure 2. *IAV profiles (in L) according to IAP (in mmHg) for men (top panel) and women (bottom panel) in the sample.*

3. Logistic growth mixed-effects modelling

Let the nonlinear mixed-effects model for the response random variable IAV_{ij} that records the intra-abdominal volume value for individual i, $i = 1,\ldots,n$ with standardized intra-abdominal pressure value IAP_{ij}, $j = 1,\ldots,J_i$, defined in terms of a conditional normal

distribution as follows

$$(IAV_{ij} \mid \mu_{ij}, \sigma^2) \sim \mathrm{N}(\mu_{ij}, \sigma^2), \tag{1}$$

where μ_{ij} is the mean of the *IAV* value of a patient with IAP_{ij} value and can be expressed in terms of the conditional logistic growth function

$$(\mu_{ij} \mid a_i, b_i, c_i, IAP_{ij}) = \frac{a_i}{1+\exp\{-(b_i + c_i IAP_{ij})\}}, \tag{2}$$

with parameters a_i, b_i, and c_i determining the growth of the function, and σ^2 the unknown variance associated to the random measurement error of the normal (1).

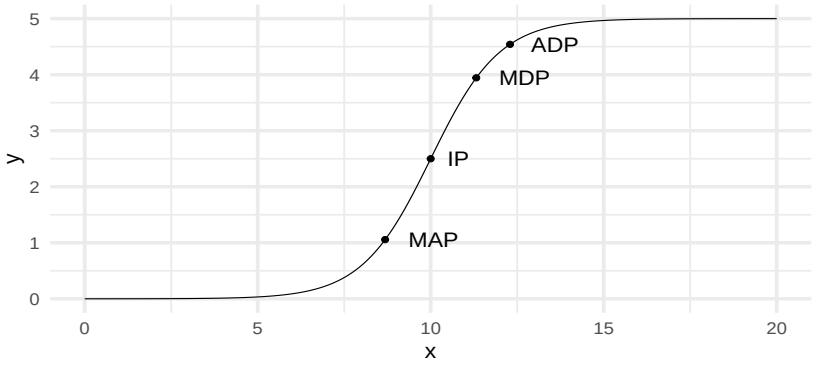


Figure 3. *Graphics of the logistic growth function* $5/[1+exp\{-(-10+x)\}]$*, the subsequent asymptotic value, and its MAP, IP, ADP, and MDP points.*

The logistic growth model for μ_{ij} has important features which are very valuable to better understand the relationship between *IAP* and *IAV* (Davidian, 2008):

- It is an increasing sigmoid function (see Figure 3), or *S*-curve, whose name comes from its shape and was introduced by the mathematician Pierre-François Verhulst in the 19th century to study the growth of populations in autocatalytic chemical reactions (Cramer, 2004).
- The asymptotic value of *IAV* when *IAP* goes to infinity is a_i.
- The inflection point (*IP*), where the curve changes from being concave downward to concave upward and therefore it is the point at which the acceleration of the process switches from positive to negative, is $-b_i/c_i$ for *IAP*. The value of *IAV* at this point is $a_i/2$.
- The maximum acceleration and deceleration point, *MAP* and *MDP* respectively, have the following *IAP* and *IAV* coordinates, $((-(\ln(2+\sqrt{3})+b_i)/c_i, a_i/(3-\sqrt{3}))$ and $(-(-\ln(2-\sqrt{3})+b_i)/c_i, a_i/(3+\sqrt{3}))$.

- The asymptotic deceleration point (*ADP*) is calculated by equalling the fourth derivative to 0. It is located after the maximum deceleration point, and it indicates the point in which the acceleration is negative but close to 0. Therefore, it is expected that the increase of the function is not of much practical interest. The *ADP* is $-(\ln(5-2\sqrt{6})+b_i)/c_i$ for *IAP*. The value of *IAV* at this point is $a_i(3+\sqrt{6})/6$.

By way of illustration, Figure 3 shows the graph of the logistic growth model $y = 5/[1+\exp\{-(-10+x)\}]$ with generic variables x and y, and the location on the graph of the special points described above.

Hierarchical modelling for parameters a_i and b_i was based on expert information and connected them with standardized age and gender covariates. Parameter c_i was associated to covariate gender. We discarded its connection to covariate age as a consequence of a previous analysis of variable selection that we will discuss later. Furthermore, a_i and b_i also included a random effect specifically associated to each individual that allow to connect all their repeated observations. We have not included any random effect in the modelling of the parameter c_i because it would generate a random interaction term with the *IAP* values that would be difficult to understand and justify. Following this reasoning, our model would be

$$a_i = \beta_0^{(a)} + u_i^{(a)} + \beta_W^{(a)} I_W(i) + \beta_A^{(a)} Age_i, \tag{3}$$

$$b_i = \beta_0^{(b)} + u_i^{(b)} + \beta_W^{(b)} I_W(i) + \beta_A^{(b)} Age_i, \tag{4}$$

$$c_i = \beta_0^{(c)} + \beta_W^{(c)} I_W(i), \tag{5}$$

where $\boldsymbol{\beta}_0 = (\beta_0^{(a)}, \beta_0^{(b)}, \beta_0^{(c)})^\intercal$ stands for the common intercept with the men group being the reference group, $I_W(i)$ is the indicator variable with value 1 if individual i is a woman and 0 otherwise, $\boldsymbol{\beta}_W = (\beta_W^{(a)}, \beta_W^{(b)}, \beta_W^{(c)})^\intercal$ and $\boldsymbol{\beta}_A = (\beta_A^{(a)}, \beta_A^{(b)})^\intercal$ are the vector of regression coefficients associated with individual i being a woman and their standardized age, respectively. Random effects $u_i^{(a)}$ and $u_i^{(b)}$, $i = 1, \ldots, n$, are assumed conditional independent given σ_a^2 and σ_b^2 and normally distributed according to $(u_i^{(a)}|\sigma_a^2) \sim \mathrm{N}(0, \sigma_a^2)$ and $(u_i^{(b)}|\sigma_b^2) \sim \mathrm{N}(0, \sigma_b^2)$.

The Bayesian model is completed with the elicitation of a prior distribution for the parameters and hyperparameters $\boldsymbol{\theta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_W, \boldsymbol{\beta}_A, \sigma, \sigma_a, \sigma_b)^\intercal$ of the model. We assume prior independence between them and select the uniform distribution $\mathrm{U}(0,10)$ for all standard deviation terms. The elicited marginal prior distributions for $\beta_0^{(a)}$ and $\beta_0^{(c)}$ are $\mathrm{U}(0,20)$ and $\mathrm{U}(0,10)$, respectively. These uniform distributions are sufficiently large to cover generously the whole range of possible values of both parameters. A normal distribution $\mathrm{N}(0, 10^2)$ is selected for $\beta_0^{(b)}$, $\beta_W^{(a)}$, $\beta_W^{(b)}$, $\beta_W^{(c)}$, $\beta_A^{(a)}$, and $\beta_A^{(b)}$ to allow the parameters to move freely between a wide range of positive and negative values.

4. Posterior inferences and predictions

4.1. Posterior distribution

The relevant quantities in the inferential process are the parametric vector $\boldsymbol{\theta}$ and the set of random effects associated to the individuals in the sample $\boldsymbol{u} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)^\top$, where $\boldsymbol{u}_i = (u_i^{(a)}, u_i^{(b)})$. The posterior distribution $\pi(\boldsymbol{\theta}, \boldsymbol{u} \mid \mathcal{D})$, where $\mathcal{D}$ represents the observed *IAV* and *IAP* data of all individuals in the sample as well as their age and gender, contains all the relevant information of the problem and it is usually the starting point of all relevant inferences. It was approximated by means of Markov Chain Monte Carlo (MCMC) simulation methods through the JAGS software (Plummer, 2003). For the estimated model, we ran three parallel chains with 1,000,000 iterations and a burn-in of 500,000. Chains were also thinned by storing every 1,000th iteration to reduce autocorrelation in the sample. Convergence to the joint posterior distribution was guaranteed by visualising every autocorrelation function plot by means of `mcmcplot` package for the R software and assuring an effective number of independent simulation draws greater than 100. For the sake of reproducibility we have generated a fictitious databank, which together with the R code for the analyses is available as supplementary material here https://github.com/gcalvobayarri/intra_abdominal_volume_model.git.

Table 1. *Posterior summaries (mean, standard deviation and* 95% *credible interval) for the parameters and hyperparameters of the logistic growth model with covariates gender and standardized age.*

	Logistic growth model		
Parameters	mean	sd	$CI_{0.95}$
$\beta_0^{(a)}$	5.729	0.377	(4.968, 6.452)
$\beta_W^{(a)}$	−0.418	0.259	(−0.927, 0.095)
$\beta_A^{(a)}$	0.101	0.124	(−0.145, 0.349)
σ_a	1.670	0.090	(1.501, 1.860)
$\beta_0^{(b)}$	1.080	0.181	(0.730, 1.440)
$\beta_W^{(b)}$	−0.270	0.125	(−0.517, −0.028)
$\beta_A^{(b)}$	0.134	0.054	(0.026, 0.241)
σ_b	0.650	0.041	(0.572, 0.736)
$\beta_0^{(c)}$	2.260	0.120	(2.029, 2.503)
$\beta_W^{(c)}$	−0.264	0.082	(−0.431, −0.101)
σ	0.490	0.011	(0.469, 0.513)

Table 1 summarizes $\pi(\boldsymbol{\theta}, \boldsymbol{u} \mid \mathcal{D})$. The posterior mean of $\beta_0^{(a)}$ and $\beta_0^{(b)}$ provides an approximate overall assessment of the baseline values of a_i and b_i for male patients. In

the case of the asymptotic value a_i, it decreases by about 0.418 in the female group (although this estimation has a lot of uncertainty), and shows a slight positive trend with age. Differences between individuals are relevant as it can be seen from the estimation of the standard deviation of the random effect in a_i, 1.67. The parameter b_i has an approximate basal value of 1.08 in the men group, which decreases by -0.27 units in the women group. Age also has a positive estimation and the random effect associated to individuals are also important for b_i, especially because this term appears on an exponential scale and negative sign in the quotient of the growth curve. Finally, the posterior mean for the c_i term is about 2.26 in the men group and decreases in 0.264 units in the group of women. The posterior mean of the standard deviation associated to the measurement error is not very large but it does have a very high accuracy. The fact that the *IAP* value of the *IP*, *ADP*, *MAP* and *MDP* of individual i depends on b_i and c_i proportionally to $-b_i/c_i$, and that the estimated coefficient associated to age is positive for b_i implies that the relationship of the *IP*, *ADP*, *MAP* and *MDP* for *IAP* coordinate with age is negative but barely important.

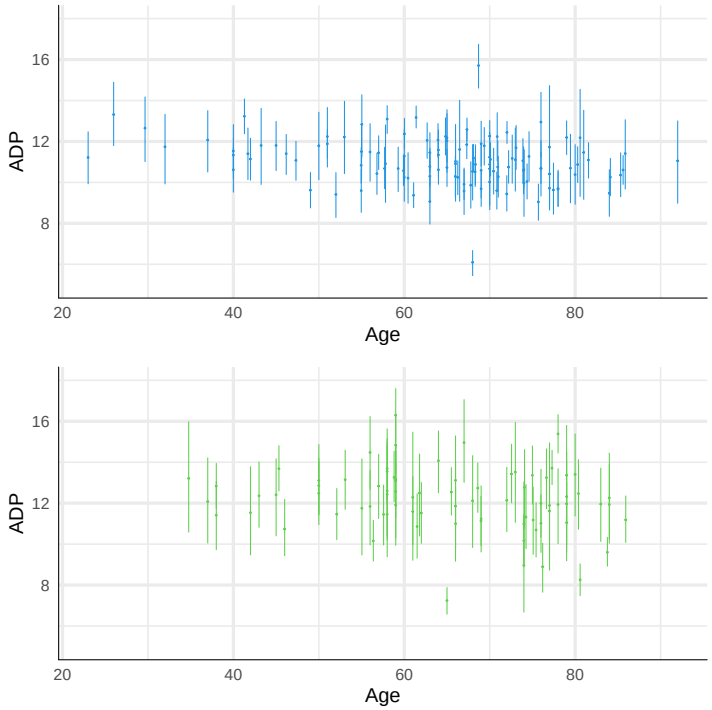


Figure 4. *Posterior mean and 95% credible interval of the ADP (IAP value) of the men (top panel) and the women (bottom panel) in the sample. Patients are ordered in the x-axis in terms of their age.*

As mentioned above, the posterior distribution is the starting point for the analysis of the different outcomes of interest in the study. In the following, we will present different

results that may be useful to better understand the relationship between *IAV* and *IAP* at both the individual and population level and thus be able to answer the scientific questions raised by the study. But first we would like to make a brief comment on the variable selection process discussed above for parameter c_i of the growth model. In this context, we considered different modelling approaches for c_i with regard to covariate gender. The Deviance Information Criterion (Spiegelhalter et al., 2002) was used for model comparison and according to this criterion the best model was the one with only the gender covariate and a common population term in parameter c_i as stated before.

4.2. Posterior individual outcomes

The basic inferential process allows the Bayesian methodology to obtain information both individually and in terms of the target population. In the following we focus on *ADP*. The mean of the *IAP* value of *ADP* for individual i, ADP_i, depends on b_i and c_i, which in turn depends on $(\boldsymbol{\theta}, \boldsymbol{u}_i)$. Consequently, we can compute the posterior distribution of the true ADP_i of each individual i in the sample from the subsequent posterior distribution $\pi(\boldsymbol{\theta}, \boldsymbol{u}_i \mid \mathcal{D})$. Figure 4 shows the posterior *ADP* mean and a 95% credible interval for the individuals in the sample ranked by age. The first thing that is striking in both graphs is the great difference in both the men and women groups in the range of credibility intervals, which is mainly explained by the differences in the number of repeated observations for each of them.

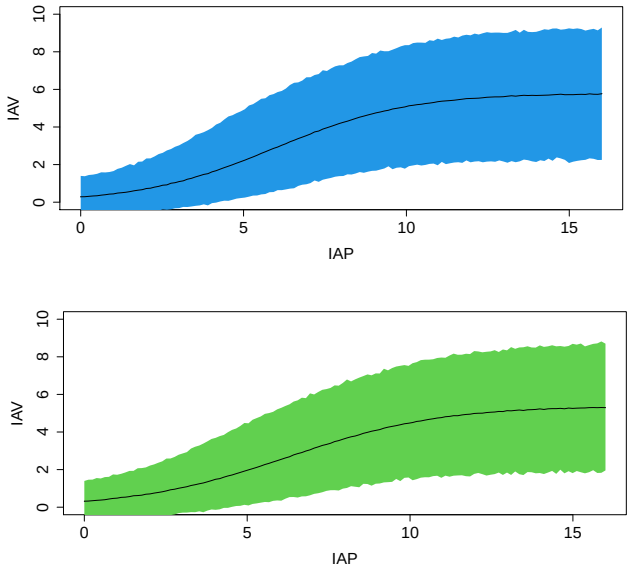


Figure 5. *Posterior predictive mean of the IAV and 95% predictive interval with regard to IAP values for a man (top panel) and a woman (bottom panel) aged 64.65 years (the sample mean).*

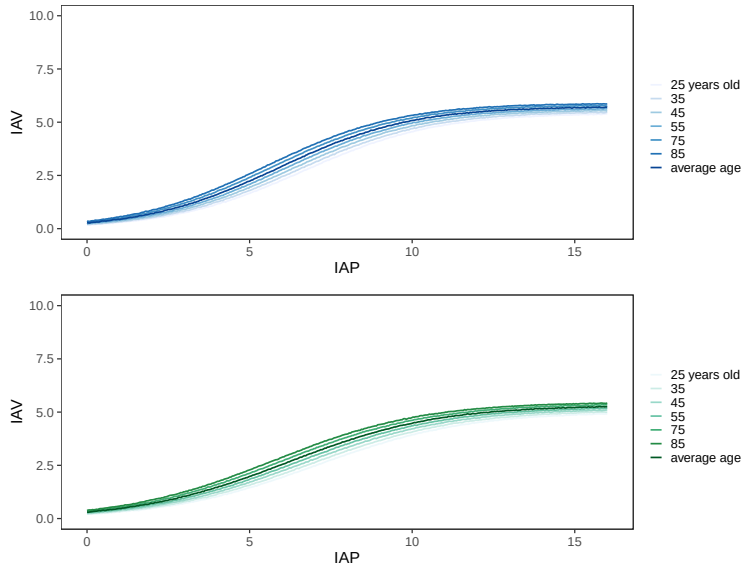


Figure 6. *Posterior predictive mean of the IAV with regard to IAP values for a man (top panel) and a woman (bottom panel) stratified by age.*

Prediction of observations for new individuals of the target population is an important issue that Bayesian statistics approaches in a natural way. The posterior predictive distribution of the random variable $Y_{n+1,j}$ that records the *IAV* value for a new individual, $n+1$, of the population with regard to their *IAP*, standardized age and gender values, which from now on we will denote by $\boldsymbol{x}_{n+1,j}$, depends on the conditional model in (1) and the posterior distribution $\pi(\boldsymbol{\theta}, \boldsymbol{u}_{n+1} \mid \mathcal{D})$, where $\boldsymbol{u}_{n+1}$ are the random effects associated to that individual $n+1$. It is computed as follows

$$f(y_{n+1,j} \mid \boldsymbol{x}_{n+1,j}, \mathcal{D}) = \int f(y_{n+1,j} \mid \boldsymbol{x}_{n+1,j}, \boldsymbol{\theta}, \boldsymbol{u}_{n+1})\, \pi(\boldsymbol{\theta}, \boldsymbol{u}_{n+1} \mid \mathcal{D})\, \mathrm{d}(\boldsymbol{\theta}, \boldsymbol{u}_{n+1}), \qquad (6)$$

where the posterior $\pi(\boldsymbol{\theta}, u_{n+1} \mid \mathcal{D})$ factorizes in terms of the marginal posterior distribution $\pi(\boldsymbol{\theta} \mid \mathcal{D})$ and the conditional distributions for the random effects $(u^a_{n+1} \mid \sigma^2_a) \sim N(0, \sigma^2_a)$ and $(u^b_{n+1} \mid \sigma^2_b) \sim N(0, \sigma^2_b)$. Figure 5 shows the posterior predictive mean and a 95% predictive interval for the *IAV* value of a new individual of the target population with age 64.65, the sample mean of the data, with respect to their *IAP* value and their gender. Both groups behave very similarly. The stabilisation of the values of *IAV* in both groups can be clearly seen, as well as the variability associated with the predictive processes, which is always greater in comparison with the estimation processes themselves. Finally, Figure 6 shows the posterior predictive mean for the response *IAV* variable with regard to *IAP* values of men and women with different ages. Of course, as we observed with the approximate posterior distribution of $\beta^{(b)}_A$ in the Table 1, a positive relationship

between *IAV* and age can be observed in the graphic, but it is very mild and possibly not very relevant for practical purposes in clinical scenarios.

4.3. Posterior population outcomes

Random effects connect the different repeated measures of the same individual in the statistical model and allow for the computation of individual-specific outcomes. We would also like to be able to have not only that individual information, but also outcomes that can provide general information about the target population. This aim implies to work with the marginal formulation of the model in (1) and (2) which we would obtain by integrating out the random effects of the conditional modelling as follows

$$f(y_{ij} \mid \boldsymbol{x}_{ij}, \boldsymbol{\theta}) = \int f(y_{ij} \mid \boldsymbol{x}_{ij}, \boldsymbol{\theta}, \boldsymbol{u})\, f(\boldsymbol{u} \mid \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{u} = \int \mathrm{N}(\mu_{ij}, \sigma^2)\, f(\boldsymbol{u} \mid \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{u}. \qquad (7)$$

This marginal formulation only depends on the parameter and hyperparameters of the model $\boldsymbol{\theta}$ and is the basis for the computation of any feature of this marginal model. For simplicity, we only focus on the true asymptotic *IAV* value and the true asymptotic deceleration point *ADP* for a patient with an average age.

Figure 7 shows the posterior distribution of the asymptotic *IAV* for men and women aged 64.65 years (the mean of the sample). There is not much difference between the two distributions. An estimation of the asymptotic *IAV* in the group of men is 5.60 L. while in the group of women it is 5.25 L. Figure 8 shows the joint posterior distribution, in terms of contour lines, of the *ADP* pressure point and the subsequent volume value for men and women aged 64.65 years (the sample mean) as well as the marginal distributions of both quantities. Posterior mean for the *ADP*'s pressure and volume is 10.06 mmHg. and 5.05 L. in men aged 64.65, and 10.87 mmHg. and 4.74 L. in the group of women with the same age, respectively. A similar analysis is possible for *MAP*, *IP* and *MDP*. However, their posterior results for both coordinates (*IAP* and *IAV*) are proportional to those of *ADP* and their information would be repetitive.

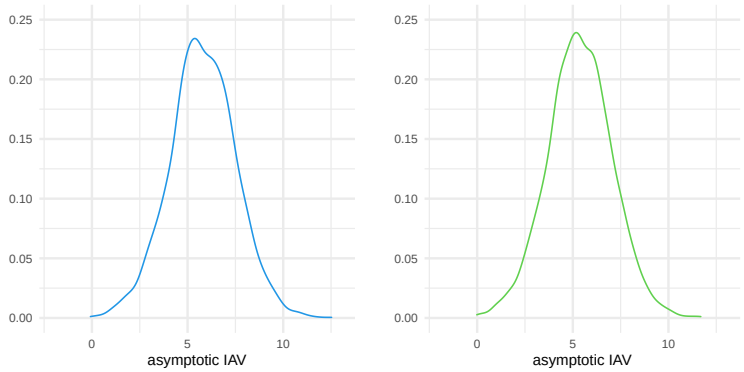


Figure 7. *Posterior distribution of the asymptotic IAV for men (on the left) and women (on the right).*

5. Model checking

Model checking is an essential component of any statistical analysis which has generated an extensive literature within the Bayesian reasoning (Vehtari and Ojanen, 2012). Our interest in this subject focuses on assessing, following the philosophy in Gelman et al (2014), whether the possible shortcomings of our model have a relevant effect on the derived results. We approach model checking via posterior predictive distributions following the ideas by Box (1980), who states that prediction (and not estimation) enables "criticism of the entertained model in light of current data".

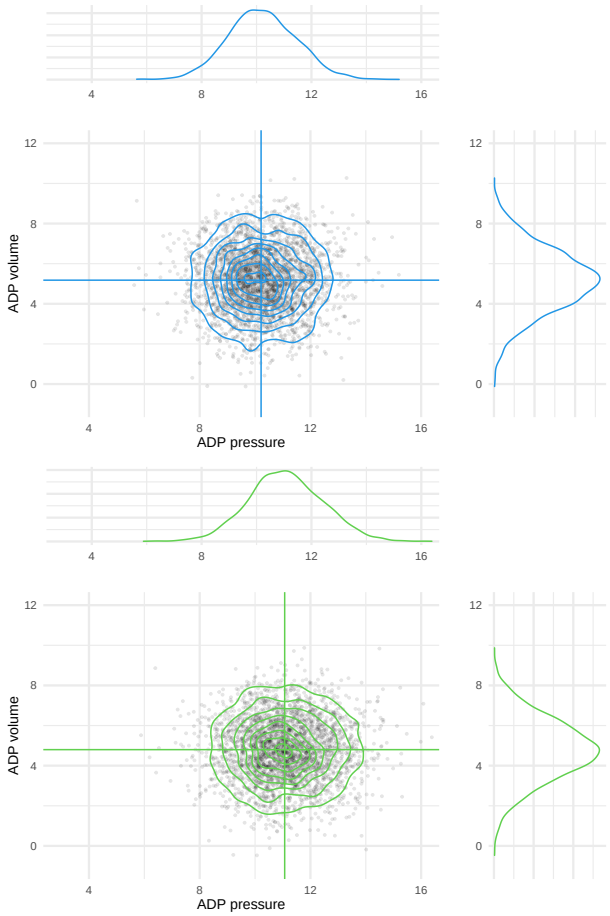


Figure 8. *Joint posterior distribution and contour lines of the IAP and IAV coordinates for the true ADP and posterior marginal distribution for each of both quantities for men (top panel) and women (bottom panel) aged 64.65 years (the sample mean). The horizontal and vertical lines represent the approximate posterior mean of IAV − ADP and the approximate posterior mean of IAP − ADP, respectively.*

We check our model through the cross-validated predictive density Gelfand, Dey and Chang (1992) defined as the conditional posterior density of a future *IAV* value $y_{ij}^{(rep)}$ for individual i, $i = 1,\ldots,n$ with standardized *IAP* value x_{ij} of a replicate experiment

$$f(y_{ij}^{(rep)} \mid \mathcal{D}^{-(ij)}) = \int f(y_{ij}^{(rep)} \mid \boldsymbol{x}_{ij}, \boldsymbol{\theta}, \boldsymbol{u})\, \pi(\boldsymbol{\theta}, \boldsymbol{u} \mid \mathcal{D}^{-(ij)})\, d(\boldsymbol{\theta}, \boldsymbol{u}),$$

where $\mathcal{D}^{-(ij)}$ are all the data in $\mathcal{D}$ except for the observation y_{ij} (leave-one-out (LOO) procedure).

The fundamental idea underlying this proposal assumes that if the estimated model is correct, each observation can be considered as a random value from the cross-validated predictive density Chen, Shao, and Ibrahim (2000). In this framework, we consider two complementary characteristics of such predictive distribution assessed at each observed value y_{ij}. These quantities are the conditional predictive ordinate (CPO) and the cross-validated probability integral (PIT), and are defined as:

$$\begin{aligned} CPO_{ij} &= f(y_{ij} \mid \mathcal{D}^{-(ij)}), \\ PIT_{ij} &= P(Y_{ij}^{(rep)} \leq y_{ij} \mid \mathcal{D}^{-(ij)}). \end{aligned}$$

CPO_{ij} values correspond to the ordinates in the y_{ij} of the cross-validated predictive density. Large CPO_{ij} values support the selected model because indicate a good tuning between the data and the model. PIT_{ij} is the posterior probability that the replicated $(ij)th$ observation is less or equal the subsequent observed value. When the model is well calibrated these probabilities follow a uniform distribution in the unit interval.

The direct implementation of these quantities is computationally very expensive because we would need to approximate as many posterior distributions as we have elements in $\mathcal{D}$. This is not necessary because the application of self-normalized importance sampling allows CPOs and PITs to be approximated from draws of the posterior distribution $\pi(\boldsymbol{\theta}, \boldsymbol{u} \mid \mathcal{D})$ computed with the complete data $\mathcal{D}$ (Gelfand, 1996; Ntzoufras, 2011). Computation of CPOs and PITs was done by means of the R software from the posterior outputs obtained with JAGS.

Figure 9 shows the histogram of CPOs and PITs respectively. The information provided in both cases suggests that the model used has some shortcomings that can be improved. We have some values of the CPO that are small and the PITs do not seem to be uniformly distributed mainly due to a remarkable abundance of values close to zero.

Figure 10 shows how PIT values are distributed along *IAP*. Theoretically, $PIT's$ should be uniformly distributed between 0 and 1 at each *IAP* point. However, from $IAP \approx 6.5$ (vertical red line) to $IAP \approx 14$ PIT values do in general do not exceed 0.5. This behaviour indicates that our model performs well when we work with small values of *IAP*, overpredicts observations of *IAV* for medium and medium-high values of the covariate *IAP*, and finally, it seems to improve with large IAP values.

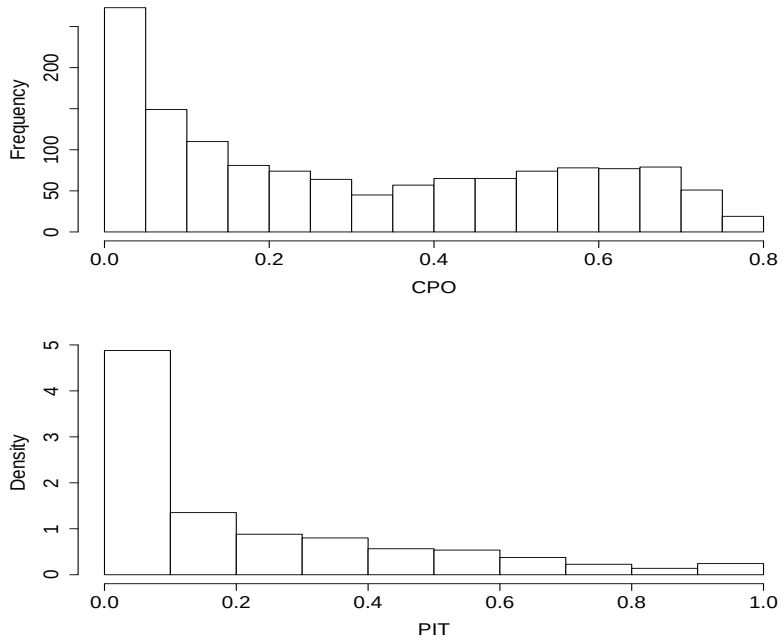


Figure 9. *Histogram of the approximate CPO (on the top) and density histogram of the approximate PIT (in the bottom) quantities for all the observations.*

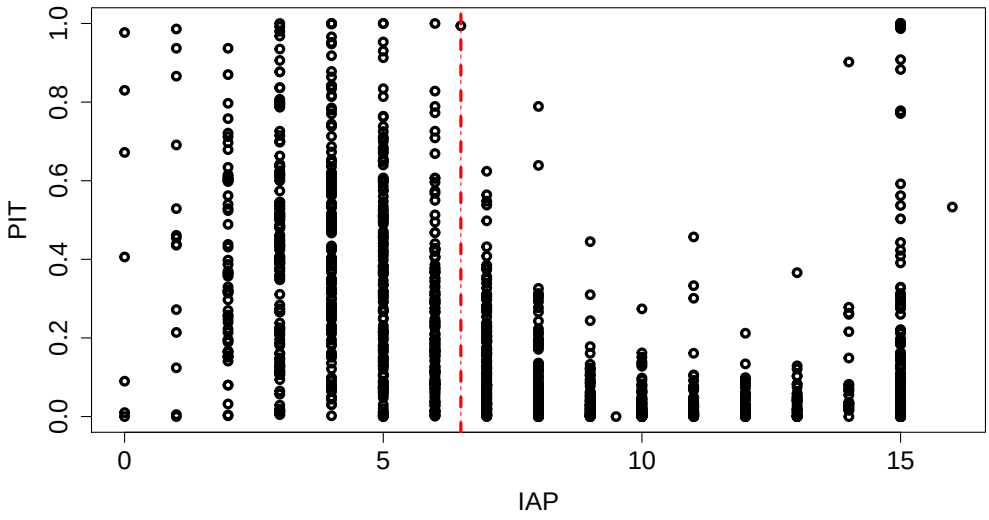


Figure 10. *Distribution of PIT values along the different values of IAP.*

Conclusions

Precision medicine tenets are that different interventions have distinct effects in different people and that this variability can, at least in part, be characterized and predicted (Senn, 2016). In this study we have tried to lay the foundation for the mathematical modeling of the abdomen behaviour during pneumoperitoneum insufflation. We have also parameterized such model to achieve predictive capability based on a few simple baseline characteristics. This is the first step in a precision medicine approach to pneumoperitoneum insufflation for laparoscopic surgery. This process can be potentially scaled up and recursively performed throughout the duration of the surgical intervention to ensure that even if conditions change, we could be able to provide an optimal surgical field to the surgeon while exposing the patient the lowest possible pressure.

With this procedure, we would like to achieve an optimal surgical workspace while minimizing the pressure administered to the patient. In other words, each subject would receive a titrated pressure according to her/his characteristics. Also, the ability to predict where the marginal gain in volume diminishes by deriving critical points on the parameterized curve have an especially interesting clinical potential.

Bayesian inference can provide a suitable inferential framework in this context. First of all, Bayesian hierarchical models are useful to elicit and formulate the different sources of variation and uncertainty of the problem and incorporate suitable terms into the model to account for them. In this particular case, the model includes nonlinear effects through a logistic growth function. As model fitting relies on MCMC methods, inference about particular elements of interest in the model becomes feasible. For example, the logistic growth model has a known parametric form from which some crucial critical points can be derived analytically but inference on these points is far from straightforward. However, the output produced by MCMC during model fitting can be exploited to compute the posterior marginals of these particular points as well as those of the other model parameters. This provides extra information that can be used during the laparoscopic surgery. Inference about these critical points under other inferential frameworks would not be so straightforward.

The most important critical point in our study is ADP, as this controls how much air is insufflated during surgery. From a clinical point of view, when operating on new patients, ADP's predictive distribution can help physicians provide adequate insufflation during laparoscopic surgery.

As we have illustrated, model checking is critical for health applications of statistical methodology as this will allow potential bias to be detected. We believe that the use of the model checking techniques should be widely adopted when relying on statistical models for medical practice to detect and avoid systematic biases in medical treatment (Obermeyer et al., 2019).

The study presented in this paper illustrates a preliminary analysis in which 198 patients have been enrolled. In the future, we aim to conduct a larger trial so that a wider range of patients is represented. Furthermore, other covariates will be recorded and in-

cluded into the model to reduce the uncertainty about the estimates and predictions, and increase the accuracy of insufflation. We plan to refine our model with anthropometric measurements. We are recording not only height and weight, but also waist and hip circumference and abdomen sagittal height to have abdominal, volume and body mass surface and update our model with these new data. Furthermore, data from medical imaging such as abdominal volume estimation based on routine preoperative computerized tomography images or ultrasonic assessment of the abdominal wall thickness and fat component can be explored as covariate alternatives. We will also record the number of previous open and laparoscopic abdominal surgeries, as well as, in the case of women, the number of pregnancies. Finally, models with different assumptions will be considered such as non-homoscedastic models with increasing variability, different types of curves such as the Gompertz curve (Funatogawa and Funatogawa, 2018), or even the inclusion of random effects to assess the possible variability among the different studies.

Acknowledgements

We would like to highlight and thank the work of the editor of the journal and the three anonymous referees who reviewed the paper because their comments and suggestions have greatly improved its quality and clarity. This paper was supported by research grant PID2019-106341GB-I00 funded by Ministerio de Ciencia e Innovación (Spain) and the Project MECESBAYES (SBPLY/17/180501/000491) funded by the Consejería de Educación, Cultura y Deportes, Junta de Comunidades de Castilla-La Mancha (Spain). Gabriel Calvo is also supported by grant FPU18/03101 funded by the Ministerio de Ciencia e Innovación (MCI, Spain). Merck Sharp & Dohme funded the IPPColLapse II study (Protocol Code No. 53607). This is an investigator-initiated study in which the sponsors and funders have no roles in study design, analysis of data, or reporting.

References

Becker, C., Plymale, M. A., Wennergren, J., Totten, C., Stigall, K., Roth, J. S. (2017) Compliance of the abdominal wall during laparoscopic insufflation. *Surgical Endoscopy*, 31, 1947–1951

Box, G. E. P. (1980). Sampling and Bayes's inference in scientific modelling and robustness (with discussion). *Journal of the Royal Statistical Society, Series A*, 143, 383-430.

Chen, M.-H., Shao, Q.-M., and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. New York: Springer.

Colon Cancer Laparoscopic or Open Resection Study Group; Buunen, M., Veldkamp, R., Hop, W. C., Kuhry, E., Jeekel, J., Haglind, E., et al. (2009). Survival after laparoscopic surgery versus open surgery for colon cancer: long-term outcome of a randomised clinical trial. *The Lancet Oncology* 10(1), 44–52.

Cramer, J. S. (2004). The early origins of the logit model. *Studies in History and Philosophy of Biological and Biomedical Sciences* 35, 613–626.

Davidian, M. (2008). Non-linear mixed-effects model. In *Longitudinal data analysis*. Chapman and Hall/CRC. p. 121–156.

Diaz-Cambronero, O., Flor Lorente, B., Mazzinari, G., Vila Montañés M, García Gregorio, N,, Robles Hernández, D. et al. (2020). A multifaceted individualized pneumoperitoneum strategy for laparoscopic colorectal surgery: a multicenter observational feasibility study. *Surgical Endoscopy*, 33, 252–260.

Diaz-Cambronero, O., Mazzinari, G., Flor Lorente, B., García Gregorio, N., Robles-Hernández, D., Olmedilla Arnal, L. E. et al. (2020) Effect of an individualized versus standard pneumoperitoneum pressurestrategy onpostoperative recovery: a randomized clinical trial in laparoscopic colorectal surgery. *British Journal of Surgery*, 107, 1605–1614.

Fang, Z. and Bailey, R. L. (2001. Nonlinear Mixed Effects Modeling for Slash Pine Dominant Height Growth Following Intensive Silvicultural Treatments. *Forest Science*, 47, 287–300.

Forstemann, T., Trzewik, J., Holste, J., Batke, B., Konerding, M. A., Wolloscheck, T., Hartung C. (2011). Forces and deformations of the abdominal wall–a mechanical and geometrical approach to the linea alba. *Journal of Biomechanics*, 44, 600–606.

Funatogawa, I., Funatogawa, T. (2018). Nonlinear Mixed Effects Models, Growth Curves, and Autoregressive Linear Mixed Effects Models. In: JSS Research Series in Statistics (ed). *Longitudinal Data Analysis*, pp 99–117. Springer, Singapore.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. In *Bayesian Statistics* 4 (Eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith). Oxford: Oxford University Press, 165–180.

Gelfand, A. E. (1996). Model determination using sampling-based methods. In *Markov Chain Monte Carlo in Practice* 4 (Eds. W. Gilks, S. Richardson, and D. Spiegelhalter). Chapman & Hall, 145–161.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. (2014). *Bayesian Data Analysis*, Third Edition. Boca Raton: Chapman and Hall.

Giltinan, D. M. (2006). Pharmacokinetics and pharmacodynamics. In P. Armitageand T. Colton (eds), *Encyclopedia of Biostatistics*, 2nd ed,, pp. 600–606. Wiley, Hoboken, NJ.

Laird, N. M. and Ware, J. H. (1982). Random-Effects for Longitudinal Data. *Biometrics*, 38, 963–974.

Lee, Y. and Nelder, J. A. (2004). Conditional and Marginal Models: Another View . *Statistical Science*, 19(2), 219–238.

Lindsey, J. K. (2001). *Nonlinear Models in Medical Statistics*. Oxford University Press, Oxford.

Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear Mixed Effects Models for Repeated Measures Data. *Biometrics*, 46(3), 673–687.

Loredo, T. J. (1989) From laplace to supernova sn 1987a: Bayesian inference in astrophysics. In: Fougère PF (ed). *Maximum entropy and Bayesian methods*, pp 81–142. Kluwer Academic publishers, Dordrecht.

Loredo, T. J. (1992) Promise of Bayesian inference for astrophysics. In: Feigelson E, Babu G (eds). *Statistical challenges in modern astronomy*, pp 275–297. Springer, New York.

Mazzinari, G., Diaz-Cambronero, O., Alonso-Iñigo, J. M., García Gregorio, N., Ayas-Montero, B. et al. (2020). Intraabdominal pressure targeted positive end-expiratory pressure during laparoscopic surgery: an open-label, nonrandomized, crossover, clinical trial. *Anesthesiology*, 132, 667–677.

Mazzinari, G., Diaz-Cambronero, O., Serpa Neto, A., Martínez Cañada, A., and Rovira, L., Argente Navarro, M. P., et al. (2021). Modeling intra-abdominal volume and respiratory driving pressure during pneumoperitoneum insufflation − a patient-level data meta-analysis. *Journal of Applied Physiology* 130(3), 721–728.

Mulier, J., Dillemans, B., Crombach, M., Missant, C., Sels, A. (2009). On the abdominal pressure volume relationship. *The Internet Journal of Anesthesiology* 21, 1–7. 5221.

Neudecker, J., Sauerland, S., Neugebauer, E. A. M. et al. (2002) The EAES clinical practice guidelines on the pneumoperitoneum for laparoscopic surgery. *Surgical Endoscopy*, 16(7), 1121–43

Neugebauer, E. A. M, Becker, M., Buess, G. F. et al. (2010). EAES recommendations on methodology of innovation management in endoscopic surgery. *Surgical Endoscopy*, 24(7), 1594–1615.

Ntzoufras, I. (2011). *Bayesian modeling using WinBUGS*. John Wiley & Sons.

Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 6464 (366), 447–453.

Pache, B., Hübner, M., Jurt, J., Demartines, N., and Grass, F. (2017). Minimally invasive surgery and enhanced recovery after surgery: the ideal combination? *Journal of Surgical Oncology*, 116(5), 613–616.

Peek, M. S., Russek-Cohen, E., Wait, D. A. and Forseth, I. N. (2002). Physiological response curve analysis using nonlinear mixed models. *Oecologia*, 132, 175–180.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*. p. 1–10.

Senn, S. (2016). Mastering variation: variance components personalised medicine. *Statistics in Medicine*, 30(7), 966–977.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 583-639.

Vehtari, A. and Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Journal of Biomechanics*, 6, 142–228.

SORT 45 (2) July-December 2021, 163-180 DOI: 10.2436/20.8080.02.114

Median bilinear models in the presence of extreme values

Miguel Santolino*

Abstract

Bilinear regression models involving a nonlinear interaction term are applied in many fields (e.g., Goodman's RC model, Lee-Carter mortality model or CAPM financial model). In many of these contexts data often exhibit extreme values. We propose the use of bilinear models to estimate the median of the conditional distribution in the presence of extreme values. The aim of this paper is to provide alternative methods to estimate median bilinear models. A calibration strategy based on an iterative estimation process of a sequence of median linear regression is developed. Mean and median bilinear models are compared in two applications with extreme observations. The first application deals with simulated data. The second application refers to Spanish mortality data involving years with atypical high mortality (Spanish flu, civil war and HIV/AIDS). The performance of the median bilinear model was superior to that of the mean bilinear model. Median bilinear models may be a good alternative to mean bilinear models in the presence of extreme values when the centre of the conditional distribution is of interest.

MSC: 62H12, 62H17, 62JO2, 62L12.

Keywords: Outliers, quantile regression, single factor models, nonlinear, multiplicative.

1. Introduction

In regression analysis the effect of the interaction between two explanatory variables on the dependent variable is often of great interest. Two-way analysis of variance (ANOVA) models have been widely applied in linear regression analysis when a measurement dependent variable is regressed on two categorical independent variables, and the aim is to assess the main effect of the two nominal variables but also the interaction effect between them (Yates and Cochran, 1938). Two-way ANOVA models are linear

* Riskcenter-IREA, Dept. Econometrics, University of Barcelona, Spain. E-mail: msantolino@ub.edu
Received: February 2021
Accepted: November 2021

regression models where the joint interaction effect is included as an additional regressor. So, linear regression techniques may be directly applied to estimate parameters, such as least squares or maximum likelihood methods.

A more flexible modelling in two-way tables are the regression models in which the multiplicative interaction structure is specified as a nonlinear term. These models are usually named bilinear models (Gabriel, 1978), although other names are often in use for these models, such as biadditive models (Denis and Pázman, 1999) or additive main effects and multiplicative interaction (AMMI) models (Van Eeuwijk, 1992, 1995). The unknown parameters of bilinear models may be also estimated by least squares or maximum likelihood. Least squares estimators of the nonlinear term are derived using singular value decomposition of the matrix of residuals (Gabriel, 1978; Lee and Carter, 1992). Maximum likelihood estimators may be obtained by an iterative process (Goodman, 1979, 1981).

Bilinear regression models involving multiplicatively structured interactions are widely applied. Many models used in social sciences fits to this setting, including the row-column association model for two-way tables (Goodman, 1979, 1981), the uniform difference (UNIDIFF) or layer effect model for three-way tables (Erikson and Goldthorpe, 1992; Xie, 1992), generalized additive main effects and multiplicative interaction effects (GAMMI) models for crop yields (Van Eeuwijk, 1992, 1995), the one-dimensional Rasch-type model for binary responses (Turner, Firth and Kosmidis, 2013) or the stereotype regression model for ordered multinomial data (Anderson, 1984). In time series analysis, statistical factor models can be understood as multiplicative interaction models (Croux et al., 2003). Factor models are widely applied in finance for calculating the investment risk in asset pricing theory, such as the capital asset pricing model (CAPM) model or the Fama-French model (Black, Jensen and Scholes, 1972). In demography and actuarial science, factor models are used to predict the future mortality. In fact, most of mortality projections models, such as Lee-Carter and Renshaw-Haberman models can be understood as multiplicative interaction models (Lee and Carter, 1992; Renshaw and Haberman, 2006; Macias and Santolino, 2018; Moyano-Silva et al., 2020).

In many of these contexts data often show extreme values. When the centre of the conditional distribution is of interest, a common practice is to consider extreme values as outliers and remove them from the dataset prior to estimation. Formally, an outlier is a data point that deviates so far from the other observations because it was generated by a totally different mechanism or simply by error (Hawkins, 1980; Justel, Peña and Tay, 2001). Deleting outliers is important because those values can increase error variance and influence estimates. However, this strategy should be taken very cautiously when data points are extreme values but not outliers. Extreme values are events that might happen, so we should be very cautious before deleting these values from datasets.

A different approach is here followed. It is well known that the median is a robust measure of central tendency. Median bilinear models may be a good alternative to mean bilinear models in the presence of extreme values (Gabriel and Odoroff, 1984). In this article we propose the use of the bilinear regression setting to model the median of the

conditional distribution as a nonlinear function of predictors. The aim of this article is twofold: 1) to show how the parameters of the median bilinear model can be estimated and, 2) to compare the performance of the conditional median bilinear regression and the conditional mean bilinear regression in the presence of extreme values.

The main contribution of the paper is to review alternative methods to estimate median bilinear models. Bilinear models are nonlinear regressions. The techniques available for estimating nonlinear regression models for the conditional median are not as well developed as those for the conditional mean estimation. Koenker and Park (1996) proposed to calibrate median nonlinear regression models by means of the linearization of the objetive function. Here we propose an alternative calibration approach based on an iterative estimation process of a sequence of median linear regressions. This second alternative is novel. It was first used by Moyano-Silva, Pérez-Marín and Santolino (2020) to estimate the Lee-Carter stochastic mortality model. We here generalize this strategy to estimate median bilinear models with two main factors. To solve the underlying linear optimization problems, we use interior point methods (Koenker and Park, 1996; Portnoy and Koenker, 1997) and the maximum likelihood approach (Sánchez, Labros and Labra, 2013). This paper focuses on the evaluation of goodness-of-fit of mean and median bilinear models in presence of extreme values. However, bootstrapping techniques can be used to estimate standard errors when inference on coefficient estimates is of interest (Buchinsky, 1995).

Two applications are illustrated for the comparison of the median and mean bilinear models. The first application is based on simulated data. In the second application real Spanish mortality data are used to estimate the median and mean (log)bilinear stochastic mortality models. In both applications, bilinear models are calibrated using the whole sample. The performance of the fitted models is then evaluated computing a series of goodness-of-fit measures for the whole sample and when extreme values are removed.

The article is structured as follows. Section 2 introduces the mean and median bilinear regression models. Section 3 shows the parameter estimation methods of the mean bilinear regression model. Section 4 describes the calibration strategies of the median bilinear regression model. The two applications are illustrated in Section 5. Main conclusions are summarized in Section 6.

2. Bilinear regression model

Let Y be a continuous random variable with finite expectation and cumulative distribution function F_Y defined by $F_Y(y) = P(Y \leq y)$. The inverse function of F_Y is known as *quantile function*, Q. The quantile of order α is defined as $Q_\alpha(Y) = F_Y^{-1}(\alpha) = \inf\{y \mid F_Y(y) \geq \alpha\}$ where $\alpha \in (0,1)$. The quantile is a left-continuous increasing function. If F_Y is continuous and strictly increasing, the mathematical expectation can be represented as $\mathbb{E}(Y) = \int_0^1 Q_{1-u}(Y)du$. The median is the quantile of order 0.5.

Let consider two categorical variables. The first factor has I levels ($i \in \{1,\ldots,I\}$) and the second has J levels ($j \in \{1,\ldots,J\}$). The sample size is N such as $N = I \cdot J$.

Let y_{ij} be the random variable conditional on the levels i and j. The bilinear regression model in two-way tables is defined as:

$$y_{ij} = a_i + b_j + c_i \cdot d_j + \varepsilon_{ij} \tag{1}$$

where a_i is the main effect of the level i and b_j is the main effect of the level j. The coefficients of the nonlinear term are c_i and d_j capturing the interaction effect of the two levels. Finally, ε_{ij} is the error random variable. Note that infinite solutions exist in (1). For any scalars z, u and v, the following transformations $\{\tilde{a}_i, \tilde{b}_j, \tilde{c}_i, \tilde{d}_j\} = \{a_i - z \cdot c_i, b_j - d_j \cdot v - z \cdot v, \frac{c_i + v}{u}, u \cdot (d_j + z)\}$ give unaltered outcome values. To overcome the lack of identifiability and to help in the interpretation, the following two constraints are often set: $\sum_i c_i = 1$ and $\sum_j d_j = 0$.

In the case of independent and identically zero-mean distributed random errors, the conditional expected value of y_{ij} may be expressed as

$$\mathbb{E}(y_{ij}) = a_i + b_j + c_i \cdot d_j \tag{2}$$

Analogously, in case of independent and identically zero-median distributed random errors, the median of y_{ij} may be expressed as (Bassett and Koenker, 1978),

$$Q_{0.5}(y_{ij}) = a_i + b_j + c_i \cdot d_j \tag{3}$$

Sections 3 and 4 are devoted to estimate the vectors of coefficients, $a = (a_1, \ldots, a_I)$, $b = (b_1, \ldots, b_J)$, $c = (c_1, \ldots, c_I)$ and $d = (d_1, \ldots, d_J)$ in (2) and (3), respectively.

3. Mean bilinear model: calibration

Two widely used techniques to estimate the parameters of (2) are the least squares and the maximum likelihood methods.

Least squared errors

The expectation is the value that minimizes the sum of squared deviations. One strategy for estimating the parameters is to minimize the sum of squared errors, as follows:

$$\min_{\theta \in \mathbb{R}^{2 \cdot (I+J)}} \sum_{i,j} (y_{ij} - a_i - b_j - c_i \cdot d_j)^2 \tag{4}$$

where θ is the set of parameters to estimate, $\theta = (a, b, c, d)$. Coefficients in (4) cannot be directly estimated by ordinary least squares because the right-hand side of equation (2) is not linear with the parameters. To estimate the coefficients, Gabriel (1978) proposed to fit the bilinear models in a two-stage process: (1) fit the linear part of the model, then take residuals, and (2) fit the bilinear part to the residuals.

Stage 1 uses linear least squares to solve the least squares problem. The resulting vectors $\hat{a} = (\hat{a}_1, \ldots, \hat{a}_I)$ and $\hat{b} = (\hat{b}_1, \ldots, \hat{b}_J)$ are then introduced into the joint fitting

problem. The $(I \times J)$-matrix A, where the (i,j)-element is $a_{ij} = y_{ij} - \hat{a}_i - \hat{b}_j$, is decomposed by singular value decomposition, $\text{svd}(A) = U\Lambda V^{\intercal}$. The vector of estimates $\hat{c} = (\hat{c}_1, \ldots, \hat{c}_I)$ is the first column of U, $\hat{c} = (u_{1,1}, \ldots, u_{I,1})$, and the vector of estimates $\hat{d} = (\hat{d}_1, \ldots, \hat{d}_J)$ is the first column of V multiplied by the first eigenvalue $\lambda_{1,1}$, $\hat{d} = \lambda_{1,1} \cdot (v_{1,1}, \ldots, v_{J,1})$.

Maximum likelihood

Goodman (1979) proposed to use a iterative method for estimating bilinear models by maximum likelihood. Suppose the log-likelihood function is given by $l(\theta) = \sum_{i,j} log\,(f(y_{ij}))$, where f is the density function of y_{ij}. The function l may be maximized by an iterative process in which the elementary newton method is applied for the score functions of each set of parameters. In the mean bilinear model we have three sets of parameters. Denote the vector of initial values $\hat{\theta}^0 = (\hat{a}^0, \hat{b}^0, \hat{c}^0, \hat{d}^0)$ and $l^0 = l(\hat{\theta}^0)$. In the iteration step v, parameters are updated as follows:

1. Given $\hat{\theta}^v$, $\hat{a}^{v+1} = \hat{a}^v - \frac{\partial l^v / \partial a}{\partial^2 l^v / \partial a^2}$, $\hat{b}^{v+1} = \hat{b}^v - \frac{\partial l^v / \partial b}{\partial^2 l^v / \partial b^2}$, $\hat{c}^{v+1} = \hat{c}^v$ and $\hat{d}^{v+1} = \hat{d}^v$.

2. Given $\hat{\theta}^{v+1}$, $\hat{c}^{v+2} = \hat{c}^{v+1} - \frac{\partial l^{v+1} / \partial c}{\partial^2 l^{v+1} / \partial c^2}$ and $\hat{a}^{v+2} = \hat{a}^{v+1}$, $\hat{b}^{v+2} = \hat{b}^{v+1}$ and $\hat{d}^{v+2} = \hat{d}^{v+1}$.

3. Given $\hat{\theta}^{v+2}$, $\hat{d}^{v+3} = \hat{d}^{v+2} - \frac{\partial l^{v+2} / \partial d}{\partial^2 l^{v+2} / \partial d^2}$, $\hat{a}^{v+3} = \hat{a}^{v+2}$, $\hat{b}^{v+3} = \hat{b}^{v+2}$ and $\hat{c}^{v+3} = \hat{c}^{v+2}$.

4. If $l^{v+3} - l^v \leq \eta$ then *stop*, where η is the tolerance value, otherwise, $\hat{\theta}^v = \hat{\theta}^{v+3}$ and move to step 1.

An application of this method to model the Poisson distributed number of deaths is shown by Brouhns, Denuit and Vermunt (2002).

4. Median bilinear model: Least absolute errors

The mean minimizes the sum of squared deviations and the median is the value that minimizes the sum of absolute deviations. The parameters of the median regression are estimated minimizing absolute errors, as follows:

$$\min_{\theta \in \mathbb{R}^{2\cdot(I+J)}} \sum_{i,j} y_{ij} - a_i - b_j - c_i \cdot d_j \tag{5}$$

The expression (5) can be rewritten as the following minimization problem:

$$\min_{\theta \in \mathbb{R}^{2\cdot(I+J)}, u \geq 0, v \geq 0} 0.5\Big(\sum_{i,j} u_{ij} + \sum_{i,j} v_{ij}\Big) \tag{6}$$

subject to

$$y_{ij} - a_i - b_j - c_i \cdot d_j - u_{ij} + v_{ij} = 0.$$

where $u_{ij} = \varepsilon_{ij}$ if $\varepsilon_{ij} > 0$ or 0 otherwise, and $v_{ij} = |\varepsilon_{ij}|$ if $\varepsilon_{ij} < 0$ or 0 otherwise, and $\varepsilon_{ij} = y_{ij} - a_i - b_j - c_i \cdot d_j$. Let us use the following notation $\varepsilon_{ij}(\theta) = y_{ij} - f_{ij}(\theta)$, with $f_{ij}(\theta) = a_i + b_j + c_i \cdot d_j$, to indicate that ε_{ij} depends on the set of parameters θ. Two alternative strategies are adopted to estimate the parameters in (6) based on the conversion of the original nonlinear optimization problem in a sequence of linear problems.

4.1. Strategy A: Linearization of the objective function

Strategy A transforms the nonlinear problem (6) in a sequence of linear problems. Provided that the functions $\varepsilon_{ij}(\theta)$ are continuously derivable in θ, the Lagrangian function may be expressed as $L(s,t,w) = u^{\mathsf{T}}(0.5\mathbf{1}_N - s - t) + v^{\mathsf{T}}(0.5\mathbf{1}_N + s - w) + \varepsilon(\theta)^{\mathsf{T}}s$, where $\mathbf{1}_N^{\mathsf{T}}$ is a N-column vector of 1's, $\varepsilon(\theta) = (\varepsilon_{1,1}(\theta), ..., \varepsilon_{ij}(\theta))^{\mathsf{T}}$ and s, t and w are the multipliers of Lagrange with t and w are non-negative vectors. Taking partial derivatives with respect to the model parameters θ and the decision variables u and v, we obtain the dual feasibility conditions. The dual version of (6) can be then expressed as,

$$\max_{s\in[-0.5,0.5]^N} \varepsilon(\theta)^{\mathsf{T}}s \qquad s.t \quad J(\theta)^{\mathsf{T}}s = 0, \tag{7}$$

where $J(\theta)$ is the vector of first derivatives of $f_{ij}(\theta)$ with respect to θ (El-Attar, Vidyasagar and Dutta, 1979).

4.1.1. Calibration: Affine scale method

Let us consider the locally linearized approximation $\varepsilon(\theta+\Delta) \approx \varepsilon(\theta) - J(\theta)\cdot\Delta$. Koenker and Park (1996) propose to replace $\varepsilon(\theta)$ by the linear approximation $\varepsilon(\theta+\Delta)$ and, then, to apply iteratively the affine scaling method to solve the dual optimization problem (7). Consider the set of initial values $\hat{\theta}^0 = (\hat{a}^0, \hat{b}^0, \hat{c}^0, \hat{d}^0)$. In the iteration step v, parameters are updated as follows:

1. Refine s with Meketon algorithm and estimate Δ which depends on s and $J(\hat{\theta}^v)$, and $\varepsilon(\hat{\theta}^v)$.

2. To ensure that the linearized approximation generates feasible steps, update $\hat{\theta}$ as $\hat{\theta}^{v+1} = \hat{\theta}^v + \lambda\hat{\Delta}$, where $\hat{\Delta}$ is the direction step and λ the length of the step. The length of the step $\lambda \in [0,1]$ is estimated minimizing the primal optimization problem (6) for $\varepsilon(\hat{\theta}^v + \lambda\hat{\Delta})$.

3. If $\sum_{i,j}\left(|\varepsilon_{ij}(\hat{\theta}^{v+1})| - |\varepsilon_{ij}(\hat{\theta}^v)|\right) \leq \eta$ then *stop*, where η is the tolerance value. Otherwise, move to step 4.

4. Project the refined s in the null space of the updated $J(\hat{\theta}^{v+1})$ and rescale to ensure that it is bounded in $[-0.5, 0.5]$, and move to the next iteration.

4.2. Strategy B: Sequence of median linear regressions

Under the strategy B, coefficients in (6) are also estimated by means of an iterative process of a sequence of linear optimization problems. Strategy B draws inspiration from Wilmoth (1993) who replied the method described by Goodman (1979) to the case of minimum least square estimators. Wilmoth (1993) proposed an iterative process to estimate the parameters of the mean bilinear model sequentially by least square techniques. Santolino (2020) adopted this strategy to estimate the parameters of the Lee-Carter quantile mortality model by least absolute techniques. We here describe this strategy for the median bilinear regression. Like the median polish for additive models (Emerson and Hoaglin, 1983), our method relies on the properties of homogeneity and translation invariance satisfied by the median, i.e., for any constant $k \in \mathbb{R}$, the following two equalities are satisfied, $Q_{0.5}(k \cdot y_{ij}) = k \cdot Q_{0.5}(y_{ij})$ and $Q_{0.5}(y_{ij}+k) = Q_{0.5}(y_{ij}) + k$.

Let consider the set of initial values $\hat{\theta}^0 = (\hat{a}^0, \hat{b}^0, \hat{c}^0, \hat{d}^0)$. In the iteration v, parameters are updated as follows:

1. Given $\hat{\theta}^v$, estimate the parameters γ_{a_i} and γ_{b_j} fitting $Q_{0.5}\left(y_{ij}^v\right) = \gamma_{a_i} \cdot \hat{a}_i^v + \gamma_{b_j} \cdot \hat{b}_j^v$, where $y_{ij}^v = y_{ij} - \hat{c}_i^v \cdot \hat{d}_j^v$. Update $\hat{a}_i^{v+1} = \hat{\gamma}_{a_i} \cdot \hat{a}_i^v$ and $\hat{b}_j^{v+1} = \hat{\gamma}_{b_j} \cdot \hat{b}_j^v$, $\hat{c}^{v+1} = \hat{c}^v$ and $\hat{d}^{v+1} = \hat{d}^v$.

2. Given $\hat{\theta}^{v+1}$, estimate the parameter γ_{c_i} fitting $Q_{0.5}\left(y_{ij}^{v+1}\right) = \gamma_{c_i} \cdot \hat{c}_i^{v+1}$, where $y_{ij}^{v+1} = \frac{y_{ij} - \hat{a}_i^{v+1} - \hat{b}_j^{v+1}}{\hat{d}_j^{v+1}}$. Update $\hat{c}_i^{v+2} = \hat{\gamma}_{c_i} \cdot \hat{c}_i^{v+1}$, $\hat{a}^{v+2} = \hat{a}^{v+1}$, $\hat{b}^{v+2} = \hat{b}^{v+1}$ and $\hat{d}^{v+2} = \hat{d}^{v+1}$.

3. Given $\hat{\theta}^{v+2}$, estimate the parameter γ_{d_j} fitting $Q_{0.5}\left(y_{ij}^{v+2}\right) = \gamma_{d_j} \cdot \hat{d}_j^{v+2}$, where $y_{ij}^{v+2} = \frac{y_{ij} - \hat{a}_i^{v+2} - \hat{b}_j^{v+2}}{\hat{c}_i^{v+2}}$. Update $\hat{d}_j^{v+3} = \hat{\gamma}_{d_j} \cdot \hat{d}_j^{v+2}$, $\hat{a}^{v+3} = \hat{a}^{v+2}$, $\hat{b}^{v+3} = \hat{b}^{v+2}$ and $\hat{c}^{v+3} = \hat{c}^{v+2}$.

4. If $\sum_{i,j}\left(|\varepsilon_{ij}(\hat{\theta}^{v+3}| - |\varepsilon_{ij}(\hat{\theta}^v|\right) \leq \eta$ then *stop*, where η is the tolerance value. Otherwise, $\hat{\theta}^v = \hat{\theta}^{v+3}$ and move to step 1.

4.2.1. Calibration of a median linear regression

With the application of this strategy, the problem of estimating a median bilinear regression is converted into a problem of estimating iteratively a sequence of three median linear regressions. A median linear regression in matrix notation may be expressed as $Q_{0.5}(Y) = X^{\intercal}\gamma$, where Y is the response vector, γ is the set of parameters to estimate and X is the design matrix. At each step, the following optimization problem has to be resolved:

$$\min_{\gamma, u \geq 0, v \geq 0} 0.5\mathbf{1}_N^{\intercal} u + 0.5\mathbf{1}_N^{\intercal} v \qquad s.t \quad X^{\intercal}\gamma + u - v = Y. \tag{8}$$

Different methods may be applied to estimate the parameters. We briefly describe two estimation methods which are the Mehrotra's Predictor-Corrector method (Port-

noy and Koenker, 1997) and the likelihood-based approach (Machado and Silva, 2011; Sánchez et al., 2013).

Mehrotra's Predictor-Corrector method

Alternative algorithms for linear programs with bounded variables may be used to solve (8). A widely used algorithm is the Mehrotra's Predictor-Corrector (MPC) method described in Mehrotra (1992). As the affine scale algorithm, the MPC algorithm belongs to the class of point interior methods. The MPC method is an appropriate algorithm to solve the canonical linear program: $\min\{c^{\intercal}x : Ax = b, \quad x \geq 0\}$, where $A \in \mathbb{R}^{mxN}$, y, $b \in \mathbb{R}^m$ and c, x, $s \in \mathbb{R}^N$, and its dual problem, $\max\{b^{\intercal}y : A^{\intercal}y + s = c, \quad s \geq 0\}$. The MPC method finds the joint solution of the primal and dual equations (Salahi, Peng and Terlaky, 2008).

The dual optimization problem of (8) is $\max\left\{y^{\intercal}s : X^{\intercal}s = 0, \quad s \in [-0.5, 0.5]^N\right.$, where s are the multipliers of Lagrange. Setting $a = s + 0.5$, the maximization problem is converted to $\max\left\{y^{\intercal}a : X^{\intercal}a = (0.5)X^{\intercal}\mathbf{1}_N, \quad a \in [0,1]^N\right.$. Changing the sign of y, it becomes a minimization problem which fits in the setting of the canonical linear program in which the use of MPC method is appropriate.

Maximum likelihood

The likelihood-based approach is based on the asymmetric Laplace distribution to replicate the optimization problem (8). Suppose that the response variable y_l follows an asymmetric Laplace distribution with location parameter $x_l^{\intercal}\gamma$, scale parameter σ and skewness parameter α, where x_l is the l row of the design matrix, with $l = 1, \ldots, N$. The likelihood function is

$$L(\gamma, \sigma) = \frac{\alpha^N (1-\alpha)^N}{\sigma^N} \exp\left\{ -\sum_{l=1}^{N} \rho_\alpha \left(\frac{y_l - x_l^{\intercal}\gamma}{\sigma} \right) \right\},$$

where the loss function is defined as $\rho_\alpha(r_l) = r_l(\alpha - I_{r_l})$ for $\alpha \in (0,1)$, I_{r_l} is an indicator function such that $I_{r_l} = 1$ if $r_l < 0$ and zero otherwise. Note that for $\alpha = 0.5$, if σ is considered a nuisance parameter, the maximization of the $L(\gamma, \sigma)$ is equivalent to minimize the objective function (8). Sánchez et al. (2013) describe the steps to obtain the ML estimates based on the expectation-maximization (EM) algorithm.

5. Results

In this section it is compared the performance of mean bilinear models and median bilinear models in presence of extreme values in two different contexts. We illustrate the use of these models with a simulated database and in a real application to the Spanish mortality data. The parameters of the mean bilinear regression model were estimated by least squares (Mean SVD) and also by maximum likelihood (Mean MV). Median bilinear regression models were estimated by the method A and the method B. In the case

of the method A, the parameters were estimated by the affine scaling method (Med A-AS). In the case of the method B, we apply the interior point method with the Mehrotra's Predictor-Corrector algorithm (Med B-MP) and the maximum likelihood approach based on the asymmetric Laplace distribution (Med B-MV) to calibrate the model.

All results were calculated in R (R Core Team, 2020). The estimates of the mean bilinear model by maximum likelihood were obtained by means of the *gnm* package (Turner and Firth, 2018) that applies the iterative process in which the linear terms are updated by reweighted least squares (Turner and Firth, 2018; Dutang, 2017). The *nls* function in the *stats* package may be also used to fit a mean bilinear model by a iterative process to minimize least square errors. Median regression models may be estimated by interior point methods with R package *quantreg* (Koenker, 2019), but some implemented functions can only deal with full-rank design matrices. We use the function *rq.fit.fnb* of the package *quantreg* and a version of the function *nlrq* available in Koenker (2020) to calibrate median regression models based on the MPC method and on the Meketon algorithm. Finally, the R package *ALDqr* can be used to estimate median linear models by maximum likelihood (Sánchez et al., 2013). We modify the function *EM.qr* of this package to deal with sparse matrices. The data and code used in the data analysis are publicly available on GitHub (FMBM, 2021)

5.1. Simulation

For illustrative purposes a simulated dataset with extreme values is used for the estimation of median bilinear models. We simulate a database generated by the model (1) in case that the error is normally distributed and there are shocks involving extreme outcomes. Let consider the response variable y_{ij} is generated by $y_{ij} = a_i + b_j + c_i \cdot d_j + \varepsilon_{ij}$, where $\varepsilon_{ij} \sim N(0, 0.05)$. The first factor a has 50 levels, $(i \in \{1, \ldots, 50\})$, and the second b has 40 levels, $(j \in \{1, \ldots, 40\})$. The description of coefficients used in the simulation are shown in Table (1).

Table 1. *Descriptive statistics of simulated data.*

	Min.	1st quartile	Median	Mean	3rd quartile	Max.
a_i	1.34	2.95	3.78	3.72	4.85	5.97
b_j	0.00	0.16	0.27	0.31	0.36	0.72
c_i	-2.36	-1.02	-0.42	-0.02	0.2	5.48
d_j	-0.03	-0.02	0.00	0.00	0.01	0.08
y_{ij}	1.32	2.88	4.10	4.04	5.18	7.08

Now we incorporate the extreme outcomes (shocks) to the simulated data. Suppose that the response variable y_{ij} is affected by shocks as follows, $y^s_{ij} = y_{ij} + B \cdot U$, where B is a bernoulli variable that takes 1 with probability p and U is a discrete random variable that takes values $\{-8, -6, 6, 8\}$ with probability 0.25 each one. We consider five different scenarios in relation to the frequency of shocks, $p = (0, 0.01, 0.025, 0.05, 0.1)$,

that is, scenario without shocks ($p = 0$), scenario in which the 1% of observations are extreme values ($p = 0.01$), scenario with 2.5% of extreme values ($p = 0.025$), scenario with 5% of extreme values ($p = 0.05$) and scenario with 10% of extreme values ($p = 0.1$).

The mean bilinear and the median bilinear models are fitted to the simulated data in the five scenarios. The sum of squared errors (SSE) and the sum of absolute errors (SAE) are computed for each fitted bilinear model in the scenarios with extreme observations. In order to evaluate the influence of extreme values on estimates, the bilinear models calibrated in the scenarios with extreme observations are also used to compute the statistics of fit for the simulated data without shocks. Results are shown in Table 2.

Table 2. *Statistics of fit of the mean and the median bilinear models.*

		Without shocks	1% shocks		2.5% shocks		5% shocks		10% shocks	
			y^s_{ij}	y_{ij}	y^s_{ij}	y_{ij}	y^s_{ij}	y_{ij}	y^s_{ij}	y_{ij}
Mean SVD	SAE	75.83	338.38	277.75	925.01	670.65	1620.63	1082.47	2803.60	1825.38
	SSE	4.54	730.60	245.61	2686.90	652.01	5620.11	1313.34	10734.41	2905.55
Mean MV	SAE	75.83	338.38	277.75	925.01	670.65	1616.85	1081.19	2803.27	1825.07
	SSE	4.54	730.60	245.61	2686.90	652.01	5600.56	1313.74	10734.41	2905.64
Med A-AS	SAE	75.80	306.88	238.60	922.89	667.86	1504.16	887.85	2707.61	1543.91
	SSE	4.55	852.07	318.87	2687.99	647.77	6102.31	1167.75	12326.93	2613.01
Med B-MP	SAE	74.21	193.53	74.38	487.85	74.79	932.92	75.55	1767.64	77.59
	SSE	4.74	961.50	4.77	3320.13	4.77	6893.69	4.89	13581.87	5.20
Med B-MV	SAE	75.53	195.57	76.91	493.22	82.09	939.44	84.18	1712.79	193.47
	SSE	4.73	956.70	5.01	3300.06	5.64	6872.17	6.01	12848.81	646.84

If we focus on the performance of fitted models in the scenario without shocks (second column of Table 2), as expected, the lowest SSE is observed for the mean bilinear models, and the lowest SAE for the median bilinear models. In fact, this behaviour is repeated in the scenarios with extreme observations when the statistics of fit were computed for the simulated data with the shocks (y^s_{ij}).

However, this conclusion varies when the estimated bilinear models fitted in the scenarios with extreme values are analysed in the scenarios without shocks (y_{ij}). A lower SSE associated to mean bilinear models is not longer observed when the shocks are removed from the simulated data and the statistics of fit are computed again. Now, the fitted median bilinear models show a lower SAE in all scenarios and also a lower SSE in almost all scenarios in comparison with the fitted mean bilinear models. In particular, the performance of the two median bilinear models fitted by method B is clearly better than the performance of the fitted mean bilinear models, and it is also higher than that of the median bilinear model estimated by method A. Comparing between the two median bilinear models fitted by method B, the MPC method seems to provide more stable estimates when the number of extreme values increases. Finally, almost identical outcomes were obtained with the two methods of coefficient estimation for the mean bilinear model.

5.2. Application to mortality data

The second example uses Spanish mortality data to illustrate the application of bilinear models in presence of extreme values. One of the most influential approaches to the stochastic modelling of mortality rates is the parametric nonlinear regression model introduced by Lee and Carter (1992). The Lee-Carter model proposes estimating the conditional mean mortality rate as the nonlinear combination of age and calendar year parameters. Santolino (2020) adopts the Lee-Carter framework to estimate the conditional quantile mortality rate.

The Lee-Carter modelling fits in the setting of the bilinear models defined in (1) in which the main effect of level j (calendar year) is equal to zero, i.e. the response variable (log of the mortality rate) is regressed by the main effect of level i (age) and the interaction between levels i and j. We here estimate the Lee-Carter mean mortality model and the Lee-Carter median mortality model for the Spanish male population. The number of deaths observed, exposures and central mortality rates for the Spanish population by gender were obtained directly from the Human Mortality Database (HMD, 2020). Mortality information is available for ages between 0 and 110, but the number of observations is ineluctably small at the extreme ages and patterns at very advance ages are difficult to observe (Robine et al., 2007). We select ages between 0 and 100 years, which is a common practice in demographics.

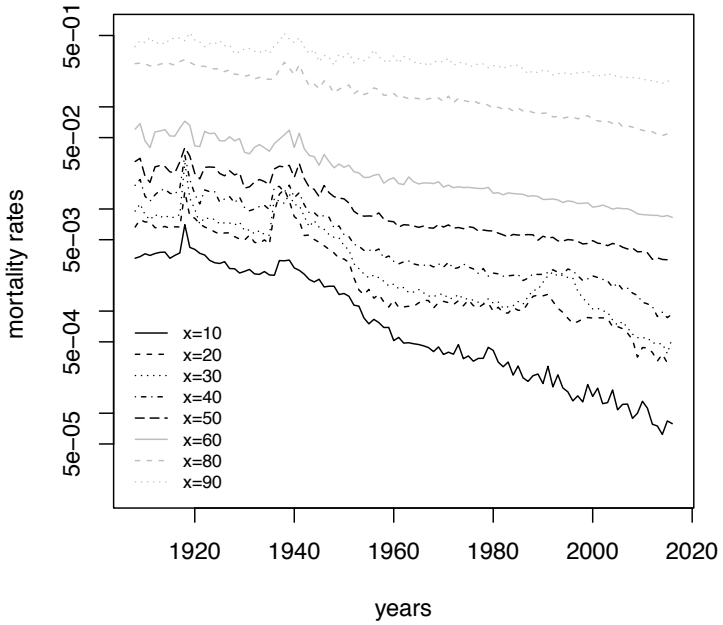


Figure 1. *Mortality rates of Spanish male population at different ages over 1908–2016 period.*

The mortality data cover the observation period between the years 1908 and 2016. Social progress has led to a notable reduction in mortality of the Spanish population

through this period. However, there are three spans of time in which the decreasing trend is disrupted, namely, the Spanish flu, the Spanish Civil War and HIV/AIDS. The Spanish flu was a severe influenza pandemic with deadly consequences in 1918 and the following four years (Carreras and Tafunell, 2005). The Spanish Civil War took place between 1936 and 1939. The postwar era formally ended in 1953 with the signing of the US economic agreement (Pact of Madrid). During the war and the first half of the postwar period, poverty and malnutrition affected remarkably the mortality (Jiménez Lucena, 1994). Finally, mortality associated with HIV dramatically increased during the late 80s and 90s, particularly in middle-aged population (CNE, 2011). Figure 1 shows Spanish male mortality rates at different ages in the period 1908–2016, in which these three peaks in the mortality rate are sharply appreciated.

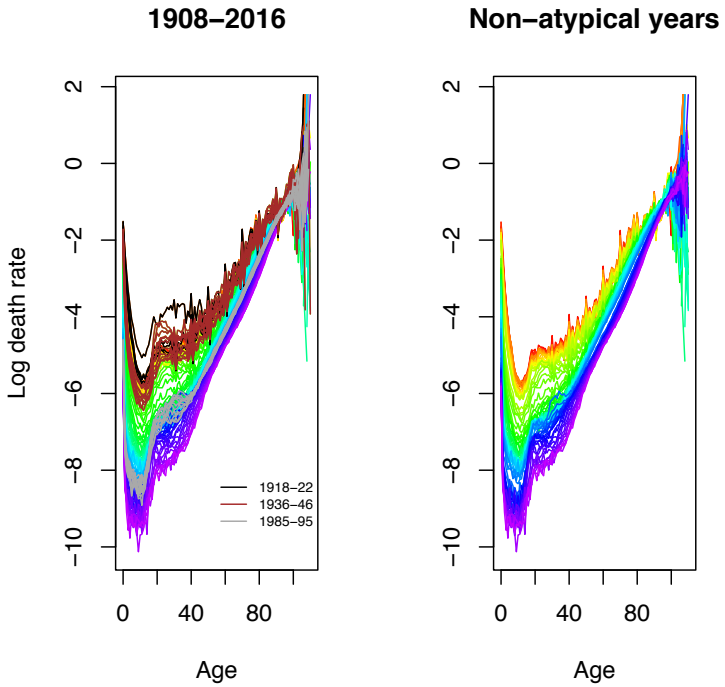


Figure 2. *Mortality rates at different years of Spanish male population.*

The impact of years involving atypically high mortality rate values may be observed in Figure 2. Mortality rates (in log scale) at ages between 0 and 100 are showed for all years in Figure 2 (left). Each line corresponds to the log mortality rates at 0-100 ages in a particular calendar year, and calendar years are differentiated by colors. In case of a continuous reduction in mortality rates over time through the 1908-2016 period, the colored lines should not overlap themselves and they do it. In Figure 2 (right) the years belonging to the time intervals 1918–1922, 1936–1946 and 1985–1995 are not represented. Note that, when these atypical years are removed, the lines seldom overlap themselves.

The mean and median bilinear models are fitted to the Spanish mortality data. The models are calibrated with data involving all calendar years. The measures of goodness-of-fit are computed for the whole sample and when years belonging to the time intervals 1918–1922, 1936–1946 and 1985–1995 are excluded. The sum of squared errors and absolute errors are shown for each fitted bilinear model in Table 3.

Table 3. *Statistics of fit of bilinear models fitted to Spanish male mortality data.*

	All years		Without atypical years	
	SAE	SSE	SAE	SSE
Mean SVD	1279.07	272.76	908.44	172.35
Mean MV	1279.07	272.76	908.44	172.35
Med A-AS	1226.87	300.96	843.84	171.63
Med B-MP	1227.25	301.49	842.90	170.79
Med B-MV	1235.09	288.29	857.19	169.49

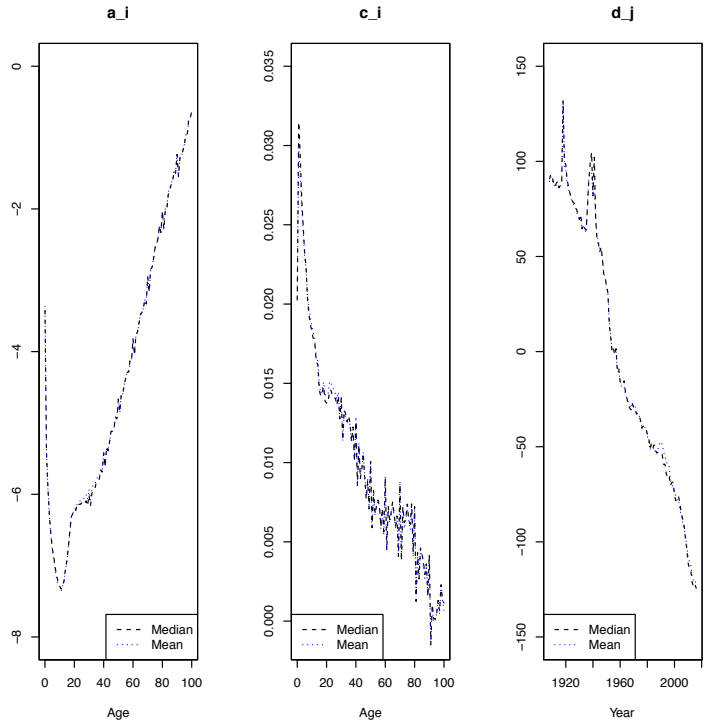


Figure 3. *Coefficient estimates of the Mean MV and Med A-AS models.*

When the goodness-of-fit statistics are computed for the whole sample, the fitted mean bilinear models have lower SSE values and higher SAE values compared to the fitted median bilinear models. Whether or not atypical years are considered in the computation of the statistics of fit, the fitted median bilinear models show lower SSE and SAE values than the fitted mean bilinear models. The performance of the median bilin-

ear models calibrated by the affine scaling method and the MPC method is very similar. Among the fitted median bilinear models, the median bilinear model calibrated by maximum likelihood shows the highest SAE value and the lowest SSE value. Comparing the calibration methods of the mean bilinear model, the same results are obtained with the two methods of coefficient estimation.

A comparison of coefficient estimates of Mean MV and Med A-AS models is provided in Figure 3. Small differences in coefficient estimates are observed for ages in the 20–40 interval and years in time intervals 1936–1946 and 1985–1995.

In mortality applications it is also important to evaluate the prediction power of models. Backtesting is applied to evaluate the prediction accuracy of mean and median models for annual periods up to five years. Alternatively, resampling methods could be used to analyse the prediction power of stochastic mortality models (Atance, Debón and Navarro, 2020). The sum of absolute prediction errors (SAPE) and the sum of squared prediction errors (SSPE) are shown for each bilinear model in Table 4. Median bilinear models show lower SAPE values in all cases and also lower SSPE values in the four-year and five-year forecasting periods (2013–2016 and 2012–2016, respectively).

Table 4. *Backtesting to evaluate prediction power of bilinear models for different periods of forecasting.*

	2016		2015–2016		2014–2016		2013–2016		2012–2016	
	SAPE	SSPE	SAPE	SSPE	SAPE	SSPE	SAPE	SSPE	SAPE	SSPE
Mean SVD	15.14	3.82	28.02	6.90	45.60	11.76	67.07	17.31	82.39	21.04
Mean MV	15.14	3.82	28.02	6.90	45.60	11.76	67.07	17.31	82.39	21.04
Med A-AS	14.52	4.52	27.47	8.89	42.98	13.35	59.46	17.22	73.31	20.83
Med B-MP	14.44	4.51	27.51	8.69	42.89	13.27	59.74	17.30	73.07	20.88
Med B-MV	14.29	4.20	26.99	8.11	42.75	12.75	60.96	17.16	75.00	21.16

6. Conclusions

Conditional mean bilinear regression models have been broadly used in many research fields. In many of the contexts that mean bilinear models are applied, data have extreme observations. It is know that in presence of extreme values the mean may be an inaccurate statistic to reflect the centre of the conditional distribution. In this article we have compared the performance of the mean bilinear model and the median bilinear model in different contexts involving extreme observations.

In the bilinear modelling the multiplicatively interaction structure is specified as a nonlinear term. Alternative methods of parameter estimation for nonlinear regressions are applied. The mean bilinear model is estimated by lest squares and maximum likelihood. The method of parameter estimation for nonlinear median regression involving the linearization of the objective function is compared with the calibration strategy of

the median bilinear model in which coefficients are estimated by an iterative process of a sequence of median linear regressions. This second calibration strategy was first used by Santolino (2020) and here it is generalized to the median bilinear model setting.

Mean and median bilinear models are compared in two applications involving extreme values. The first application deals with simulated data with extreme values. The second application is illustrated by means of mortality data of the Spanish population over the 1908–2016 period. During this period, there were a set of years with a particular high mortality (Spanish flu, civil war and HIV/AIDS). Statistics of goodness-of-fit were compared. The fitted median bilinear models showed the lowest sum of absolute errors and the fitted mean bilinear models the lowest sum of square errors. However, when observations with extreme values were removed, the fitted median bilinear models showed the lowest values in the two statistics of goodness-of-fit. This result would confirm that the estimated median is a more appropriate statistic to reflect the centre of the conditional distribution than the estimated mean in these two applications. In the context of COVID-19 using median rather mean approaches when estimating mortality models may be relevant due to the unusual data points arising in 2020 and 2021.

Analysing the two calibration strategies of the median bilinear regression model, we found that the strategy involving the sequence of median linear regressions performed clearly better than the strategy associated to the linearization of the objective function in the application with simulated data and similarly in the application with mortality data.

We conclude that the application of the median bilinear model may be more appropriate than the mean bilinear model in presence of extreme values, whether the centre of the conditional distribution is of interest. Parameters of the median bilinear model may be easily estimated by means of calibrating sequentially median linear regressions. These concluding remarks are relevant in fields such as the stochastic mortality modelling in which researchers have to deal often with data involving extreme observations (wars, pandemics, natural disasters, famines, etc.), and, in general, in any context of application of bilinear models in which the presence of extreme values is frequent.

Acknowledgments

The author thanks Daniel Peña, Montserrat Guillen and participants of the 2020 IREA-UB Annual Seminar for their useful comments. The author acknowledges support received from the Spanish Ministry of Science and Innovation under Grant PID2019-105986GB-C21 and from the Catalan Government under Grant 2020-PANDE-00074.

References

Anderson, J.A. (1984). Regression and ordered categorical variables (with Discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, 47, 203–210.

Atance, D., Debón, A., and Navarro, E. (2020) A comparison of forecasting mortality models using resampling methods. *Mathematics*, 8, 1550.

Bassett, G. and Koenker, R. (1978). Theory of least absolute error regression. *Journal of the American Statistical Association*, 73, 618–622.

Black, F., Jensen, M.C. and Scholes, M. (1972). The capital asset pricing model: some empirical tests. *Studies in the theory of capital markets*, New York, Praeger: 79–121.

Brouhns, N., Denuit, M., and Vermunt, J. (2002). A Poisson log-bilinear regression approach to the construction of projected life table. *Insurance: Mathematics and Economics*, 31, 373–393.

Buchinsky, M. (1995). Estimating the asymptotic covariance matrix for quantile regression models a Monte Carlo study. *Journal of Econometrics*, 68, 303–338.

Carreras, A. and Tafunell, X.(2005). *Estadísticas Históricas de España: siglos XIX-XX*. Fundación BBVA, 2a Edic., Bilbao.

CNE (2011). *Área de vigilancia de VIH y conductas de riesgo. Mortalidad por VIH/Sida en España, año 2009. Evolución 1981-2009*. Centro Nacional de Epimediología, Secretaría del Plan Nacional Sobre el Sida, Gobierno de Espaã.

Croux, C., Filzmoser, P., Pison, G. and Rousseeuw, P.J. (2003). Fitting multiplicative models by robust alternating regressions. *Statistics and Computing* , 13, 23–36.

Denis, J.B. and Pázman, A.(1999). Bias of LS estimators in nonlinear regression models with constraints. Part II: Biadditive models. *Applications of Mathematics*, 44, 375–403.

Dutang, C. (2017). Some explanations about the IWLS algorithm to fit generalized linear models. *Technical report*, hal-01577698.

El-Attar, R. A., Vidyasagar, M., and Dutta, S. R. K. (1979). An Algorithm for l_1-norm minimization with application to nonlinear l_1-approximation. *SIAM Journal on Numerical Analysis*, 16, 70–86.

Emerson, J.D. and Hoaglin. D.C. (1983). Analysis of two-way tables by medians. In D. C. Hoaglin, F. Mosteller and J. W. Tukey (Eds.), *Understanding Robust and Exploratory Data Analysis*, 165–210, New York City: John Wiley and Sons.

Erikson, R. and Goldthorpe, J.H. (1992). *The Constant Flux: A Study of Class Mobility in Industrial Societies*. Oxford: Clarendon Press. .Ch. 3.

FMBM (2021). *Fitting Median Bilinear Model*. Available at https://github.com/msantolino/Median-Bilinear-Models. Accessed 8 November 2021.

Gabriel, K.R. (1978). Least squares approximation of matrices by additive and multiplicative models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40, 186–196.

Gabriel, K. R. and Odoroff, L. (1984). Resistant lower rank approximation of matrices. In E. Diday, M. Jambu, L. Lebart, J. Pages and R. Tomassone (Eds.), *Data analysis and informatics III*, 23–40, Amsterdam: North-Holland.

Goodman, L.A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association*, 74, 537–552.

Goodman, L.A. (1981). Association models and canonical correlation in the analysis of cross- classifications having ordered categories. *Journal of the American Statistical Association*, 76, 320–334.

Hawkins, D.M. (1980). *Identification of outliers*. Monographs on applied probability and statistics, Chapman & Hall.

HMD (2020). *Human Mortality Database. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany)*. Available at www.mortality.org. University of California, Accessed 15 January 2020.

Jiménez Lucena, I. (1994). El tifus exantemático de la posguerra española (1939–1943): el uso de una enfermedad colectiva en la legitimación del 'Nuevo estado'. *Dynamis : Acta Hispanica ad Medicinae Scientiarumque. Historiam Illustrandam*, 14, 185–198.

Justel, A., Peña, D. and Tay, R.S (2001). Detection of outlier patches in autoregressive time series. *Statistica Sinica*, 11, 651–673.

Koenker, R. (2019). *quantreg: Quantile regression*. R package version 5.42.

Koenker, R. (2020). Non linear quantile regression. http://www.econ.uiuc.edu/~roger/research/nlrq/nlrq.html, Accessed 16 February 2021.

Koenker, R. and Park, B. J. (1996). An interior point algorithm for nonlinear quantile regression. *Journal of Econometrics*, 71, 265–283.

Lee, R. D. and Carter, L. R. (1992). Modeling and forecasting U. S. mortality. *Journal of the American Statistical Association*, 87, 659–671.

Machado, J. and Silva, J. S. (2011). MSS: Stata module to perform heteroskedasticity test for quantile and OLS regressions. Statistical Software Components, Boston College Department of Economics.

Macias, Y. and Santolino, M. (2018). Application of Lee-Carter and Renshaw-Haberman models in life insurance products. *Anales del Instituto de Actuarios Españoles*, 24, 53–78.

Mehrotra, S. (1992). On the implementation of a primal−dual interior point method. *SIAM Journal on Optimization*, 2, 575–601.

Moyano-Silva, P.A., Pérez-Marín, A.M. and Santolino, M. (2020). Estimation of stochastic mortality models for Chile. *Anales del Instituto de Actuarios Españoles*, 4, 225–256.

Portnoy, S. and Koenker, R. (1997). The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators. *Statistical Science*, 12, 279–300.

R Core Team, (2020) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Renshaw, A. and Haberman, S. (2006). A cohort-based extension to the Lee-Carter model for mortality Reduction Factors. *Insurance: Mathematics and Economics*, 38:556–570.

Robine, J.M, Crimmins, E.M, Horiuchi S. and Zeng Yi, Z.(2007). *Human Longevity, Individual Life Duration, and the Growth of the Oldest-Old Population*. International Studies in Population, Springer.

Salahi, M., Peng, J. and Terlaky, T. (2008) On Mehrotra-type predictor-corrector algorithms. *SIAM Journal on Optimization*, 18, 1377–1397.

Sánchez, B., Labros, H. and Labra, V. (2013). Likelihood based inference for quantile regression using the asymmetric Laplace distribution. *Journal of Statistical Computation and Simulation*, 81, 1565–1578

Santolino, M. (2020). The Lee-Carter quantile mortality model. *Scandinavian Actuarial Journal*, 7, 614–633

Turner, H. and Firth, D. (2018). *Generalized nonlinear models in R: an overview of the gnm package*. R package version 1.1-0.

Turner, H., Firth, D. and Kosmidis, I. (2013). Generalized nonlinear models in R. *6th International Conference of the ERCIM WG on Computational and Methodological Statistics, ERCIM*, London.

Van Eeuwijk, Fred A. (1992). Multiplicative models for genotype-environment interaction in plant breeding. *Statistica Applicata*, 4, 393–406.

Van Eeuwijk, Fred A. (1995). Multiplicative interaction in generalized linear models. *Biometrics* , 51, 1017–032.

Wilmoth, J. (1993). *Computational Methods for Fitting and Extrapolating the Lee-Carter Model of Mortality Change*. Technical Report. Department of Demography. University of California.

Xie, Y. (1992). The log-multiplicative layer effect model for comparing mobility tables. *American Sociological Review*, 57, 380–395.

Yates, F. and Cochran, W.G. (1938). The analysis of groups of experiments. *The Journal of Agricultural Science*, 28, 556–580.

SORT 45 (2) July-December 2021, 181-200 DOI: 10.2436/20.8080.02.115

Exponentiated power Maxwell distribution with quantile regression and applications

Francisco A. Segovia[1], Yolanda M. Gómez[2] and Diego I. Gallardo[3]

Abstract

In this paper we introduce an extension of the power Maxwell distribution. We also discuss a reparametrized version of this model applied to quantile regression. Some properties of the model and estimation based on the maximum likelihood estimation method are studied. We also present a simulation study to assess the performance of estimators in such finite samples, and two applications to real data sets to illustrate the model.

MSC: 62E10, 62J02.

Keywords: Maxwell distribution, exponentiated distributions, maximum likelihood, quantile regression.

1. Introduction

Lehmann (1953) and Durrans (1992) introduced a family of distributions named exponentiated distributions. Their cumulative distribution function (CDF) is defined as

$$\varphi_F(w;\gamma) = F(w)^\gamma, \quad w \in \mathbb{R}, \gamma > 0 \tag{1}$$

where $F(w)$ is the CDF for a certain random variable. It follows directly that the probability density function (PDF) is

$$\varphi_f(w;\gamma) = \gamma f(w) F(w)^{\gamma-1}, \tag{2}$$

where $f(w)$ is the PDF related to $F(w)$. Durrans (1992) considered this methodology by using the normal distribution, i.e., $F = \Phi$ and $f = \phi$, the normal CDF and PDF of

[1] Departamento de Matemática, Facultad de Ingeniería, Universidad de Atacama, Copiapó, Chile.

[2] Departamento de Matemática, Facultad de Ingeniería, Universidad de Atacama, Copiapó, Chile, e-mai: yolanda.gomez@uda.cl

[3] Departamento de Matemática, Facultad de Ingeniería, Universidad de Atacama, Copiapó, Chile. e-mail: diego.gallardo@uda.cl

Received: March 2021
Accepted: November 2021

the standard normal distribution, respectively. This model was also discussed in more detail in Gupta and Gupta (2007), Pewsey, Gómez and Bolfarine (2012) and Rêgo, Cintra and Cordeiro (2012). Gupta and Kundu (1999) used this methodology to introduce the generalized exponential distribution, setting $F(w)$ as the CDF of the exponential model. Gómez and Bolfarine (2015) consider the case where $F(w)$ is the CDF of a half-normal distribution, resulting in a distribution which belongs to the family of beta generalized half-normal distributions. Other extensions using this methodology include the exponentiated Weibull (Mudholkar and Srivastava, 1993; Mudholkar, Srivastava and Freimer, 1995), the exponentiated Pareto (Gupta, Gupta and Gupta, 1998), exponentiated Gumbel (Nadarajah, 2005), exponentiated log-normal (Kakde and Shirle, 2006), exponentiated gamma (Nadarajah and Gupta, 2007) and power piecewise exponential (Gómez, Gallardo and Arnold, 2017). The Maxwell (M) distribution was proposed by Maxwell (1860) in order to model velocities among gas molecules. Maxwell's research was generalized by Boltzmann (1871a,c,b), to develop the distribution of energies among molecules. This distribution has diverse applications in the areas of physics, chemistry, and physical chemistry, (see Dunbar (1982)). Singh et al. (2018) introduced the power Maxwell (PM) distribution, based on taking the power of a random variable that has Maxwell distribution. Segovia et al. (2020) introduced the slashed power Maxwell (SPM) distribution and use it for outlier data modelling. However they do not use those extensions of the PM distribution considering a regression structure. We consider the specific parametrization considered in Huang and Chen (2015), where the CDF and PDF of the variable are given by

$$F_W(w;\psi,\beta) = G\left(\frac{w^{2\beta}}{2\psi^2}, \frac{3}{2}\right), \quad w \geq 0 \tag{3}$$

$$f_W(w;\psi,\beta) = \frac{4\beta}{(2\psi^2)^{3/2}\sqrt{\pi}} w^{3\beta-1} \exp\left\{-\frac{1}{2\psi^2} w^{2\beta}\right\},$$

respectively, where $\psi, \beta > 0$, and $G(\cdot, a)$ denotes the CDF for the gamma distribution with shape and scale parameters equal to a and 1, respectively. On the other hand, Galarza et al. (2017) used the skewed distributions family (SKD) in order to introduce quantile regression, where one parameter represents the quantile of the distribution. Gómez et al. (2019) introduced the Gamma-sinh Cauchy (GSC) distribution aiming at applying the model to quantile regression. The resulting model can be either unimodal or bimodal depending on the combinations of two parameters, where one of them is fixed and depends on the modelled quantile. Gallardo et al. (2020a) introduced a novel parametric quantile regression model for asymmetric response variables, where the response variable follows a power skew-normal distribution. Gallardo, Gómez-Déniz and Gómez (2020b) presented a discrete distribution by discretizing a generalized half-normal distribution, which can be reparametrized for use in a regression model based on the median. Sánchez et al. (2020) use a model based on the Birnbaum-Saunders distribution in order to perform quantile regression.

The aim of this paper is to introduce an extension of the PM distribution using the methodology presented in equation (1), aiming to perform quantile regression. The resulting PDF can be either strictly increasing or unimodal. The manuscript is organized as follows. In Section 2 we introduce the exponentiated power Maxwell (EPM) distribution, and we propose the reparametrized EPM (REPM) distribution with some properties such as its CDF, hazard function (HF) and moments. In Section 3, we discuss the inference for the REPM regression model based on the maximum likelihood (ML) estimation. In Section 4 we present a simulation study in finite samples, focusing our attention on parameter recovery. In Section 5 we present two applications to real data, fitting the REPM distribution to two real data sets. Finally, in Section 6 we present the main conclusions of the work.

2. Exponentiated power Maxwell distribution

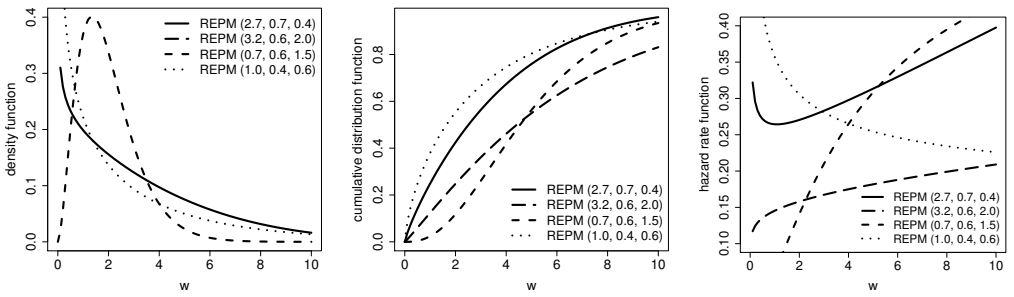
Following the methodology related to equation (1), we introduce the following extension of the PM model.

Definition 1. *A random variable W follows an exponentiated power Maxwell distribution with scale parameter ψ and shape parameters β and γ, if its CDF, PDF and HF are given, respectively, by:*

$$\begin{aligned}
F_Y(w;\psi,\beta,\gamma) &= \left[G\left(\frac{w^{2\beta}}{2\psi^2},\frac{3}{2}\right)\right]^{\gamma}, \quad w>0 \\
f_Y(w;\psi,\beta,\gamma) &= \gamma\left[G\left(\frac{w^{2\beta}}{2\psi^2},\frac{3}{2}\right)\right]^{\gamma-1}\frac{\beta w^{2\beta-1}}{\psi^2}\, g\left(\frac{w^{2\beta}}{2\psi^2},\frac{3}{2}\right), \quad w>0, \\
h_W(w;\psi,\beta,\gamma) &= \frac{\gamma\left[G\left(\frac{w^{2\beta}}{2\psi^2},\frac{3}{2}\right)\right]^{\gamma-1}\beta w^{2\beta-1}\, G\left(\frac{w^{2\beta}}{2\psi^2},\frac{3}{2}\right)}{\psi^2\left\{1-\left[G\left(\frac{w^{2\beta}}{2\psi^2},\frac{3}{2}\right)\right]^{\gamma}\right\}}, \quad w>0
\end{aligned} \tag{4}$$

where $\psi,\beta,\gamma>0$ and $g(\cdot,a)$ is the PDF related to $G(\cdot,a)$.

In Figure 1, we illustrate the PDF, CDF, and HF of the REPM distribution. It is interesting to point out that the HF can be strictly increasing, strictly decreasing, or have a bathtub shape. The equation for finding the mode is immediately obtained from calculating the first derivative of the density. However, we consider a parametrization for this model based on (μ,β,γ), where $\mu=\psi^{\frac{1}{\beta}}$. We denote this as REPM$(\mu,\beta,\gamma)$. The main object of this parametrization will be justified later.



(a) **(b)** **(c)**

Figure 1. *Plots of the PDF (a), CDF (b) and HF (c) for different combinations of parameters of the REPM*(ψ,β,γ) *distribution.*

Proposition 1. *If* $W \sim REPM(\mu,\beta,\gamma)$*, the rth non-central moment of* W *can be calculated as*

$$E(W^r) = \int_0^1 \frac{1}{\beta g\left(\dfrac{w^{2\beta}}{2\mu^{2\beta}}, \dfrac{3}{2}\right)} r w^{r-2\beta} \mu^{2\beta} (1-u)^{\gamma} du$$

for $r \geq 1$*, where* $w = [2\mu^{2\beta} G^{-1}(u, 3/2)]^{1/(2\beta)}$*,* G^{-1} *is the inverse function of* $G(\cdot, a)$*.*

Proof. By using the definition of expectation and making the substitution $u = G\left(\dfrac{w^{2\beta}}{2\psi^2}, \dfrac{3}{2}\right)$, the result is immediate ■.

The gamma distribution is very useful to express both the CDF and the PDF of the REPM distribution. However, usual quantities of interest such as he mean and mode of the model do not have closed form. Therefore, in order to perform regression analysis in the model, other alternatives should be studied, as we illustrate in the following proposition.

Proposition 2. *If* $W \sim REPM(\mu,\beta,\gamma)$*, then* $100 \times \rho$*-th, the* ρ*-th quantile* $0 < \rho < 1$*, is given by*

$$p_\rho = \left[2\mu^{2\beta} G^{-1}\left(\rho^{1/\gamma}, \frac{3}{2}\right)\right]^{1/2\beta}, \tag{5}$$

Proof. It is immediate using the definition of quantile ■.

Corollary 1. *From proposition 2, it follows directly that the median of the REPM distribution is*

$$Me(w) = \left[2\mu^{2\beta} G^{-1}\left(0.5^{1/\gamma}, \frac{3}{2}\right)\right]^{1/2\beta}.$$

Table 1 shows the mean, variance, median and mode for different values of μ, β and γ. Note that the mean, variance and median increase as γ increases; all four quantities increase as μ increases. It is also interesting to point out that the variance grows extremely as β decreases ($\beta < 1$). On the other hand

Table 1. *Mean, variance, median and mode for the REPM model with different combination of parameters.*

(μ, β, γ)	Mean	Variance	Median	Mode
$(1.3, 1.5, 0.5)$	1.403	1.386	1.365	0.347
$(1.3, 1.5, 1.0)$	1.738	1.732	1.724	0.254
$(1.3, 1.5, 1.5)$	1.912	1.904	1.891	0.207
$(1.3, 1.5, 2.0)$	2.024	2.015	1.997	0.179
$(2.3, 0.5, 1.5)$	8.566	7.230	4.545	34.909
$(2.3, 1.0, 1.5)$	4.189	4.078	3.848	2.154
$(2.3, 1.5, 1.5)$	3.382	3.369	3.346	0.648
$(2.3, 2.0, 1.5)$	3.055	3.063	3.081	0.305
$(0.6, 1.5, 1.5)$	0.882	0.879	0.872	0.044
$(1.0, 1.5, 1.5)$	1.471	1.465	1.455	0.122
$(1.3, 1.5, 1.5)$	1.912	1.904	1.891	0.207
$(1.6, 1.5, 1.5)$	2.353	2.344	2.328	0.313

$$F_W(\mu; \mu, \beta, \gamma) = \left[G\left(\frac{1}{2}, \frac{3}{2}\right)\right]^{\gamma} = C^{\gamma}, \tag{6}$$

with $C = G(1/2, 3/2) = 2\Phi(1) - 2\phi(1) - 1 \approx 0.199$. In equation (6), we note that the CDF evaluated in μ depends only on the value of γ. As C^{γ} is a strictly decreasing function for γ and $0 < C < 1$, the equation $F_W(\mu; \mu, \beta, \gamma) = \rho$, (for $0 < \rho < 1$) has a unique solution for γ. Specifically,

$$F_W(\mu; \mu, \beta, \gamma) = \rho \Leftrightarrow \gamma = \frac{\log(\rho)}{\log(C)}.$$

For a fixed ρ, if we set $\gamma = \gamma(\rho) = \log(\rho)/\log(C)$ as fixed, then μ represents directly the ρth quantile of the distribution. Table 2 shows some values for $\gamma(\rho)$ with different values for ρ. Henceforth, we will use the notation REPM(μ, β, γ) to refer to this alternative parametrization. This is a very useful result, because in practice many characteris-

Table 2. *Value of* $\gamma(\rho)$ *for some values of* ρ.

ρ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$\gamma(\rho)$	1.425	0.996	0.745	0.567	0.429	0.316	0.221	0.138	0.065

tics inherent to each observation are available. For this reason, we introduce a regression framework for applying the model to any quantile of the distribution. This also allows a more detailed relation among the covariates and the response variable than is possible using the regression in a single measure such as mean or median. To be more specific, for a non-homogeneous population, we consider that $w_i(\rho)$, the ρ-th quantile of the response variable, are independent and are such that $w_i(\rho) \sim \text{REPM}(\mu_i(\rho), \beta(\rho), \gamma(\rho))$, $i = 1, \ldots, n$, where the quantile of such variable is related to a set of covariates, say $\mathbf{x}_i^{\top} = (x_{i1}, ..., x_{ip})$, through the logarithmic link as

$$\log \mu_i(\rho) = \mathbf{x}_i^{\top} \boldsymbol{\tau}(\rho), \quad i = 1, \ldots, n, \tag{7}$$

where $\boldsymbol{\tau}(\rho) = (\tau_0(\rho), \ldots, \tau_p(\rho))^{\top}$ are the regression coefficients. These can be interpreted as follows: $\exp(\tau_0(\rho))$ represents the value of the ρ-th quantile of the response variable when all covariates are fixed at 0; and $\exp(\tau_j(\rho))$, $j = 1, \ldots, p$, represents the percentage increment (or decrement) in the ρ-th quantile for the response variable when the j-th covariate is increased by one unit and the rest of the covariates are fixed.

To avoid overloading the notation, hereinafter we use simply μ_i, β and γ instead of $\mu_i(\rho), \beta(\rho), \gamma(\rho)$ to specify the parameters, but understanding that in a regression model context, we are interested in modelling the ρ-th quantile.

3. Inference

In this section, we discuss the ML estimation for the REPM regression model under a classical approach. Let $W_i(\rho) \sim \text{REPM}(\mu_i, \beta, \gamma)$ independent variables, where the ith observation is related to a set of covariates $\mathbf{x}_i$ as in equation (7) and $\gamma = \gamma(\rho) =$

$\log(\rho)/\log(C)$ is fixed. The log-likelihood function for $\boldsymbol{\theta} = (\boldsymbol{\tau}^\top, \beta, \gamma)^\top$ is

$$\ell(\boldsymbol{\theta}) = n\log(\gamma) + (\gamma-1)\sum_{i=1}^{n}\log\left(G\left(\frac{w_i^{2\beta}}{2\mu_i^{2\beta}}, \frac{3}{2}\right)\right) + n\log(\beta) - 2\sum_{i=1}^{n}\log(\mu_i^{\beta}) + \\ + (2\beta-1)\sum_{i=1}^{n}\log(w_i) - n\log\left(\frac{\Gamma(3/2)}{\sqrt{2}}\right) - \sum_{i=1}^{n}\log(\mu_i^{\beta}) + \beta\sum_{i=1}^{n}\log(w_i) - \frac{1}{2^{2\beta}}\sum_{i=1}^{n}\left(\frac{w_i}{\mu_i}\right)^{2\beta}. \tag{8}$$

The ML estimators can be obtained by maximizing equation (8), using numerical procedures such as the Newton-Raphson algorithm. As an alternative, we use the `optim` routine in the R software (R Core Team, 2021) for the `L-BFGS-B` method, which is a limited memory modification for the traditional Broyden-Fletcher-Goldfarb-Shanno algorithm (BFGS), a constrained Quasi-Newton type algorithm which avoids the computation of the hessian matrix for the objective function and its respective inverse. The asymptotic variance of the ML estimators (say $\widehat{\boldsymbol{\theta}}$) can be estimated as follows $\widehat{\mathrm{Var}(\boldsymbol{\theta})} = \mathrm{diag}(-\mathbf{I}(\widehat{\boldsymbol{\theta}})^{-1})$, where $\mathbf{I}(\widehat{\boldsymbol{\theta}})$ is observed Fisher information evaluated in $\widehat{\boldsymbol{\theta}}$, that is

$$\mathbf{I}(\widehat{\boldsymbol{\theta}}) = -\left.\frac{\partial\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}.$$

Details about the components of this matrix can be found in appendix A. The asymptotic distribution of $\widehat{\boldsymbol{\theta}}$ is $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \sim \mathbf{N}(0, \mathbf{I}(\widehat{\boldsymbol{\theta}})^{-1})$, as $n \to +\infty$.

In order to perform a residual analysis, we can use the quantile residuals (see Dunn and Smith (1996)) defined as

$$r_i = \Phi^{-1}[F_W(w_i; \widehat{\boldsymbol{\theta}})], \quad i = 1, 2, ..., n,$$

where $F_W(w_i; \widehat{\boldsymbol{\theta}})$ is the CDF of the REPM model evaluated in the ML estimate of $\boldsymbol{\theta}$. As the ML estimator is a consistent estimator (when $n \to +\infty$), and if the model is appropriate for the data, $r_1, r_2, ..., r_n$ should be a random sample from the standard normal distribution. Also note that the independent observation assumption implies that the quantile residuals are also independent. The normality assumption can be tested, for instance, by a normality test such as the Kolmogorov-Smirnov (KS) (see Kolmogorov (1993) and Smirnov (1939)), Shapiro-Wilks (SW) (see Shapiro and Wilks (1965)) and Anderson-Darling (AD) (see Anderson and Darling (1952)) tests.

Table 3. *Simulation study for the REPM model.*

true values					$n = 50$				$n = 100$				$n = 200$			
ρ	β	τ_0	τ_1	estimator	mean	SE	CP	RMSE	mean	SE	CP	RMSE	mean	SE	CP	RMSE
0.50	2.0	2.0	0.5	β	2.049	0.176	1.000	0.086	2.017	0.081	0.686	0.168	2.007	0.057	0.642	0.120
				τ_0	2.049	0.176	1.000	0.086	2.023	0.120	1.000	0.065	2.010	0.087	1.000	0.045
				τ_1	0.505	0.150	0.940	0.136	0.503	0.101	0.969	0.110	0.502	0.073	0.945	0.072
			1.5	β	2.038	0.115	0.655	0.263	2.017	0.081	0.686	0.168	2.007	0.057	0.674	0.115
				τ_0	2.049	0.174	1.000	0.093	2.020	0.123	1.000	0.065	2.011	0.085	1.000	0.045
				τ_1	1.529	0.170	0.975	0.154	1.520	0.125	0.969	0.110	1.504	0.084	0.976	0.075
	2.0	0.5	2.0	β	2.096	0.116	0.658	0.264	2.041	0.082	0.675	0.167	2.019	0.057	0.696	0.114
				τ_0	0.498	0.102	0.98	0.089	0.499	0.079	0.979	0.066	0.499	0.052	0.981	0.044
				τ_1	2.005	0.184	0.985	0.140	2.003	0.133	0.989	0.103	2.002	0.092	0.985	0.070
		1.5		β	2.082	0.116	0.633	0.270	2.043	0.081	0.661	0.175	2.018	0.057	0.658	0.121
				τ_0	1.494	0.157	0.988	0.110	1.503	0.104	0.997	0.066	1.500	0.073	0.997	0.043
				τ_1	2.005	0.199	0.977	0.166	1.998	0.128	0.992	0.102	2.000	0.092	0.986	0.072
	0.5	2.0	2.0	β	0.525	0.116	0.995	0.069	0.512	0.082	0.999	0.043	0.506	0.057	1.000	0.03
				τ_0	1.983	0.350	0.915	0.386	2.011	0.237	0.932	0.248	2.004	0.169	0.940	0.175
				τ_1	2.034	0.638	0.933	0.651	1.987	0.441	0.956	0.415	1.999	0.309	0.974	0.284
	1.5			β	1.575	0.116	0.748	0.205	1.531	0.081	0.807	0.127	1.518	0.057	0.797	0.089
				τ_0	1.997	0.163	0.991	0.116	2.003	0.123	0.991	0.094	1.999	0.084	0.992	0.060
				τ_1	2.004	0.188	0.956	0.175	1.999	0.15	0.957	0.143	2.003	0.108	0.965	0.102
0.75	2.0	2.0	0.5	β	2.096	0.126	0.645	0.287	2.052	0.088	0.671	0.190	2.019	0.062	0.667	0.129
				τ_0	1.987	0.202	0.988	0.135	1.989	0.149	0.995	0.104	1.997	0.101	0.994	0.064
				τ_1	0.498	0.218	0.928	0.241	0.509	0.165	0.942	0.169	0.500	0.107	0.962	0.102
			1.5	β	2.106	0.126	0.604	0.306	2.05	0.088	0.663	0.187	2.023	0.062	0.702	0.122
				τ_0	1.986	0.209	0.992	0.140	1.996	0.144	0.993	0.093	1.998	0.103	0.994	0.069
				τ_1	1.502	0.258	0.958	0.245	1.497	0.167	0.957	0.161	1.497	0.123	0.952	0.117
	2.0	0.5	2.0	β	2.117	0.126	0.605	0.310	2.050	0.088	0.637	0.192	2.029	0.062	0.660	0.131
				τ_0	0.484	0.150	0.951	0.149	0.488	0.103	0.947	0.102	0.497	0.074	0.968	0.069
				τ_1	2.008	0.272	0.972	0.25	2.007	0.181	0.963	0.164	1.999	0.130	0.972	0.115
		1.5		β	2.103	0.126	0.636	0.289	2.039	0.088	0.677	0.187	2.022	0.062	0.685	0.125
				τ_0	1.483	0.18	0.985	0.134	1.491	0.126	0.986	0.094	1.500	0.091	0.991	0.067
				τ_1	2.006	0.257	0.973	0.222	2.000	0.181	0.969	0.160	1.992	0.133	0.973	0.118
	0.5	2.0	2.0	β	0.528	0.126	0.997	0.074	0.511	0.088	1.000	0.047	0.507	0.062	0.999	0.032
				τ_0	1.961	0.532	0.939	0.575	1.971	0.355	0.936	0.367	1.990	0.250	0.939	0.257
				τ_1	1.947	0.987	0.931	1.001	1.983	0.632	0.952	0.627	2.009	0.448	0.951	0.436
	1.5			β	1.568	0.126	0.778	0.216	1.540	0.088	0.776	0.143	1.519	0.062	0.783	0.101
				τ_0	1.994	0.198	0.978	0.161	1.99	0.151	0.976	0.126	1.994	0.104	0.976	0.087
				τ_1	1.974	0.295	0.939	0.300	2.002	0.225	0.951	0.224	2.000	0.157	0.961	0.153

4. Simulation study

In this section, we present a simulation study in order to assess the performance of the ML estimators for the REPM regression model. We considered one covariate, i.e., $\mu_i = \tau_0 + \tau_1 x_i$, $\gamma(\rho)$ as fixed, and the covariates $x_1, \ldots, x_n$ were simulated from the standard uniform distribution. We considered six vectors for (β, τ_0, τ_1): (2, 2, 0.5), (2, 2, 1.5), (2, 0.5, 2), (2, 1.5, 2), (0.5, 2, 2), (1.5, 2, 2); three values for the sample size: 50, 100 and 200; and two values for the modelled quantile: 0.50 and 0.75, totalling 36 combinations of parameters, sample size and quantile. Each scenario was replicated 1,000 times. To simulate values from the REPM model, we can use the following algorithm based on the inverse transform method:

- Generate $U_i \sim U(0,1), \quad i = 1, 2, ..., n.$
- Compute $W_i = \left[2\mu^{2\beta} G^{-1}\left(U_i^{1/\gamma}, \frac{3}{2} \right) \right]^{1/2\beta}$.

For each sample, we compute the ML estimates and the estimated standard errors based on the estimated hessian matrix. Table 3 summarizes the results, considering the mean of the ML estimations, their standard errors (SE), the 95% coverage probability (CP) based on the asymptotic normality for the ML estimators and the estimated root mean squared error (RMSE). Note that as the sample size increases, the mean of the ML estimators is closer to the true value of the parameters, while the RMSE decreases, suggesting than the estimators are consistent for the REPM model even in a finite sample size. Results also suggest that the CP terms converge to the nominal values with which they were built, suggesting that the asymptotic normality of the estimators is also reasonable in finite samples for the REPM model.

5. Application

In this section we illustrate our proposal with two real data sets, comparing it with other proposals in the literature. In the first application we fit the REPM model without covariates. We compare the results with the M, PM and gamma (G) distributions. In the second application we fit our proposal considering covariates, comparing results with the GSC, skewed Laplace (SKL) and skewed Student-t (SKT) models. Codes in R software (R Core Team, 2021) are avaliable as supplementary material.

5.1. Reinfection time data

In certain populations the occurrence of sexually transmitted diseases blue is a major problem. Even those that are not lethal represent a threat that must be taken into account. Specifically, gonorrhea and chlamydia are a focus of investigation because they are often asymptomatic in females. As a result they are often left untreated, which can lead to complications such as sterility. The following data set corresponds to the time to reinfection of 887 individuals by either gonorrhea or chlamydia, where the subject had already been infected with one of these diseases previously (see Klein and Moeschberger (2003)). This data set can be found in the std data included in the KMSurv R package (Klein, Moeschberger and Yan, 2012).

Table 4. *Descriptive analysis for the reinfection time data.*

mean	s.d.	median	interquartile range	min.	max.	skewness	kurtosis
369.5	370.1	247.0	501.0	1.0	1529.0	1.2	3.5

Table 4 shows a descriptive analysis for this data set. Note that 50% of the individuals were reinfected within the first 8 months. The times also present a positive skewness and a kurtosis slightly greater than normal distribution. Figure 5 shows the ML estimates for the parameters of the M, PM, G and REPM distributions. For each model we also present the AIC criteria, which suggest that the REPM model gives a better fit than the rest of the models. Figure 2 depicts the histogram with the estimated PDF and comparing the empirical CDF with the estimated CDF for the models discussed, showing that the REPM model presents a better fit for this data. Finally, Figure 3 shows the quantile-quantile (QQ) plots for the REPM, PM and G distributions. Note that the QQ plots suggest that, of the three models tested, the REPM is the most appropriate for this data set.

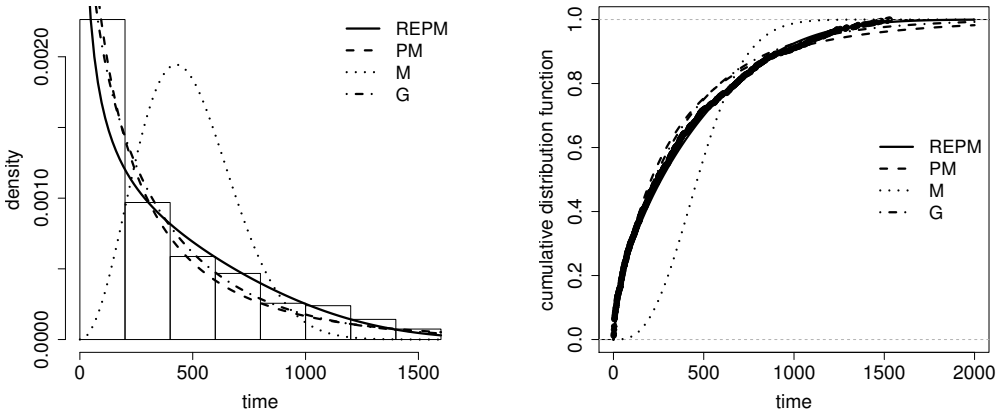
5.2. Clotting data

This data set presents measurements of the clotting time of blood (`time`, in seconds) for normal plasma diluted to nine different percentage concentrations with prothrombin-free plasma (`lconc`, in logarithm scale) for 18 patients. It must also be considered that the clotting time was induced by two lots of thromboplastin (`lot2`, categorized as 0 and 1). The data (see MLGdata R package) are available in McCullagh and Nelder (1989) (p. 302) (see R code below).

```
clotting<-data.frame(time=c(118, 58, 42, 35, 27, 25, 21, 19,
    18, 69, 35, 26, 21, 18, 16, 13, 12, 12),
    lconc=c(1.609, 2.303, 2.708, 2.996, 3.401, 3.689, 4.094,
    4.382, 4.605, 1.609, 2.303, 2.708, 2.996, 3.401, 3.689,
    4.094, 4.382, 4.605),
    lot=factor(c(rep(0, 9), rep(1, 9))))
```

Table 5. *Maximum likelihood estimates for the data with it's respective standard deviation in parenthesis for the infection time data*

Parameter	M	PM	G	REPM
α	$< 0.001(0.028)$	0.038 (0.004)	0.796 (0.027)	—
β	—	0.321 (0.009)	$0.002(< 0.001)$	**1.079 (0.158)**
μ	—	—	—	**578.576(0.150)**
γ	—	—	—	**0.177(0.195)**
log-likelihood	-7593.0	-6053.0	-6033.8	$-$**6013.3**
AIC	15188.0	12109.9	12071.6	**12032.7**

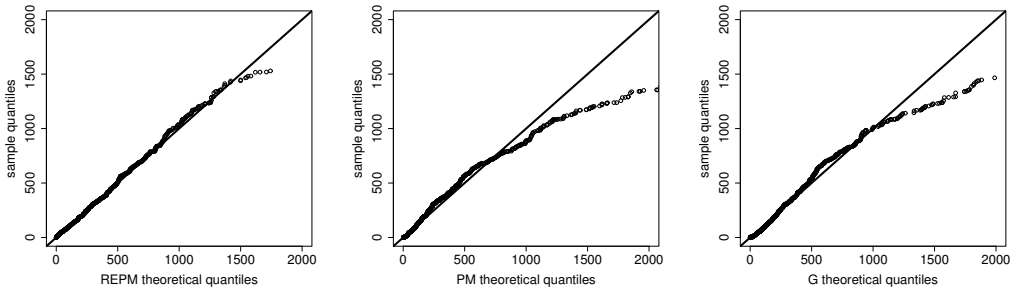


(a) *Histogram* **(b)** *Empirical plot*

Figure 2. *Histogram and empirical plot for the reinfection time data.*

We aim to model the clotting `time` for the i-th individual using `lconc`, `lot2` and the interaction between those covariates. We considered $\texttt{time}(\rho) \sim \text{REPM}(\mu_i, \beta, \gamma)$, where $\gamma = \gamma(\rho) = \log(\rho)/\log(C)$ is fixed and

$$\mu_i = \mu_i(\rho) = \exp\left(\tau_0 + \tau_1 \texttt{lconc}_i + \tau_2 \texttt{lot2}_i + \tau_3 \texttt{lconc}_i \times \texttt{lot2}_i\right), \quad i = 1, ..., 18,$$



(a) *QQ-plot, REPM.* **(b)** *QQ-plot, PM.* **(c)** *QQ-plot, G.*

Figure 3. *Q-Q plot for the REPM, PM and G models for the reinfection time data.*

Table 6 presents a descriptive analysis for the `global time`, `time` for `lot` $= 0$ ($\texttt{time}_0$), `time` for `lot` $= 1$ ($\texttt{time}_1$) and `lconc` for the clotting data set. We can verify that the global `time` has a significant standard deviation and is positively skewed, with a considerable kurtosis coefficient. Moreover, Figure 4 shows the plots for `time` versus `lconc` separated by `lot`.

Table 6. *Descriptive analysis for the clotting data.*

variable	mean	s.d.	median	interquartile range	min.	max.	skewness	kurtosis
`global time`	32.500	26.440	23.000	17.000	12.000	118.000	2.127	7.185
$\texttt{time}_0$	40.333	31.851	27.000	21.000	18.000	118.000	1.805	5.078
$\texttt{time}_1$	24.667	18.248	18.000	13.000	12.000	69.000	1.780	5.012

Table 7 shows the AIC values and p-values obtained in the K-S test for the quantile residuals, for the SKL, SKT, GSC and REPM quantile regression models different quantile values. Note that the AIC for the REPM is the lowest value of all the models (except for $\rho = 0.1$); the K-S test does not reject the null hypothesis that quantile residuals for this model are a random sample from the standard normal distribution (except for $\rho = 0.9$) with any significance level, suggesting that the model is appropriate for all the modelled quantiles (except for $\rho = 0.90$).

Figure 5 shows the ML estimator for the regression coefficients with their their respective asymptotic 95% confidence intervals. Note that `lconc` and `lot2` are significant in explaining all the quantiles modelled. Figure 6 shows the profile density for the ρ-th quantile of `time` for $\rho = 0.5$ and $\rho = 0.75$. Note how the distribution of the `time` according to our model seems to differ from the other distributions, showing a better representation of the population. Regarding the interpretation of the coefficients,

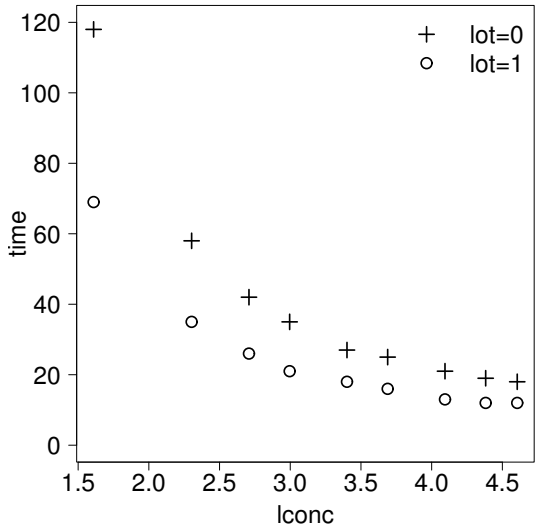


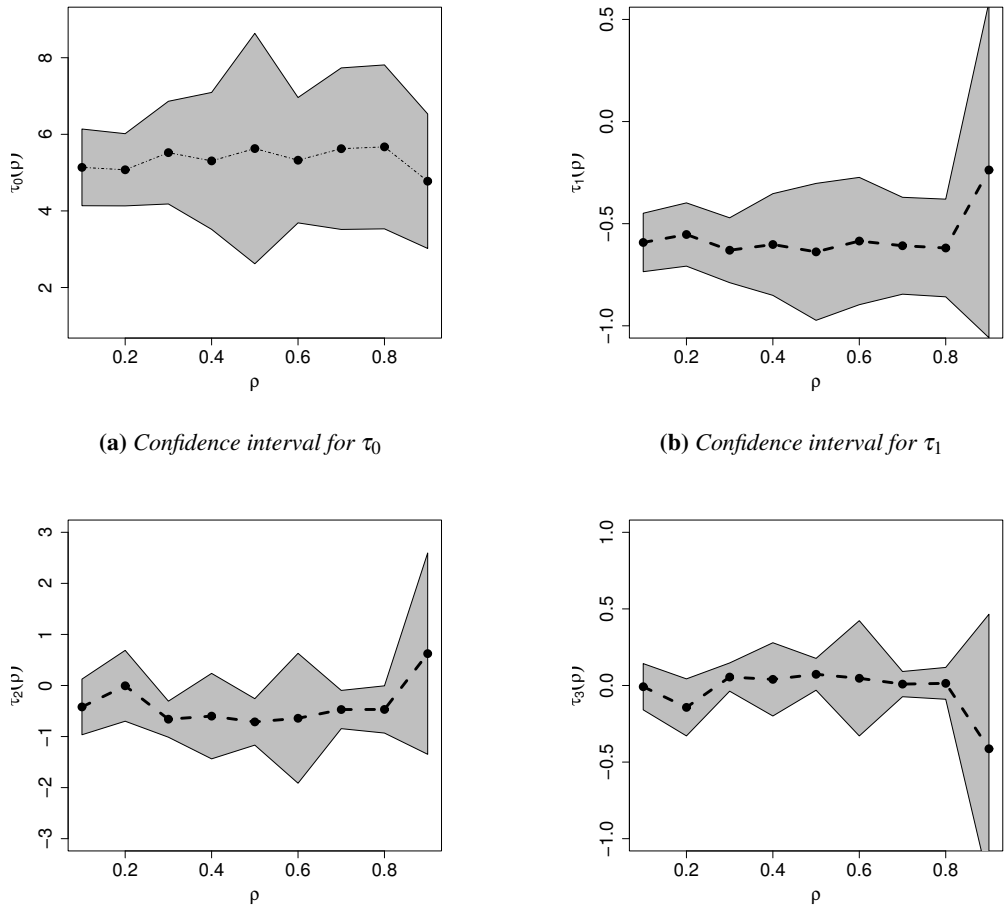
Figure 4. *Plot for clotting data.*

for example, we can conclude that

- For $\rho = 0.5$ (the median case) we obtain $\exp(\widehat{\tau}_1) = 0.528$. This means that for a fixed type of thromboplastin, the median of the clotting `time` decreases by 47.2% for each unit increase in the `lnonc`.
- For $\rho = 0.5$ (the median case) $\exp(\widehat{\tau}_2) = 0.490$. This implies that for a fixed `lnonc`, the median of the clotting `time` decreases by 51.0% when the type of thromboplastin is changed from `lot2` $= 1$ to `lot2` $= 0$.

Table 7. *AIC and p-values for the K-S test of SKT, SKL, GSC, and REPM model for the clotting data.*

	AIC				K-S			
ρ	SKT	SKL	GSC	**REPM**	SKT	SKL	GSC	**REPM**
0.10	121.130	127.916	110.820	**111.367**	0.003	0.003	0.186	**0.431**
0.25	125.554	132.958	118.335	**109.253**	0.004	0.001	0.119	**0.428**
0.50	133.049	143.110	129.903	**111.556**	0.002	< 0.001	0.250	**0.247**
0.75	151.402	155.568	144.733	**113.330**	0.092	0.500	0.018	**0.190**
0.90	149.034	150.596	167.269	**130.857**	0.125	0.200	< 0.001	< 0.001



(a) *Confidence interval for* τ_0

(b) *Confidence interval for* τ_1

(c) *Confidence interval for* τ_2

(d) *Confidence interval for* τ_3

Figure 5. *ML estimation of the regression coefficientes (with their respectively asymptotic* 95% *confidence interval), for the different values of the* ρ*-th quantile for the clotting data*

6. Conclusions

Exponentiated distributions have been used to extend a variety of well-known distribution models, resulting in flexible distributions that can be applied in a greater diversity of scenarios. This paper proposes the REPM distribution as an alternative model by which to introduce covariates, obtaining interpretations related to the quantile of the distribution. Nowadays there is a reasonable set of classic distributions with positive support, such as the exponential, gamma, Weibull, log-normal (LN), etc. So the question naturally arises "Why consider the REPM model instead of the common distribution that works well?". While it is true that models like LN and G have proved to be flexible

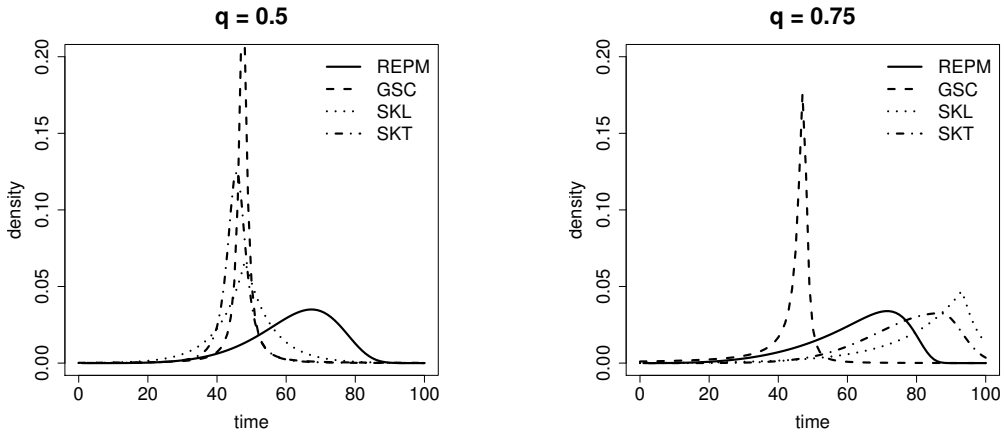


Figure 6. *Distribution for 0.5 (a) and 0.75 (b) quantiles of time considering* `lconc` *and* `lot2` *equal to 2.3 and 0, respectively. Curves in solid, dashed, dotted and dot-dash line represent the density functions estimated by the REPM, GSC, SKL and SKT models, respectively, for the clotting data*

enough to cover many situations, there are a few factors that must be borne in mind. For example, the LN distribution has a hazard rate function that may be unrealistic in some contexts, such as lifetimes data sets, since it is decreasing for long values. On the other hand, the G distribution, although it has a less strict hazard rate function, is not as flexible as the corresponding REPM model; moreover it does not have a closed function for the ρ-th quantile, i.e. quantile regression cannot be applied simply in this model. The real data applications above show that the REPM is a competent alternative to such traditional models.

Funding. *The research of Francisco A. Segovia was supported by Vicerrectoría de Investigación y Postgrado de la Universidad de Atacama. The research of Yolanda M. Gómez was supported by proyecto DIUDA programa de inserción NÂº 22367 of the Universidad de Atacama.*

A. Appendix: Score function and observed Fisher information

We devote this section to express the components of $\mathbf{I}(\widehat{\boldsymbol{\theta}})$ discussed in Section 3.

If $W \sim REPM(\boldsymbol{\theta})$, with $\boldsymbol{\theta} = (\mu, \beta, \gamma)^{\intercal}$, then we can $\partial^2 \log f_W(w;\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\intercal}$, as follows

$$
\begin{aligned}
\frac{\partial^2 \log f_W(w;\boldsymbol{\theta})}{\partial\beta^2} &= (\gamma-1)\left\{-\left[\frac{1}{G(\cdot)}\frac{\partial G(\cdot)}{\partial\beta}\right]^2+\frac{1}{g(\cdot)}\left[\frac{\partial \log G(\cdot)}{\partial\beta}\right]\left[\log\left(\frac{w}{\mu}\right)g(\cdot)+\frac{\partial g(\cdot)}{\partial\beta}\right]\right\},\\
\frac{\partial^2 \log f_W(w;\boldsymbol{\theta})}{\partial\mu^2} &= (1-\gamma)\left\{\left[\frac{1}{G(\cdot)}\frac{\partial G(\cdot)}{\partial\mu}\right]^2+\frac{1}{g(\cdot)}\left[\frac{\partial \log G(\cdot)}{\partial\mu}\right]\left[\frac{\partial g(\cdot)}{\partial\mu}-(2\beta-1)u^{-1}g(\cdot)\right]\right\},\\
\frac{\partial^2 \log f_W(w;\boldsymbol{\theta})}{\partial\gamma^2} &= -\frac{1}{\gamma^2},\\
\frac{\partial^2 \log f_W(w;\boldsymbol{\theta})}{\partial\beta\partial\mu} &= -(\gamma-1)w^{2\beta}\mu^{-2\beta-1}\left[1+2\beta\log\left(\frac{w}{\mu}\right)\right]\left\{\left(\frac{w}{\mu}\right)^{2\beta}\log\left(\frac{w}{\mu}\right)+\frac{g(\cdot)}{G(\cdot)}\right\},\\
\frac{\partial^2 \log f_W(w;\boldsymbol{\theta})}{\partial\beta\partial\gamma} &= \frac{1}{G(\cdot)}\frac{\partial G(\cdot)}{\partial\beta},\\
\frac{\partial^2 \log f_W(w;\boldsymbol{\theta})}{\partial\mu\partial\gamma} &= \frac{1}{G(\cdot)}\frac{\partial G(\cdot)}{\partial\mu},
\end{aligned}
$$

where $G(\cdot) = G(w^{2\beta}/2\mu^{2\beta}, 3/2)$, $g(\cdot) = g(w^{2\beta}/2\mu^{2\beta}, 3/2)$, and

$$
\begin{aligned}
\frac{\partial G(\cdot)}{\partial\beta} &= g(\cdot)\left(\frac{w}{\mu}\right)^{2\beta}\log\left(\frac{w}{\mu}\right),\\
\frac{\partial G(\cdot)}{\partial\mu} &= -\beta\mu^{-2\beta-1}w^{2\beta}g(\cdot),\\
\frac{\partial g(\cdot)}{\partial\beta} &= g(\cdot)\log\left(\frac{w}{\mu}\right)\left[1-\left(\frac{w}{\mu}\right)^{2\beta}\right],\\
\frac{\partial g(\cdot)}{\partial\mu} &= \frac{\beta}{\mu}g(\cdot)\left[\left(\frac{w}{\mu}\right)^{2\beta}-1\right].
\end{aligned}
$$

References

Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics*, 23(2), 193–212.

Boltzmann, L. (1871a). Analytischer beweis des zweiten haubtsatzes der mechanischen wärmetheorie aus den sätzen über das gleichgewicht der lebendigen kraft. *Wiener Berichte*, 63, 712–732.

Boltzmann, L. (1871b). Einige allgemeine sätze über wärmegleichgewicht. *Wiener Berichte*, 63, 679–711.

Boltzmann, L. (1871c). Über das wärmegleichgewicht zwischen mehratomigen gasmolekülen. *Wiener Berichte*, 63, 397–418.

Dunbar, R. (1982). Deriving the Maxwell distribution. *Journal of Chemical Education*, 59, 22–23.

Dunn, K. P. and Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5, 236–244.

Durrans, R. (1992). Distributions of fractional order statistics in hydrology. *Water Resources Research*, 28, 1649–1655.

Galarza, C., Lachos, V., Cabral, C. and Castro, L. (2017). Robust quantile regression using a generalized class of skewed distributions. *Stat*, 6, 113–130.

Gallardo, D., Bourguignon, M., Galarza, C. and Gómez, H. (2020a). A parametric quantile regression model for asymmetric response variables on the real line. *Symmetry*, 12, 1938.

Gallardo, D., Gómez-Déniz, E. and Gómez, H. (2020b). Discrete generalized half-normal distribution and its applications in quantile regression. *SORT*, 44(2), 265–284.

Gómez, Y. and Bolfarine, H. (2015). Likelihood-based inference for the power half-normal distribution. *Journal of Statistical Theory and Applications*, 14(4), 383–398.

Gómez, Y., Gallardo, D. and Arnold, B. (2017). The power piecewise exponential model. *Journal of Statistical Computation and Simulation*, 88(5), 825–840.

Gómez, Y., Gómez-Déniz, E., Venegas, O., Gallardo, D. and Gómez, H. (2019). An asymmetric bimodal distribution with application to quantile regression. *Symmetry,* 11, 899.

Gupta, R. and Gupta, R. (2008). Analyzing skewed data by power normal model. *TEST*, 17, 197–210.

Gupta, R. and Kundu, D. (1999). Generalized exponential distributions. *Australian New Zealand Journal of Statistics*, 41(2), 173–188.

Gupta, R., Gupta, P. and Gupta, R. (1998). Modeling failure time data by Lehmann alternatives. *Communications in Statistics - Theory and Methods*, 27(4), 887–904.

Huang, J. and Chen, S. (2015). Tail behavior of the generalized Maxwell distribution. *Communications in Statistics - Theory and Methods*, 45(14), 4230–4236.

Kakde, C. and Shirle, D. (2006). On exponentiated lognormal distribution. *International Journal of Agricultural and Statistical Sciencies,* 2, 319–326.

Klein, J. and Moeschberger, M. (2003). *Survival Analysis Techniques for Censored and Trun Cated Data*, Springer-Verlag, New York, Vol. 2.

Klein, J., Moeschberger, M. and Yan, J. (2012). KMsurv: Data sets from Klein and Moeschberger (1997), Survival Analysis. https://CRAN.R-project.org/package= KMsurv, R package version 0.1-5.

Kolmogorov, A. (1993). Sulla determinazione empirica di uma legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4, 83-91.

Lehmann, E. (1953). The power of rank tests. *Annals of Mathematical Statistics*, 24(1), 23–43.

Maxwell, J. (1860). Illustrations of the dynamical theory of gases. Part I. On the motions and collisions of perfectly elastic spheres. *Philosophical Magazine*, Series 4, 19, 19–32.

McCullagh, P. and Nelder, J. (1989). Generalized Linear Models. *Monographs on Statistics and Applied Probability*, 37. Chapman & Hall.

Mudholkar, G. and Srivastava, D. (1993). Exponentiated Weibull family for analyzing bathtub failure-rate date. *IEEE Transactions on Reliability*, 42, 299–302.

Mudholkar, G., Srivastava, D. and Freimer, M. (1995). The exponentiated Weibull family: a reanalysis of the bus-motor-failure data. *Technometrics*, 37, 436–445.

Nadarajah, S. (2005) The exponentiated Gumbel distribution with climate application. *Environmetrics*, 17, 13–23.

Nadarajah, S. and Gupta, A. (2007). The exponentiated gamma distribution with application to drought data. *Calcutta Statistical Association Bulletin*, 59, 29–54.

Pewsey, A., Gómez, H. and Bolfarine, H. (2012). Likelihood based inference for distributions of fractional order statistics. *TEST*, 21, 775–789.

R Core Team (2021). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/

Rêgo, L., Cintra, R. and Cordeiro, G. (2012). On some properties of the beta normal distribution. *Communications in Statistics - Theory and Methods*, 41, 3722–3738.

Sánchez, L., Leiva, V., Galea, M. and Saulo, H. (2020). Birnbaum-Saunders quantile regression models with application to spatial data. *Mathematics*, 8, 1000.

Segovia, F., Gómez, Y., Venegas, O. and Gómez, H. (2020). A power Maxwell distribution with heavy tails and applications. *Mathematics*, 8, 1116.

Shapiro, S. and Wilks, M. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52, 591–611.

Singh, A., Bakouch, H., Kumar, S. and Singh, U. (2018). Power Maxwell distribution: Statistical properties, Estimation and Application. arXiv, 1807.01200v1.

Smirnov, N. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bulletin Mathématique de l'Université de Moscow*, 2, 3-11.

Information for authors and subscribers

Author Guidelines

SORT accepts for publication only original articles that have not been submitted simultaneously to any other journal in the areas of statistics, operations research, official statistics or biometrics. Furthermore, once a paper is accepted it must not be published elsewhere in the same or similar form.

SORT is an **Open Access** journal which **does not** charge publication **fees**.

Articles should be preferably of an applied nature and may include computational or educational elements. Publication will be exclusively in English. All articles will be forwarded for systematic peer review by independent specialists and/or members of the Editorial Board.

Submission of papers must be in electronic form only at our **RACO** (Revistes Catalanes en Accés Obert) submission site. Initial submission of the paper should be a single document in **PDF** format, including all **figures and tables** embedded in the main text body. **Supplementary material** may be submitted by the authors at the time of submission of a paper by uploading it with the main paper at our RACO submission site. **New authors**: please register. Upon successful registration you will be sent an e-mail with instructions to verify your registration.

The article should be prepared in **double-spaced** format, using a **12-point** typeface. SORT strongly recommends the use of its LaTeX template.

The **title page** must contain the following items: title, name of the author(s), professional affiliation and complete address, and an abstract (75–100 words) followed by the keywords and MSC2010 Classification of the American Mathematical Society.

Before submitting an article, the author(s) would be well advised to ensure that the text uses **correct English**. Otherwise the article may be returned for language improvement before entering the review process.

Bibliographic references within the text must follow one of these formats, depending on the way they are cited: author surname followed by the year of publication in parentheses [e.g., Mahalanobis (1936) or Rao (1982b)])]; or author surname and year in parentheses, without comma [e.g. (Mahalanobis 1936) or (Rao 1982b) or (Mahalanobis 1936, Rao 1982b)]. The complete reference citations should be listed alphabetically at the end of the article, with multiple publications by a single author listed chronologically. Examples of reference formats are as follows:

- Article: Casella, G. and Robert, C. (1998). Post-processing accept-reject samples: recycling and rescaling. Journal of Computational and Graphical Statistics, 7, 139–157.

- Book: Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2003). Bayesian Data Analysis, 2nd Ed. Chapman & Hall / CRC, New York.

- Chapter in book: Engelmann, B. (2006). Measures of a rating's discriminative power-applications and limitations. In: Engelmann, B. and Rauhmeier, R. (eds), The Basel Risk Parameters: Estimation, Validation, and Stress Testing. Springer, New York.

- Online article (put issue or page numbers and last accessed date): Marek, M. and Lesaffre, E. (2011). Hierarchical generalized linear models: The R package HGLMMM. Journal of Statistical Software, 39 (13). http://www.jstatsoft.org/v39/i13. Last accessed 28 March 2011.

Explanatory footnotes should be used only when absolutely necessary. They should be numbered sequentially and placed at the bottom of the corresponding page. **Tables and figures** should also be numbered sequentially.

Papers should not normally exceed about **25 pages** of the **PDF** format (**40 pages** of the format provided by the SORT **LaTeX** template) including all figures, tables and references. Authors should consider transferring content such as long tables and supporting methodological details to the online supplementary material on the journal's web site, particularly if the paper is long.

Once the article has positively passed the first review round, the executive editor assigned with the evaluation of the paper will send comments and suggestions to the authors to improve the paper. At this stage, the executive editor will ask the authors to submit a revised version of the paper using the SORT **LaTeX** template.

Once the article has been accepted, the journal editorial office will **contact the authors** with further instructions about this final version, asking for the source files.

Submission Preparation Checklist

As part of the submission process, authors are required to check off their submission's compliance with all of the following items, and submissions may be returned to authors that do not adhere to these guidelines.

1. The submitted manuscript follows the guidelines to authors published by SORT
2. Published articles are under a Creative Commons License BY-NC-ND
3. Font size is 12 point
4. Text is double-spaced
5. Title page includes title, name(s) of author(s), professional affiliation(s), complete address of corresponding author
6. Abstract is 75-100 words and contains no notation, no references and no abbreviations
7. Keywords and MSC2010 classification have been provided
8. Bibliographic references are according to SORT's prescribed format
9. English spelling and grammar have been checked
10. Manuscript is submitted in PDF format

Copyright notice and author opinions

SORT Statistics and Operations Research Transactions
Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58 - 08003 Barcelona. SPAIN
Tel. +34-93.557.30.76 – Fax +34-93.557.30.01
sort@idescat.cat

How to cite articles published in SORT

Commenges, D. (2003). Likelihood for interval-censored observations from multi-state models. *SORT*, 27 (1), 1-12.

Subscription form

SORT *(Statistics and Operations Research Transactions)*

Name _____

Organisation _____

Street Address _____

Zip/Postal code _____ City _____

State/Country _____ Tel. _____

Fax _____ NIF/VAT Registration Number _____

E-mail _____

Date _____

Signature

I wish to subscribe to **SORT *(Statistics and Operations Research Transactions)***) from now on

Annual subscription rates:

– Spain: €42 (4 % VAT included)
– Other countries: €46 (4 % VAT included)

Price for individual issues (current and back issues):

– Spain: €15/issue (4 % VAT included)
– Other countries: €17/issue (4 % VAT included)

Please send this subscription form (or a photocopy) to:

SORT *(Statistics and Operations Research Transactions)*
Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona
SPAIN
Fax: +34-93-557 30 01

Or by e-mail to:
sort@idescat.cat