## CONTENTS

**Metodološki zvezki, Vol. 13, No. 1, 2016**

Metodološki zvezki, Vol. 13, No. 1, 2016, 1-15

# Interval Prediction of Order Statistics Based on Records by Employing Inter-Record Times: A Study Under Two Parameter Exponential Distribution

Morteza Amini[1] and S.M.T.K. MirMostafaee[2]

**Abstract**

In this note, we propose a parametric inferential procedure for predicting future order statistics, based on record values, which takes inter-record times into account. We utilize the additional information contained in inter-record times for predicting future order statistics on the basis of observed record values from an independent sample. The two parameter exponential distribution is assumed to be the underlying distribution.

## 1 Introduction

Suppose $Y_1, \ldots, Y_m$ are independent and identically distributed (iid) observations from an absolutely continuous cumulative distribution function (cdf) $F$, possessing probability density function (pdf) $f$. The order statistics of the sample $Y_1, \ldots, Y_m$, represented by $Y_{1:m} < \cdots < Y_{m:m}$, are obtained by arranging the sample in an increasing order. Order statistics have been used in a wide range of applications, including robust statistical estimation, detection of outliers, characterization of probability distributions, goodness-of-fit tests, entropy estimation, analysis of censored samples, reliability analysis, quality control and strength of materials. A useful survey of available results until 2003 is given in the book of David and Nagaraja (2003).

Let $X_1, X_2, \ldots$ be a sequence of iid random variables, independent of and identically distributed to $Y_1$. An observation $X_j$ is called an upper (lower) record value if its value exceeds (resp. falls below) those of all the previous observations, that is the $n^{\text{th}}$ upper (resp. lower) record value, $U_n$ (resp. $L_n$), is defined as $X_{T_n}$, where $T_1 = 1$, with probability 1, and $T_n = \min\{j : \ j > T_{n-1}, \ X_j > X_{T_{n-1}}\}$ (resp. $T_n = \min\{j : \ j > T_{n-1}, \ X_j < X_{T_{n-1}}\}$), for $n > 1$. Throughout this paper we

[1]Department of Statistics, School of Mathematics, Statistics and computer Science, College of Science, University of Tehran, P.O. Box 14155-6455, Tehran, Iran; morteza.amini@ut.ac.ir

[2] Department of Statistics, Faculty of Mathematical Sciences, University of Mazandaran, P.O. Box 47416-1467, Babolsar, Iran; m.mirmostafaee@umz.ac.ir

deal with upper record values for a predictive inference. Similar results can be obtained for the case of lower record values. The inter-record time statistic, defined as

$$\Delta_s = T_{s+1} - T_s,\ s \geq 1,$$

is the number of observations between $s^{\text{th}}$ and $(s+1)^{\text{th}}$ record values. For more details we refer the reader to Arnold et al. (1998). Record data arise in a wide variety of practical situations including industrial stress testing, finance, meteorological analysis, hydrology, seismology, sporting and athletic events, and mining surveys.

The problem of predicting future observations has been extensively studied in the literature and several parametric and non-parametric procedures are developed for prediction. In many practical data-analytic situations, one is interested in constructing a prediction interval on the basis of available observations. There are situations in which the available observations and the predictable future observation are of the same type. The prediction of future records on the basis of observed records from the same distribution and prediction of order statistics based on order statistics are studied, among others, by Dunsmore (1983), Nagaraja (1984), Chou (1988), Awad and Raqab (2000), Raqab and Balakrishnan (2008) and the references therein.

Recently, Ahmadi and Balakrishnan (2010), Ahmadi and MirMostafaee (2009), Ahmadi et al. (2010) and MirMostafaee and Ahmadi (2011), discussed the prediction of future records from a $Y$-sequence based on the order statistics observed from an independent $X$-sequence, and vice versa.

In predicting future order statistics on the basis of observed record statistics, sometimes the available observations also include inter-record times which can be utilized as additional information to improve the predictive inference. In other words, when both record values and the inter-record times are available, it would be nice to employ the information included in both records and record times. Feuerverger and Hall (1998) emphasized that "However, the record times and record values jointly contain considerably more information about $F$ than the record values alone." Actually, applying the additional information about record times is not a new subject and several authors focused on inference based on both record values and record times, see for example Samaniego and Whitaker (1986), Lin et al. (2003), Doostparast (2009), Doostparast and Balakrishnan (2013), Kızılaslan and Nadar (2014) and MirMostafaee et al. (2016).

In this paper, a two parameter exponential distribution, $Exp(\mu, \sigma)$, with pdf

$$f(x; \mu, \sigma) = \frac{1}{\sigma} e^{-(x-\mu)/\sigma}, \quad x > \mu, \quad \mu \in \mathbb{R},\ \sigma > 0, \tag{1.1}$$

is considered as the underlying distribution. We write $Z \sim Exp(\mu, \sigma)$ if the pdf of $Z$ can be expressed as (1.1). Note that $\mu$ and $\sigma$ are the location and scale parameters, respectively. Throughout this paper we assume that both parameters, $\mu$ and $\sigma$, are unknown.

Now, suppose that $Y_1, \cdots, Y_m$ constitute a future random sample from a two parameter exponential distribution, i.e. $Y_1, \cdots, Y_m \overset{iid}{\sim} Exp(\mu, \sigma)$ and $Y_{1:m} < \cdots < Y_{j:m}$ are the corresponding order statistics of this sample. In addition, $\bar{Y}_m = m^{-1} \sum_{i=1}^{m} Y_{i:m}$ denotes the mean of this future sample. If $Y_1, \cdots, Y_m$ denote the times to failure of $m$ independent units in a lifetime test, then $\bar{Y}_m$ can be interpreted as the mean time on test of these failed units. We assume that the available data include the observed upper record

values, $U_1, \cdots, U_n$, given the inter-record times, $(\Delta_1, \ldots, \Delta_{n-1})$. We emphasize that these record values are assumed to be extracted from a sequence of iid random variables $\{X_j, j = 1, 2, \cdots\}$ where $X_j \sim Exp(\mu, \sigma)$ for $j = 1, 2, \cdots$. Moreover, the sequence $\{X_j, j = 1, 2, \cdots\}$ and the sample $\{Y_i, i = 1, \cdots, m\}$ are statistically independent. Note that $n$ is the number of the observed record values and depends on the experiment, however, $m$ is the sample size of the future observations and it can be considered arbitrary. In addition, $n$ and $m$ are unrelated. The problem of interest is to obtain conditional prediction intervals for $j^{\text{th}}$ future order statistic, $Y_{j:m}$, as well as for the mean, $\bar{Y}_m$, in a future sample on the basis of the available data. We compare our conditional prediction intervals with the unconditional ones proposed by Ahmadi and MirMostafaee (2009) and observe an improvement over the predictive inference without inter-record times. Therefore, we consider two cases: (a) The informative data contain only the upper record values, (b) The informative data contain the upper record values and the inter-record times, and then we observe that case (b) has some predictive inferential improvement in comparison with case (a).

The rest of the paper is organized as follows. Some general preliminaries are presented in Section 2. Conditional prediction intervals for the future $j^{\text{th}}$ order statistic, $Y_{j:m}$, and the mean of the future sample, $\bar{Y}_m$, based on record values of given inter-record times for the two parameter exponential distribution are studied in Sections 3 and 4. An illustrative example and some concluding remarks are involved in Sections 5 and 6. The R codes for computing some results of the paper are given in the appendix.

# 2 Preliminaries

In this section, we present some general preliminary results used in future sections. Given upper record values $u_1, \ldots, u_{n-1}$, which are observed and extracted from the sequence $\{X_j; j \geq 1\}$, inter-record times $\Delta_1, \ldots, \Delta_{n-1}$ are independent geometrically distributed random variables with success probabilities $\bar{F}(u_i),\ i = 1, \ldots, n-1$. Furthermore, the record values $U_1, \ldots, U_n$ form a Markov Chain with adjacent transition pdf equal to the left truncated pdf of the underlying distribution, see Arnold et al. (1998). Thus, the joint distribution of $\mathbf{U}_n = (U_1, \ldots, U_n)$ and $\boldsymbol{\Delta}_n = (\Delta_1, \ldots, \Delta_{n-1})$ is

$$f_{\mathbf{U}_n, \boldsymbol{\Delta}_n}(\mathbf{u}_n, \boldsymbol{\delta}_n) = \prod_{i=1}^{n-1} f(u_i)[F(u_i)]^{\delta_i - 1} f(u_n), \tag{2.1}$$

where $\mathbf{u}_n = (u_1, \ldots, u_n) \in \mathbb{X}^n$, in which $\mathbb{X}$ is the support of $X$ and $\boldsymbol{\delta}_n = (\delta_1, \ldots, \delta_{n-1}) \in \mathbb{N}^{n-1}$, see Samaniego and Whitaker (1986) and Arnold et al. (1998) page 169. We emphasize that $\boldsymbol{\Delta}_n$ contains $n-1$ positive integer-valued discrete random variables and $\boldsymbol{\delta}_n$ is the observed vector of $\boldsymbol{\Delta}_n$. By integrating (2.1) with respect to (w.r.t.) $u_1, \ldots, u_n$, we can easily prove the following result.

**Lemma 1** *The joint probability mass function of* $\Delta_1, \ldots, \Delta_{n-1}$ *is*

$$P_{\boldsymbol{\Delta}_n}(\boldsymbol{\delta}_n) = \Pr(\boldsymbol{\Delta}_n = \boldsymbol{\delta}_n) = \sum_{j=1}^{n-1} c_j(n, \boldsymbol{\delta}_n)[(a_1(n, j, \boldsymbol{\delta}_n) + 1)(a_1(n, j, \boldsymbol{\delta}_n) + a_n(n, j, \boldsymbol{\delta}_n) + 2)]^{-1},$$

*where*

$$c_j(n,\boldsymbol{\delta}_n)=(-1)^{n-j-1}\left[\prod_{j_1=0}^{j-2}\left(\sum_{t=n-j+1}^{n-j_1-1}\delta_t\right)\prod_{j_2=0}^{n-j-2}\left(\sum_{t=j_2+2}^{n-j}\delta_t\right)\right]^{-1},$$

$$a_1(n,j,\boldsymbol{\delta}_n)=\sum_{t=1}^{n-j}\delta_t-1,\quad a_n(n,j,\boldsymbol{\delta}_n)=\sum_{t=n-j+1}^{n-1}\delta_t,$$

*in which we assume for* $a>b$, $\sum_{t=a}^{b}\delta_t=0$ *and* $\prod_{t=a}^{b}\delta_t=1$.

In this paper, we need the conditional distribution of $U_1$ and $U_n$ given by $\boldsymbol{\Delta}_n=\boldsymbol{\delta}_n$ as follows.

**Lemma 2** *The conditional pdf of* $U_1$ *and* $U_n$ *given* $\boldsymbol{\Delta}_n=\boldsymbol{\delta}_n$ *is*

$$f_{U_1,U_n|\,\boldsymbol{\Delta}_n}(u_1,u_n|\,\boldsymbol{\delta}_n)=[P_{\boldsymbol{\Delta}_n}(\boldsymbol{\delta}_n)]^{-1}\sum_{j=1}^{n-1}c_j(n,\boldsymbol{\delta}_n)[F(u_1)]^{a_1(n,j,\boldsymbol{\delta}_n)}[F(u_n)]^{a_n(n,j,\boldsymbol{\delta}_n)}f(u_1)f(u_n),$$

*where* $c_j(n,\boldsymbol{\delta}_n)$, $a_1(n,j,\boldsymbol{\delta}_n)$, $a_n(n,j,\boldsymbol{\delta}_n)$ *and* $P_{\boldsymbol{\Delta}_n}(\boldsymbol{\delta}_n)$ *are as in Lemma 1.*

The proof of Lemma 2 is straightforward by integrating (2.1) w.r.t. $u_2,\ldots,u_{n-1}$ and dividing the obtained equation by $P_{\boldsymbol{\Delta}_n}(\boldsymbol{\delta}_n)$.

# 3 Conditional prediction intervals for order statistics

In this section, the goal is to find a conditional prediction interval for $Y_{j:m}$ when the observed $U_1,\ldots,U_n$ are available given $\boldsymbol{\Delta}_n=\boldsymbol{\delta}_n$ for the two parameter exponential distribution.

To this end, we consider the pivotal quantity

$$W_j=\frac{Y_{j:m}-U_1}{U_n-U_1}. \tag{3.1}$$

Note that the pivotal quantity $W_j$ is the same as the one considered by Ahmadi and MirMostafaee (2009). This quantity is location and scale invariant namely it is free of both unknown parameters i.e. the location parameter $\mu$ and the scale parameter $\sigma$. It is also a simple function of both observed and future statistics, so that the future statistic can be derived from it easily. Ahmadi and MirMostafaee (2009) found the unconditional distribution of $W_j$ while we present the conditional distribution of $W_j$ given $\boldsymbol{\Delta}_n=\boldsymbol{\delta}_n$, (i.e. the inter-record times are assumed to be known and fixed) in the following theorem.

**Theorem 1** *The conditional cdf of* $W_j$ *in* (3.1) *given* $\boldsymbol{\Delta}_n=\boldsymbol{\delta}_n$ *is for* $w>0$

$$\begin{aligned}F_{W_j|\,\boldsymbol{\Delta}_n}(w|\boldsymbol{\delta}_n) &= \sum_{l=j}^{m}\sum_{j_1=1}^{n-1}\sum_{j_2=0}^{l}\sum_{j_3=0}^{a_1(n,j_1,\boldsymbol{\delta}_n)}\sum_{j_4=0}^{a_n(n,j_1,\boldsymbol{\delta}_n)}\frac{\binom{m}{l}\binom{a_1(n,j_1,\boldsymbol{\delta}_n)}{j_3}\binom{a_n(n,j_1,\boldsymbol{\delta}_n)}{j_4}\binom{l}{j_2}}{(-1)^{j_2+j_3+j_4}P_{\boldsymbol{\Delta}_n}(\boldsymbol{\delta}_n)}\\ &\quad\times c_{j_1}(n,\boldsymbol{\delta}_n)[(j_2+m-l+j_3+j_4+2)((j_2+m-l)w+j_4+1)]^{-1},\end{aligned}$$

*and for* $w < 0$

$$F_{W_j|\,\boldsymbol{\Delta}_n}(w|\boldsymbol{\delta}_n) = \sum_{l=j}^{m}\sum_{j_1=1}^{n-1}\sum_{j_2=0}^{l}\sum_{j_3=0}^{a_1(n,j_1,\,\boldsymbol{\delta}_n)}\sum_{j_4=0}^{a_n(n,j_1,\,\boldsymbol{\delta}_n)} \frac{\binom{m}{l}\binom{a_1(n,j_1,\,\boldsymbol{\delta}_n)}{j_3}\binom{a_n(n,j_1,\,\boldsymbol{\delta}_n)}{j_4}\binom{l}{j_2}}{(-1)^{j_2+j_3+j_4}P_{\boldsymbol{\Delta}_n}(\boldsymbol{\delta}_n)}$$

$$\times c_{j_1}(n,\boldsymbol{\delta}_n)[(j_2+m-l+j_3+j_4+2)(j_4+1-w(j_3+j_4+2))]^{-1},$$

*where* $a_1(n,j_1,\ \boldsymbol{\delta}_n)$ *and* $a_n(n,j_1,\ \boldsymbol{\delta}_n)$ *are defined in Lemma 1 and* $P_{\boldsymbol{\Delta}_n}(\boldsymbol{\delta}_n)$ *is the joint mass function of* $\Delta_1,\ldots,\Delta_{n-1}$ *which is also given in Lemma 1.*

**Proof:** Letting $J^*_{n,1} = (U_n - U_1)/\sigma$, $U^*_1 = (U_1-\mu)/\sigma$ and $Y^*_{j:m} = (Y_{j:m}-\mu)/\sigma$, we may write

$$F_{W_j|\,\boldsymbol{\Delta}_n}(w|\boldsymbol{\delta}_n) = \int_0^\infty\int_0^\infty F_{Y^*_{j:m}}(vw+u) f_{U^*_1,J^*_{n,1}|\,\boldsymbol{\Delta}_n}(u,v|\boldsymbol{\delta}_n)\,\mathrm{d}u\,\mathrm{d}v. \tag{3.2}$$

For $t>0$, we have

$$F_{Y^*_{j:m}}(t) = \sum_{l=j}^{m}\binom{m}{l}(1-e^{-t})^l e^{-(m-l)t}. \tag{3.3}$$

Also, from Lemma 2, we obtain

$$f_{U^*_1,J^*_{n,1}|\,\boldsymbol{\Delta}_n}(u,v|\ \boldsymbol{\delta}_n) = [P_{\boldsymbol{\Delta}_n}(\boldsymbol{\delta}_n)]^{-1}\sum_{j=1}^{n-1} c_j(n,\boldsymbol{\delta}_n)[1-e^{-u}]^{a_1(n,j,\boldsymbol{\delta}_n)}[1-e^{-(u+v)}]^{a_n(n,j,\boldsymbol{\delta}_n)}e^{-(2u+v)}. \tag{3.4}$$

Hence, by substituting (3.4) and (3.3) in (3.2) and using the binomial expansions, we have for $w>0$,

$$F_{W_j|\,\boldsymbol{\Delta}_n}(w|\boldsymbol{\delta}_n) = \sum_{l=j}^{m}\sum_{j_1=1}^{n-1}\sum_{j_2=0}^{l}\sum_{j_3=0}^{a_1(n,j_1,\,\boldsymbol{\delta}_n)}\sum_{j_4=0}^{a_n(n,j_1,\,\boldsymbol{\delta}_n)} \frac{\binom{m}{l}\binom{a_1(n,j_1,\,\boldsymbol{\delta}_n)}{j_3}\binom{a_n(n,j_1,\,\boldsymbol{\delta}_n)}{j_4}\binom{l}{j_2}c_{j_1}(n,\boldsymbol{\delta}_n)}{(-1)^{j_2+j_3+j_4}P_{\boldsymbol{\Delta}_n}(\boldsymbol{\delta}_n)}$$

$$\times\int_0^\infty\int_0^\infty e^{-(j_2+m-l+j_3+j_4+2)u}e^{-((j_2+m-l)w+j_4+1)v}\,\mathrm{d}u\,\mathrm{d}v,$$

and therefore we naturally arrive at the desired expression. Similarly, we may attain the expression for $F_{W_j|\,\boldsymbol{\Delta}_n}(w|\boldsymbol{\delta}_n)$ when $w<0$ after substituting (3.4) and (3.3) in (3.2) by noting that the integral w.r.t. $u$ must be taken from $-vw$ to $\infty$. □

Let $w_\gamma(n,m,j;\boldsymbol{\delta}_n)$ be the $\gamma^{\text{th}}$ conditional quantile of $W_j$ given $\boldsymbol{\Delta}_n = \boldsymbol{\delta}_n$, i.e.

$$\Pr(W_j < w_\gamma(n,m,j;\boldsymbol{\delta}_n)|\ \boldsymbol{\Delta}_n = \boldsymbol{\delta}_n) = \gamma.$$

To find $100(1-\alpha)\%$ two-sided conditional prediction intervals for $Y_{j:m}$ based on record values given $\boldsymbol{\Delta}_n = \boldsymbol{\delta}_n$, we have to find the conditional quantiles $w_{\alpha_1}(n,m,j;\boldsymbol{\delta}_n)$ and $w_{1-\alpha_2}(n,m,j;\boldsymbol{\delta}_n)$, for $\alpha_1+\alpha_2=\alpha$, $0<\alpha_i<1$, $i=1,2$, numerically.

Now, a $100(1-\alpha)\%$ conditional prediction interval for $Y_{j:m}$ based on record values given $\boldsymbol{\Delta}_n = \boldsymbol{\delta}_n$, is given by

$$(U_1 + w_{\alpha_1}(n,m,j;\boldsymbol{\delta}_n)(U_n-U_1), U_1 + w_{1-\alpha_2}(n,m,j;\boldsymbol{\delta}_n)(U_n-U_1)). \tag{3.5}$$

**Table 1:** The values of $w_{0.025}(3,m,j)$, $w_{0.975}(3,m,j)$, $w_{0.975}(3,m,j)-w_{0.025}(3,m,j)$, $w_{0.025}(3,m,j;\boldsymbol{\delta}_n)$, $w_{0.975}(3,m,j;\boldsymbol{\delta}_n)$, $w_{0.975}(3,m,j;\boldsymbol{\delta}_n)-w_{0.025}(3,m,j;\boldsymbol{\delta}_n)$, for $m=10,20$, $j=5,7,10$ (for $m=10$), $j=12,17,20$ (for $m=20$) and different values of $\boldsymbol{\delta}_n$.

| | $m$ | 10 | | | 20 | | |
|---|---|---|---|---|---|---|---|
| | $j$ | 5 | 7 | 10 | 12 | 17 | 20 |
| Unconditional | $w_{0.025}$ | -3.671 | -2.814 | -0.907 | -3.140 | -1.760 | -0.380 |
| | $w_{0.975}$ | 1.278 | 2.635 | 9.761 | 1.827 | 4.767 | 12.186 |
| | $w_{0.975}-w_{0.025}$ | **4.949** | **5.449** | **10.668** | **4.967** | **6.527** | **12.566** |
| $\boldsymbol{\delta}_n=(1,2)$ | $w_{0.025}$ | -1.097 | -0.500 | 0.249 | -0.651 | 0.055 | 0.464 |
| $P_{\boldsymbol{\Delta}_n}(\boldsymbol{\delta}_n)=0.0833$ | $w_{0.975}$ | 1.766 | 3.502 | 11.868 | 2.459 | 6.108 | 14.586 |
| | $w_{0.975}-w_{0.025}$ | **2.863** | **4.002** | **11.619** | **3.110** | **6.053** | **14.122** |
| $\boldsymbol{\delta}_n=(1,3)$ | $w_{0.025}$ | -1.288 | -0.652 | 0.201 | -0.827 | -0.025 | 0.420 |
| $P_{\boldsymbol{\Delta}_n}(\boldsymbol{\delta}_n)=0.05$ | $w_{0.975}$ | 1.290 | 2.627 | 9.481 | 1.786 | 4.675 | 11.690 |
| | $w_{0.975}-w_{0.025}$ | **2.578** | **3.279** | **9.280** | **2.613** | **4.700** | **11.270** |
| $\boldsymbol{\delta}_n=(1,4)$ | $w_{0.025}$ | -1.427 | -0.774 | 0.160 | -0.965 | -0.098 | 0.386 |
| $P_{\boldsymbol{\Delta}_n}(\boldsymbol{\delta}_n)=0.0333$ | $w_{0.975}$ | 1.022 | 2.106 | 7.984 | 1.398 | 3.793 | 9.872 |
| | $w_{0.975}-w_{0.025}$ | **2.449** | **2.880** | **7.824** | **2.363** | **3.891** | **9.486** |
| $\boldsymbol{\delta}_n=(2,3)$ | $w_{0.025}$ | -2.181 | -1.267 | 0.045 | -1.538 | -0.320 | 0.324 |
| $P_{\boldsymbol{\Delta}_n}(\boldsymbol{\delta}_n)=0.0167$ | $w_{0.975}$ | 1.027 | 2.413 | 10.212 | 1.502 | 4.669 | 12.787 |
| | $w_{0.975}-w_{0.025}$ | **3.208** | **3.680** | **10.167** | **3.040** | **4.989** | **12.463** |
| $\boldsymbol{\delta}_n=(2,4)$ | $w_{0.025}$ | -2.330 | -1.409 | -0.008 | -1.697 | -0.415 | 0.289 |
| $P_{\boldsymbol{\Delta}_n}(\boldsymbol{\delta}_n)=0.0119$ | $w_{0.975}$ | 0.823 | 1.976 | 8.880 | 1.193 | 3.896 | 11.163 |
| | $w_{0.975}-w_{0.025}$ | **3.153** | **3.385** | **8.888** | **2.890** | **4.311** | **10.874** |

Conditionally on $\boldsymbol{\delta}_n$, we get more information about the unknown parameters $\mu$ and $\sigma$, or generally more information about $F$, which leads to better prediction intervals for $Y_{j:m}$. It is noted that conditioning on inter-record times does not decrease the length of the prediction interval necessarily and increase or decrease in the location and scale of the interval depend on the values of $\boldsymbol{\delta}_n$. For the purpose of illustration, consider the conditional quantiles of $W_j$, which are computed and tabulated in Table 1, for $\alpha=0.05$, $n=3$, $m=10,20$, $j=5,7,10$ ($m=10$), $j=12,17,20$ ($m=20$) and some values of $\boldsymbol{\delta}_n$. The values of unconditional quantiles of $W_j$ in Table 1 are taken from Ahmadi and MirMostafaee (2009), Tables 3 and 4. By comparing the entries of Table 1, one can observe that for a few cases, the conditional prediction intervals have bigger lengths, especially when we predict the biggest future order statistic, i.e. $Y_{m:m}$. But note that in the most cases the conditional intervals are shorter than the unconditional ones for different values of $\boldsymbol{\delta}_n$, so we may conclude that generally the conditional prediction approach leads to shorter (and hence better) prediction intervals in average for different values of $\boldsymbol{\delta}_n$ and this can be considered as an improvement.

# 4 Conditional Prediction Intervals for the mean of future sample

The problem of constructing a conditional prediction interval for $\bar{Y}_m$ on the basis of observed $U_1, \ldots, U_n$, given $\boldsymbol{\Delta}_n = \boldsymbol{\delta}_n$, using the pivotal quantity

$$V_m = \frac{\bar{Y}_m - U_1}{U_n - U_1}, \tag{4.1}$$

is considered for the two parameter exponential distribution in this section. Note that the pivotal quantity $V_m$ has been also considered by Ahmadi and MirMostafaee (2009) and its unconditional distribution has been obtained by them. Moreover, $V_m$ is also location and scale invariant and therefore is free of the unknown location and scale parameters. The following theorem presents the conditional distribution function of $V_m$ given $\boldsymbol{\Delta}_n = \boldsymbol{\delta}_n$.

**Theorem 2** *The conditional distribution function of* $V_m$ *in* (4.1) *given* $\boldsymbol{\Delta}_n = \boldsymbol{\delta}_n$ *is*

$$\begin{aligned} F_{V_m|\,\boldsymbol{\Delta}_n}(x|\boldsymbol{\delta}_n) &= 1 - \sum_{l=0}^{m-1}\sum_{j_1=1}^{n-1}\sum_{j_2=0}^{l}\sum_{j_3=0}^{a_1(n,j_1,\,\boldsymbol{\delta}_n)}\sum_{j_4=0}^{a_n(n,j_1,\,\boldsymbol{\delta}_n)} \frac{\binom{a_1(n,j_1,\,\boldsymbol{\delta}_n)}{j_3}\binom{a_n(n,j_1,\,\boldsymbol{\delta}_n)}{j_4}\binom{l}{j_2}}{(-1)^{j_3+j_4}P_{\boldsymbol{\Delta}_n}(\boldsymbol{\delta}_n)l!} \\ &\quad \times \frac{c_{j_1}(n,\boldsymbol{\delta}_n)x^{j_2}m^l\Gamma(l-j_2+1)\Gamma(j_2+1)}{(m+j_3+j_4+2)^{l-j_2+1}(mx+j_4+1)^{j_2+1}}, \end{aligned}$$

*for* $x > 0$, *and*

$$\begin{aligned} F_{V_m|\,\boldsymbol{\Delta}_n}(x|\boldsymbol{\delta}_n) &= \sum_{j_1=1}^{n-1}\sum_{j_3=0}^{a_1(n,j_1,\,\boldsymbol{\delta}_n)}\sum_{j_4=0}^{a_n(n,j_1,\,\boldsymbol{\delta}_n)} \frac{(-1)^{j_3+j_4}\binom{a_1(n,j_1,\,\boldsymbol{\delta}_n)}{j_3}\binom{a_n(n,j_1,\,\boldsymbol{\delta}_n)}{j_4}c_{j_1}(n,\boldsymbol{\delta}_n)}{P_{\boldsymbol{\Delta}_n}(\boldsymbol{\delta}_n)(2+j_3+j_4)[j_4+1-(2+j_3+j_4)x]} \\ &\quad - \sum_{l=0}^{m-1}\sum_{j_1=1}^{n-1}\sum_{j_2=0}^{l}\sum_{j_3=0}^{a_1(n,j_1,\,\boldsymbol{\delta}_n)}\sum_{j_4=0}^{a_n(n,j_1,\,\boldsymbol{\delta}_n)}\sum_{j_5=0}^{l-j_2} \frac{\binom{a_1(n,j_1,\,\boldsymbol{\delta}_n)}{j_3}\binom{a_n(n,j_1,\,\boldsymbol{\delta}_n)}{j_4}\binom{l}{j_2}}{(-1)^{j_3+j_4+j_5}P_{\boldsymbol{\Delta}_n}(\boldsymbol{\delta}_n)l!} \\ &\quad \times \frac{c_{j_1}(n,\boldsymbol{\delta}_n)x^{j_2+j_5}m^l\Gamma(l-j_2+1)\Gamma(j_2+j_5+1)}{j_5!(m+j_3+j_4+2)^{l-j_2-j_5+1}[j_4+1-(j_3+j_4+2)x]^{j_2+j_5+1}}, \end{aligned}$$

*for* $x < 0$, *where* $a_1(n, j_1,\ \boldsymbol{\delta}_n)$ *and* $a_n(n, j_1,\ \boldsymbol{\delta}_n)$ *are given in Lemma 1*

**Proof:** Let $J^*_{n,1} = (U_n - U_1)/\sigma$, $U^*_1 = (U_1 - \mu)/\sigma$ and $\bar{Y}^*_m = (\bar{Y}_m - \mu)/\sigma$. Note that

$$F_{V_m|\,\boldsymbol{\Delta}_n}(x|\boldsymbol{\delta}_n) = \int_0^\infty \int_0^\infty F_{\bar{Y}^*_m}(vx+u) f_{U^*_1, J^*_{n,1}|\,\boldsymbol{\Delta}_n}(u,v|\boldsymbol{\delta}_n)\,\mathrm{d}u\,\mathrm{d}v, \tag{4.2}$$

where $f_{U^*_1, J^*_{n,1}|\,\boldsymbol{\Delta}_n}(u,v|\ \boldsymbol{\delta}_n)$ is given in (3.4). Since $m\bar{Y}^*_m \sim \Gamma(m,1)$, that is for $t > 0$

$$F_{\bar{Y}^*_m}(t) = 1 - \sum_{l=0}^{m-1}\frac{(mt)^l e^{-mt}}{l!}, \tag{4.3}$$

so by substituting (3.4) and (4.3) in (4.2) and using the binomial expansions, we get for $x < 0$

$$F_{V_m|\,\boldsymbol{\Delta}_n}(x|\boldsymbol{\delta}_n) = \sum_{j_1=1}^{n-1} \sum_{j_3=0}^{a_1(n,j_1,\,\boldsymbol{\delta}_n)} \sum_{j_4=0}^{a_n(n,j_1,\,\boldsymbol{\delta}_n)} \frac{\binom{a_1(n,j_1,\,\boldsymbol{\delta}_n)}{j_3}\binom{a_n(n,j_1,\,\boldsymbol{\delta}_n)}{j_4} c_{j_1}(n,\boldsymbol{\delta}_n)}{(-1)^{j_3+j_4} P_{\boldsymbol{\Delta}_n}(\boldsymbol{\delta}_n)}$$

$$\times \int_0^\infty \int_{-vx}^\infty e^{-(j_3+j_4+2)u} e^{-(j_4+1)v}\, \mathrm{d}u\, \mathrm{d}v$$

$$- \sum_{l=0}^{m-1} \sum_{j_1=1}^{n-1} \sum_{j_2=0}^{l} \sum_{j_3=0}^{a_1(n,j_1,\,\boldsymbol{\delta}_n)} \sum_{j_4=0}^{a_n(n,j_1,\,\boldsymbol{\delta}_n)} \frac{\binom{a_1(n,j_1,\,\boldsymbol{\delta}_n)}{j_3}\binom{a_n(n,j_1,\,\boldsymbol{\delta}_n)}{j_4}\binom{l}{j_2}}{(-1)^{j_3+j_4} P_{\boldsymbol{\Delta}_n}(\boldsymbol{\delta}_n) l!}$$

$$\times c_{j_1}(n,\boldsymbol{\delta}_n) x^{j_2} m^l \int_0^\infty \int_{-vx}^\infty e^{-(m+j_3+j_4+2)u} e^{-(mx+j_4+1)v} u^{l-j_2} v^{j_2}\, \mathrm{d}u\, \mathrm{d}v$$

$$= \sum_{j_1=1}^{n-1} \sum_{j_3=0}^{a_1(n,j_1,\,\boldsymbol{\delta}_n)} \sum_{j_4=0}^{a_n(n,j_1,\,\boldsymbol{\delta}_n)} \frac{(-1)^{j_3+j_4}\binom{a_1(n,j_1,\,\boldsymbol{\delta}_n)}{j_3}\binom{a_n(n,j_1,\,\boldsymbol{\delta}_n)}{j_4} c_{j_1}(n,\boldsymbol{\delta}_n)}{P_{\boldsymbol{\Delta}_n}(\boldsymbol{\delta}_n)(2+j_3+j_4)[j_4+1-(2+j_3+j_4)x]}$$

$$- \sum_{l=0}^{m-1} \sum_{j_1=1}^{n-1} \sum_{j_2=0}^{l} \sum_{j_3=0}^{a_1(n,j_1,\,\boldsymbol{\delta}_n)} \sum_{j_4=0}^{a_n(n,j_1,\,\boldsymbol{\delta}_n)} \sum_{j_5=0}^{l-j_2} \frac{\binom{a_1(n,j_1,\,\boldsymbol{\delta}_n)}{j_3}\binom{a_n(n,j_1,\,\boldsymbol{\delta}_n)}{j_4}\binom{l}{j_2}}{(-1)^{j_3+j_4+j_5} P_{\boldsymbol{\Delta}_n}(\boldsymbol{\delta}_n) l!}$$

$$\times \frac{c_{j_1}(n,\boldsymbol{\delta}_n) x^{j_2+j_5} m^l \Gamma(l-j_2+1)}{j_5!(m+j_3+j_4+2)^{l-j_2-j_5+1}} \int_0^\infty e^{-(j_4+1-(j_3+j_4+2)x)v} v^{j_2+j_5}\, \mathrm{d}v$$

and therefore we naturally attain the desired result. Similarly, we may deduce the desired expression for $F_{V_m|\,\boldsymbol{\Delta}_n}(x|\boldsymbol{\delta}_n)$ when $x > 0$. □

To find conditional prediction interval for $\bar{Y}_m$ based on records given $\boldsymbol{\Delta}_n = \boldsymbol{\delta}_n$, we have to find the conditional quantiles of $V_m$ given $\boldsymbol{\Delta}_n = \boldsymbol{\delta}_n$, $v_{\alpha_1}(n,m;\boldsymbol{\delta}_n)$ and $v_{1-\alpha_2}(n,m;\boldsymbol{\delta}_n)$, for $\alpha_1+\alpha_2=\alpha$, $0<\alpha_i<1$, $i=1,2$, numerically, where

$$\Pr(V_m < v_\gamma(n,m;\boldsymbol{\delta}_n)|\ \boldsymbol{\Delta}_n = \boldsymbol{\delta}_n) = \gamma.$$

A $100(1-\alpha)\%$ conditional prediction interval for $\bar{Y}_m$ based on record values given $\boldsymbol{\Delta}_n = \boldsymbol{\delta}_n$ then is

$$(U_1 + v_{\alpha_1}(n,m;\boldsymbol{\delta}_n)(U_n - U_1), U_1 + v_{1-\alpha_2}(n,m;\boldsymbol{\delta}_n)(U_n - U_1)). \tag{4.4}$$

An illustrative example has been presented in Section 5.

# 5 An illustrative example

In this section, we illustrate the proposed procedures by considering a real data set. A rock crushing machine has to be reset if, at any operation, the size of rock being crushed

**Table 2:** 95% CPIs and UPIs for $Y_{12:20}$, $Y_{20:20}$ and $\bar{Y}_{20}$ for Example 1.

| | CPI | UPI |
|---|---|---|
| $Y_{12:20}$ | (0, 24.17836) | (0, 54.061745) |
| $Y_{20:20}$ | (13.290315, 183.67385) | (0, 307.85602) |
| $\bar{Y}_{20}$ | (0, 26.233175) | (0, 61.32183) |

is larger than any that has been crushed before. The following data given by Dunsmore (1983) are the sizes dealt with up to the third time that the machine has been reset:

9.3, 0.6, 24.4, 18.1, 6.6, 9.0, 14.3, 6.6, 13.0, 2.4, 5.6, 33.8.

The record values were the sizes at the operation when resetting was necessary. Dunsmore (1983) assumed that these data follow an $Exp(0, \sigma)$ distribution. Clearly, we have

$$U_1 = 9.3, \;\; U_2 = 24.4, \;\; U_3 = 33.8,$$

$$T_1 = 1, \;\; T_2 = 3, \;\; T_3 = 12,$$

$$\Delta_1 = 2, \quad \text{and} \quad \Delta_2 = 9.$$

Consider a future sample of size $m = 20$. We want to find equi-tailed 95% conditional prediction intervals (CPIs) for $Y_{12:20}$, $Y_{20:20}$ and $\bar{Y}_{20}$ using (3.5) and (4.4) and compare these intervals with unconditional ones (UPIs). The results are given in Table 2. Note that some lower bounds have got negative values, which were replaced by zero. We can see that the conditional prediction intervals are shorter than the corresponding unconditional ones.

# 6 Concluding remarks

In this paper, we found prediction intervals for the future order statistics based on record values, given record time statistics, when the underlying distribution is two parameter exponential. These intervals have the advantage of utilizing more information embedded in the observed sequence in comparison with their corresponding unconditional ones obtained by Ahmadi and MirMostafaee (2009). These ideas can be extended to the non-parametric and the Bayesian context. The conditional point predictors are also of interest. Work on these problems is currently under process and we hope to report these findings in future papers.

# Acknowledgement

We are very grateful to the respected editor and the respected referees for their insightful comments and suggestions which have led to this improved version.

# References


[1] Ahmadi, J. and Balakrishnan, N. (2010): Prediction of order statistics and record values from two independent sequences, *Statistics*, **44**, 417–430.

[2] Ahmadi, J. and MirMostafaee, S.M.T.K. (2009): Prediction intervals for future records and order statistics coming from two parameter exponential distribution, *Statistics and Probability Letters*, **79**, 977–983.

[3] Ahmadi, J. and MirMostafaee, S.M.T.K. and Balakrishnan, N. (2010): Nonparametric prediction intervals for future record intervals based on order statistics, *Statistics and Probability Letters*, **80**, 1663–1672.

[4] Arnold, B.C., Balakrishnan, N., and Nagaraja, H.N. (1998): *Records*, John Wiley & Sons, New York.

[5] Awad, A.M. and Raqab, M.Z. (2000): Prediction intervals for the future record values from exponential distribution: comparative study. *Journal of Statistical Computation and Simulation*, **65**, 325–340.

[6] Chou, Youn-Min. (1988): One-sided simultaneous prediction intervals for the order statistics of $l$ future samples from an exponential distribution. *Communications in Statististics-Theory and Methods*, **17**, 3995–4003.

[7] David, H.A. and Nagaraja, H.N. (2003): *Order Statistics*, Third edition, John Wiley & Sons, New York.

[8] Dunsmore, I.R. (1983): The future occurrence of records. *Annals of the Institute of Statistical Mathematics*, **35**, 276–277.

[9] Doostparast, M. (2009): A note on estimation based on record data. *Metrika*, **69**, 69–80.

[10] Doostparast, M. and Balakrishnan, N. (2013): Pareto analysis based on records. *Statistics*, **47**, 1075–1089.

[11] Feuerverger, A. and Hall, P. (1998): On statistical inference based on record values. *Extremes*, **1**, 169–190.

[12] Kızılaslan, F. and Nadar, M. (2015): Estimation with the generalized exponential distribution based on record values and inter-record times. *Journal of Statistical Computation and Simulation*, **85**, 978–999.

[13] Lin, C.T., Wu, S.J.S and Balakrishnan, N. (2003): Parameter estimation for the linear hazard rate distribution based on records and inter-record times. *Communications in Statististics-Theory and Methods*, **32**, 729–748.

[14] MirMostafaee, S.M.T.K. and Ahmadi, J. (2011): Point prediction of future order statistics from exponential distribution, *Statistics and Probability Letters*, **81**, 360–370.

[15] MirMostafaee, S.M.T.K., Amini, M. and Balakrishnan, N. (2016): Exact nonparametric conditional inference based on k-records, given inter k-record times. *Journal of the Korean Statistical Society*, Accepted.

[16] Nagaraja, H.N. (1984): Asymptotic linear prediction of extreme order statistics. *Annals of the Institute of Statistical Mathematics*, **36**, 289–299.

[17] Raqab, M.Z. and Balakrishnan, N. (2008): Prediction intervals for future records. *Statistics and Probability Letters*, **78**, 1955-1963.

[18] Samaniego, F.J. and Whitaker, L.R. (1986): On estimating population characteristics from record-breaking observations. I. parametric results. *Naval Research Logistics Quarterly*, **33**, 531–543.

# Appendix

Here, we present the R codes for computing the conditional cumulative distribution functions of $W_j$, (see Theorem 1) and $V_m$ (see Theorem 2). R functions for computing the unconditional cumulative distribution functions of $W_j$ and $V_m$ (see Ahmadi and MirMostafaee, 2009) are also given.

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%  cjn function %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

cjn=function(n,j,delta){
z=(-1)^(n-j-1)
z1=n-j+1
z2=j-2
z4=n-j-2
z5=n-j
s=1
if(z2>=0 & z1>=0){
for(j1 in 0:z2){
z3=n-j1-1
ss=ifelse(z3>=z1,sum(delta[z1:z3]),0)
s=s*ss
}}
t=1
if(z4>=0){
for(j2 in 0:z4){
z6=j2+2
tt=ifelse(z5>=z6,sum(delta[z6:z5]),0)
t=t*tt
}}
return(z/t/s)
```

```
}



%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%% Mass probability of Delta %%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

pdelta=function(n,delta){
n1=n-1
pdel=0
for(jj in 1:n1){
nj=n-jj
nj1=n-jj+1
A=cjn(n,jj,delta)
a1=ifelse(nj>=1,sum(delta[1:nj]),0)-1
an=ifelse(n1>=nj1,sum(delta[nj1:n1]),0)
C=(a1+1)*(a1+an+2)
pdel=pdel+A/C
}
return(pdel)
}

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%% conditional cdf of W      %%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

Fw=function(n,j,m,w,delta){
n1=n-1
pw=0
for(l in j:m){
for(j1 in 1:n1){
for(j2 in 0:l){
nj1=n-j1+1
nj=n-j1
a1=ifelse(nj>=1,sum(delta[1:nj]),0)-1
an=ifelse(n1>=nj1,sum(delta[nj1:n1]),0)
for(j3 in 0:a1){
for(j4 in 0:an){
A=choose(m,l)*choose(a1,j3)*choose(an,j4)*choose(l,j2)
*((-1)^(j2+j3+j4))*cjn(n,j1,delta)/pdelta(n,delta)
B=j2+m-l+j3+j4+2
if(w<0) C=B*(j4+1-w*(j3+j4+2))
if(w>=0) C=B*(w*(j2+m-l)+j4+1)
pw=pw+A/C
}}}}}
```

```
return(pw)
}


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%% conditional cdf of V      %%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

Fv=function(n,m,v,delta){
pv=0
n1=n-1
m1=m-1
if(v>=0){
for(l in 0:m1){
for(j1 in 1:n1){
for(j2 in 0:l){
nj1=n-j1+1
nj=n-j1
a1=ifelse(nj>=1,sum(delta[1:nj]),0)-1
an=ifelse(n1>=nj1,sum(delta[nj1:n1]),0)
for(j3 in 0:a1){
for(j4 in 0:an){
A=choose(a1,j3)*choose(an,j4)*choose(l,j2)/factorial(l)
/pdelta(n,delta)*((-1)^(j3+j4))
B=cjn(n,j1,delta)*(v^j2)*(m^l)*gamma(l-j2+1)*gamma(j2+1)
/((m+j3+j4+2)^(l-j2+1))/((m*v+j4+1)^(j2+1))
pv=pv+A*B
}}}}}}
if(v>=0) pv=1-pv
pv1=0
pv2=0
if(v<0){
for(j1 in 1:n1){
nj1=n-j1+1
nj=n-j1
a1=ifelse(nj>=1,sum(delta[1:nj]),0)-1
an=ifelse(n1>=nj1,sum(delta[nj1:n1]),0)
for(j3 in 0:a1){
for(j4 in 0:an){
A=((-1)^(j3+j4))*choose(a1,j3)*choose(an,j4)*cjn(n,j1,delta)
/pdelta(n,delta)/(2+j3+j4)/(j4+1-v*(2+j3+j4))
pv1=pv1+A
}}}
for(l in 0:m1){
for(j1 in 1:n1){
for(j2 in 0:l){
```

```
nj1=n-j1+1
nj=n-j1
a1=ifelse(nj>=1,sum(delta[1:nj]),0)-1
an=ifelse(n1>=nj1,sum(delta[nj1:n1]),0)
for(j3 in 0:a1){
for(j4 in 0:an){
lj2=l-j2
for(j5 in 0:lj2){
A=choose(a1,j3)*choose(an,j4)*choose(l,j2)/factorial(l)
/pdelta(n,delta)*((-1)^(j3+j4+j5))
B=cjn(n,j1,delta)*(v^(j2+j5))*(m^l)*gamma(l-j2+1)
*gamma(j2+j5+1)/factorial(j5)/((m+j3+j4+2)^(l-j2-j5+1))
/((j4+1-v*(j3+j4+2))^(j2+j5+1))
pv2=pv2+A*B
}}}}}}
pv=pv1-pv2
}
return(pv)
}
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%% unconditional cdf of W    %%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

FwU=function(n,j,m,w){
pw=0
if(w<0) pw=(m-j+1)*((1-w)^(1-n))/(m+1)
if(w>=0){
ss=0
j1=j-1
for(i in 0:j1){
ss=ss+choose(j1,i)*((-1)^i)*((1+w*(m-j+i+1))^(1-n))
/(m-j+i+1)/(m-j+i+2)
}
pw=1-j*choose(m,j)*ss
}
return(pw)
}

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%% unconditional cdf of V    %%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

FvU=function(n,m,v){
pv=0
if(v<0) pv=((1-v)^(1-n))/((1+1/m)^m)
if(v>=0){
```

```
m1=m-1
s1=0
s2=0
for(i in 0:m1){
nn=n+i-2
s1=s1+choose(nn,i)*((1-1/(m*v+1))^i)*((1/(m*v+1))^(n-1))
*((m/(m+1))^(m-i))
s2=s2+choose(nn,i)*((1-1/(m*v+1))^i)*((1/(m*v+1))^(n-1))
}
pv=s1+1-s2
}
return(pv)
}
```

Metodološki zvezki, Vol. 13, No. 1, 17-25

# Series of Incomplete Row-Column Designs with Two Units per Cell

**Anindita Datta[1], Seema Jaggi[1], Cini Varghese[1] and Eldho Varghese[1]**

**Abstract**

Here, two series of incomplete row-column designs with two units per cell have been developed that are structurally complete, i.e. all the cells corresponding to the intersection of row and column receive two distinct treatments. Properties of these classes of designs have been studied and the methods result in designs in which the elementary contrasts of treatment effects are estimated with same variance.

## 1. Introduction

Row-column designs are used for controlling heterogeneity in the experimental material in two directions. Most of the row-column designs developed in the literature have one unit corresponding to the intersection of row and column. Row-column designs with more than one unit per cell are used when the number of treatments is substantially large with limited number of replicates. For example, (Bailey and Monod, 2001), to conduct an experiment for comparing 4 treatments using 4 plants with leaves at 2 different heights, the following row-column design with complete rows and columns having two units per cell can be used:

| | Plants | | | |
|---|---|---|---|---|
| Leaf Height | 1 2 | 2 3 | 3 4 | 4 1 |
| | 3 4 | 4 1 | 1 2 | 2 3 |

These designs are termed as semi-Latin squares in the literature. An $(n \times n)/k$ semi-Latin square is an arrangement of $nk$ symbols (treatments) in an $(n \times n)$ square array such that

[1] ICAR-Indian Agricultural Statistics Research Institute, Library Avenue, New Delhi-110 012. India. Tel.: +91-11-25847284, E-mail address: seema.iasri@outlook.com, seema@iasri.res.in

each row-column intersection contains k symbols and each symbol occurs once in each row and each column. Trojan squares are a special class of semi-Latin squares based on sets of mutually orthogonal superimposed Latin squares and have been shown to be maximally efficient for pair-wise treatment comparisons in the plots-within-blocks stratum (Bailey, 1992). Following is an example of a Trojan square design of size (4 × 4)/2 for 8 treatments constructed by superimposing two mutually orthogonal Latin squares of size 4, one with 1, 2, 3 and 4 treatments and the other with 5, 6, 7 and 8 treatments.

| | Columns | | | |
|---|---|---|---|---|
| Rows | 1 5 | 2 6 | 3 7 | 4 8 |
| | 2 7 | 1 8 | 4 5 | 3 6 |
| | 3 8 | 4 7 | 1 6 | 2 5 |
| | 4 6 | 3 5 | 2 8 | 1 7 |

This arrangement could be extended for 12 treatments of size (4 × 4)/3 by superimposing the third orthogonal Latin square of size 4 but no further Trojan extension is possible, there being only three mutually orthogonal Latin squares of size 4.

Complete Trojan squares of size (n × n)/k have $n^2$ blocks of size k and require n replicates of nk treatments. Sometimes, design or cost constraints make complete Trojan squares impossible and then Incomplete Trojan squares of size [(n-1) × n]/ k or of size [n × (n-1)]/ k can be useful. Such incomplete Trojan squares can be constructed by omitting any complete row or any complete column from any Trojan design of size (n x n)/k. Trojan squares were first discussed by Harshbarger and Davis (1952) but then it was named as Latinized Near Balanced Rectangular Lattices having k = n-1. Later, Darby and Gilbert (1958) discussed the general case for k < n and introduced the name Trojan square designs where k > 2. However, all designs of the Latinized Rectangular Lattice type are now commonly described as Trojan squares for any 1 < k < n. Williams (1986) generalized the notion and called semi-Latin squares as Latinized incomplete-block designs. Andersen and Hilton (1980) called semi-Latin squares as (1, 1, k) Latin rectangles.

Preece and Freeman (1983) discussed the combinatorial properties of semi-Latin squares and related designs. Bailey (1988) discussed further construction for a range of semi-Latin and Trojan square designs. Bailey (1992) gave methods of constructing a range of semi-Latin and Trojan square designs, studied their efficiencies and showed that the Trojan squares are the optimal choice of semi-Latin squares for pair-wise comparisons of treatment means. These are particularly suitable for crop research experiments either in field or in the glasshouse. Trojan squares are normally the best choice of semi-Latin squares for crop research (Edmondson, 1998). Bedford and Whitaker (2001) have given several methods of construction of semi-Latin squares.

Dharmalingam (2002) gave an application of Trojan square designs and used it to obtain partial triallel crosses. Edmondson (2002) constructed generalized incomplete Trojan square designs, denoted by $(m \times n)/k$ where m denotes the number of replicates of nk treatments, based on a set of k cyclic generators.

There existed three optimal $(4 \times 4)/4$ semi-Latin squares (Bailey and Chigbu, 1997) for sixteen treatments in blocks of size four. Since these squares do not have the same concurrences, there was a need for distinguishing one square from the others and determining the most preferred square in a given context. Chigbu (2003) obtained the best of the three optimal $(4 \times 4)/4$ semi-Latin squares by finding and comparing the variances of elementary contrasts of treatments for the squares.

Jaggi et al. (2010) defined Generalized Incomplete Trojan-Type Designs and developed method of constructing these designs. Varghese and Jaggi (2011) obtained generalized row-column designs and showed their application in obtaining mating plans. Datta et al. (2014) obtained some methods of constructing row-column designs with multiple units per cell that are structurally incomplete i.e. corresponding to the intersection of any row and column, there is at least one cell which does not contain any treatment. Datta et al. (2015) developed row-column designs with multiple units per cell with equal/ unequal cell sizes. Jaggi et al. (2016) obtained another series of generalized incomplete Trojan-type designs for number of treatments $v= sm+1$.

Most of the methods available in the literature are for complete rows and complete columns. Here, two methods of constructing row-column designs with two units per cell in incomplete rows or columns are obtained that are balanced for estimating elementary contrasts of treatment effects.

## 2. Experimental Setup and Information Matrix

We consider a row-column design with v treatments arranged in m rows, n columns and in each row-column intersection (i.e. cells) there are k units or plots resulting in total mnk experimental units or observations. The following four-way classified model with treatments, rows, columns and cells as the four classifications, is considered:

$$\mathbf{Y} = \mathbf{X}_1\theta_1 + \mathbf{X}_2\theta_2 + \mathbf{e},$$

where

$$\mathbf{X_1} = [\ \boldsymbol{\Delta}'\ ]\ , \quad \mathbf{X_2} = [\mathbf{1} \quad \mathbf{D}_1' \quad \mathbf{D}_2' \quad \mathbf{D}_3']',$$

$\theta_1 = (\boldsymbol{\tau})$ is the vector of parameters of interest and $\theta_2 = (\mu \quad \boldsymbol{\beta} \quad \boldsymbol{\gamma} \quad \boldsymbol{\eta})$ is the vector of nuisance parameters. $\mathbf{Y}$ is a mnk × 1 vector of observations, $\mu$ is the grand mean, $\mathbf{1}$ is the mnk × 1 vector of ones, $\boldsymbol{\Delta}'$ is mnk × v matrix of observations versus treatments, $\boldsymbol{\tau}$ is a v × 1 vector of treatment effects, $\mathbf{D}_1'$ is mnk × m matrix of observations versus rows, $\boldsymbol{\beta}$ is m × 1 vector of row effects, $\mathbf{D}_2'$ is mnk × n matrix of observations versus columns, $\boldsymbol{\gamma}$ is n × 1 vector of column effects, $\mathbf{D}_3'$ is mnk × mn matrix of observations versus cells, $\boldsymbol{\eta}$ is mn × 1 vector of cell effects and $\mathbf{e}$ is mnk × 1 vector of random errors with $E(\mathbf{e}) = 0$ and $D(\mathbf{e}) = \sigma^2\,\mathbf{I}$.

The information matrix of row-column design with multiple units per cell for treatment effects is obtained as

$$\begin{aligned}\mathbf{C} = \mathbf{R}_\tau - \{&\left(\mathbf{N}_1\mathbf{K}_{11}\mathbf{N}_1' + \mathbf{N}_2\mathbf{K}_{21}\mathbf{N}_1' + \mathbf{N}_3\mathbf{K}_{31}\mathbf{N}_1' + \mathbf{N}_1\mathbf{K}_{12}\mathbf{N}_2' + \mathbf{N}_2\mathbf{K}_{22}\mathbf{N}_2'\right) \\ &+ (\mathbf{N}_3\mathbf{K}_{23}\mathbf{N}_2' + \mathbf{N}_1\mathbf{K}_{13}\mathbf{N}_3' + \mathbf{N}_2\mathbf{K}_{23}\mathbf{N}_3' + \mathbf{N}_3\mathbf{K}_{33}\mathbf{N}_3')\}\end{aligned} \qquad \ldots(2.1)$$

where

$$\mathbf{K}_{11} = \mathbf{K}_\beta^- + \left[\mathbf{K}_\beta^- (\mathbf{M}_1\mathbf{K}_{22}\mathbf{M}_1' + \mathbf{M}_2\mathbf{K}_{23}'\mathbf{M}_1' + \mathbf{M}_1\mathbf{K}_{23}\mathbf{M}_1' + \mathbf{M}_2\mathbf{K}_{33}\mathbf{M}_2')\mathbf{K}_\beta^-\right]$$

$$\mathbf{K}_{12} = -\mathbf{K}_\beta^- \left(\mathbf{M}_1\mathbf{K}_{22} + \mathbf{M}_2\mathbf{K}_{23}\right)$$

$$\mathbf{K}_{13} = \mathbf{K}_\beta^- \left(\mathbf{M}_1\mathbf{K}_{23} + \mathbf{M}_2\mathbf{K}_{33}\right)$$

$$\mathbf{K}_{22} = \mathbf{K}_\gamma^- + \mathbf{K}_\gamma^-\mathbf{M}_3\,(\mathbf{K}_\eta - \mathbf{M}_3'\mathbf{K}_\gamma^-\mathbf{M}_3)^-\,\mathbf{M}_3'\mathbf{K}_\gamma^-$$

$$\mathbf{K}_{23} = -\mathbf{K}_\gamma^-\mathbf{M}_3\,(\mathbf{K}_\eta - \mathbf{M}_3'\mathbf{K}_\gamma^-\mathbf{M}_3)^-$$

$$\mathbf{K}_{33} = (\mathbf{K}_\eta - \mathbf{M}_3'\mathbf{K}_\gamma^-\mathbf{M}_3)^-$$

$\mathbf{R}_\tau$ is the diagonal matrix of replication of treatments, $\mathbf{K}_\beta$ is the diagonal matrix of row-sizes, $\mathbf{K}_\gamma$ is the diagonal matrix of column-sizes and $\mathbf{K}_\eta$ is the diagonal matrix of cell-sizes. $\mathbf{N}_1$ is the incidence matrix of treatments versus rows, $\mathbf{N}_2$ is the incidence matrix of treatments versus columns, $\mathbf{N}_3$ is the incidence matrix of treatments versus cells, $\mathbf{M}_1$ is the incidence matrix of rows versus columns, $\mathbf{M}_2$ is the incidence matrix of rows versus cells and $\mathbf{M}_3$ is the incidence matrix of columns versus cells.

The $v \times v$ matrix $\mathbf{C}$ is symmetric, non-negative definite with zero row and column sums.

# 3. Methods of Construction

We present here two methods of constructing row-column designs with two units per cell. The first method is for odd number of treatments and the second is for even number.

**Method 3.1:** For $v = 2t + 1$ ($t > 1$), obtain the following initial column having two units per cell:

| | |
|---|---|
| 1 | 2t + 1 |
| 2 | 2t |
| 3 | 2t - 1 |
| . | . |
| . | . |
| . | . |
| t | 2t - (t - 2) |

Develop 2t more columns horizontally from the initial column by adding 1,2,...,2t consecutively reducing mod (2t+1). The resulting design is a row-column design with two units per cell and with m = t rows of size 2(2t+1), n = (2t+1) columns of size 2t, k = 2 and r = 2t replications. The design is complete row-wise and column-wise it is incomplete. The information matrix of the design for treatment effects is obtained from (2.1) as

$$\mathbf{C} = (t + 0.5)\mathbf{I} - 0.5\mathbf{J}\,.$$

It is seen that all the elementary treatment contrasts are estimated with same variance. This method would give designs for all odd number of treatments.

**Example 3.1.1:** Let t = 3, so v = 7. The contents of the initial column are obtained as follows:

| | |
|---|---|
| 1 | 7 |
| 2 | 6 |
| 3 | 5 |

Developing this column by adding 1,2,…,6 reducing mod 7 would result in the following row-column design in three rows of size 14, 7 columns of size 6 with 2 units per cell and replication of each treatment being 6:

| | Columns | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Rows | 1 | 7 | 2 | 1 | 3 | 2 | 4 | 3 | 5 | 4 | 6 | 5 | 7 | 6 |
| | 2 | 6 | 3 | 7 | 4 | 1 | 5 | 2 | 6 | 3 | 7 | 4 | 1 | 5 |
| | 3 | 5 | 4 | 6 | 5 | 7 | 6 | 1 | 7 | 2 | 1 | 3 | 2 | 4 |

The information matrix for estimating treatment effects is $\mathbf{C} = 3.5\mathbf{I} - 0.5\mathbf{J}$ .

**Example 3.1.2:** For t = 4, the contents of the initial column for v = 9 are obtained as follows:

| | |
|---|---|
| 1 | 9 |
| 2 | 8 |
| 3 | 7 |
| 4 | 6 |

The row-column design in 4 rows of size 18, 9 columns of size 8 and 8 replications is constructed as given.

| | Columns | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Rows | 1 | 9 | 2 | 1 | 3 | 2 | 4 | 3 | 5 | 4 | 6 | 5 | 7 | 6 | 8 | 7 | 9 | 8 |
| | 2 | 8 | 3 | 9 | 4 | 1 | 5 | 2 | 6 | 3 | 7 | 4 | 8 | 5 | 9 | 6 | 1 | 7 |
| | 3 | 7 | 4 | 8 | 5 | 9 | 6 | 1 | 7 | 2 | 8 | 3 | 9 | 4 | 1 | 5 | 2 | 6 |
| | 4 | 6 | 5 | 7 | 6 | 8 | 7 | 9 | 8 | 1 | 9 | 2 | 1 | 3 | 2 | 4 | 3 | 5 |

Here, the design is complete row-wise with each treatment occurring twice and column-wise it is incomplete.

**Method 3.2:** For v even, obtain the following initial column with 2 units per cell:

| | |
|---|---|
| 1 | v |
| v | 2 |
| 2 | v-1 |
| v-1 | 3 |
| . | . |
| . | . |
| . | . |
| $v - (\frac{v}{2}-2)$ | $v - \frac{v}{2}$ |
| $\frac{v}{2}$ | $\frac{v}{2}+1$ |

Develop $\frac{v}{2}-1$ more columns horizontally from the initial column by adding $1,2,...,(\frac{v}{2}-1)$ consecutively reducing mod v. The resulting design is a row-column design with 2 units per cell in incomplete rows and complete columns. The parameters of the design are v, m = (v-1) rows of size v, n = $\frac{v}{2}$ columns of size 2(v-1), k = 2 and r = (v-1). From (2.1), the information matrix for treatment effects is obtained as

$$C = \frac{v}{2}\mathbf{I} - 0.5\mathbf{J}.$$

Thus, all the elementary contrasts of treatment effects are estimated with same variance.

**Example 3.2.1:** For v = 8, following is a row-column design with cells containing 2 units in 7 rows of size 8 each and 4 columns of size 14 each:

| | Columns | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Rows | 1 | 8 | 2 | 1 | 3 | 2 | 4 | 3 |
| | 8 | 2 | 1 | 3 | 2 | 4 | 3 | 5 |
| | 2 | 7 | 3 | 8 | 4 | 1 | 5 | 2 |
| | 7 | 3 | 8 | 4 | 1 | 5 | 2 | 6 |
| | 3 | 6 | 4 | 7 | 5 | 8 | 6 | 1 |
| | 6 | 4 | 7 | 5 | 8 | 6 | 1 | 7 |
| | 4 | 5 | 5 | 6 | 6 | 7 | 7 | 8 |

The canonical efficiency factor of the above two series of designs is obtained as

$$E = \frac{H}{r}, \text{ where } H = \left( \frac{1}{v\text{-}1} \sum_{i=1}^{v-1} \lambda_i^{-1} \right)^{-1},$$

$\lambda_i$ are the eigen-values of **C**- matrix of the designs obtained and r is the number of replications of the treatments. It is assumed that $\sigma^2$ is same for the developed design and the orthogonal design to which it is compared. The canonical efficiency factor of the developed designs was worked out and was found to be fairly good.

## Acknowledgements

The authors are grateful to the Editor and the reviewers for the constructive comments that have led to considerable improvement in the paper.

## References

[1] Andersen, L.D. and Hilton, A.J.W. (1980): Generalized Latin rectangles I: Constructions and decomposition, *Discrete Maths*., **31**, 125-152.

[2] Bailey, R.A. (1988): Semi Latin squares, *J. Statist. Plan. Inf*., **18**, 299-312.

[3] Bailey, R.A. (1992): Efficient semi-Latin squares, *Statistica Sinica,* **2**, 413-437.

[4] Bailey, R.A. and Chigbu, P.E. (1997): Enumeration of semi-Latin squares, *Discrete Maths.*, **167/168**, 73-84.

[5] Bailey, R.A. and Monod, H. (2001): Efficient semi-Latin rectangles: Designs for plant disease experiments, *Scand. J. Statist*., **28**, 257-270.

[6] Bedford, D. and Whitaker, R.M. (2001): A new construction for efficient semi-Latin squares, *J. Statist. Plan. Inf.*, **98**, 287-292.

[7] Chigbu, P.E. (2003): The "best" of the three optimal (4 x 4)/4 semi-Latin squares, Sankhya*: The Indian Journal of Statistics*, **65**(3), 641–648.

[8] Darby, L.A. and Gilbert, N. (1958): The Trojan Square, *Euphytica*, **7**, 183-188.

[9] Datta, A., Jaggi, S., Varghese, C. and Varghese, E. (2014): Structurally incomplete row-column designs with multiple units per cell. *Statistics and Applications*, **12**, (1&2): 71-79.

[10] Datta, A., Jaggi, S., Varghese, C. and Varghese, E. (2015): Some series of row-column designs with multiple units per cell. *Calcutta Statistical Association Bulletin*, **67**, (265-266), 89-99.

[11] Dharmalingam, M. (2002): Construction of partial triallel crosses based on Trojan square design, *J. App. Statist*., **29**(5), 675-702.

[12] Edmondson, R.N. (1998): Trojan square and incomplete Trojan square design for crop research, *J. Agric. Sci.* **131**, 135-142.

[13] Edmondson, R.N. (2002): Generalized incomplete Trojan designs, *Biometrika*, **89**(4), 877-891.

[14] Harshbarger, B. and Davis, L.L. (1952): Latinized rectangular lattices, *Biometrics*, **8**, 73-84.

[14] Jaggi, Seema, Varghese, Cini, Varghese, Eldho and Sharma, V.K. (2010): Generalized incomplete Trojan-type designs, *Statistics and Probability Letters*, **80**, 706-710.

[15] Jaggi, Seema, Varghese, Cini, and Varghese Eldho (2016): A series of generalized incomplete Trojan-type designs. *Journal of Combinatorics, Information and System Sciences: American Journal*, **40**(1-4), 53-60.

[16] Preece, D.A. and Freeman, G.H. (1983): Semi-Latin squares and related designs, *J. R. Statist. Soc. B* **45**, 267-277.

[17] Varghese, Cini and Jaggi, Seema (2011): Mating designs using generalized incomplete Trojan-type designs, *Journal of Statistics and Applications*, **6**(3-4), 85-93.

[18] Williams, E.R. (1986): Row and column designs with contiguous replicates, *Australian J. Statist*., **28**, 154-163.

Metodološki zvezki, Vol. 13, No. 1, 2016, 27-58

# Adjustment of Recall Errors in Duration Data Using SIMEX

Jose Pina-Sánchez[1]

**Abstract**

It is widely accepted that due to memory failures retrospective survey questions tend to be prone to measurement error. However, the proportion of studies using such data that attempt to adjust for the measurement problem is shockingly low. Arguably, to a great extent this is due to both the complexity of the methods available and the need to access a subsample containing either a gold standard or replicated values. Here I suggest the implementation of a version of SIMEX capable of adjusting for the types of multiplicative measurement errors associated with memory failures in the retrospective report of durations of life-course events. SIMEX is a method relatively simple to implement and it does not require the use of replicated or validation data so long as the error process can be adequately specified. To assess the effectiveness of the method I use simulated data. I create twelve scenarios based on the combinations of three outcome models (linear, logit and Poisson) and four types of multiplicative errors (non-systematic, systematic negative, systematic positive and heteroscedastic) affecting one of the explanatory variables. I show that SIMEX can be satisfactorily implemented in each of these scenarios. Furthermore, the method can also achieve partial adjustments even in scenarios where the actual distribution and prevalence of the measurement error differs substantially from what is assumed in the adjustment, which makes it an interesting sensitivity tool in those cases where all that is known about the error process is reduced to an educated guess.

## 1 Introduction

Applied quantitative searchers commonly assume that variables included in their models are measured perfectly. This often implicit assumption is, however, difficult to maintain when using survey data as interviewer effects, interviewee fatigue, social desirability bias, lack of cooperation, or plain deceit inevitably introduce measurement error (ME). This is especially true for surveys using a retrospective design, which collect information about past events

[1] School of Law, University of Leeds, J.PinaSanchez@leeds.ac.uk

from a single contact with respondents. The advantages of retrospective designs, in comparison with prospective studies[2], are well known: a) immune to problems of attrition; b) cheaper to administer; and c) more capable of detecting transitions occurring in short periods. Retrospective questions are however prone to ME as they require respondents to both interpret the question correctly and recall events that took place in the past.

The consequences of using data affected by ME are both difficult to estimate and potentially disastrous (Nugent, Graycheck & Basham, 2000; and Vardeman et al, 2010)3. Unfortunately, the latter is rarely acknowledged, and in certain cases it is directly misunderstood For example, Carroll et al (2006) point at the widespread belief that ME affecting an explanatory variable will only attenuate the regression estimate of that variable[4]. Even amongst researchers that acknowledge the potential consequences of ME very little is done to tackle the problem besides mentioning it as a caveat. There are two reasons for this: the requirement of additional data and the complexity of the adjustment methods available.

Generally, methods for the adjustment of ME need to be informed about the true unobserved values using additional data. For example, multiple imputation (Rubin, 1987, and Cole, Chu & Greenland, 2006) requires access to a validation subsample where the true values are observed. Regression calibration (Carroll & Stefanski, 1990; and Glesjer, 1990) needs at least repeated measurements, while two stage least squares (Theil, 1953) requires instrumental variables. However, researchers' access to this type of data tends to be the exception rather than the norm. In addition, these three methods belong to the family of functional methods – i.e. those that do not make any assumptions about the distribution of the true values. A second group of methods known as structural methods are technically more complex, amongst other things because they require specifying the probability function of the unobserved true values. Examples of structural methods are likelihood based adjustments, either Bayesian or Frequentist. These methods account directly for the ME mechanism in place, which tend to involve ad hoc specifications, in turn increasing the complexity of the adjustment.

---

[2] See Solga (2001) for a comparison of data quality derived from prospective and retrospective questions.

[3] *"Measurement error is, to borrow a metaphor, a gremlin hiding in the details of our research that can contaminate the entire set of estimated regression parameters."* (Nugent, et al. 2000: 60). *"Even the most elementary statistical methods have their practical effectiveness limited by measurement variation."* (Vardeman et al., 2010: 46).

[4] *"Despite admonitions of Fuller (1987) and others to the contrary, it is a common perception that the effect of ME is always to attenuate the line. In fact, attenuation depends critically on the classical additive ME model."* (Carroll et al., 2006: 46).

In this paper I will use simulated data to study the effectiveness of multiplicative Simulation Extrapolation Method (SIMEX) (Carroll et al. 2006; and Biewen, Nolte & Rosemann, 2008). This is an extension of the standard SIMEX method (Cook & Stefanski, 1994) capable of adjusting for the recall errors that are typically observed in the retrospective reports of life-course events. SIMEX implementation is relatively simple in that only requires an estimate of the reliability of the variable affected by ME. This is normally obtained using a subsample of replicated data. Here I will assume that such information is not available to the researcher, as it is often the case. Instead I will use this technique to show its potential to carry out sensitivity analysis when the reliability ratios have to be assumed.

That is, I will demonstrate how the problem of recall errors so ubiquitous in retrospective data can be effectively dealt with by researchers who do not have neither the technical background to carry out complex adjustments, nor access to additional sources of data. In so doing my ultimate goal is to encourage a wider audience of survey researchers both to reflect about the implications of relying variables affected by ME and to consider the possibility of assessing the robustness of their findings.

In the following section I review the theory regarding the types of errors that can be expected from retrospective questions and the models that have been normally used to specify them. In Section 3, I present the simulated data that will be used in the analysis and illustrate the implications of using an explanatory variable affected by multiplicative errors in different outcome models. Section 4 lays out the functioning of the standard SIMEX and the extension considered to accommodate multiplicative ME. In Section 5 the results of the analysis are presented, and in Section 6 I conclude with a discussion of the relevance of the main findings.

# 2 Modelling Memory Failures in Retrospective Questions on Life-Course Events

Most studies aiming to assess the implications of ME or to adjust for them assume a simple error mechanism known as the classical ME model. This model was first formally defined by Novick (1966) as follows,

$$X_i^* = X_i + V_i \tag{2.1}$$

where $X^*$ is the observed variable, equal to the true variable $X$, plus the ME term $V$, which fulfils six important assumptions:

$$Classical\ Model \begin{cases} E(V) = 0\,; & null\ expectancy \\ Var(V_i) = Var(V)\,; & homoscedasticity \\ V \sim N\big(0, Var(V)\big)\,; & normally\ distributed \\ Cov(X, V) = 0\,; & independence\ error\ and\ true\ value \\ Cov\big(V_i, V_j\big) = 0\,; & independence\ between\ errors \\ Cov(\in, V) = 0\,; & non - differentiality \end{cases} \tag{2.2}$$

1. Null expectancy refers to the assumption that the error term is non-systematic, or in other words, the expected value of the error term is zero, $E(V) = 0$.

2. The assumption of homoscedasticity indicates that the variance of the error term is assumed to remain constant across subjects, $Var(V_i) = Var(V) = \sigma_V^2$.

3. In addition to having an expectation of zero and constant variance the error term is normally distributed, $V \sim N(0, \sigma_V^2)$.

4. The correlation between the true value and the error term is assumed to be zero, $Cov(X, V) = 0$.

5. Furthermore, the correlation between different values of the error term is also assumed to be zero, $Cov\big(V_i, V_j\big) = 0$, where $V_i$ and $V_j$ represent any two values of the error term, for subjects $i$ and $j$.

6. The last assumption, non-differentiality, only becomes relevant when $X^*$ is used in a regression model to specify a response variable, $Y$. It indicates that, given the true value, the ME is not associated with the residual term from the regression model, $\in$. That is, $E(Y|X, X^*) = E(Y|X)$, or alternatively, $Cov(\in, V) = 0$.

The second and sixth assumptions were not originally established by Novick (1966), but they have been included here because they are often required in the application of the adjustment methods. Assumptions 1 and 4 (null expectancy and independence of the error and true value) can be used to define the expected value and the variance of the true value as follows,

$$E(X^*) = E(X) + E(V) = E(X) \tag{2.3}$$

and,

$$Var(X^*) = Var(X) + Cov(X, V) + Var(V) = Var(X) + Var(V) \quad (2.4)$$

which can in turn be used to define the reliability of an observed variable affected by classical ME, $\rho_{X^*}$, as the ratio of the true to observed variance,

$$\rho_{X^*} = \frac{Var(X)}{Var(X^*)} = \frac{Var(X)}{Var(X) + Var(V)} \quad (2.5)$$

Notice that in order to calculate the reliability ratio either one of the unobserved variances ($Var(X)$ or $Var(V)$) needs to have been previously estimated.

The classical model reflects nicely the type of ME that we can expect to find in measurement processes that are prone to random errors. However, it is not the most appropriate model that we can use to reflect the memory failures associated with retrospective reports of developmental milestones, transitions of events, or ages at onset (Pickles, Pickering & Taylor, 1996), such as the recall of age at menarche, or the date since last employed. Those reports convey duration (or time-to-event) data, which by definition cannot be negative, a possible outcome in the classical model when $V_i$ is negative and bigger than $X_i$.

Furthermore, although determining the prevalence of ME in these types of questions is not always straightforward, it makes sense to think that the distance from the interview has something to do with the magnitude of the ME. Pickles et al (1998) point at different views on this issue. On the one hand, different authors (Golub, Johnson & Labouvie, 2000; Johnson & Schultz, 2005) have detected telescoping effects in reports of dates of onset of a particular event. The term telescoping was coined by Neter and Waksberg (1964) to refer to the temporal displacement of an event whereby people perceive recent events as being more remote than they are (known as backward telescoping or time expansion) and distant events as being more recent than they are (forward telescoping or time compression).[5] Golub, Johnson & Labouvie, (2000) and Johnson & Schultz (2005) have detected telescoping effects in reports of dates of onset of a particular event.

Conversely, another group of researchers (Huttenlocher, Hedges & Prohaska, 1988; Rubin and Baddeley, 1989; or Bradburn, Huttenlocher & Schwarz, 1994) have argued that rather than distorted time perceptions recall errors take the form of non-systematic ME around the reported date with its size being proportional to the distance between the day of the interview

[5] For a review of the cognitive processes resulting in telescoping see Janssen and Chessa (2006).

and the reported date. That is, the further away the date of the event to be reported the harder its recall and therefore the bigger the ME. Both these findings question the appropriateness of applying the additive ME model to retrospective designs.

ME induced by memory failures can instead be better represented as a multiplicative ME model (Holt, McDonald & Skinner, 1991; Pickles et al., 1996 and 1998; Skinner & Humphreys, 1999; Augustin, 1999; Glewwe, 2007; and Dumangane, 2007) that builds upon the classical ME model as follows,

$$X_i^* = X_i \cdot V_i \tag{2.6}$$

Here the multiplicative relation between $X$ and $V$ reflects that the effect of $V$ on the observed variable $X^*$ is proportional to the value of $X$. In addition, for the specific case of non-systematic multiplicative ME, most of the assumptions about the error term described in equation 2.2 apply. I will refer to this type of error as classical multiplicative from here on. The only exceptions are items 1 and 3. Here, $V$ follows a log-normal distribution bounded from 0 to $\infty$, and with mean equal to 1. This way the ME has a relatively symmetric effect across the true values and maintains the scale used in duration data. Note as well that, this same model can also be used to account for backward and forward telescoping effects by shifting the distribution of $V$ to the right or left, so its mean goes below or above 1.

# 3 The Impact of Classical Multiplicative Measurement Error in Regression Analyses

Gustafson (2003) traced out analytically the impact of classical multiplicative ME affecting an explanatory variable, $X^*$, in a linear model where a second explanatory variable, $Z$, is measured without error,

$$Y = \beta_0 + \beta_1 X^* + \beta_2 Z + \in \tag{3.1}$$

Classical multiplicative ME produces attenuation in the $\beta_1$ regression coefficient directly proportional to the variance of the ME term, increasing the more skewed to the right $X$ is, and as the correlation between the true variable $X$ and $Z$ grows stronger. Gustafson (2003) does not evaluate, however, the impact on the $\beta_2$ error-free regression coefficient or consider non-linear outcome models.

To address these issues, this section explores the impact of different types of multiplicative ME on three different generalised linear model specifications – linear, logit and Poisson regressions - containing error-prone and perfectly measured explanatory variables, $X^*$ and $Z$. A simulated dataset of 1000 observations allows sufficient statistical power to detect moderate parameter estimates while keeping a low computational burden given the number of scenarios that will be explored. To reflect the features of duration data, the true variable $X$ is taken to be exponentially distributed with mean 8.75 and range (1.15, 44.45); $Z$ follows a standard normal distribution, and both of them are associated with the response variable of the linear model, $Y$, which is also shaped as a standard normal distribution. In addition, to generate the response variables for the logit and Poisson models, $Y$ is recoded as a binary variable, $Y_{ca}$,

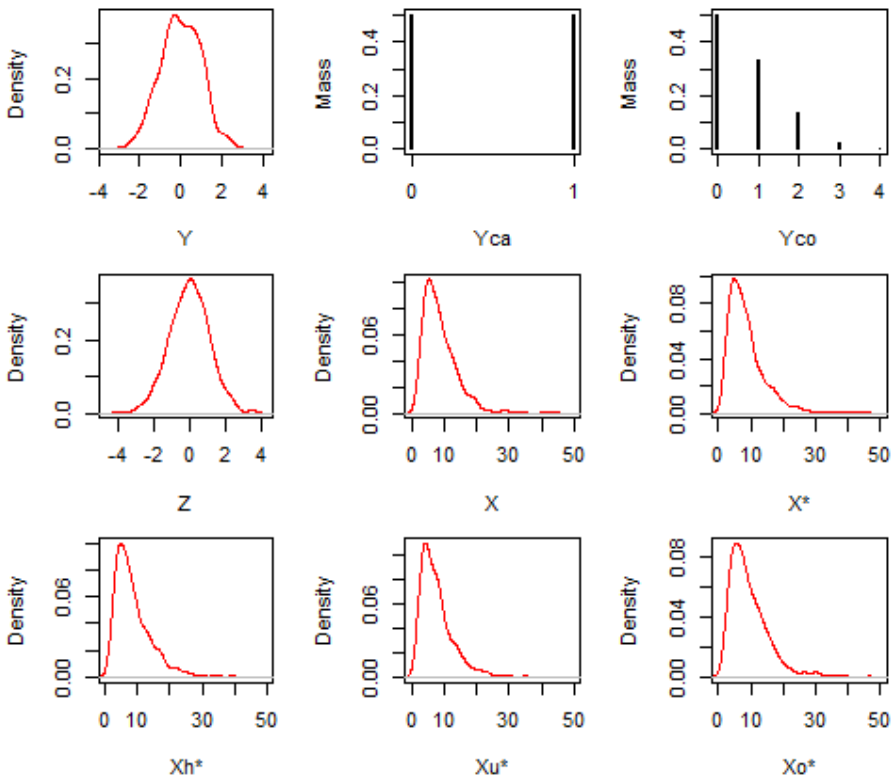
$$Y_{ca} \begin{cases} 1, & Y \geq 0 \\ 0, & Y < 0 \end{cases}$$

and as a count variable, $Y_{co}$,

$$Y_{co} \begin{cases} 0, & Y < 0 \\ 1, & 0 \leq Y > 1 \\ 2, & 1 \leq Y > 2 \\ 3, & 2 \leq Y > 3 \\ 4, & Y \geq 3 \end{cases}$$

The four simulated ME scenarios are represented by the variables, $X^*$, $X_h^*$, $X_u^*$ and $X_o^*$. In each of these scenarios $X$ is subject to normally distributed classical multiplicative ME. I choose to simulate normal instead of log-normal errors (as explained in equation 2.6) to ensure that they are perfectly symmetric around their mean (the latter are skewed to the right to a certain extent). Figure 1 shows the probability and mass functions for each of the variables simulated, while the specific code used in R is shown in Appendix I.

**Figure 1:** Probability Density and Mass Functions of the Simulated Variables

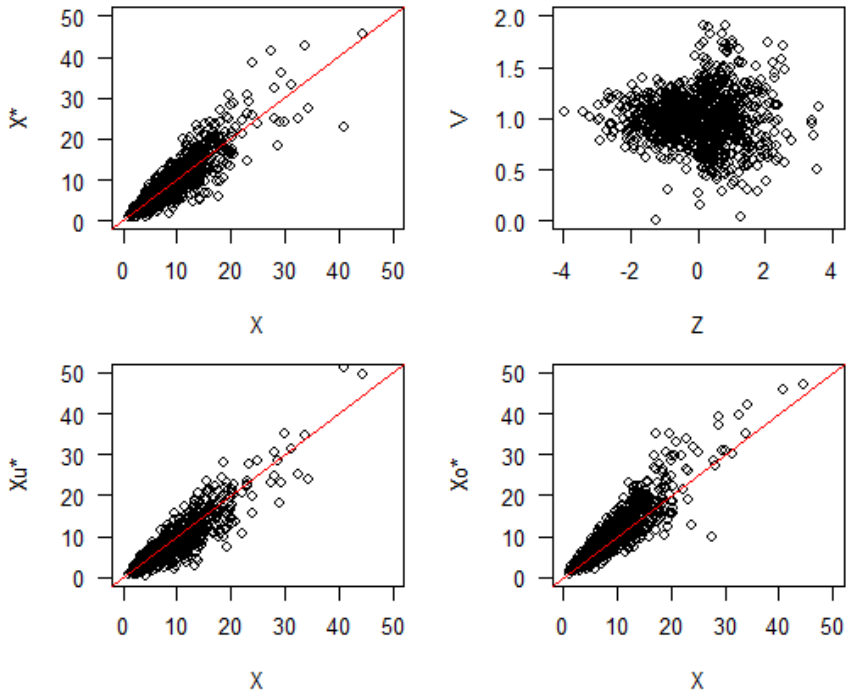



In the first ME scenario I simulate non-systematic errors distributed as a $N(1,.25)$. The multiplicative effect of these errors results in a new variable $X^*$ with a reliability ratio (RR) of .816. In the second scenario I explore the effect of heteroscedastic ME by changing the distribution of the errors from $N(1,.15)$ to $N(1,.35)$ when $Z > 0$. This is a type of ME that could take place when different survey modes are used. For example, Roberts (2007) - after reviewing the literature - concluded that telephone interviews place a higher cognitive demand on the interviewee than face-to-face interviews, which tend to make them more prone to measurement error. In the third scenario I study the effect of systematically underreported durations by simulating errors distributed as $N(.9,.25)$. These are the types of errors that could be expected in the presence of forward telescoping bias (e.g. Golub et al., 2000, and Johnson and Schultz, 2005, found evidence of these types of errors in reports of onset of drug usage and smoking, respectively), but also in the report of durations of socially undesirable events (e.g. Pina-Sánchez, Koskinen & Plewis, 2013 and Pina-Sánchez, Koskinen & Plewis, 2014, found an increased tendency to underreport the longer spells of unemployment). Lastly, I explore the opposite scenario, one where the errors are distributed

as $N(1.1, .25)$ to reflect overreported durations, which could be expected in the presence of backward telescoping or in reports of socially desirable events.

The effects of this four types of simulated ME are shown using scatter-plots in Figure 2. Notice that the top-right plot uses $Z$ instead of $X$.in the y-axis.

**Figure 2:** Scatterplots of the effect of the different types of measurement error considered




To assess the impact that these types of errors have on the regression coefficient of a linear, a logit, and a Poisson model, I compare the results from each of these models when $X^*$ is used (the naïve model) instead of $X$ (the true model). Specifically, I focus on the bias in the regression coefficients,

$$BIAS = \hat{\beta}_n - \hat{\beta}_t \tag{3.2}$$

where the subscript $n$ stands for the naïve model and $t$ for the true model. In addition, to compare the impact of ME across models and across regression coefficients using different scales, I calculate a relative measure of the bias as follows,

$$R.BIAS = \frac{|(\hat{\beta}_n - \hat{\beta}_t)100|}{|\hat{\beta}_t|} \tag{3.3}$$

Results for the different models studied and the impact generated by the different types of ME are presented in Table 1. In all of the scenarios studied the effect of ME was reflected in a downward bias for $\beta_1$ (the coefficient for the variable $X^*$), and in upward biases for $\beta_0$ and $\beta_2$ (the coefficients for the constant and $Z$). In addition to the observed differences in the direction of the biases across coefficients there are also strong differences in their intensity. The size of the bias for $\beta_2$ is about twice as large as the bias for $\beta_0$ and $\beta_1$, reaching levels as alarming as 94.8% for the logit model with heteroscedastic ME, although the average size of the bias across all the scenarios is 39.5%.

**Table 1:** Impact of Measurement Error in the Regression Estimates

| | | Linear | | | | Logit | | | | Poisson | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Coef | SE | Bias | R.Bias | Coef | SE | Bias | R.Bias | Coef | SE | Bias | R.Bias |
| True model | $\beta_0$ | -1.297 | .035 | | | -5.997 | .388 | | | -1.362 | .069 | | |
| | $\beta_1$ | .150 | .003 | | | .768 | .050 | | | .092 | .004 | | |
| | $\beta_2$ | .111 | .016 | | | .210 | .099 | | | .082 | .038 | | |
| Naïve: multi. | $\beta_0$ | -1.013 | .038 | .284 | 21.9% | -3.810 | .258 | 2.187 | 36.5% | -1.198 | .065 | .284 | 20.8% |
| | $\beta_1$ | .118 | .004 | -.032 | 21.2% | .494 | .034 | -.275 | 37.5% | .075 | .004 | -.032 | 34.6% |
| | $\beta_2$ | .156 | .019 | .045 | 40.9% | .275 | .084 | .065 | 30.9% | .157 | .037 | .045 | 55.1% |
| Naïve: hetero. | $\beta_0$ | -.998 | .039 | .372 | 28.7% | -3.649 | .248 | 2.372 | 39.5% | -1.178 | .065 | .372 | 27.4% |
| | $\beta_1$ | .116 | .004 | -.043 | 28.8% | .471 | .032 | -.299 | 38.9% | .074 | .004 | -.043 | 47.0% |
| | $\beta_2$ | .169 | .020 | .067 | 60.7% | .377 | .085 | .199 | 94.8% | .141 | .037 | .067 | 81.7% |
| Naïve: under. | $\beta_0$ | -.937 | .038 | .325 | 25.1% | -3.441 | .233 | 1.962 | 32.7% | -1.054 | .059 | .325 | 23.9% |
| | $\beta_1$ | .121 | .004 | -.024 | 15.9% | .490 | .033 | -.183 | 23.8% | .069 | .004 | -.024 | 26.1% |
| | $\beta_2$ | .166 | .020 | .052 | 46.9% | .321 | .084 | .124 | 58.9% | .149 | .037 | .052 | 63.2% |
| Naïve: over. | $\beta_0$ | -1.028 | .038 | .245 | 18.9% | -4.029 | .266 | 1.842 | 30.7% | -1.110 | .061 | .245 | 18.0% |
| | $\beta_1$ | .108 | .003 | -.039 | 26.1% | .470 | .031 | -.284 | 37.0% | .061 | .003 | -.039 | 42.7% |
| | $\beta_2$ | .150 | .019 | .053 | 48.0% | .284 | .087 | .152 | 72.4% | .134 | .037 | .053 | 64.6% |

While the different ME scenarios clearly show attenuated coefficients, none of the coefficients actually became statistically non-significant or changed their sign in comparison to the naïve models. This is partly due to the small effect that ME had on the standard errors, which were underestimated by a third of their size in the true logit model, and only slightly underestimated and overestimated when using a Poisson and a linear model, respectively.

These results are consistent with Biewen et al. (2008) who, in a simulated probit model with one predictor, find an upward bias in the constant and a downward bias in the slope

induced by classical multiplicative ME. These results obtained here serve to reinforce these findings. In the presence of a type of ME different than classical additive or for a model different than simple linear regression the direction of the bias is not always towards the null. The difficulty to anticipate the direction and size of these biases – even in scenarios with moderate prevalence of ME – makes the implementation of adjustment methods an indispensable part analysing survey data prone to these types of ME.

# 4 Standard SIMEX and Extensions to Account for Classical Multiplicative Measurement Error

The study of the adjustment of multiplicative errors dates back to the decade of the 80s. Fuller (1984) and Hwang (1986) developed a method-of-moments correction for multiplicative ME in the explanatory variables of a linear model. This method assumes that the value of ME variance is known – or that it can be estimated - and is limited to applications where the ME mechanism is affecting one of the explanatory variables only in the context of a linear model. Lyles and Kupper (1997) compared the effectiveness of this method with others such as regression calibration, and a quasi-likelihood approach, which could be applied to other non-linear outcome models.

These methods, as mentioned above, are however of limited use to applied researchers in that they either require additional data in the form of replicated measures or validation subsamples, or are complex to implement. Regression calibration requires additional data in the form of replicated measures or a validation subsample. Quasi-likelihood approaches only need an estimate of the variance of the ME, and much like those relying on Bayesian statistics can be applied when a full likelihood approach is not feasible due to computational intractability. However, their implementation is relatively complex, starting from the need to use specialised software (such as WinBUGS when considering Bayesian adjustments), which discourages many analysts from attempting the implementation of the necessary adjustment.

Due to only requiring an estimate of the variance of the ME, the simplicity of its application, and its generalizability to any other outcome model regardless of its complexity[6], SIMEX represents a very convenient alternative. SIMEX was first presented by Cook and Stefanski (1994) and refined in the following years by Stefanski and Cook (1995) and

[6] See for example He et al. (2007) who applied SIMEX to an Accelerated Failure Time models with one of the explanatory variable affected by classical ME, or Battauz et al. (2008) who adjusted for a similar type of ME problem but for an ordinal probit model as the outcome model.

Carroll, Küchenhoff, Lombard, and Stefanski (1996). *"The key idea underlying SIMEX is the fact that the effect of measurement error on an estimator can be determined experimentally via simulation"* (Carroll et al., 2006: 98).

In particular, SIMEX exploits the relationship between the size of the ME affecting a variable and the size of the bias in the regression estimates in the outcome model. Following Fuller (1987) we know that the unadjusted estimator of the slope, $\hat{\beta}_1$, does not converge asymptotically to the parameter $\beta_1$ but to:

$$\hat{\beta}_1^* = \beta_1\left(\sigma_X^2/(\sigma_X^2 + \sigma_V^2)\right) \tag{4.1}$$

where $\sigma_X^2$ and $\sigma_V^2$ represent the variance of the true explanatory variable and the error term. In other words, the estimator of the slope is biased downwards in absolute terms by a factor equal to the reliability ratio, $\rho_{X^*}$ (defined in equation 2.5), of the observed variable, $X^*$. In this situation, and if $\rho_{X^*}$ or $\sigma_V^2$ is known, it would not be practical nor efficient to use SIMEX, since the adjustment would simply be achieved by substituting the variance terms in equation 4.1. However, I will use this simple setting for illustrative purposes.

To facilitate the understanding of the method, the steps involved in its implementation are outlined below using a simple example of bias in the slope of a simple linear regression, where explanatory variable, $X^*$, is prone to classical additive ME (equations 2.1 and 2.2). The implementation of SIMEX is divided into six phases:

1) The first step involves simulating additional explanatory variables with increasing levels of ME. These new variables are generated in a way that emulates the classical ME model, but with successively larger values of $\sigma_V^2$ affecting $X$. Specifically, $K$ new explanatory variables $X_k^*(\lambda_k)$ are generated by the rule:

$$X_k^*(\lambda_k) = X^* + \sqrt{(\lambda_k)}V \tag{4.2}$$

with $k = 0,1,\ldots,K$, the simulated error normally distributed, $V \sim N(0, \sigma_V^2)$, and, $\lambda_0 < \lambda_1 < \ldots < \lambda_K$, a set of parameters used to amplify the ME variance (often these are (.5, 1, 1.5, 2)).

2) Once the different variables with added ME have been generated, the outcome model is re-estimated using this new data, and the values of the estimator of interest (i.e. $\beta_1$) for the different levels of ME $(\lambda_k)$ are saved. In particular, for the case of a simple linear model with

the explanatory variable affected by classical ME, and using the data-generating rule described in equation 4.2, the estimator of the slope will now converge to:

$$\hat{\beta}_{1k}^{*} = \beta_1 \sigma_X^2 / (\sigma_X^2 + (1 + \lambda_k)\sigma_V^2) \tag{4.3}$$

where the bias increases monotonically as $\lambda_k$ increases.

3) In order to reduce the Monte Carlo error associated with the simulation procedure steps 1 and 2 are repeated *B* times so a mean estimate of $\hat{\beta}_{1k}^{*}$ for $b = 1, \dots, B$ can be computed, where the rule of thumb[7] is to use $B = 100$ iterations.

4) At this stage the $\bar{\hat{\beta}}_{1k}^{*}$ and $\lambda_k$ values can be paired considering the former as a function of the latter, $G\left(\bar{\hat{\beta}}_{1k}^{*}, \lambda_k\right)$, known as the extrapolation function, which should be plotted in order to obtain a first insight of its shape.

5) The extrapolation function is estimated using a regression model, with data $\left(\bar{\hat{\beta}}_{1k}^{*}, \lambda_k\right)$. Carroll et al. (2006) recommend the use of one of three types of simple functional forms.

a) linear, $G\left(\bar{\hat{\beta}}_{1k}^{*}, \lambda_k\right) = \zeta_1 + \zeta_2 \lambda_k$

b) quadratic, $G\left(\bar{\hat{\beta}}_{1k}^{*}, \lambda_k\right) = \zeta_1 + \zeta_2 \lambda_k + \zeta_3 \lambda_k^2$

c) non-linear or ratio-linear, $G\left(\bar{\hat{\beta}}_{1k}^{*}, \lambda_k\right) = \zeta_1 + \zeta_2 / (\zeta_3 + \lambda_k)$

For the example presented here, and if the extrapolation function is well approximated by the chosen functional form, we would find the following function,

$$E\left(\bar{\hat{\beta}}_{1k}^{*} | \lambda_k\right) = G\left(\bar{\hat{\beta}}_{1k}^{*}, \lambda_k\right) = \beta_1 \sigma_X^2 / (\sigma_X^2 + (1 + \lambda_k)\sigma_V^2) \tag{4.4}$$

6) From here, the SIMEX estimate, $\hat{\beta}_{SIMEX}$, can be calculated by extrapolating $G\left(\bar{\hat{\beta}}_{1k}^{*}, \lambda_k\right)$ to $G\left(\bar{\hat{\beta}}_{1k}^{*}, \lambda_k = -1\right)$. Note that from equation 4.4 when $\lambda_k = -1$ the bias is cancelled out.

Figure 3 represents the SIMEX process graphically. The solid line denotes the part of the extrapolation function that can be approximately observed through the regression estimates resulting after the outcome model is specified using simulated predictors with increasing

[7] This is the number of iterations used by default in the SIMEX packages in STATA and R.

levels of ME, and the dashed line represents the extrapolation to the case of no ME, which gives the adjusted estimate.

**Figure 3:** Extrapolation function

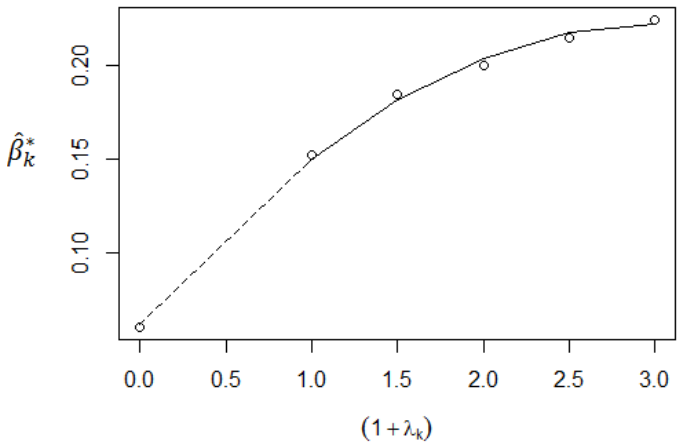
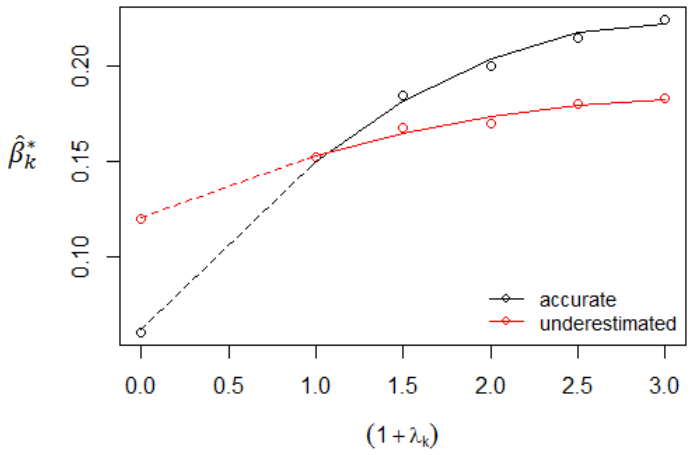



Figure 3 also shows some of the limitations of SIMEX. The entire extrapolation function cannot be observed, hence, it is hard to assess the quality of the adjustment. In addition, the extrapolation function needs to be approximated using a simple functional form. Therefore adjustments are only approximated, and their effectiveness depends on how well the extrapolation function is estimated, for which the choice of the right functional form is crucial. In the case depicted by Figure 3 it makes sense to think of the quadratic function as the better approximation, but it might not always be so clear.

Another cause of concern stems from the accuracy of the estimate of $\sigma_V^2$ that is used in the simulations. For example, considering the case depicted in Figure 3, if $\sigma_V^2$ is underestimated, the extrapolation function will have a flatter slope and the adjustment would only be partial. That is, for an underestimated $\sigma_V^2$, lower values of $\hat{\beta}_k^*$ would have been generated for $(1 + \lambda_k) = (1.5, 2, 2.5, 3)$, which would have made the estimated extrapolation function shallower, and produced a bigger - and still biased - adjusted estimate when extrapolating to $\lambda_k = -1$. Such suboptimal adjustment is illustrated in Figure 4 where I compare the extrapolation function shown in Figure 3 with a similar one that would be obtained if $\sigma_V^2$ had been underestimated.

**Figure 4:** Comparison of Extrapolation Functions




Interestingly the application of SIMEX to any other regression model affected by a problem of error-in-variables would follow the same logic than the example of a simple linear regression model presented here, regardless of the complexity of the outcome model under study. Even more interestingly – at least for this paper's topic of study – is the fact that SIMEX can be applied to ME problems different than the standard classical additive model so long as the ME-generating process can be simulated via Monte Carlo methods (Carroll, 2006). Two remarkable extensions that have proven to be robust in the literature are: 1) MC-SIMEX (Küchenhoff et al. 2006), the application of the SIMEX methodology to problems of misclassification of either the response or an explanatory variable in the outcome model; and 2) SIMEX for classical multiplicative ME (Carroll et al., 2006, and Biewen et al. 2008).

In order to accommodate the standard SIMEX method to account for the classical multiplicative ME setting Carroll et al. (2006) propose a change in the way the simulated variables are increasingly affected by ME, $X_k^*$, in step 1. In particular, equation 4.3 is substituted by

$$X^*(\lambda_k) = exp\{\log(X^*) + \sqrt{\lambda_k} log(V)\} \qquad (4.5)$$

to represent the multiplicative relationship between the observed durations and the simulated noise, while the rest of the six-step algorithm is implemented as before. However, the expression in equation 4.5 cannot be used to generate negative errors, which in a setting like the one assumed here, where errors are normally (instead of lognormally) distributed, could create complications. To avoid this problem I will use the following error-generating rule suggested by Biewen et al. (2008),

$$X^*(\lambda_k) = X^* \cdot V^{\lambda_k} \tag{4.6}$$

Lastly, to estimate the standard errors of $\hat{\beta}_{SIMEX}$ I use the bootstrapping pairs algorithm[8], where entire cases covering the response and explanatory variables are resampled with replacement, and for each new sample the SIMEX process is rerun. Bootstrap is only one of the different options available to estimate the variance of the SIMEX estimator. Carroll et al. (1996) suggest using a method based on the sandwich estimator and on the theory of M-estimators to obtain an asymptotic covariance estimator. Specifically, this method is based on the asymptotic equivalence of $\hat{\beta}(\lambda_k)$ and an M-estimator, producing a closed form equation from which the standard errors of $\hat{\beta}_{SIMEX}$ can be directly derived, which avoids the computationally intensive process of replicating the SIMEX procedure for a number of new samples. However, this method has been developed for the specific case of classical additive ME. For extensions of SIMEX to different types of ME processes non-parametric methods such as bootstrap or jackknife become a natural alternative.

# 5 Effectiveness of the Adjustments for Different Types of Errors and Estimates of their Variance

SIMEX requires an estimate of the variance of the ME, $\hat{\sigma}_V^2$. Ideally the actual parameter is known[9], although in the presence of replicated measures $\hat{\sigma}_V^2$ can also be estimated. However, access to such type of data tends to be the exception rather than the norm. Thus, here I study the effectiveness of SIMEX as a sensitivity tool. That is, I assume that the researcher suspects the presence of ME in one of the variables being used but can only provide an educated guess[10] of the distribution and prevalence of that ME.

For each of the different ME scenarios presented in Section 3, I explore the effectiveness of SIMEX in reducing the bias found in the naïve models when different values of $\sigma_V^2$ are tried. In particular I review the extent of the adjustments assuming that $E(V) = 0$, and $\sigma_V^2$ is equal to .25, .176, .336, and .412 which are equivalent to assuming reliability ratios of $X^*$

[8] See Keele (2008) for a description of the differences between bootstrapping algorithms.
[9] For example, Biewen et al. (2008) suggested using multiplicative ME as an strategy to anonymised data, while offering the value of $\sigma_V^2$ to data users so they can adjust for the implications of the artificially created ME in their analyses.
[10] Ideally based on validation studies looking at similar types of questions available in the literature.

equal to .816, .9, .7, and .6, respectively. I refer to each of these scenarios as assuming a "correct", "overestimated", "underestimated", and "highly underestimated" reliability ratio. In addition, to assess the effectiveness of the adjustment in the presence of forward or backward telescoping effects when the ME process is correctly estimated I take the "correct" adjustments for the systematic negative and positive ME scenarios to use the right ME process, that is $V \sim N(.9, .25)$ and $V \sim N(1.1, .25)$, respectively.

For consistency's sake I use the linear extrapolation function across all of the adjustments[11]. This extrapolation function was chosen instead of the more commonly used quadratic extrapolation to generate more conservative adjustments in scenarios using $\hat{\sigma}_V^2 > \sigma_V^2$, which in my analysis would be the cases of "underestimated" and "very underestimated" reliability ratios. For the estimation of the standard errors of $\hat{\beta}_{SIMEX}$ I run the bootstrapping pairs algorithm for 100 iterations (just like in Biewen et al., 2008)[12], and within each of these iterations I run the six steps of the SIMEX process another 100 times[13]. Lastly, to calculate the effectiveness of the adjustments I use measures of absolute and relative bias as in Section 3. The only differences are in terms of notation: I now substitute $\hat{\beta}_S$ by $\hat{\beta}_{SIMEX}$ in equation 3.2 and 3.3, and for the R.BIAS I also substitute the denominator by the BIAS in the survey. Results are shown in Table 2.

[11] The adequacy of the linear extrapolation can be assessed in Figure A2 (Appendix III), where I show the extrapolation functions for the adjustment of non-systematic ME using the correct $\sigma_V^2$.

[12] This is less than what would be recommendable to obtain precise estimates of the standard errors but it is sufficiently good considering that the SIMEX process is also computationally intensive and that a compromise needs to be reached.

[13] Figure A1 (Appendix III) includes scatterplots that reflect the simulation of increasing levels of the ME generation process (step 1 of the SIMEX algorithm) when the correct $\sigma_V^2$ is used.

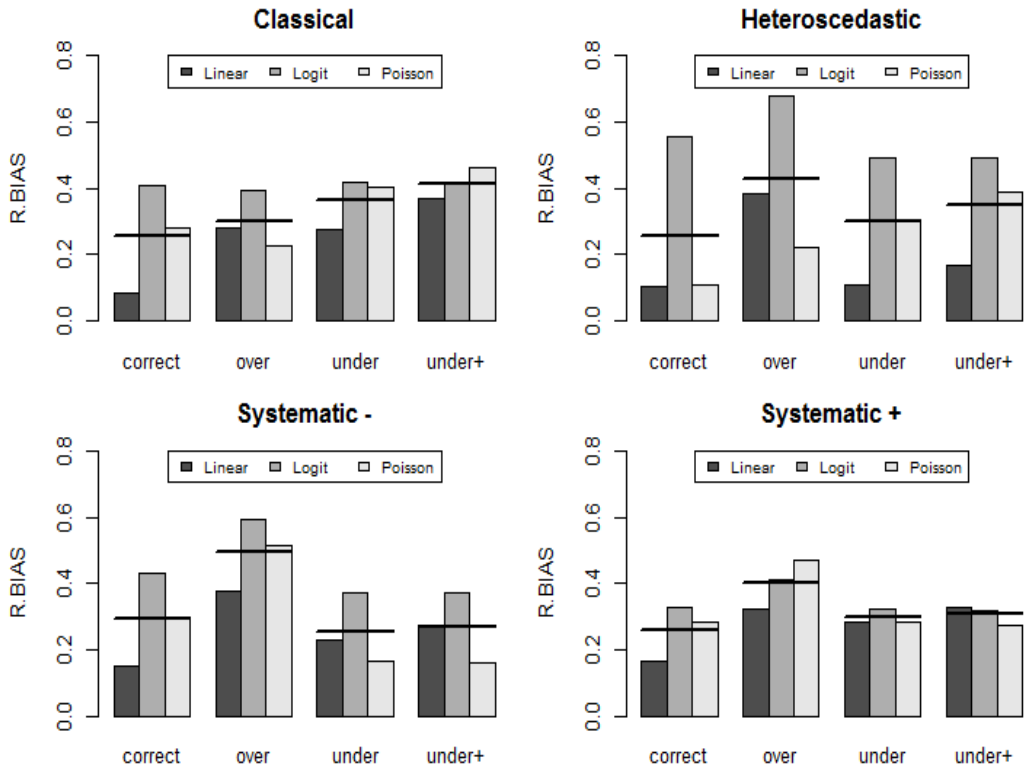
**Table 2:** Results of the Adjustments

| | | | Linear | | | | Logit | | | | Poisson | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | RR | Param. | Coef. | SE | Bias | R.Bias | Coef. | SE | Bias | R.Bias | Coef. | SE | Bias | R.Bias |
| Classical | Correct | $\beta_0$ | -1.283 | .017 | .014 | 4.8% | -5.030 | .100 | .967 | 44.2% | -1.400 | .026 | -.038 | 23.1% |
| | | $\beta_1$ | .152 | .002 | .002 | 7.5% | .662 | .014 | -.106 | 38.6% | .097 | .003 | .005 | 31.6% |
| | | $\beta_2$ | .105 | .006 | -.006 | 12.9% | .184 | .018 | -.026 | 39.8% | .104 | .007 | .022 | 29.9% |
| | Over-estimated | $\beta_0$ | -1.198 | .013 | .098 | 34.7% | -4.714 | .087 | 1.282 | 58.6% | -1.343 | .021 | .019 | 11.7% |
| | | $\beta_1$ | .141 | .002 | -.009 | 26.8% | .617 | .012 | -.151 | 54.9% | .090 | .002 | -.001 | 7.3% |
| | | $\beta_2$ | .121 | .004 | .010 | 23.2% | .207 | .016 | -.003 | 4.4% | .119 | .005 | .037 | 48.9% |
| | Under-estimated | $\beta_0$ | -1.339 | .018 | -.043 | 15.1% | -5.148 | .095 | .849 | 38.8% | -1.436 | .027 | -.074 | 44.9% |
| | | $\beta_1$ | .160 | .002 | .010 | 32.1% | .681 | .014 | -.087 | 31.6% | .101 | .003 | .010 | 58.1% |
| | | $\beta_2$ | .095 | .006 | -.016 | 35.5% | .174 | .020 | -.035 | 54.6% | .096 | .009 | .014 | 18.2% |
| | Very under-estimated | $\beta_0$ | -1.361 | .018 | -.064 | 22.7% | -5.134 | .076 | .863 | 39.4% | -1.451 | .030 | -.089 | 54.4% |
| | | $\beta_1$ | .163 | .002 | .014 | 42.7% | .681 | .011 | -.087 | 31.6% | .103 | .003 | .012 | 71.0% |
| | | $\beta_2$ | .090 | .007 | -.020 | 44.7% | .175 | .020 | -.034 | 53.1% | .092 | .009 | .010 | 13.2% |
| Heteroscedastic | Correct | $\beta_0$ | -1.265 | .019 | .032 | 10.6% | -4.790 | .111 | 1.207 | 51.4% | -1.378 | .034 | -.016 | 8.6% |
| | | $\beta_1$ | .149 | .002 | <.001 | 1.1% | .629 | .016 | -.140 | 46.9% | .096 | .003 | .004 | 22.8% |
| | | $\beta_2$ | .122 | .005 | .011 | 18.8% | .324 | .020 | .114 | 68.1% | .082 | .009 | <.001 | 0.8% |
| | Over-estimated | $\beta_0$ | -1.182 | .014 | .114 | 38.3% | -4.481 | .093 | 1.515 | 64.5% | -1.317 | .027 | .045 | 24.3% |
| | | $\beta_1$ | .139 | .002 | -.011 | 32.3% | .585 | .013 | -.183 | 61.6% | .089 | .003 | -.003 | 15.4% |
| | | $\beta_2$ | .137 | .004 | .026 | 44.4% | .340 | .014 | .130 | 77.4% | .098 | .006 | .016 | 26.3% |
| | Under-estimated | $\beta_0$ | -1.320 | .019 | -.023 | 7.8% | -4.925 | .090 | 1.071 | 45.6% | -1.414 | .034 | -.052 | 28.5% |
| | | $\beta_1$ | .157 | .002 | .007 | 21.2% | .650 | .013 | -.118 | 39.8% | .100 | .003 | .008 | 47.7% |
| | | $\beta_2$ | .112 | .007 | .002 | 3.0% | .315 | .020 | .105 | 62.8% | .073 | .010 | -.009 | 15.9% |
| | Very under-estimated | $\beta_0$ | -1.342 | .018 | -.045 | 15.1% | -4.923 | .083 | 1.073 | 45.7% | -1.427 | .035 | -.065 | 35.5% |
| | | $\beta_1$ | .160 | .002 | .011 | 31.1% | .651 | .011 | -.117 | 39.3% | .102 | .003 | .010 | 58.3% |
| | | $\beta_2$ | .108 | .008 | -.002 | 3.5% | .314 | .019 | .104 | 62.1% | .069 | .009 | -.013 | 22.6% |
| Systematic Negative | Correct | $\beta_0$ | -1.201 | .019 | .096 | 26.6% | -4.565 | .085 | 1.432 | 56.0% | -1.214 | .031 | .148 | 47.9% |
| | | $\beta_1$ | .153 | .003 | .003 | 11.4% | .649 | .013 | -.119 | 42.9% | .086 | .003 | -.006 | 25.2% |
| | | $\beta_2$ | .115 | .005 | .004 | 7.7% | .243 | .017 | .033 | 29.9% | .093 | .009 | .011 | 16.5% |
| | Over-estimated | $\beta_0$ | -1.106 | .013 | .191 | 53.0% | -4.207 | .072 | 1.790 | 70.1% | -1.157 | .020 | .205 | 66.6% |
| | | $\beta_1$ | .144 | .002 | -.005 | 18.6% | .606 | .011 | -.163 | 58.5% | .081 | .002 | -.010 | 44.5% |
| | | $\beta_2$ | .134 | .004 | .023 | 41.5% | .265 | .012 | .055 | 49.5% | .111 | .007 | .029 | 43.3% |
| | Under-estimated | $\beta_0$ | -1.236 | .022 | .060 | 16.7% | -4.640 | .094 | 1.357 | 53.1% | -1.235 | .031 | .127 | 41.3% |
| | | $\beta_1$ | .164 | .003 | .014 | 48.1% | .677 | .015 | -.091 | 32.9% | .092 | .003 | <.001 | 0.9% |
| | | $\beta_2$ | .109 | .006 | -.002 | 3.6% | .239 | .020 | .029 | 26.1% | .087 | .010 | .005 | 7.2% |
| | Very under-estimated | $\beta_0$ | -1.257 | .018 | .040 | 11.1% | -4.644 | .091 | 1.352 | 52.9% | -1.247 | .031 | .115 | 37.3% |
| | | $\beta_1$ | .167 | .003 | .017 | 60.5% | .680 | .015 | -.088 | 31.8% | .094 | .004 | .002 | 9.6% |
| | | $\beta_2$ | .105 | .007 | -.006 | 10.7% | .239 | .017 | .029 | 26.1% | .083 | .010 | .001 | 1.0% |
| Systematic Positive | Correct | $\beta_0$ | -1.278 | .019 | .018 | 6.9% | -5.252 | .097 | .744 | 37.8% | -1.263 | .033 | .099 | 39.2% |
| | | $\beta_1$ | .141 | .002 | -.009 | 21.9% | .632 | .012 | -.136 | 45.7% | .079 | .003 | -.013 | 41.4% |
| | | $\beta_2$ | .102 | .005 | -.008 | 20.3% | .199 | .018 | -.011 | 14.4% | .080 | .009 | -.002 | 4.0% |
| | Over-estimated | $\beta_0$ | -1.209 | .014 | .087 | 32.4% | -4.959 | .087 | 1.038 | 52.7% | -1.221 | .017 | .141 | 56.1% |
| | | $\beta_1$ | .129 | .002 | -.021 | 50.4% | .584 | .011 | -.185 | 61.9% | .072 | .002 | -.019 | 63.4% |
| | | $\beta_2$ | .116 | .004 | .005 | 13.7% | .217 | .014 | .007 | 8.9% | .093 | .007 | .011 | 21.8% |
| | Under-estimated | $\beta_0$ | -1.355 | .023 | -.058 | 21.5% | -5.450 | .110 | .547 | 27.8% | -1.307 | .038 | .055 | 21.7% |
| | | $\beta_1$ | .146 | .003 | -.004 | 8.7% | .649 | .014 | -.119 | 40.0% | .082 | .003 | -.010 | 32.7% |
| | | $\beta_2$ | .089 | .006 | -.022 | 55.2% | .188 | .022 | -.022 | 29.2% | .066 | .010 | -.016 | 30.3% |
| | Very under-estimated | $\beta_0$ | -1.378 | .020 | -.082 | 30.4% | -5.444 | .102 | .553 | 28.1% | -1.322 | .034 | .040 | 16.1% |
| | | $\beta_1$ | .149 | .002 | .000 | 0.7% | .650 | .013 | -.118 | 39.6% | .084 | .003 | -.008 | 26.6% |
| | | $\beta_2$ | .084 | .008 | -.027 | 66.5% | .189 | .019 | -.021 | 27.6% | .062 | .011 | -.021 | 39.7% |

A first point to notice is that compared to results from the true models presented in Table 1, the standard errors are underestimated by a half. This might be due to the small size of the true standard errors (expressed in the second or third decimal point), but it also illustrates that the variance of $\hat{\beta}_{SIMEX}$ using bootstrap can only be approximated.

Regarding the adjustment in terms of the reduction of the biases found in the naïve analyses we can observe varying levels of success. The effectiveness of the adjustment ranged from being able to reduce it to .8% of its size (for $\beta_2$ in the Poisson model affected by heteroscedastic ME and using the correct estimate of $\sigma_V^2$) to a less impressive figure of 77.4% (for $\beta_2$ in the logit model affected by heteroscedastic ME and using an over-estimated reliability ratio). In spite of this variability, the adjustments explored could be considered quite successful since on average they managed to reduce the biases found in the naïve models to 32.8% of their original size.

**Figure 5:** Adjustments in Terms of R.BIAS[*,**]




**[*]The category of barplots "under +" represents a very underestimated reliability ratio of .6.**
**[**]The flat lines indicate the average R.BIAS for the three types of models.**

To elucidate some trends about the relative effectiveness of SIMEX for the different scenarios explored I have grouped some of the results in Figure 5. Each bar represents the average R.BIAS for the three regression coefficients comprised in each of the models and for

each of the scenarios studied, while the black horizontal lines represent the bars averages over the three outcome models studied.

On average the adjustments are most effective on the linear model, reducing the R.BIAS to 24.3% of its original size, whereas for the logit models the average adjustment is 43.7%, and for the Poisson is 30.3%. This better performance for the linear outcome model might be related to the linear extrapolation function used across all of the adjustments, which would be the most appropriate function when the changes in bias are proportional to the increased levels of ME. Directly from Figure 5 we can also see that considering the three coefficients of each model, the most successful adjustment was for the linear model when the correct ME process is used, which reduced the bias to just 8.4% of its original size. But even in the worst scenario where heteroscedastic ME is simulated and the reliability ratio of the measure is overestimated the bias can still be reduced to 38.3% of its size. On the other hand, the least promising results were obtained for the adjustments of the heteroscedastic ME when a logit outcome model is used. Here, regardless of the assumed reliability ratio no adjustment could reduce the bias by more than 50.6%.

# 6 Conclusion

The presence of ME in retrospective questions is widely acknowledged, however, little is done to tackle this problem. A majority of studies using this type of data deal with the ME problem by adding caveats to their findings, whereas those attempting to implement the necessary adjustments are very uncommon. The implications of this problem are truly daunting. Here I have simulated moderate levels of different types of multiplicative errors, which could be expected to arise as a result of memory failures in the report of dates of onset or end of spells, to explore the impact that such ME could have on the regression coefficients of different models. Across the scenarios studied I found an average bias of about 40% the size of the true coefficients, reaching up to 95% in certain cases.

I pointed at two fundamental barriers limiting the implementation of adjustment methods. First, most of these methods require access to additional sources of data in the form of replicated measures, validation subsamples, or instrumental variables, which are rarely available to a majority of researchers. Second, most methods are relatively complex to implement, discouraging researchers from using them. A very illustrative case is that of

adjustments relying on Bayesian statistics; their reliance on MCMC and prior probabilities offers remarkable flexibility to deal with complex ME processes, even in the absence of replicated or validation data. However, it is this complexity together with other practical matters such as the need to use specialised statistical packages that tend to dissuade researchers from using them.

To deal with recall errors in the report of dates of onset when no data to inform about the ME process is available I have suggested the implementation of the more practical SIMEX method. SIMEX is relatively simple to implement, it can be easily replicated to different outcome models regardless of their complexity, and it only requires - in its standard form - knowledge about the variance of the error term. Furthermore, although SIMEX was initially created to adjust for classical additive ME, it can also be extended to account for different types of ME processes so long as these can be simulated using Monte Carlo methods. I have used this feature to assess the effectiveness of SIMEX in the presence of multiplicative errors. In particular, I have explored the application of SIMEX to classical, heteroscedastic, systematic positive and systematic negative multiplicative errors. The types of errors that could be expected from general memory failures, but also those seen when different survey modes are used, or in the presence of backward and forward telescoping effects, respectively.

In the presence of these types of errors, SIMEX adjustments where the distribution of the ME is known have shown satisfactory results, managing to reduce the size of the biases found in the estimates of the naïve models to less than one third of their size on average. But perhaps more interesting is the fact that SIMEX also achieved reasonably good results even when the ME process is assumed to be non-systematic and its variance is only approximated. The quality of the adjustments varied substantially – as it could not be any different given the several scenarios explored – but in each of the 144 estimates studied SIMEX managed to produce positive adjustments, with the worst of them all achieving a reduction of the bias found in the naïve model of 22.6%.

This capacity to obtain partial adjustments even when the type of multiplicative error is only approximated, together with the relative simplicity with which it is implemented, makes SIMEX an ideal method to be used as a sensitivity tool. Researchers concerned of using duration data affected by recall errors could obtain an estimate of the magnitude of the impact, which would allow them to provide more informative caveats regarding the degree to which the validity of their findings is affected. To do that they can use the multiplicative SIMEX process presented in Appendix II. The only two alterations needed would be the re-

specification of the outcome model and the choice of the size of the variance of the error term.

When the latter is not known the method could be run using an educated guess for the reliability ratio of the variable prone to ME. Alternatively, for those questions where previous studies of the validity and reliability of responses are available, the researcher could use an average of the estimates obtained in the literature. The opportunity to use such sensitivity analyses even when no replicated measures or a validation subsample is available also illustrates the importance of studies aiming to assess the prevalence of ME in different types of survey questions. The more we know about the ME processes affecting survey responses the better adjustments could be achieved and the higher the validity of studies using survey data will be.

## Acknowledgements

I thank my colleague and friend Albert Varela for his useful comments, which have substantially improved the quality of this manuscript.

## References

[1] Augustin, T. (1999): Correcting for Measurement Error in Parametric Duration Models by Quasi-likelihood. *Munchen Institut fur Statistik*, from: http://epub.ub.uni-muenchen.de/1546/1/paper_157.pdf.

[2] Battauz, M., Bellio, R. & Gori, E. (2008): Reducing Measurement Error in Student Achievement Estimation. *Psychometrika*, **73**, 289-302.

[3] Biewen, E., Nolte, S. & Rosemann, M. (2008): Perturbation by Multiplicative Noise and The Simulation Extrapolation Method. *Advances in Statistical Analysis*, **92**, 375-389.

[4] Bradburn, N. M., Huttenlocher, J. & Hedges, L. (1994): Telescoping and Temporal Memory. In N. Schwarz et al. (Ed): *Autobiographical Memory and The Validity of Retrospective Reports*, 203-215. New York: Springer.

[5] Carroll, R. J. & Stefanski, L. A. (1990): Approximate Quasilikelihood Estimation in Models with Surrogate Predictors. *Journal of the American Statistical Association*, **85**, 652-663.

[6] Carroll, R., Küchenhoff, H., Lombard, F. & Stefanski, L. (1996): Asymptotics for the SIMEX Estimator in Nonlinear Measurement Error Models. *Journal of the American Statistical Association*, **91**, 242-250.

[7] Carroll, R., Ruppert, D., Stefanski, L. & Crainiceanu, C. (2006): *Measurement Error in Nonlinear Models; a Modern Perspective*, Boca Raton: Chapman and Hall.

[8] Cook, J. & Stefanski, L. (1994): A Simulation Extrapolation Method for Parametric Measurement Error Models. *Journal of the American Statistical Association*, **89**, 1314–1328.

[9] Cole, S., Chu, H. & Greenland, S. (2006): Multiple-Imputation for Measurement-Error Correction. *International Journal of Epidemiology*, **35**, 1074-1081.

[10] Da Silva, D. & Skinner, Ch. (2014): The Use of Accuracy Indicators to Correct for Survey Measurement Error. *Journal of the Royal Statistical Society: Series C*, **62**, 303-319.

[11] Dumangane, M. (2007): Measurement Error Bias Reduction in Unemployment Durations. *Centre for Microdata Methods and Practice*, 3, from: http://www.cemmap.ac.uk/wps/cwp0603.pdf.

[12] Efron, B. & Tibshirani, R. J. (1993): *An Introduction to the Bootstrap*. Boca Raton: CRC press.

[13] Fuller, W. (1987): *Measurement Error Models*. New York: John Wiley and Sons.

[14] Glesjer, L. (1990): Improvements of the Naive Approach to Estimation in Nonlinear Errors-in-Variables Regression Models. In P. Brown & W. Fuller (Ed): *Statistical Analysis of Error Measurement Models and Application*, 99-114. Providence: American Mathematics Society.

[15] Glewwe, P. (2007): Measurement Error Bias in Estimates of Income and Income Growth among the Poor: Analytical Results and a Correction Formula. *Economic Development and Cultural Change*, **56**, 163-189.

[16] Golub, A., Johnson, B. D. & Labouvie, E. (2000): On Correcting Biases in Self-Reports of Age at First Substance Use with Repeated Cross-Section Analysis. *Journal of Quantitative Criminology*, **16**, 45-68.

[17] Gustafson, P. (2003): *Measurement Error and Misclassification in Statistics and Epidemiology*. Boca Raton: Chapman and Hall.

[18] He, W., Yi, G. & Xiong, J. (2007): Accelerated Failure Time Models with Covariates Subject to Measurement Error. *Statisitcs in Medicine*, **26**, 4817-4832.

[19] Holt, D., McDonald, J.W. & Skinner, C.J. (1991): The Effect of Measurement Error on Event History Analysis. In P. Biemer (Ed): *Measurement Error in Surveys*, 665-685. New York: John Wiley.

[20] Huber, P. J. (1964): Robust Estimation of a Location Parameter. *Annals of Mathematical Statistics*, **35**, 73-101.

[21] Huttenlocher, J., Hedges, L. & Prohaska, V. (1988): Hierarchical Organization in Ordered Domains: Estimating the Dates of Events. *Psychological Review*, **95**, 471-484.

[22] Hwang, J. T. (1986): Multiplicative Errors-in-Variables Models with Applications to Recent Data Released by the US Department of Energy. *Journal of the American Statistical Association*, **81**, 680-688.

[23] Janssen, S. M., Chessa, A. G. & Murre, J. M. (2006): Memory for Time: How People Date Events. *Memory & Cognition*, **34**, 138-147.

[24] Johnson, E. O. & Schultz, L. (2005): Forward Telescoping Bias in Reported Age of Onset: An Example from Cigarette Smoking. *International Journal of Methods in Psychiatric Research*, **14**, 119-129.

[25] Küchenhoff, H., Mwalili, S.M. & Lesaffre, E. (2006): A General Model for Dealing with Misclassification in Regression: The Misclassification SIMEX. *Biometrics*, **62**, 85-96.

[26] Lyles, R. H. & Kupper, L. L. (1997): A Detailed Evaluation of Adjustment Methods for Multiplicative Measurement Error in Linear Regression with Applications in Occupational Epidemiology. *Biometrics*, 1008-1025.

[27] Neter, J. & Waksberg, J. (1964): A Study of Response Errors in Expenditures Data from Household Interviews. *Journal of the American Statistical Association*, **59**, 18-55.

[28] Novick, M.R. (1966): The Axioms and Principal Results of Classical Test Theory. *Journal of Mathematical Psychology*, **3**, 1-18.

[29] Nugent, W., Graycheck, L. & Basham, R. (2000): A Devil Hidden in the Details: The Effects of Measurement Error in Regression Analysis. *Journal of Social Service Research*, **27**, 53-75.

[30] Pickles, A., Pickering, K. & Taylor, C. (1996): Reconciling Recalled Dates of Developmental Milestones, Events and Transitions: A Mixed Generalized Linear Model with Random Mean and Variance Functions. *Journal of the Royal Statistical Society. Series A*, 225-234.

[31] Pickles, A., Pickering, K., Simonoff, E., Silberg, J., Meyer, J. & Maes, H. (1998): Genetic "Clocks" and "Soft" Events: A Twin Model for Pubertal Development and Other Recalled Sequences of Developmental Milestones, Transitions, or Ages at Onset. *Behavior Genetics*, **28**, 243-253.

[32] Pina-Sánchez, J., Koskinen, J. & Plewis, I. (2013): Implications of Retrospective Measurement Error in Event History Analysis. *Metodología de Encuestas*, **15**, 5-25.

[33] Pina-Sánchez, J., Koskinen, J. & Plewis, I. (2014): Measurement Error in Retrospective Work Histories. *Survey Research Methods*, **8**, 43-55.

[34] Poterba, J. M. & Summers, L. H. (1984): Response variation in the CPS: Caveats for the unemployment analyst. *Monthly Labor Review*, **107**, 37-43.

[35] Prentice, R. (1982): Covariate Measurement Errors and Parameter Estimation in a Failure Time Regression Model. *Biometrika*, **69**, 331-342.

[36] Rappaport, S. M. (1991): Assessment of Long-Term Exposures to Toxic Substances in Air. *Annals of Occupational Hygiene*, **35**, 61-122.

[37] Rappaport, S. M., Kromhouta, H. & Symanski, E. (1993): Variation of Exposure Between Workers in Homogeneous Exposure Groups. *The American Industrial Hygiene Association Journal*, **54**, 654-662.

[38] Roberts, C. (2007): Mixing modes of data collection in surveys: A methodological review. *Economic and Social Research Council - National Centre for Research Methods*, from: http://eprints.ncrm.ac.uk/418/.

[39] Rubin, D. C. (1987): *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley and Sons.

[40] Rubin, D. C. & Baddeley, A. D. (1989): Telescoping Is Not Time Compression: A Model. *Memory & Cognition*, **17**, 653-661.

[41] Skinner, C. & Humphreys, K. (1999): Weibull Regression for Lifetimes Measured with Error. *Lifetime Data Analysis*, **5**, 23-37.

[42] Skinner, C. (2000): Dealing with Measurement Error in Panel Analysis. In D. Rose (Ed): *Researching Social and Economic Change*, 113-125. New York: Routledge.

[43] Solga, H. (2001): Longitudinal Survey and the Study of Occupational Mobility: Panel and Retrospective Design in Comparison. *Quality and Quantity*, **35**, 291-309.

[44] Stefanski, L. & Cook, J. (1995): Simulation-Extrapolation: The Measurement Error Jackknife. *Journal of the American Statistical Association*, **90**, 1247-1256.

[45] Theil, H. (1953): *Repeated Least Squares Applied to Complete Equation Systems*. The Hague: Central Planning Bureau.

[46] Valaste, M., Lehtonen, R. & Vehkalahti, K. (2010): Multiple Imputation for Measurement Error Correction in Survey Data. *Q2010 European Conference on Quality in Official Statistics*, 89.

[47] Vardeman, S. B., Wendelberger, J. R., Burr, T., Hamada, M. S., Moore, L. M., Jobe, J. M., Morris, M. D. & Wu, H. (2010): Elementary statistical methods and measurement error. *The American Statistician*, **64**, 46-51.

# Appendix I. R Script

## Data Simulations

```
set.seed(10)

#The simulated variables

Y = rnorm(1000,0,1)
Yca = ifelse(Y>=0,1,0)
Yco = ifelse(Y<0,0,ifelse(Y>=0&Y<1,1,ifelse(Y>=1&Y<2,2,ifelse(Y>=2&Y<3,3,4))))
X1 = exp((Y+4)*.5 + rnorm(1000,0,.25))
X2 = Y*.4 + rnorm(1000,0,1)

#The classical multiplicative measurement error

X1star = X1*rnorm(1000,1,.25)

#The heteroscedastic measurement error

U = seq(1:1000)
for(i in 1:1000){
        U[i] = ifelse(X2[i]*rnorm(1,1,.5)<0, rnorm(1,1,.15), rnorm(1,1,.30))
        }
U = ifelse(U<=0, .001, U)
plot(X2,U)
X1star = X1*U

#The systematic negative measurement error

X1star = X1*rnorm(1000,.9,.25)

#The systematic positive measurement error

X1star = X1*rnorm(1000,.1.1,.25)
```

## The SIMEX Process

```
##Example assuming classical multiplicative measurement error, a linear outcome model, and known variance of the error term##############################################################################

#The outcome models.

lm.true = lm(Y ~ X1 + X2)
summary(lm.true)
lm.naive = lm(Y ~ X1star + X2)
summary(lm.naive)

#Estimates of the impact of measurement error

sum.n = summary(lm.naive)
sum.n = summary(lm.naive)
biases = coef(lm.naive) - coef(lm.true)

#Matrices to save results from the SIMEX process.

results = matrix(c(0),nrow=12,ncol=12,byrow=TRUE)
colnames(results) = c("lin.coef", "lin.bias", "lin.rbias")
rownames(results) = c("right.cons", "right.X1", "right.X2", "over.cons", "over.X1", "over.X2", "under.cons", "under.X1", "under.X2", "very.cons", "very.X1", "very.X2")
SE = matrix(c(0),nrow=100,ncol=3,byrow=TRUE)
```

```
##The SIMEX process#############################################################################

#noise.5

noise.5 = matrix(c(0),nrow=1000,ncol=3,byrow=TRUE)
for(i in 1:1000){
        ME.5 = rnorm(1000,1,.25)^.5
        X1star.5 = X1star * ME.5
        lm.noise.5 = lm(Y ~ X1star.5 + X2)
        noise.5[i,1] = coef(lm.noise.5)[1]
        noise.5[i,2] = coef(lm.noise.5)[2]
        noise.5[i,3] = coef(lm.noise.5)[3]
        }
avg.5_cons = mean(noise.5[,1])
avg.5_X1 = mean(noise.5[,2])
avg.5_X2 = mean(noise.5[,3])

#noise1

noise1 = matrix(c(0),nrow=1000,ncol=3,byrow=TRUE)
for(i in 1:1000){
        ME1 = rnorm(1000,1,.25)^1
        X1star1 = X1star * ME1
        lm.noise1 = lm(Y ~ X1star1 + X2)
        noise1[i,1] = coef(lm.noise1)[1]
        noise1[i,2] = coef(lm.noise1)[2]
        noise1[i,3] = coef(lm.noise1)[3]
        }
avg1_cons = mean(noise1[,1])
avg1_X1 = mean(noise1[,2])
avg1_X2 = mean(noise1[,3])

#noise1.5

noise1.5 = matrix(c(0),nrow=1000,ncol=3,byrow=TRUE)
for(i in 1:1000){
        ME1.5 = rnorm(1000,1,.25)^1.5
        X1star1.5 = X1star * ME1.5
        lm.noise1.5 = lm(Y ~ X1star1.5 + X2)
        noise1.5[i,1] = coef(lm.noise1.5)[1]
        noise1.5[i,2] = coef(lm.noise1.5)[2]
        noise1.5[i,3] = coef(lm.noise1.5)[3]
        }
avg1.5_cons = mean(noise1.5[,1])
avg1.5_X1 = mean(noise1.5[,2])
avg1.5_X2 = mean(noise1.5[,3])

#noise2

noise2 = matrix(c(0),nrow=1000,ncol=3,byrow=TRUE)
for(i in 1:1000){
        ME2 = rnorm(1000,1,.25)^2
        X1star2 = X1star * ME2
        lm.noise2 = lm(Y ~ X1star2 + X2)
        noise2[i,1] = coef(lm.noise2)[1]
        noise2[i,2] = coef(lm.noise2)[2]
        noise2[i,3] = coef(lm.noise2)[3]
        }
avg2_cons = mean(noise2[,1])
avg2_X1 = mean(noise2[,2])
avg2_X2 = mean(noise2[,3])

#I put the mean regression estimates from each level of simulated measurement error in a dataset
```

```
avg_noiseADJ_cons = NA
avg_noiseADJ_X1 = NA
avg_noiseADJ_X2 = NA
lambda = c(-1, 0, .5, 1, 1.5, 2)
addi1 = c(avg_noiseADJ_cons, coef(lm.naive)[1], avg.5_cons, avg1_cons, avg1.5_cons, avg2_cons)
addi2 = c(avg_noiseADJ_X1, coef(lm.naive)[2], avg.5_X1, avg1_X1, avg1.5_X1, avg2_X1)
addi3 = c(avg_noiseADJ_X2, coef(lm.naive)[3], avg.5_X2, avg1_X2, avg1.5_X2, avg2_X2)
SIMEX = data.frame(lambda, addi1, addi2, addi3)
names(SIMEX) = c("lambda","cons","X1","X2")

#I obtain the adjusted SIMEX estimates using a linear extrapolation function

SIMEXna = SIMEX[-1,]
SIMEX_cons = lm(SIMEXna$cons ~ SIMEXna$lambda)
SIMEX[1,2] = coef(SIMEX_cons)[1] + (-1)*coef(SIMEX_cons)[2]
SIMEX_X1 = lm(SIMEXna$X1 ~ SIMEXna$lambda)
SIMEX[1,3] = coef(SIMEX_X1)[1] + (-1)*coef(SIMEX_X1)[2]
SIMEX_X2 = lm(SIMEXna$X2 ~ SIMEXna$lambda)
SIMEX[1,4] = coef(SIMEX_X2)[1] + (-1)*coef(SIMEX_X2)[2]

#I save the adjusted estimates and the remaining bias

results[1,1] = SIMEX[1,2]
results[2,1] = SIMEX[1,3]
results[3,1] = SIMEX[1,4]
bias1 = SIMEX[1,2]-coef(lm.true)[1]
bias2 = SIMEX[1,3]-coef(lm.true)[2]
bias3 = SIMEX[1,4]-coef(lm.true)[3]
results[1,2] = bias1
results[2,2] = bias2
results[3,2] = bias3

#I calculate the R.BIAS

R.BIAS.1 = (abs(coef(lm.naive)[1]-coef(lm.true)[1])*100)/abs(coef(lm.true)[1])
R.BIAS.adj.1 = (abs(SIMEX[1,2]-coef(lm.true)[1])*100)/abs(coef(lm.true)[1])
results[1,3] = R.BIAS.adj.1 / R.BIAS.1
R.BIAS.2 = (abs(coef(lm.naive)[2]-coef(lm.true)[2])*100)/abs(coef(lm.true)[2])
R.BIAS.adj.2 = (abs(SIMEX[1,3]-coef(lm.true)[2])*100)/abs(coef(lm.true)[2])
results[2,3] = R.BIAS.adj.2 / R.BIAS.2
R.BIAS.3 = (abs(coef(lm.naive)[3]-coef(lm.true)[3])*100)/abs(coef(lm.true)[3])
R.BIAS.adj.3 = (abs(SIMEX[1,4]-coef(lm.true)[3])*100)/abs(coef(lm.true)[3])
results[3,3] = R.BIAS.adj.3 / R.BIAS.3


##The bootstrap process to calculate the standard errors obtained from the SIMEX adjustment##############

#The double-loop

for(l in 1:100){
        boot = data[sample(1:nrow(data), 1000, replace=TRUE),]

        #noise.5

        noise.5 = matrix(c(0),nrow=100,ncol=3,byrow=TRUE)
        for(i in 1:100){
                boot$ME.5 = rnorm(1000,1,.25)^.5
                boot$X1star.5 = boot$X1star * boot$ME.5
                lm.noise.5 = lm(Y ~ X1star.5 + X2, data=boot)
                noise.5[i,1] = coef(lm.noise.5)[1]
                noise.5[i,2] = coef(lm.noise.5)[2]
                noise.5[i,3] = coef(lm.noise.5)[3]
                }
        avg.5_cons = mean(noise.5[,1])
        avg.5_X1 = mean(noise.5[,2])
        avg.5_X2 = mean(noise.5[,3])
```

```
#noise1

noise1 = matrix(c(0),nrow=100,ncol=3,byrow=TRUE)
        for(i in 1:100){
        boot$ME1 = rnorm(1000,1,.25)^1
        boot$X1star1 = boot$X1star * boot$ME1
        lm.noise1 = lm(Y ~ X1star1 + X2, data=boot)
        noise1[i,1] = coef(lm.noise1)[1]
        noise1[i,2] = coef(lm.noise1)[2]
        noise1[i,3] = coef(lm.noise1)[3]
        }
avg1_cons = mean(noise1[,1])
avg1_X1 = mean(noise1[,2])
avg1_X2 = mean(noise1[,3])

#noise1.5

noise1.5 = matrix(c(0),nrow=100,ncol=3,byrow=TRUE)
for(i in 1:100){
        boot$ME1.5 = rnorm(1000,1,.25)^1.5
        boot$X1star1.5 = boot$X1star * boot$ME1.5
        lm.noise1.5 = lm(Y ~ X1star1.5 + X2, data=boot)
        noise1.5[i,1] = coef(lm.noise1.5)[1]
        noise1.5[i,2] = coef(lm.noise1.5)[2]
        noise1.5[i,3] = coef(lm.noise1.5)[3]
        }
avg1.5_cons = mean(noise1.5[,1])
avg1.5_X1 = mean(noise1.5[,2])
avg1.5_X2 = mean(noise1.5[,3])

#noise2

noise2 = matrix(c(0),nrow=100,ncol=3,byrow=TRUE)
        for(i in 1:100){
        boot$ME2 = rnorm(1000,1,.25)^2
        boot$X1star2 = boot$X1star * boot$ME2
        lm.noise2 = lm(Y ~ X1star2 + X2, data=boot)
        noise2[i,1] = coef(lm.noise2)[1]
        noise2[i,2] = coef(lm.noise2)[2]
        noise2[i,3] = coef(lm.noise2)[3]
        }
avg2_cons = mean(noise2[,1])
avg2_X1 = mean(noise2[,2])
avg2_X2 = mean(noise2[,3])

#I save the adjusted estimates and the remaining bias

avg_noiseADJ_cons = NA
avg_noiseADJ_X1 = NA
avg_noiseADJ_X2 = NA
lambda = c(-1, 0, .5, 1, 1.5, 2)
addi1 = c(avg_noiseADJ_cons, coef(lm.naive)[1], avg.5_cons, avg1_cons, avg1.5_cons, avg2_cons)
addi2 = c(avg_noiseADJ_X1, coef(lm.naive)[2], avg.5_X1, avg1_X1, avg1.5_X1, avg2_X1)
addi3 = c(avg_noiseADJ_X2, coef(lm.naive)[3], avg.5_X2, avg1_X2, avg1.5_X2, avg2_X2)
SIMEX = data.frame(lambda, addi1, addi2, addi3)
names(SIMEX) = c("lambda","cons","X1","X2")

#I obtain the adjusted estimate using a linear extrapolation function

SIMEXna = SIMEX[-1,]
SIMEX_cons = lm(SIMEXna$cons ~ SIMEXna$lambda)
SIMEX[1,2] = coef(SIMEX_cons)[1] + (-1)*coef(SIMEX_cons)[2]
SIMEX_X1 = lm(SIMEXna$X1 ~ SIMEXna$lambda)
SIMEX[1,3] = coef(SIMEX_X1)[1] + (-1)*coef(SIMEX_X1)[2]
```

```
        SIMEX_X2 = lm(SIMEXna$X2 ~ SIMEXna$lambda)
        SIMEX[1,4] = coef(SIMEX_X2)[1] + (-1)*coef(SIMEX_X2)[2]
        #I save the SIMEX adjustment for each of the 100 bootstrap iterations
        SE[l,1] = SIMEX[1,2]
        SE[l,2] = SIMEX[1,3]
        SE[l,3] = SIMEX[1,4]
        }
#I obtain the standard errors
SE1 = sd(SE[,1])
SE2 = sd(SE[,2])
SE3 = sd(SE[,3])
```

# Appendix II. Illustrations of the SIMEX Process

Figure A1 shows the effect of the increased levels of simulated measurement error on $X$ using $\sigma_V^2 \sim (0, .25)$ and $\lambda_k = (0.5, 1, 1.5, 2)$.

**Figure A1:** Scatterplots of $\boldsymbol{X_1}$ and increasing levels of measurement error

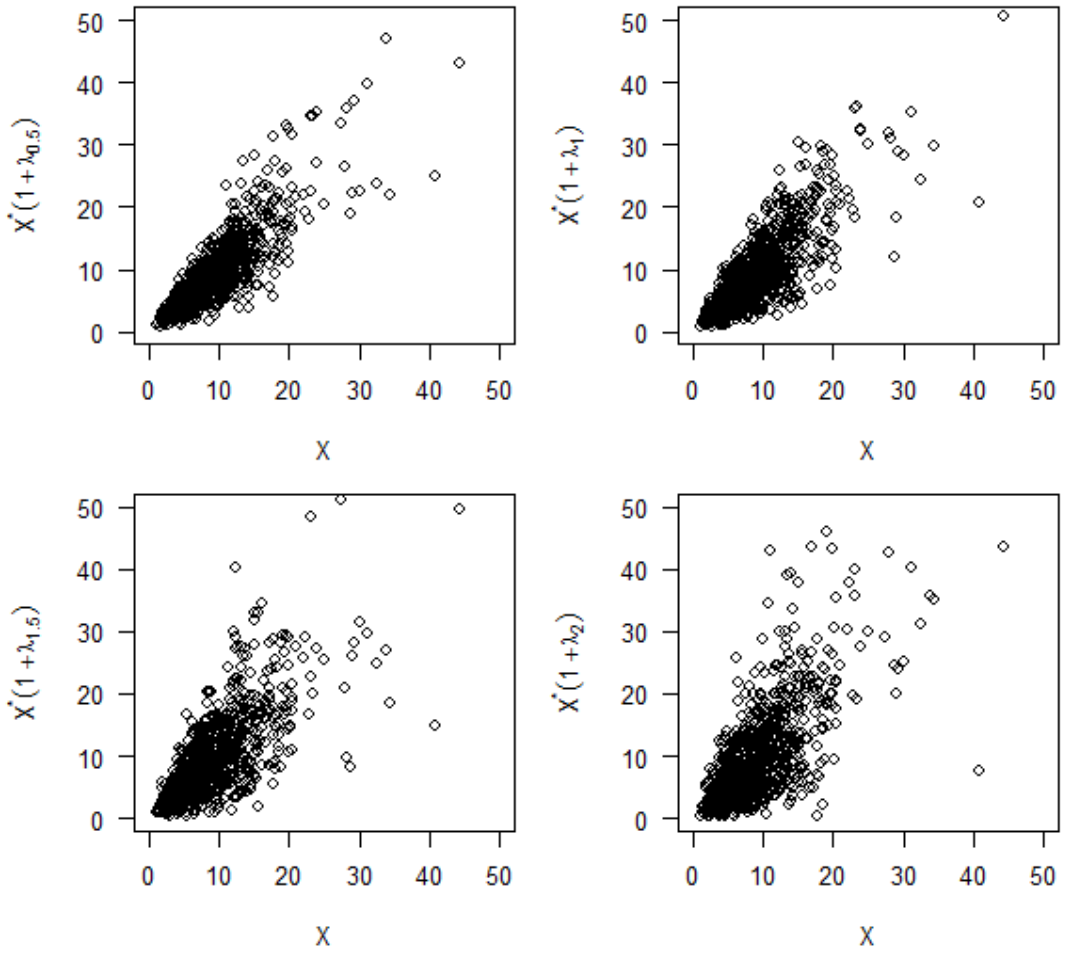
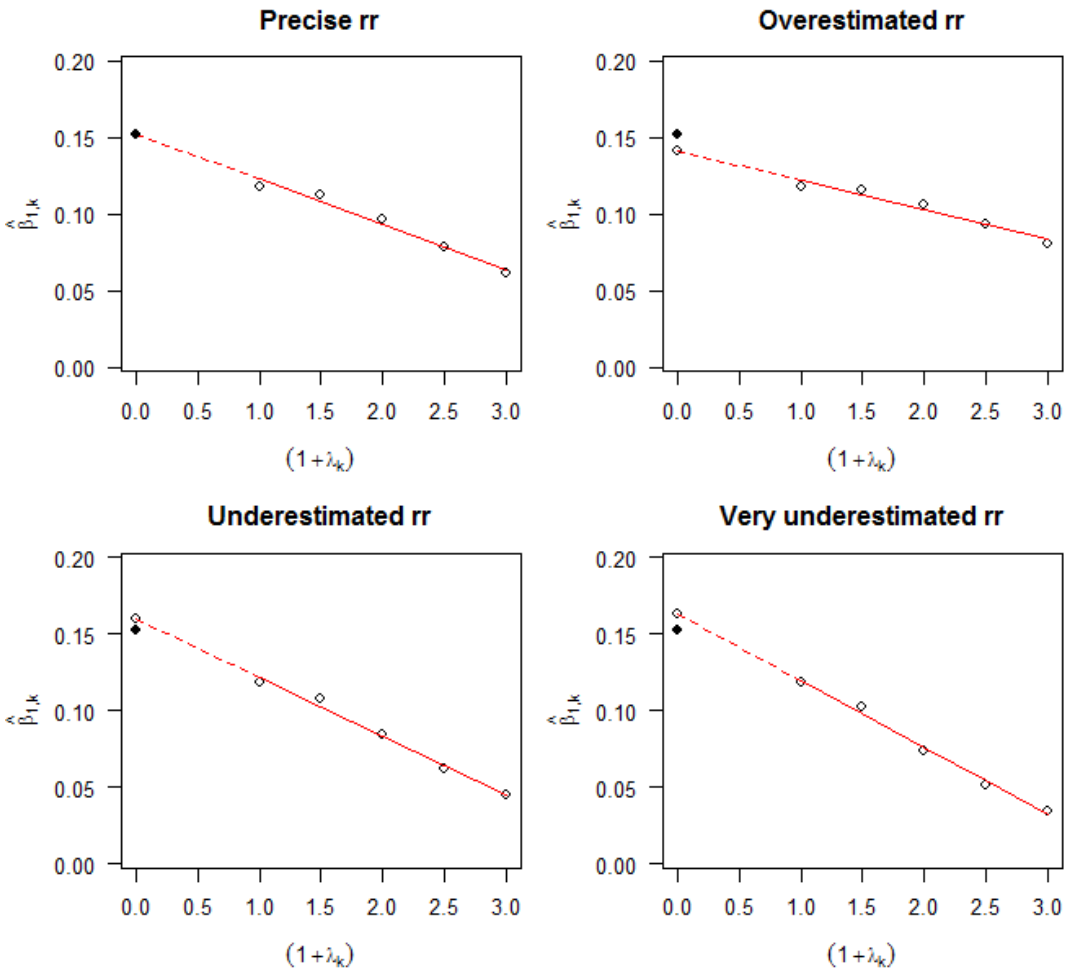



Figure A2 shows the extrapolation functions for $\beta_1$ when the outcome model is linear and $X_1$ is affected by classical multiplicative measurement. Each of the plots represents one of the four scenarios where different reliability ratios were assumed.

**Figure A2:** Extrapolation functions for the linear model



Metodološki zvezki, Vol. 13, No. 1, 2016, 59-67

# Odds Ratio, Hazard Ratio and Relative Risk

Janez Stare[1]
Delphine Maucort-Boulch[2]

**Abstract**

Odds ratio (OR) is a statistic commonly encountered in professional or scientific medical literature. Most readers perceive it as relative risk (RR), although most of them do not know why that would be true. But since such perception is mostly correct, there is nothing (or almost nothing) wrong with that. It is nevertheless useful to be reminded now and then what is the relation between the relative risk and the odds ratio, and when by equating the two statistics we are sometimes forcing OR to be something it is not. Another statistic, which is often also perceived as a relative risk, is the hazard ratio (HR). We encounter it, for example, when we fit the Cox model to survival data. Under proportional hazards it is probably "natural" to think in the following way: if the probability of death in one group is at every time point *k*-times as high as the probability of death in another group, then the relative risk must be *k*, regardless of where in time we are. This could be hardly further from the truth and in this paper we try to dispense with this blunder.

# 1 Introduction

## 1.1 Relative risk

In medical studies, probability of seeing a certain event in some group is usually called risk, while epidemiologists might prefer the term incidence (Savitz, 1992). For comparison of risks between groups, the ratio of risks, or the relative risk, is a statistic of choice. Formally, if $\pi_1$ is the probability of the event in group 1, and $\pi_2$ is the probability of the event in group 2, then the **relative risk** is

$$RR = \frac{\pi_1}{\pi_2}.$$

The reason of preferring relative risk over the **difference** of risks

$$RD = \pi_1 - \pi_2$$

lies in the fact that the population risks of most diseases are rather small and so differences less dramatic (Walter, 2000). For example, if the probability of some cancer in one group is 0.001, and in the other 0.009, the difference is 0.008 (same as between 0.419 and 0.411), but the relative risk is 9!

[1] Department of Biostatistics and Medical Informatics, University of Ljubljana, Slovenia; janez.stare *at* mf.uni-lj.si

[2] Service de Biostatistique, Hospices Civils de Lyon, Lyon, France; delphine.maucort-boulch *at* chu-lyon.fr

**Table 1:** Probability of death among men and women on the Titanic.

| Sex | Died | Survived | Risk |
|---|---|---|---|
| men | 1364 | 367 | $1364/1731 = 0.79$ |
| women | 126 | 344 | $126/470 = 0.27$ |

Table 1 provides an example where the event, unfortunately, was not rare. The relative risk of death of men compared to women is

$$RR = \frac{0.79}{0.27} = 2.93.$$

## 1.2 Odds ratio

The other statistics, commonly encountered in medical literature, is the **odds ratio** (Bland and Altman, 2000). Odds are the ratio of the probability of an event occurring in a group, divided by the probability of that event not occurring

$$\text{odds} = \frac{\pi}{1-\pi}.$$

For example, if probability of death in a group is 0.75, the odds are equal to 3, since the probability of death is three times higher than the probability of surviving. Table 2 gives the odds among men and women on the Titanic.

**Table 2:** Odds for death among men and women on the Titanic, $\pi$ denotes the probability of death.

| | Death | Survival | |
|---|---|---|---|
| Sex | $\pi$ | $1-\pi$ | Odds |
| men | 0.79 | 0.21 | 3.76 |
| women | 0.27 | 0.73 | 0.37 |

If risk was the same in both groups, the odds would be equal. A comparison of odds, the **odds ratio**, might then make sense.

$$OR = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_2}{1-\pi_2}}$$

Odds ratio for the Titanic example is

$$OR = \frac{3.76}{0.37} = 10.16.$$

This is very different from the relative risk calculated on the same data and may come as a surprise to some readers who are accustomed of thinking of odds ratio as of relative risk (Greenland, 1987).

Since we already have relative risk, why would we want to calculate the odds ratio? The answer is not obvious and it is best explained via an example (Nurminen, 1995).

Case-control studies are quite common in medical studies. In these we select a sample of patients and a sample of controls, and study occurrence of some factor, hopefully predictive, in the two groups. The reason for collecting data in such a way is that it takes a long time and big sample sizes to do a follow up study, that is a study in which two groups, with and without a factor, are followed long enough for a disease to appear in numbers large enough to do statistical tests with acceptable power.

Table 3 shows fictional data on prostate cancer and baldness. We see that of the 129 cases, 72 were bald, and 55 were not, while among the 139 controls 82 were bald. Let us remind ourselves that in order to calculate the relative risk between the two groups we would need probabilities of cancer occurring, so probability to have cancer for bald and not bald people. It may seem natural to estimate these probabilities as $\frac{72}{154}$ and $\frac{55}{112}$, and so $RR$ as

**Table 3:** Prostate cancer and baldness

| | Case | Control | total |
|---|---|---|---|
| bald | 72 | 82 | 154 |
| not bald | 55 | 57 | 112 |
| total | 129 | 139 | 268 |

$$RR = \frac{\frac{72}{154}}{\frac{55}{112}} = 0.95,$$

but is this correct?

It is very important to understand that this is not correct. Since we randomly chose cases and controls, we can estimate probabilities of observing baldness (or not) among them; but NOT the probabilities of observing cancer among the bald (and not bald) people.

This means that in a study like this we CANNOT calculate the relative risk.

# 2 Relative risk and odds ratio (RR in OR)

The literature dealing with the relation between relative risk and odds ratio is quite extensive (some examples are (Davies et al., 1998; Deeks, 1998; Newman, 2001; Nurminen, 1995; Pearce, 1993; Savitz, 1992; Zhang and Yu, 1998)). We still hope that the derivation below will be useful.

Table 4 gives a 2x2 table in general notation.

Using this notation we have

$$RR = \frac{\frac{n_{11}}{n_{11}+n_{12}}}{\frac{n_{21}}{n_{21}+n_{22}}} = \frac{n_{11}}{n_{21}} \cdot \frac{n_{21}+n_{22}}{n_{11}+n_{12}}$$

and

$$OR = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_2}{1-\pi_2}} = \frac{\frac{n_{11}/(n_{11}+n_{12})}{n_{12}/(n_{11}+n_{12})}}{\frac{n_{21}/(n_{21}+n_{22})}{n_{22}/(n_{21}+n_{22})}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

**Table 4:** A 2x2 table in general notation.

| | **Outcome** | | |
|---|---|---|---|
| **Factor** | Death (Case) | Survival (Control) | Total |
| yes | $n_{11}$ | $n_{12}$ | $n_{11}+n_{12}$ |
| no | $n_{21}$ | $n_{22}$ | $n_{21}+n_{22}$ |
| total | $n_{11}+n_{21}$ | $n_{12}+n_{22}$ | $n$ |

Let us now multiply one column, say cases, by $k$.
Then we have

$$RR = \frac{kn_{11}}{kn_{21}} \cdot \frac{kn_{21}+n_{22}}{kn_{11}+n_{12}} = \frac{n_{11}}{n_{21}} \cdot \frac{kn_{21}+n_{22}}{kn_{11}+n_{12}}$$

and

$$OR = \frac{kn_{11}n_{22}}{n_{12}kn_{21}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}.$$

We see that the 'relative risk' is now different, but the odds ratio does not change if we change the ratio of cases versus controls.
Until now we have learned the following:

1. we can calculate relative risk IF we can estimate probabilities of an outcome in EACH group.

2. we can't do that in case control studies.

3. we can calculate the odds ratio even if we don't know the probabilities in the groups.

**It would then be nice, if odds ratio was close to relative risk.**

Let us now look at the relation between the relative risk and the odds ratio (Zhang and Yu, 1998).

$$OR = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_2}{1-\pi_2}} = \frac{\pi_1}{\pi_2} \cdot \frac{1-\pi_2}{1-\pi_1} = RR \cdot \frac{1-\pi_2}{1-\pi_1} \tag{2.1}$$

From this we see that OR is always further away from 1 than RR. But, more importantly, we see that the odds ratio is close to the relative risk **if probabilities of the outcome are small** (Davies et al., 1998). And it is this fact that enables us, most of the time, to approximate the relative risk with the odds ratio. Table 5 below illustrates the relationship between RR and OR for some probabilities of the outcome.

# 3 Relation between RR and HR

If one searches the Internet for the relation between the hazard ratio and the relative risk, one will one will predominantly find statements that tell us that these two statistics are more or less equal (Nurminen, 1995). For example, the Glossary at the British Medical Journal site http://clinicalevidence.bmj.com/ceweb/resources/glossary.jsp says

**Table 5:** Examples of RR and OR for different probabilities.

| $\pi_1$ | $\pi_2$ | RR | OR |
|---|---|---|---|
| .4 | .1 | 4 | 6 |
| .2 | .3 | .67 | .58 |
| .04 | .01 | 4 | 4.125 |
| .02 | .03 | .67 | .66 |

> **Hazard ratio (HR)**
>
> Broadly equivalent to relative risk (RR); useful when the risk is not constant with respect to time. It uses information collected at different times. The term is typically used in the context of survival over time. If the HR is 0.5 then the relative risk of dying in one group is half the risk of dying in the other group.

The same site has the following definition for Relative Risk

> **Relative risk (RR)**
>
> The number of times more likely ($RR > 1$) or less likely ($RR < 1$) an event is to happen in one group compared with another. It is the ratio of the absolute risk (AR) for each group. It is analogous to the odds ratio (OR) when events are rare.
>
> Relative risk is calculated as the absolute risk (AR) in the intervention group divided by the AR in the control group.

It would seem that the claim above about HR and RR is generally accepted as correct, although we couldn't find any derivation supporting it. Some of the confusion might be caused by esteemed authors who, in trying to avoid a somewhat unfortunate name *proportional hazards model*, call such models *relative risk models* (Kalbfleisch and Prentice, 2002). It is of course obvious that by risk they are referring to the conditional probability of dying in a small interval, so $r(t) = P(t \leq T < t + \Delta t | T \geq t)$, but the ratio of such risks is not what people usually understand under the term *relative risk*, since relative risk is about absolute and not conditional probabilities..

So most of the confusion, or wrong perception, probably comes from this 'natural' line of thought: if hazard ratio is $k$ at all times, then the relative risk must be $k$ at all times. And this is of course wrong.

Relative risk (RR) is a ratio of two probabilities: probability of an event in one group divided by the probability of the same event in the other group. When studying survival, we have to explicitly state in which time interval we are calculating this probability. So, for a given time $t$, the relative risk is

$$RR(t) = \frac{P(T \leq t | X = x_1)}{P(T \leq t | X = x_2)}$$

where $x_1$ and $x_2$ are values of the covariate $X$ defining the two groups (male, female for example).

Hazard ratio is a ratio of two hazard functions

$$HR(t) = \frac{\lambda_1(t, x_1)}{\lambda_2(t, x_2)} \tag{3.1}$$

and we remind the reader that the hazard function is defined as

$$\lambda(t, x) = \lim_{\Delta t \to 0^+} \frac{P(t \leq T < t + \Delta t | T \geq t, X = x)}{\Delta t}$$

and that hazard is connected to the survival function via the following formula

$$S(t, x) = e^{-\int_0^t \lambda(u,x)du}.$$

Since

$$S(t, x) = P(T > t | X = x) = 1 - P(T \leq t | X = x)$$

we can write

$$RR(t) = \frac{1 - S(t, x_1)}{1 - S(t, x_2)} = \frac{1 - e^{-\int_0^t \lambda(u,x_1)du}}{1 - e^{-\int_0^t \lambda(u,x_2)du}}. \tag{3.2}$$

It is difficult to argue that equations (3.1) and (3.2) are similar, but let's try. Sometimes, like in the comparison between the Kaplan-Meier and the Nelson-Aalen estimate of the survival function, the following argument is brought into play

$$e^{-x} \approx 1 - x. \tag{3.3}$$

This comes from the Taylor series expansion of the function $e^{-x}$ around the value 0

$$e^{-x} = 1 - x + \frac{x^2}{2} - \cdots$$

Obviously, approximation (3.3) makes sense only for very small values of $x$. Note that our values of $x$ are $\int_0^t \lambda(u, x_1)du$ and $\int_0^t \lambda(u, x_2)du$, which are cumulative hazards, increasing without limits when $t$ increases. Such an approximation will never hold, except for early times in a survival study. Applying it (wrongly!) to formula (3.2) in case of proportional hazards, so when $\lambda(u, x_1) = k\lambda(u, x_2)$, would make formulas (3.1) and (3.2) equal.

# 4 Illustration

For easier understanding in Table 6 we give detailed calculations for two groups for the first two times in a possible series of discrete event times and with proportional hazards.

So we have

$$RR(t_2) = \frac{k(p_1 + p_2 - kp_1p_2)}{p_1 + p_2 - p_1p_2}$$

which is NOT equal to $k$, but can be close for small probabilities and small $k$. As time passes, RR is further and further away from HR.

Another example is illustrated in Figure 1 and Table 7. Data were simulated from two exponential distributions with $HR = 3$ and with 500 cases in each group. We see that only at the first point, close to $t = 0$, the estimate is around 3. Later it quickly diminishes and is already halved at $t = 1$.

**Table 6:** Calculation of relative risk for at two discrete time points. Hazards are proportional and equal to $k$.

| | $t_1$ | $t_2$ |
|---|---|---|
| probability of event in group 1 | $p_1$ | $p_2$ |
| probability of event in group 2 | $kp_1$ | $kp_2$ |
| probability of survival in group 1 | $1-p_1$ | $(1-p_1)(1-p_2)$ |
| probability of survival in group 2 | $1-kp_1$ | $(1-kp_1)(1-kp_2)$ |
| probability of event up to $t$ in group 1 | $p_1$ | $1-(1-p_1)(1-p_2)$ |
| probability of event up to $t$ in group 2 | $kp_1$ | $1-(1-kp_1)(1-kp_2)$ |
| RR up to given time | $\frac{kp_1}{p_1}=k$ | $\frac{1-(1-kp_1)(1-kp_2)}{1-(1-p_1)(1-p_2)}$ |

**Table 7:** *Calculation of RR at three different time points for the situation illustrated in Figure 1*

| time | RR |
|---|---|
| 0.1 | 3.15 |
| 0.5 | 2.12 |
| 1.0 | 1.51 |

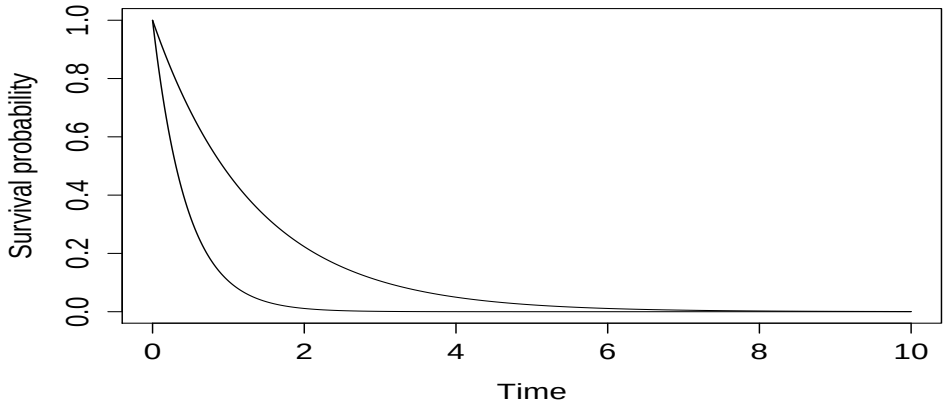



**Figure 1:** *Two exponential curves with $HR = 3$.*

# 5 Discussion

In our experience, equating odds ratios with relative risk has become too common, and results, even when probabilities of events are not small, are always interpreted as relative risks (Deeks, 1998; Greenland, 1987; Nurminen, 1995). Having odds ratios as a result of logistic regression fits of course adds to this. We believe that, in case the assumption of a rare event cannot be supported, an effort should be made to estimate relative risk correctly (if possible), or to at least give some estimates, using formula (2.1), for different values of $\pi_1$ and $\pi_2$.

It is of course possible that many of the claims about the similarity between HR and RR are made with small intervals in mind. If so, then this should be made very clear when such a statement is made (still, the question why that would be of interest, would remain). The above example from the British Medical Journal site certainly isn't clear about this.

Of course, simply stating that one has small intervals in mind is still not enough. One has to explicitly say that he/she has conditional probabilities in mind as the definition

$$RR(t) = \frac{P(t < T \leq t + \Delta t | X = x_1)}{P(t < T \leq t + \Delta t | X = x_2)}$$

is still NOT the hazard ratio, as it is not a ratio of conditional probabilities.

Maybe the easiest way to understand that a hazard ratio cannot be equal to the relative risk for any time $t$ is to realize that eventually everybody dies, so the relative risk will approach 1 with time, even though the hazard ratio is constant.

# References

[1] Beaudoin, G. (2014): Meeting the information needs of news media to increase citizens' understanding of statistical findings. Paper presented at *Work Session on the Communication of Statistics*. UNECE.

[2] Bland, J.M. and Altman, D.G. (2000): The odds ratio. *British Medical Journal*, **320**, 1468.

[3] Davies, H.T., Crombie, I.K., and Tavakoli, M. (1998): When can odds ratios mislead? *British Medical Journal*, **316**, 989–991.

[4] Deeks, J. (1998): When can odds ratios mislead? *British Medical Journal*, **317**, 1155–1156.

[5] Greenland, S. (1987): Interpretation and choice of effect measures in epidemiologic analyses. *American Journal of Epidemiology*, **125**, 761–768.

[6] Kalbfleisch, J.D. and Prentice, R.L. (2002): *The Statistical Analysis of Failure Time Data.* John Wiley & Sons.

[7] Newman, S.C. (2001): *Biostatistical Methods in Epidemiology.* John Wiley & Sons.

[8] Nurminen, M. (1995): To use or not to use the odds ratio in epidemiologic studies? *European Journal of Epidemiology*, **11**, 365–371.

[9] Pearce, N. (1993): What does the odds ratio estimate in a case-control study? *International Journal of Epidemiology*, **22**, 1189–1192.

[10] Savitz, D.A. (1992): Measurements, estimates, and inferences in reporting epidemiologic study results. *American Journal of Epidemiology*, **135**, 223–224.

[11] Walter, S.D. (2000): Choice of effect measure for epidemiological data. *Journal of Clinical Epidemiology*, **53**, 931–939.

[12] Zhang, J. and Yu, K.F. (1998): What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *The Journal of the American Medical Association*, **280**, 1690–1691.

# INSTRUCTIONS TO AUTHORS

**Language**: *Metodološki zvezki – Advances in Methodology and Statistics* is published in English.

**Submission of papers**: Authors are requested to submit their articles (complete in all respects) to the Editor by e-mail (MZ@stat-d.si). Contributions are accepted on the understanding that the authors have obtained the necessary authority for publication. Submission of a paper will be held to imply that it contains original unpublished work and is not being submitted for publication elsewhere. Articles must be prepared in LaTeX or Word. Appropriate styles and example files can be downloaded from the Journal's web page (http://www.stat-d.si/mz/).

**Review procedure:** Manuscripts are reviewed by two referees. The editor reserves the right to reject any unsuitable manuscript without requesting an external review.

## Preparation of manuscripts

**Tables and figures**: Tables and figures must appear in the text (not at the end of the text). They are numbered in the following way: Table 1, Table 2,…, Figure 1, Figure 2,…

**References within the text:** The basic reference format is (Smith, 1999). To cite a specific page or pages use (Smith, 1999: 10-12). Use "et al." when citing a work by more than three authors (Smith et al., 1999). The letters a, b, c etc. should be used to distinguish different citations by the same author(s) in the same year (Smith, 1999a; Smith, 1999b).

**Notes:** Essential notes, or citations of unusual sources, should be indicated by superscript number in the text and corresponding text under line at the bottom of the same page.

**Equations:** Equations should be centered and labeled with two numbers separated by a dot enclosed by parentheses. The first number is the current section number and the second a sequential equation number within the section, e.g., (2.1)

**Author notes and acknowledgements:** Author notes identify authors by complete name, affiliation and his/her e-mail address. Acknowledgements may include information about financial support and other assistance in preparing the manuscript.

**Reference list:** All references cited in the text should be listed alphabetically and in full after the notes at the end of the article.

**References to books, part of books or proceedings:**

[1] Smith, J.B. (1999): *Title of the Book.* Place: Publisher.

[2] Smith, J.B. and White A.B. (2000): *Title of the Book.* Place: Publisher.

[3] Smith, J. (2001): Title of the chapter. In A.B. White (Ed): *Title of the Proceedings*, 14-39. Place: Publisher.

**Reference to journals:**

[4] Smith, J.B. (2002): Title of the article**.** *Name of Journal***, 2**, 46-76.

# Metodološki zvezki
## Advances in Methodology and Statistics


| | | |
|---|---|---|
| Published by | | Izdajatelj |
| **Faculty of Social Sciences University of Ljubljana, for Statistical Society of Slovenia** | | **Fakulteta za družbene vede Univerze v Ljubljani za Statistično društvo Slovenije** |
| Editors | **Valentina Hlebec**<br>**Lara Lusa** | Urednika |
| Founding Editors | **Anuška Ferligoj**<br>**Andrej Mrvar** | Prva urednika |
| Cover Design | **Bojan Senjur**<br>**Gregor Petrič** | Oblikovanje naslovnice |
| Typesetting | **Lara Lusa** | Računalniški prelom |
| Printing | **Littera Picta d.o.o.**<br>**Ljubljana, Slovenia** | Tisk |
| | **MZ** | |
| is indexed and abstracted in | **SCOPUS**<br>**EBSCO**<br>**ECONIS**<br>**STMA-Z**<br>**ProQuest** | je indeksirana in abstrahirana v |
| Home page URL | http://www.stat-d.si/mz/ | Spletna stran |

**ISSN 1854 - 0023**