Austrian Journal of Statistics
2012, Volume 52, 149–158.
http://www.ajs.or.at/
doi:10.17713/ajs.v52iSI.1771

Estimation in Linear-Rate Simple Survival Models with Measurement Errors and Censoring

Sergiy Shklyar
Taras Shevchenko National University of Kyiv

Abstract

A simple exponential regression model is considered where the rate parameter of the response variable linearly depends on the explanatory variable. We consider complications of the model: censoring of the response variable (either upper censoring or interval observations), the additive classical error or multiplicative Berkson error in the explanatory variable, or a combination of censoring with Berkson errors. We construct or use already-known estimators in the models, and verify their performance in simulations.

Keywords: exponential regression, classical generalized linear models, errors in variables, censoring.

1. Introduction

The inverse exponential regression is a regression model where the dependent variable T_n has an exponential distribution with rate parameter being a linear function of the independent variable X_n, that is $T_n \sim \mathrm{Exp}(\beta_0 + \beta_1 X_n)$, where β_0 and β_1 are regression parameters. This is a particular case of a classical generalized linear model. The log-likelihood function is easy to produce (Carroll, Ruppert, Stefanski, and Crainiceanu 2004; Nelder and Wedderburn 1972; Wainwright and Jordan 2008; Wedderburn 1974), and the maximum likelihood estimator can be obtained by means of convex optimization.

Consider the setting where the independent variable (the true regressor) X_n or the dependent variable T_n is not observed directly. For the former, consider error models. In the classical error model with additive error, the surrogate variable $X_n^{\mathrm{meas}} = X_n + \delta_n$ is observed instead of X_n. In the Berkson multiplicative model, the true regressor X_n is a random variable itself; and $X_n = X_n^0 \exp(\epsilon_n^*)$, while X_n^0 is observed. In the model with the mixture of errors of different types, the true regressor is a perturbed initial regressor, $X_n = X_n^0 \exp(\epsilon_n^*)$, and the observable surrogate X_m^{meas} is also a perturbed initial regressor — in a different way, $X_n^{\mathrm{meas}} = X_n^0 + \delta_n$.

For the classical error, we use an additive model, where the Stefanski and Carroll conditional maximum likelihood method is applicable. For the Berkson error, we use the multiplicative model; thus the rate parameter for the exponential distribution is nonnegative as long as the initial regressor and the regression coefficients are nonnegative.

For the dependent variable, we consider the cases of direct observation of T_n, censored obser-

vation of T_n (where $\min(T_n, T_n^{\text{o.p.}})$ and $\mathbf{1}(T_n \le T_n^{\text{o.p.}})$ are observed, with $T^{\text{o.p.}}$ the observation period), and interval observation of T_n. The extreme case of models with the interval observations is the binary model, where only $\mathbf{1}(T_n \le T_n^{\text{o.p.}})$ is observed.

The estimator — in particular, the estimator for the model with interval observations of the survival time — can be used in cohort studies with repeated testing. See Section 7 of Masiuk, Kukush, Shklyar, Chepurny, and Likhtarov (2017) and references therein for setup of the study.

Augustin (2004) constructed a corrected score estimator in the Cox proportional hazards model with measurement errors. The corrected empirical likelihood method for generalized linear models with measurement errors, in particular for the gamma-inverse regression and logistic regression, is developed in Yang, Li, and J. (2015).

In Section 2 we construct the estimators in the exponential regression without measurement errors, with or without censoring of the response variable. In Section 3, we consider a model with only one type of error in the explanatory variable, either an additive classical error or a multiplicative Berkson error. In Section 4 we consider a combination of the multiplicative Berkson error in the explanatory variable with censoring of the response variable. Simulations are presented in Section 5, and Section 6 concludes.

2. Models without measurement errors

2.1. Model with finite observed survival time

We assume that the survival time has exponential distribution,

$$\mathsf{P}(T_n > t) = \exp(-(\beta_0 + \beta_1 X_n)t), \qquad t > 0. \tag{1}$$

Here X_n is a regressor. The observed variables are (X_n, T_n), $n = 1, \dots, N$, while (β_0, β_1) is a parameter of interest. Throughout the paper, we assume that $\beta_1 > 0$ and $\beta_0 + \beta_1 X_n > 0$ for all n.

The regressor X_n considered non-random, the density of T_n is

$$\mathrm{pdf}_{T_n}(t) = (\beta_0 + \beta_1 X_n) \exp(-(\beta_0 + \beta_1 X_n)t).$$

With the observation T_n substituted for the argument t, and argument (b_0, b_1) substituted for the parameter (β_0, β_1), this can be considered a one-observation (elementary) likelihood function. The likelihood function is

$$\mathcal{L}(b_0, b_1) = \prod_{n=1}^{N} (b_0 + b_1 X_n) e^{-(b_0 + b_1 X_n)T_n},$$

$$-\log(\mathcal{L}(b_0, b_1)) = \sum_{n=1}^{N} (b_0 + b_1 X_n)T_n - \sum_{n=1}^{N} \log(b_0 + b_1 X_n).$$

The maximum likelihood estimator is a solution to equations

$$\sum_{n=1}^{N} T_n = \sum_{n=1}^{N} \frac{1}{\hat\beta_0 + \hat\beta_1 X_n}, \tag{2}$$

$$\sum_{n=1}^{N} X_n T_n = \sum_{n=1}^{N} \frac{X_n}{\hat\beta_0 + \hat\beta_1 X_n}. \tag{3}$$

The linear combination of equations (2) and (3) with coefficients $\hat\beta_0$ and $\hat\beta_1$ is

$$\sum_{n=1}^{N} (\hat\beta_0 + \hat\beta_1 X_n)T_n = N, \tag{4}$$

The maximum likelihood estimator satisfies (4), which is a linear equation in $\hat{\beta}$. This allows to reduce the problem of finding the solution to one-variable optimization or one-variable one-equation solving.

2.2. Model with survival times censored from above

Let (1) still hold true. However, the survival time T_n is not always observed. On each observation, we know the regressor X_n, the observation period $T_n^{\text{o.p.}}$, whether $T_n \le T_n^{\text{o.p.}}$ or $T_n > T_n^{\text{o.p.}}$. We also observe T_n if $T_n \le T_n^{\text{o.p.}}$. In this model,

$$\begin{aligned}\operatorname{pdf}_{T_n}(t) &= (\beta_0 + \beta_1 X_n)\exp(-(\beta_0+\beta_1 X_n)t),\\ \mathsf{P}(T_n > T_n^{\text{o.p.}}) &= \exp(-(\beta_0+\beta_1 X_n)T_n^{\text{o.p.}}).\end{aligned}$$

These expressions are used to construct the likelihood function, for $T_n \le T_n^{\text{o.p}}$ and $T_n > T_n^{\text{o.p}}$ respectively:

$$\mathcal{L}(b_0,b_1) = \prod_{n\,:\,T_n\le T_n^{\text{o.p.}}} (b_0+b_1X_n)e^{-(b_0+b_1X_n)T_n} \prod_{n\,:\,T_n> T_n^{\text{o.p.}}} e^{-(b_0+b_1X_n)T_n^{\text{o.p.}}}$$

The minus log-likelihood function is

$$\begin{aligned}-\log(\mathcal{L}(b_0,b_1)) &= \sum_{n\,:\,T_n\le T_n^{\text{o.p.}}} (b_0+b_1X_n)T_n - \sum_{n\,:\,T_n\le T_n^{\text{o.p.}}} \log(b_0+b_1X_n) +\\ &\quad + \sum_{n\,:\,T_n> T_n^{\text{o.p.}}} (b_0+b_1X_n)T_n^{\text{o.p.}}\\ &= \sum_{n=1}^{N}(b_0+b_1X_n)\min(T_n,T_n^{\text{o.p.}}) - \sum_{n\,:\,T_n\le T_n^{\text{o.p.}}} \log(b_0+b_1X_n).\end{aligned}$$

The maximum likelihood estimator is a solution to equations

$$\begin{aligned}\sum_{n=1}^{N}\min(T_n,T_n^{\text{o.p.}}) &= \sum_{n\,:\,T_n\le T_n^{\text{o.p.}}} \frac{1}{\hat{\beta}_0+\hat{\beta}_1X_n},\\ \sum_{n=1}^{N}X_n\min(T_n,T_n^{\text{o.p.}}) &= \sum_{n\,:\,T_n\le T_n^{\text{o.p.}}} \frac{X_n}{\hat{\beta}_0+\hat{\beta}_1X_n}.\end{aligned}$$

The linear combination of these equations with coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ is

$$\sum_{n=1}^{N}(\hat{\beta}_0+\hat{\beta}_1X_n)\min(T_n,T_n^{\text{o.p.}}) = \sum_{n=1}^{N}\mathbf{1}(T_n\le T_n^{\text{o.p.}}).$$

Similar the case where the survival times T_n are observed without censoring, this allows to reduce the calculation of the maximum-likelihood estimator to one-dimensional equation.

2.3. Model with interval observations of survival times

Let (1) hold true, the variable T_n is not observed directly. Instead, let $T_n^{\min}$ and $T_n^{\max}$ be known such that $T_n^{\min} < T_n \le T_n^{\max}$. (We might observe the stochastic process $\{\mathbf{1}(T_n \le t),\ t>0\}$ in a finite number of points. Then we choose the ends of the interval where the process jumps from 0 to 1. If $\mathbf{1}(T_n > t) = 1$ at all observations, then we choose $T_n^{\max} = \infty$.) Then

$$\begin{aligned}\mathsf{P}[T_n^{\min} < T_n \le T_n^{\max}] &= e^{-(\beta_0+\beta_1X_n)T_n^{\min}} - e^{-(\beta_0+\beta_1X_n)T_n^{\max}} \qquad \text{if } T_n^{\max}<\infty,\\ \mathsf{P}[T_n^{\min} < T_n \le T_n^{\max}] &= e^{-(\beta_0+\beta_1X_n)T_n^{\min}} \qquad\qquad\quad \text{if } T_n^{\max}=\infty.\end{aligned}$$

The likelihood function is

$$\mathcal{L}(b_0, b_1) = \prod_{n\,:\,T_n^{\max}<\infty} (e^{-(b_0+b_1X_n)T_n^{\min}} - e^{-(b_0+b_1X_n)T_n^{\max}}) \prod_{n\,:\,T_n^{\max}=\infty} e^{-(b_0+b_1X_n)T_n^{\min}}.$$

Then

$$-\log(\mathcal{L}(b_0, b_1)) = \sum_{n=1}^{N}(b_0 + b_1X_n)T_n^{\min} - \sum_{n\,:\,T_n^{\max}<\infty} \log(1 - e^{-(b_0+b_1X_n)(T_n^{\max}-T_n^{\min})}).$$

3. Models with single-type measurement errors

3.1. Model with additive classical error

Sufficiency estimator

Let this variation of (1) hold true:

$$\mathsf{P}(T_n > t) = \exp(-(\beta_0 + \beta_1X_n^{\text{true}})t), \qquad t > 0.$$

However, X^{true} is not observed directly: it is observed with an error. Assume that X^{true} is assumed with an additive error δ, which is a zero-mean Gaussian variable. The observation is

$$X_n^{\text{meas}} = X_n^{\text{true}} + \delta_n, \qquad \delta_n \sim N(0, \sigma_{\delta,n}^2).$$

Here and hereafter, $N(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2. We also assume that T_n and δ_n are independent.

On each observation, we know T_n, X^{meas} and $\sigma_{\delta,n}^2$. The parameter of interest is (β_0, β_1).

The error-free model is a classical generalized linear model. Thus, in the model with Gaussian classical measurement error, conditional score methods are applicable. The conditional distribution of T_n given $X_n^{\text{meas}} - T_n\sigma_{\delta,n}^2\beta_1$ depends only on the parameters and observed variables. Denote the density of this conditional distribution as

$$p(T;\ X_n^{\text{meas}} - T_n\sigma_{\delta,n}^2\beta_1,\ \boldsymbol{\beta}, \sigma_{\delta,n}^2).$$

The sufficiency estimator is defined from equation

$$\sum_{n=1}^{N} \frac{\partial}{\partial \mathbf{b}} \Big(\log p(T_n;\ \Delta, \mathbf{b}, \sigma_{\delta,n}^2)\Big)\Big|_{\Delta=X_n^{\text{meas}}-T_n\sigma_{\delta,n}^2\hat{\beta}_1,\ \mathbf{b}=\hat{\boldsymbol{\beta}}} = 0.$$

Denote

$$m(X_n^{\text{meas}} - T_n\sigma_{\delta,n}^2\beta_1,\ \boldsymbol{\beta}, \sigma_{\delta,n}^2) = \mathsf{E}[T_n \mid X_n^{\text{meas}} - T_n\sigma_{\delta,n}^2\beta_1].$$

Shklyar (2014, Section 5) shows that the sufficiency estimator satisfies the equations

$$\sum_{n=1}^{N} m(X_n^{\text{meas}} - T_n\sigma_{\delta,n}^2\beta_1X_n^{\text{meas}},\ \hat{\boldsymbol{\beta}}, \sigma_{\delta,n}^2) = \sum_{n=1}^{N} T_n,$$

$$\sum_{n=1}^{N}(\hat{\beta}_0 + \hat{\beta}_1X_n^{\text{meas}}) = N.$$

Sufficiency score estimators in classical generalized linear models with classical errors added are developed in Stefanski and Carroll (1987). For concrete exponential regression model, the estimator is developed in Shklyar (2014).

Conditional score estimator

The approach based on unbiased estimation functions can be used to obtain conditional score estimators and corrected score estimators. A conditional score estimator is defined as a solution to the equation

$$\sum_{k=1}^{N}(m(X_n^{\text{meas}} - T_n\sigma_{\delta,n}^2\beta_1 X_n^{\text{meas}},\ \hat{\boldsymbol{\beta}}, \sigma_{\delta,n}^2) - T_n)k(X_n^{\text{meas}} - T_n\sigma_{\delta,n}^2\beta_1 X_n^{\text{meas}},\ \hat{\boldsymbol{\beta}}, \sigma_{\delta,n}^2) = 0$$

for some vector-valued function $k(\Delta, \boldsymbol{b}, \sigma^2)$. Different versions of the conditional score estimator, such as linear estimator and optimal estimator, are obtained in Stefanski and Carroll (1987).

Corrected score estimator

In the corrected score method, first the deconvolution equation is solved. Define the function

$$g(X^{\text{meas}}, T; \boldsymbol{b}, \sigma^2) = \begin{pmatrix} \frac{\sqrt{2\pi}}{|b_1\sigma|}\exp\left(-\frac{(b_0+b_1X^{\text{meas}})^2}{2b_1^2\sigma^2}\right)\Phi\left(-\frac{b_0+b_1X^{\text{meas}}}{|b_1\sigma|}\right) - T \\ \frac{1}{b_1} - \frac{\sqrt{2\pi}\,b_0}{b_1\,|b_1\sigma|}\exp\left(-\frac{(b_0+b_1X^{\text{meas}})^2}{2b_1^2\sigma^2}\right)\Phi\left(-\frac{b_0+b_1X^{\text{meas}}}{|b_1\sigma|}\right) - TX^{\text{meas}} \end{pmatrix}.$$

The function g is a solution to the following deconvolution equation:

$$\mathsf{E}\, g(X+\delta, T; \boldsymbol{b}, \sigma^2) = \left(\frac{1}{b_0+b_1X} - T\right)\begin{pmatrix}1\\X\end{pmatrix}, \quad \delta \sim N(0, \sigma^2).$$

The corrected score estimator is a solution to the equation

$$\sum_{n=1}^{N} g(X_n^{\text{meas}}, T_n; \widehat{\boldsymbol{\beta}}, \sigma_{\delta,n}^2) = 0.$$

As the estimation equation might have multiple solutions, the appropriate one should be chosen. The corrected-score method was developed in Nakamura (1990) and Stefanski (1989). The explicit expression for the corrected score function g is obtained in Shklyar (2014).

3.2. Model with multiplicative Berkson error

Let the real regressors X_n be known up to a multiplicative measurement errors

$$X_n^{\text{true}} = X_n^0\exp(\epsilon_n^*), \qquad \epsilon_n^* \sim N(0, \sigma_{\epsilon*n}^2). \tag{5}$$

As the true regressor is random, the relation (1) should be changed into the following:

$$\mathsf{P}[T_n > t \mid X_n^{\text{true}}] = \exp(-(\beta_0 + \beta_1 X_n^{\text{true}})t), \qquad t > 0. \tag{6}$$

The conditional density of T_n given X_n^0 is

$$\begin{aligned} p_{T_n|X_n^0=x}(t) &= p(t, x; \boldsymbol{\beta}, \sigma_{\epsilon*n}^2) = \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}(\beta_0 + \beta_1 x e^{\sigma_{\epsilon*n}u})e^{-(\beta_0+\beta_1 x\exp(\sigma_{\epsilon*n}u))t - 0.5u^2}\,du. \end{aligned} \tag{7}$$

On each observation, T_n, X_n^0 and $\sigma_{\epsilon*n}^2$ are known. As for other models, $\boldsymbol{\beta} = (\beta_0, \beta_1)$ is the parameter of interest.

Maximum likelihood estimator

The maximum likelihood estimator is a point where the likelihood function

$$\mathcal{L}(\boldsymbol{b}) = \prod_{n=1}^{N} p(T_n, X_n^0; \boldsymbol{b}, \sigma_{\epsilon*n}^2)$$

attains its maximum.

The computation of the integral in (7) might be problematic. We used numerical integration. More specifically, we used the scheme

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(u)\exp(-0.5u^2)\,du \approx \frac{1}{K}\sum_{k=1}^{K} f(\zeta_{k/(K+1)}),$$

where $\zeta_{k/(K+1)}$ are quantiles of the standard normal distribution.

In the binary regression model where the odds ratio linearly depends on the explanatory variable, with multiplicative Berkson error, the full maximum likelihood estimator is constructed in Masiuk *et al.* (2017), Section 6.4.1.

Unbiased estimating equation

Another method is based on unbiased score functions. Denote

$$m(X^0;\boldsymbol{b},\sigma^2) = \mathsf{E}\,\frac{1}{b_0 + b_1 X^0\exp(\epsilon)}, \quad \epsilon\sim N(0,\sigma^2),$$

so that $\mathsf{E}\,T_n = m(X^0;\boldsymbol{\beta},\sigma^2_{\epsilon*n})$. The estimator is sought from the equation

$$\sum_{n=1}^{N}(m(X_n^0;\widehat{\boldsymbol{\beta}},\sigma^2_{\epsilon*n}) - T_n)k(X_n^0;\widehat{\boldsymbol{\beta}},\sigma^2_{\epsilon*n}) = 0, \tag{8}$$

where $k(X^0;(b),\sigma^2)$ is a vector-valued function; e.g., $k(X^0;(b),\sigma^2) = (1, X_0)$. In Masiuk, Shklyar, Kukush, Carroll, Kovgan, and Likhtarov (2016) (see Appendix C there), this method is used for the binary model.

4. Combination of multiplicative Berkson error in regressor and censoring or interval observation of the response

Assume that (5) and (6) are still true. Similarly to Section 3.2, X_n^0 and $\sigma^2_{\epsilon*n}$ are assumed to be known for all observations. However, T_n might be censored, or might be known to belong to an interval.

The distribution of the response T_n as a function of X_n^0 is the following

$$\mathsf{P}(T_n > T^{\text{o.p.}} \mid X_n^0) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-(\beta_0+\beta_1 X_n^0\exp(\sigma_{\epsilon*n}u))T^{\text{o.p.}} - 0.5u^2}\,du;$$

$$\mathsf{P}(T^{\min} < T_n \le T^{\max} \mid X_n^0) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\Big(e^{-(\beta_0+\beta_1 X_n^0\exp(\sigma_{\epsilon*n}u))T^{\min}} - e^{-(\beta_0+\beta_1 X_n^0\exp(\sigma_{\epsilon*n}u))T^{\max}}\Big)e^{-u^2/2}\,du.$$

The likelihood function has the form

$$\mathcal{L}(\boldsymbol{b}) = (2\pi)^{-N/2}\prod_{n=1}^{N}\int_{-\infty}^{\infty}\mathcal{L}_n^{\text{n.e.}}(\boldsymbol{b};\,\exp(\sigma_{\epsilon*n}u)X_n^0)e^{-u^2/2}\,du, \tag{9}$$

where $\mathcal{L}_n^{\text{n.e.}}(\boldsymbol{b};x) = (b_0 + b_1 x)\exp(-(b_0 - b_1 x)T_n)$ if the survival time is observed directly, which is true if $T_n \le T_n^{\text{o.p.}}$ in the model with censoring; $\mathcal{L}_n^{\text{n.e.}}(\boldsymbol{b};x) = \exp(-(b_0 - b_1 x)T_n^{\text{o.p}})$ for right-censored observations, i.e. where $T_n > T_n^{\text{o.p.}}$. In the model with interval observations,

$$\mathcal{L}_n^{\text{n.e.}}(\boldsymbol{b};x) = \exp(-(b_0 + b_1 x)T_n^{\min}) - \exp(-(b_0 + b_1 x)T_n^{\max}).$$

The likelihood function is obtained by substitution of the relevant expression for $\mathcal{L}_n^{\text{n.e.}}(\boldsymbol{b}; x)$ into (9).

5. Simulations

For the model without errors in the explanatory variable (Section 2), we partially use the setup of Shklyar (2014). The explanatory variable is simulated to have a lognormal distribution, $\log X \sim N(0,1)$. The dependent variable has an exponential conditional distribution, $T \mid X \sim \text{Exp}(\beta_0 + \beta_1 X)$, with the true parameters $\beta_0 = 2$ and $\beta_1 = 5$.

We consider the cases of direct observations of T, censoring of T and interval observations of T. For censoring, we consider the censoring times $T^{\text{o.p.}} = 1$ (where about 1% of observations are right-censored) and $T^{\text{o.p.}} = 0.1$ (where about 46% of observations are right-censored). For interval observations, we assume that only the integral part of T (T rounded downwards to the whole number) is used in estimation.

The sample size is $N = 10000$ or $N = 1\,000\,000$. We simulate 1000 samples $\{(X_n, Y_n), n = 1, \ldots, N\}$, with a shared sample of X and conditionally independent (given X's) samples of T. The means and standard deviations of the maximum likelihood estimates over 1000 simulations are given in Table 1. The simulations demonstrate adequate performance of the maximum likelihood estimator in these models for this set of parameters.

Table 1: Means and standard deviations of the maximum likelihood estimates in error-free models

True value:	$\beta_0 = 2$		$\beta_1 = 5$	
Sample size	Means and standard deviations of estimates over 1000 simulations			
	of β_0		of β_1	
N	mean	std	mean	std
with times observed directly				
10 000	2.0017	0.0765	5.0010	0.1020
1 000 000	1.9999	0.0076	4.9998	0.0100
with censoring, $T^{\text{o.p.}} = 1.0$				
10 000	2.0018	0.0781	5.0010	0.1027
1 000 000	1.9999	0.0077	4.9998	0.0101
with censoring, $T^{\text{o.p.}} = 0.1$				
10 000	2.0014	0.1252	4.9998	0.1250
1 000 000	1.9997	0.0123	4.9999	0.0121
with interval observations				
10 000	2.0029	0.1798	5.0463	0.5617
1 000 000	1.9995	0.0179	5.0012	0.0533

For the model with the classical additive error in the explanatory variable and directly observed response variable, the simulations of the sufficiency estimator and the corrected-score estimator are performed in Shklyar (2014). The simulations show that the sufficiency estimator is more stable and more efficient than the corrected score estimator. That result is expected, since the linear conditional estimator and the corrected score estimator were numerically compared by Stefanski (1989) in the Poisson regression model with measurement errors.

In Table 2, the simulation results for the model with Berkson error are presented. The initial regressor X^0 is simulated to have a lognormal distribution, $\log X^0 \sim N(0,1)$. 1000 conditionally independent (given X^0's) samples $\{(X_n, T_n), n = 1, \ldots, N\}$ are simulated, with $X_n = X_n^0 \exp(\epsilon_n^*)$, $\epsilon_n^* \sim N(0, \sigma_{\epsilon*}^2)$, and $[T_n \mid X_n^0, X_n] \sim \text{Exp}(\beta_0 + \beta_1 X_n)$. The parameters used during the simulations are $\beta_0 = 2$, $\beta_1 = 5$ and $\sigma_{\epsilon*}^2 = 0.6$. The statistics for the maximum

likelihood estimator and the estimator obtained from the unbiased estimating equation (8) are presented in Table 2. For comparison, the statistics for the estimator that uses a *partially unobservable* sample of (X, T) is also presented. The estimators perform adequately in the setting used, and the maximum likelihood estimator is slightly more efficient than the estimator obtained by solving an unbiased estimating equation.

Table 2: Means and standard deviations of estimates in the model with Berkson error

True value:		$\beta_0 = 2$		$\beta_1 = 5$	
Sample size	Error variance	Means and standard deviations of estimates over 1000 simulations of β_0		of β_1	
N	$\sigma^2_{\epsilon*}$	mean	std	mean	std
The maximum likelihood estimator that uses (X, T)					
10 000	*N.A.*	2.0022	0.0645	5.0006	0.0878
1 000 000	*N.A.*	1.9993	0.0061	5.0015	0.0089
The maximum likelihood estimator that uses (X^0, T)					
10 000	0.6	1.9595	0.0788	5.0114	0.1296
1 000 000	0.6	1.9595	0.0073	5.0068	0.0123
Unbiased estimation equation method					
10 000	0.6	2.0024	0.0900	5.0022	0.1639
1 000 000	0.6	2.0000	0.0085	4.9999	0.0158

We also simulate study numerically the performance of maximum likelihood estimator in presence of both Berkson error and censoring or interval observations. The sample of (X^0, X, T) is simulated as described above, in the model with multiplicative Berkson error. However, neither X nor T is observed directly. The initial values of X^0 are observed, and T is observed with either censoring or rounding, the same way as in the model without errors. For the censoring level $T^{\text{o·p}} = 1.0$ about 1.4% of observations are right-censored, while for $T^{\text{o·p·}} = 0.1$ about 45.1% of observations are right-censored. In the model with the interval observations, two integers closest to T are observed instead of T.

The results are presented in Table 3. The estimators have significant inaccuracy, as the simulations expose the bias which cannot be explained by insufficient sample size. That is more prominent for the interval observation: the parameter β_0 is 6% overestimated, and β_1 is 15% underestimated.

Simulations for another model, namely for the binary regression model with linear odds ratio, with classical and/or Berkson error, are presented in Masiuk *et al.* (2016, 2017). It was noted (Masiuk *et al.* 2017, p. 206) that the Berkson multiplicative error has an effect on the estimators for geometric standard deviation of the multiplicative error greater than 2, which corresponds to $\sigma^2_{\epsilon*} > 0.5$.

6. Conclusion

We consider a classic generalized linear model and its further developments to cover various forms of errors and incomplete data. The base model is the simple exponential regression where the rate parameter of the response variable linearly depends on the explanatory variable. This model is complicated by adding the censoring of the response variable and/or measurement errors in the explanatory variable.

For the model without measurement errors, the maximum likelihood estimator is constructed for different types of censoring, namely for upper censoring and for interval observations.

For the model with additive classic measurement error in the explanatory variable, the sufficiency estimator, conditional-score estimators and the corrected score estimator are con-

Table 3: Means and standard deviations of the maximum likelihood estimates in the model with Berkson error in the explanatory variable and censoring or interval observations of the response variable

True value:		$\beta_0 = 2$		$\beta_1 = 5$	
Sample size	Error variance	Means and standard deviations of estimates over 1000 simulations			
		of β_0		of β_1	
N	$\sigma^2_{\epsilon*}$	mean	std	mean	std
with censoring, $T^{\text{o.p.}} = 1.0$					
10 000	0.6	1.9668	0.0792	5.0029	0.1347
1 000 000	0.6	1.9636	0.0078	5.0042	0.0129
with censoring, $T^{\text{o.p.}} = 0.1$					
10 000	0.6	2.0739	0.1444	5.0132	0.1612
1 000 000	0.6	2.0708	0.0148	5.0124	0.0154
with interval observations					
10 000	0.6	2.1275	0.1747	4.2434	0.6162
1 000 000	0.6	2.1229	0.0181	4.2135	0.0638

structed in Shklyar (2014) by methods described in Stefanski and Carroll (1987); Stefanski (1989); Nakamura (1990).

In the model with multiplicative Berkson error in the explanatory variable, we construct the maximum likelihood estimator and the unbiased score function estimator by methods used in Masiuk *et al.* (2017, 2016) for another model.

We also construct the maximum likelihood estimator in the model where the censoring of the response variable is combined with Berkson error in the explanatory variable.

The performance of the constructed estimators is verified in simulations. It is adequate in case of one type of uncertainty. In the model with combination of Berkson error in the explanatory variable and censoring of survival time the estimates are inaccurate; the inaccuracy is even more prominent in the model with with the combination of Berkson error and interval observations. Improving numerical integration might help to beat the inaccuracy.

References

Augustin T (2004). "An Exact Corrected Log-Likelihood Function for Cox's Proportional Hazards Model under Measurement Error and Some Extensions." *Scandinavian Journal of Statistics*, **31**(1), 43–50. doi:10.1111/j.1467-9469.2004.00371.x.

Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM (2004). *Measurement Errors in Nonlinear Models*. Chapman and Hall, London. ISBN 978-1-58488-633-4.

Masiuk S, Kukush A, Shklyar S, Chepurny M, Likhtarov I (2017). *Radiation Risk Estimation: Based on Measurement Errors Models*. De Gruyter, Berlin. ISBN 978-3-11-044180-2.

Masiuk SV, Shklyar SV, Kukush AG, Carroll RJ, Kovgan LN, Likhtarov LA (2016). "Estimation of Radiation Risk in Presence of Classical Additive and Berkson Multiplicative Errors in Exposure Doses." *Biostatistics*, **17**(3), 422–436. doi:10.1093/biostatistics/kxv052.

Nakamura T (1990). "Corrected-Score Functions for Error-in-Variables Models: Methodology and Application to Generalized Linear Models." *Biometrika*, **77**(1), 127–137. doi:10.1093/biomet/77.1.127.

Nelder JA, Wedderburn RWM (1972). "Generalized Linear Models." *Journal of the Royal Statistical Society: Series A*, **135**(3), 370–384. doi:10.2307/2344614.

Shklyar S (2014). "Conditional Estimators in Exponential Regression with Errors in Covariates." In V Korolyuk *et al.* (eds.), *Modern Stochastics and Applications*, number 90 in Springer Optimization and Its Applications, pp. 337–349. Springer, Cham. doi:10.1007/978-3-319-03512-3_19.

Stefanski LA (1989). "Unbiased Estimation of Nonlinear Function of a Normal Mean with Application to Measurement Error Model." *Communication in Statistics: Theory and Methods*, **18**(12), 4335–4358. doi:10.1080/03610928908830159.

Stefanski LA, Carroll RJ (1987). "Conditional Scores and Optimal Scores for Generalized Linear Measurement-Error Models." *Biometrika*, **74**(4), 703–716. doi:10.1093/biomet/74.4.703.

Wainwright MJ, Jordan MI (2008). "Graphical Models, Exponential Families, and Variational Inference." *Foundations and Trends in Machine Learning*, **1**(1–2). doi:dx.doi.org/10.1561/2200000001.

Wedderburn RWM (1974). "Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss–Newton method." *Biometrika*, **61**(3), 439–447. doi:10.1093/biomet/61.3.439.

Yang YP, Li GR, Tong TJ (2015). "Corrected Empirical Likelihood for a Class of Generalized Linear Measurement Error Models." *Science China Mathematics*, **58**(7), 1523–1536. doi:10.1007/s11425-015-4976-6.

Affiliation:

Sergiy Shklyar
Department of Probability Theory, Statistics and Actuarial Mathematics
Faculty of Mechanics and Mathematics
Taras Shevchenko National University of Kyiv
Volodymyrska Street 64, Kyiv, Ukraine, 01601
E-mail: shklyar@univ.kiev.ua