Some Model-based Estimator

Danutė Krapavickaitė and Vilma Nekrašaitė

Vilnius, Lithuania

Survey sampling:

Finite population is a fixed, mammoth parameter

Sampler does a dance around it: picks this, omits that, introducing randomness

Survey design and inference – the study of that dance

Main branches of statistics:

Data is a realization of random variables

Statistician makes inference about probability laws of these random variables

Finite population $\mathcal{U} = \{1, 2, \ldots, N\}$

Study variable y: $\{y_1, y_2, \ldots, y_N\}$,

Parameter of a population, for example, total $t_y = \sum_{k=1}^{N} y_k$, fixed, unknown

Sample $\mathbf{i} \subset \mathcal{U}$ size n drawn according to SRS and SRW designs:

$$\widehat{t}_y = \frac{N}{n} \sum_{k \in \mathbf{i}} y_k$$

$$Var_{SRS}(\widehat{t}_y) = N^2 \Big(1 - \frac{n}{N}\Big) \frac{s^2}{n}, \qquad Var_{SRW}(\widehat{t}_y) = N^2 \frac{s^2}{n}$$

They differ much if $\frac{n}{N} = \frac{1}{2}$

Conditionality principle used often in statistical inference:

> statistical inference is made conditionally
> on observed random variables whose probability distribution is known
> and not dependent on parameters about which we must make inferences.

Sampling variance is a property of process of selecting the sample and
not of the outcome of that process.

Something is wrong with this principle in survey sampling!

Nothing is fixed in the world.

What statistical model can describe the process
which has generated the finite population?

Other statisticians are interested in estimating the parameters of the model

For the survey statistician when the model is choosen
estimation of model parameters plays some assistant role for the main task:

to make inference about the realized finite population itself.

Model-based approaches:

- Bayesian prediction techniques (the posterior distribution of t_y is calculated when y_k, $k \in \mathbf{i}$ are given) Bolfarine and zacks, 1992; Ericson, Ghosh, Pfeferman, Royal
- Fiducial inference (palyginamoji analizė), Kalbfleisch and Sprott, 1992
- Likelihood prediction, Bjornstadt, Royall
- Prediction based on general linear models. Valliant, Dorfman, Royall, 2000; Chambers

Randomisation is desirable but not necessary nor sufficient for statistical inference.

Valid inference can be done with and without it, and when randomisation is present, it does ot necessary creates the best probabilistic approach for inference.

Question: wether inferences should be based on distributions created by randomization or on probability models arises when interpreting data from

- designed experiments
- clinical trials
- obsevational studies

Prediction Theory and the General Linear Model

Finite population $\mathcal{U} = \{1, 2, \ldots, N\}$

The population **vector of values of the variable** y: $\mathbf{y} = (y_1, \ldots, y_N)'$,

It is treated as realization of the **random vector** $\mathbf{Y} = (Y_1, \ldots, Y_N)'$

The goal – to estimate a linear combination of $y's$:

$$\boldsymbol{\gamma}'\mathbf{Y}, \qquad \boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_N)' \text{ – vector of constants}$$

Separate cases:

$\boldsymbol{\gamma} = (1, \ldots, 1)' \Longrightarrow \boldsymbol{\gamma}'\mathbf{y} = t_y$ population total

$\boldsymbol{\gamma} = (\frac{1}{N}, \ldots, \frac{1}{N})' \Longrightarrow \boldsymbol{\gamma}'\mathbf{y} = \mu_y$ population mean

Any sample $\mathbf{i} \subset \mathcal{U}$

Let $\mathbf{i}$ consists of the first n population elements. Then:

$$\mathbf{Y} = (\mathbf{Y}_{\mathbf{i}}, \mathbf{Y}_{\mathcal{U}\setminus\mathbf{i}})'$$

$$\mathbf{y} = (\mathbf{y}_{\mathbf{i}}, \mathbf{y}_{\mathcal{U}\setminus\mathbf{i}})'$$

$$\boldsymbol{\gamma} = (\boldsymbol{\gamma}_{\mathbf{i}}, \boldsymbol{\gamma}_{\mathcal{U}\setminus\mathbf{i}})'$$

Estimation target $\quad \boldsymbol{\gamma}'\mathbf{y} = \boldsymbol{\gamma}_{\mathbf{i}}' y_{\mathbf{i}} + \boldsymbol{\gamma}_{\mathcal{U}\setminus\mathbf{i}}' y_{\mathcal{U}\setminus\mathbf{i}}$

is a realization of $\quad \boldsymbol{\gamma}'\mathbf{Y} = \boldsymbol{\gamma}_{\mathbf{i}}'\mathbf{Y}_{\mathbf{i}} + \boldsymbol{\gamma}_{\mathcal{U}\setminus\mathbf{i}}'\mathbf{Y}_{\mathcal{U}\setminus\mathbf{i}}$

Estimation of $\quad \boldsymbol{\gamma}'\mathbf{Y} \iff$ prediction of $\quad \boldsymbol{\gamma}_{\mathcal{U}\setminus\mathbf{i}}'\mathbf{Y}_{\mathcal{U}\setminus\mathbf{i}}$.

The term **"prediction"** is used in the sense of making statistical guess about the values of Y_k which we have not seen – not in the sense of forecasting the future values.

Def. 1. A linear estimator of $\theta = \boldsymbol{\gamma}'\mathbf{Y}$ is defined as $\widehat{\theta} = w_{\mathbf{i}}'\mathbf{Y}_{\mathbf{i}}$,

$w_{\mathbf{i}} = (w_1, \ldots, w_n)'$ – vector of coefficients.

Def. 2. The estimation error of an estimator $\widehat{\theta} = w_{\mathbf{i}}'\mathbf{Y}_{\mathbf{i}}$ is

$$\widehat{\theta} - \theta = w_{\mathbf{i}}'\mathbf{Y}_{\mathbf{i}}' - \boldsymbol{\gamma}'\mathbf{Y}.$$

Def. 3. General linear model M for $\mathbf{Y}$:

$$\mathbf{E}_M(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}, \quad (1)$$
$$Var_M(\mathbf{Y}) = \mathbf{V}, \quad (2)$$

$\mathbf{X} = (\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(J)})$ – matrix of auxiliary variables,

$\boldsymbol{\beta}$ – vector of unknown parameters,

$\mathbf{V}$ – positive definite variance-covariance matrix of $\mathbf{Y}$.

Def. 4. The estimator $\widehat{\theta}$ is unbiased for θ under a model M if $\mathbf{E}_M(\widehat{\theta} - \theta) = 0$.

Def. 5. The error variance of $\widehat{\theta}$ under a model M is $Var_M(\widehat{\theta}) = \mathbf{E}_M(\widehat{\theta} - \theta)^2$.

Def. 6. A linear model unbiased estimator $\widehat{\theta}$ of θ which minimizes the error variance $Var_M(\widehat{\theta})$ is called best linear unbiased predictor (BLU) of θ.

It's expression and error variance is known.

When $\mathbf{X}$ are quantitative variables the estimator $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ which gives BLU predictor of θ, is the least squares estimator.

Let $\boldsymbol{\gamma} = (1, \ldots, 1)'$: $\qquad \boldsymbol{\gamma}'\mathbf{y} = \boldsymbol{\gamma}' y_{\mathbf{i}} + \boldsymbol{\gamma}'_{\mathcal{U}\setminus\mathbf{i}} y_{\mathcal{U}\setminus\mathbf{i}}.$

is equivavlent to

$$t_y = \sum_{k\in\mathbf{i}} y_k + \sum_{k\in\mathcal{U}\setminus\mathbf{i}} y_k.$$

Estimating $\boldsymbol{\gamma}'\mathbf{y} = t_y \iff$ predicting value $\sum_{k\in\mathcal{U}\setminus\mathbf{i}} y_k$ of unobserved random variable $\sum_{k\in\mathcal{U}\setminus\mathbf{i}} Y_k$

Taking various matrices of auxiliary variables $\mathbf{X}$, matrices $\mathbf{V}$ and vectoes $\boldsymbol{\beta}$ we obtain BLU predictors coinciding with the estimators of totals known in the design-based theory:

expansion estimator,

linear regression estimator,

stratified expansion estimator,

.....

Valliant, R, Dorfman A. H., and Royall R. M. (2000) *Finite Population sampling and Inference. A prediction Approach*. John Wiley & Sons

Prediction theory under the nonlinear model

R. Valliant, *JASA*, 1985

Nonlinear model:

$$\begin{aligned} \mathbf{E}_M(Y_k) &= f(\mathbf{X}_k; \boldsymbol{\psi}), \quad k = 1, \ldots, N && (3) \\ Var_M(\mathbf{Y}) &= \mathbf{V}, && (4) \end{aligned}$$

$\boldsymbol{\psi}$ – vector of J unknown coefficients, model parameter,

$f(\cdot; \cdot)$ – nonlinear function, has at least 3 partial derivatives with respect to $\boldsymbol{\psi}$.

Remark: if model is such that

$Y_1, \ldots, Y_N$ are independent random variables $\Longrightarrow$

$\mathbf{V}$ is diagonal $\Longrightarrow$

$Var_M(Y_k)$ is only needed for it in (4).

Population total

$$t_y = \sum_{k=1}^{N} y_k = \sum_{k \in \mathbf{i}} y_k + \sum_{k \in \mathcal{U} \setminus \mathbf{i}} y_k.$$

When $\boldsymbol{\psi}$ is known, the BLU estimator of t_y is obtained by adding $\sum_{k \in \mathbf{i}} y_k$ with BLU predictor of $\sum_{k \in \mathcal{U} \setminus \mathbf{i}} y_k$.

When $\boldsymbol{\psi}$ is not known – its estimator is inserted into BLU predictor of $\sum_{k \in \mathcal{U} \setminus \mathbf{i}} y_k$.

$\widehat{t}_y$ becomes biased.

Valliant: ***If asymptotically normal and consistent estimator of the model parameters $\boldsymbol{\psi}$ is available***

1) Estimator $\widehat{t}_y$ is consistent

2) Expressions for

$$\begin{aligned} ABias(\widehat{t}_y) &\approx \mathbf{E}_M(\widehat{t}_y - t_y), \\ AVar(\widehat{t}_y) &\approx Var_M(\widehat{t}_y - t_y). \end{aligned}$$

are obtained

3) Consistent estimator of $Var_M(\widehat{t}_y)$ constructed through the consistent estimator of $\boldsymbol{\psi}$.

Models for study variable with positive values and many zeroes

Let us consider a study variable y, values of which have the properies:

$$y_k \begin{cases} = 0 \text{ or} \\ > 0, \quad k = 1, ..., N \end{cases} \quad \Rightarrow Var(y) \quad \text{large} \Rightarrow Var(\widehat{t_y}) \quad \text{large.}$$

Examples:

Area under some kind of crop in the farm, which is not grown up often (rape),

Investment of enterprise for an environmental protection

Design-based estimators are not useful

Model-assisted estimators are not useful – there are no auxiliary variables correlated with the study variable because zero values can not be known in advance.

Model-based estimator

$\mathcal{U}^*$ – superpopulation or a model for y, $\mathcal{U}$ – its ***random*** realization,

y_k, $k = 1, 2, \ldots, N$, t_y – random variables

$$t_y = \sum_{k \in \mathbf{i}} y_k + \sum_{k \in \mathcal{U} \setminus \mathbf{i}} y_k, \qquad \widehat{t}_y = \sum_{k \in \mathbf{i}} y_k + \sum_{k \in \mathcal{U} \setminus \mathbf{i}} \widehat{y}_k$$

Nonlinear prediction $\widehat{y}_k =$?

1. F. Kalberg, *Journal of Official Statistics*, 2000

$$Y_k = \xi_k \tilde{Y}_k, \quad \tilde{Y}_k = \mathbf{x}_k' \beta + \varepsilon_k, \quad \varepsilon_k \sim \mathcal{N}(0, \sigma^2)$$

$$\xi_k = \begin{cases} 1, & p_k, \\ 0, & 1 - p_k, \end{cases} \Rightarrow \quad \widehat{\beta},\ \widehat{\sigma},\ \widehat{p}_k \quad \Rightarrow \quad \widehat{y}_k = \widehat{\mathbf{E}_M Y_k}$$

Econometric models

2. *Censored regression or Tobit* model for a study variable y.

Let there are unobserved random variables

$$Y_k^* = \mathbf{x}_k'\boldsymbol{\beta} + \varepsilon_k, \quad \varepsilon_k \sim \mathcal{N}(0, \sigma^2), \quad k = 1, 2, \ldots, N$$

ε_k, ε_l independent for $k \neq l$.

Define random variables

$$Y_k = \begin{cases} Y_k^*, & \text{if} \quad Y_k^* > 0, \\ 0, & \text{if} \quad Y_k^* \leq 0, \end{cases} \quad k = 1, 2, ..., N$$

W.H.Greene. Econometric Analysis. Prentice Hall, Upper Saddle River, 2003.

$$\mathbf{E}_M(Y_k) = f(\mathbf{x}_k; \boldsymbol{\psi}) = \mathbf{x}_k'\beta\Phi\Big(\frac{\mathbf{x}_k'\beta}{\sigma}\Big) + \sigma\phi\Big(\frac{\mathbf{x}_k'\beta}{\sigma}\Big), \qquad \Phi,\ \phi \sim \mathcal{N}(0, 1)$$

If $\mathbf{x}_k = (1, x_k)'$:

$$\boldsymbol{\psi} = (\sigma, \alpha_0, \alpha_1), \quad \alpha_0 = \beta_0/\sigma, \quad \alpha_1 = \beta_1/\sigma.$$

$$Var_M(Y_k) = \sigma^2\Phi(\alpha_0 + \alpha_1 x_k)g_k(\alpha_0, \alpha_1), \quad k = 1, \ldots, N$$

T. Amemiya, *Econometrica*, 1973.

Maximum likelihood estimators $\widehat{\sigma}$, $\widehat{\alpha}_0$, $\widehat{\alpha}_1$ of tobit model parameters σ, α_0, α_1 are consistent and asymptotically normal.

We estimate

$$\widehat{y}_k^{(tobit)} = \widehat{\mathbf{E}_M Y}_k = \mathbf{x}_k' \widehat{\beta} \Phi\Big(\frac{\mathbf{x}_k' \widehat{\beta}}{\widehat{\sigma}}\Big) + \sigma \phi\Big(\frac{\mathbf{x}_k' \widehat{\beta}}{\widehat{\sigma}}\Big),$$

$$\widehat{t}_y^{(tobit)} = \sum_{k \in \mathbf{i}} y_k + \sum_{k \in \mathcal{U} \setminus \mathbf{i}} \widehat{y}_k^{(tobit)}$$

Using results of Valliant we obtain expressions for

$$ABias(\widehat{t}_y^{(tobit)})$$

$$AVar(\widehat{t}_y^{(tobit)})$$

$$\widehat{Var}(\widehat{t}_y^{(tobit)})$$

SAS procedure LIFEREG: $\widehat{\beta}$, $\widehat{\sigma}$.

The decisions can be separated:

to have positive values of Y_k or not to have?

how big will be Y_k if "yes"?

3. Censored regression model with the unobserved stochastic threshold (Heckman model)

Maddala G. S. Limited-dependent and qualitative variables in econometrics. Cambridge University Press, 1983.

Two unobserved random variables:

$$Y_k^* = \mathbf{x}_k^{(1)\prime}\boldsymbol{\beta}^{(1)} + \varepsilon_k, \tag{5}$$

$$Z_k^* = \mathbf{x}_k^{(2)\prime}\boldsymbol{\beta}^{(2)} + u_k, \tag{6}$$

$$\begin{pmatrix}\varepsilon_k \\ u_k\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix}\sigma_\varepsilon^2 & \sigma_{\varepsilon u} \\ \sigma_{\varepsilon u} & 1\end{pmatrix}\right)$$

$$I_k = \begin{cases}1, & \text{if } Z_k^* > 0, \\ 0 & \text{otherwise.}\end{cases} \qquad Y_k = \begin{cases}Y_k^\star, & \text{if } I_k = 1, \\ 0 & \text{otherwise.}\end{cases}$$

If both equations in (5) coinside $\Longrightarrow$ Heckman $\equiv$ Tobit.

$$\mathbf{E}_M(Y_k) = f\left(\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}; \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}_k^{(2)}\right) = \Phi\left(\mathbf{x}_k^{(2)\prime}\boldsymbol{\beta}^{(2)}\right)\left(\mathbf{x}_k^{(1)\prime}\boldsymbol{\beta}^{(1)} + \sigma_{\varepsilon u}\frac{\phi(\mathbf{x}_k^{(2)\prime}\boldsymbol{\beta}^{(2)})}{\Phi(\mathbf{x}_k^{(2)\prime}\boldsymbol{\beta}^{(2)})}\right)$$

Estimation of the model parameters

$$\widehat{y}_k^{(Heckman)} = \widehat{\mathbf{E}_M}(Y_K), \quad \widehat{t}_y^{(Heckman)} = \sum_{k\in\mathbf{i}} y_k + \sum_{k\in\mathcal{U}\setminus\mathbf{i}} \widehat{y}_k^{(Heckman)}$$

Other generalizations of nonlinear model (3), (4) when residuals are not normaly distributed – skew distributions in economics.

Simulation

$N = 1\,000$ size finite opulation generated

$\mathbf{x}_k = (1, x_k)', \quad x_k \sim \mathcal{N}(\mu, q^2), \quad \varepsilon_k \sim \mathcal{N}(0, \sigma^2)$

$$Y_k^* = \beta_0 + \beta_1 x_k + \varepsilon_k \Rightarrow Y_k$$

$M = 1\,000$ replicates:

Simple random sample of size n

Estimates: ***design-based*** and ***tobit model-based***

$$\widehat{t}_y^{(SRS)} = \frac{N}{n} \sum_{k \in \mathbf{i}} y_k \qquad \widehat{t}_y^{(tobit)} = \sum_{k \in \mathbf{i}} y_k + \sum_{k \in \mathcal{U} \setminus \mathbf{i}} \widehat{y}_k^{(tobit)}$$

$$\widehat{y}_k^{(tobit)} = f(x_k; \widehat{\sigma}, \widehat{\beta}_0, \widehat{\beta}_1), \quad k \in \mathcal{U} \setminus \mathbf{i}$$

Notations: $\widehat{\eta}_m$ and $\widehat{Var}(\widehat{\eta}_m)$ – replicates of the estimates $\widehat{\eta}$:

$$\widehat{\eta} = \widehat{t}_y^{(SRS)} \quad \text{or} \quad \widehat{\eta} = \widehat{t}_y^{(tobit)}$$

$$\bar{\widehat{\eta}} = \frac{1}{M}\sum_{m=1}^{M} \widehat{\eta}_m,$$

$$\widehat{Bias}_{emp}(\widehat{\eta}) = \bar{\widehat{\eta}} - t, \quad \overline{\widehat{Bias}}(\widehat{\eta}) = \begin{cases} \frac{1}{M}\sum_{m=1}^{M} \widehat{Bias}(\widehat{\eta}_m) & \text{for } \widehat{\eta} = \widehat{t}^{(tobit)}, \\ 0 & \text{for } \widehat{\eta} = \widehat{t}^{(SRS)}, \end{cases}$$

$$\widehat{Var}_{emp}(\widehat{\eta}) = \frac{1}{M-1}\sum_{m=1}^{M} (\widehat{\eta}_m - \bar{\widehat{\eta}})^2, \qquad \overline{\widehat{Var}}(\widehat{\eta}) = \frac{1}{M}\sum_{m=1}^{M} \widehat{Var}(\widehat{\eta}_m),$$

$$\widehat{MSE}_{emp}(\widehat{\eta}) = \widehat{Bias}_{emp}(\widehat{\eta})^2 + \widehat{Var}_{emp}(\widehat{\eta}) \qquad \overline{\widehat{MSE}}(\widehat{\eta}) = \overline{\widehat{Bias}}(\widehat{\eta})^2 + \overline{\widehat{Var}}(\widehat{\eta});$$

$$RMSE_{emp}(\widehat{\eta}) = \frac{\sqrt{\overline{\widehat{MSE}}_{emp}(\widehat{\eta})}}{t}, \qquad RMSE(\widehat{\eta}) = \frac{\sqrt{\overline{\widehat{MSE}}(\widehat{\eta})}}{t}.$$

Investigated:

Dependence of $RMSE$ on n, σ, percentage of zeroes in the population.

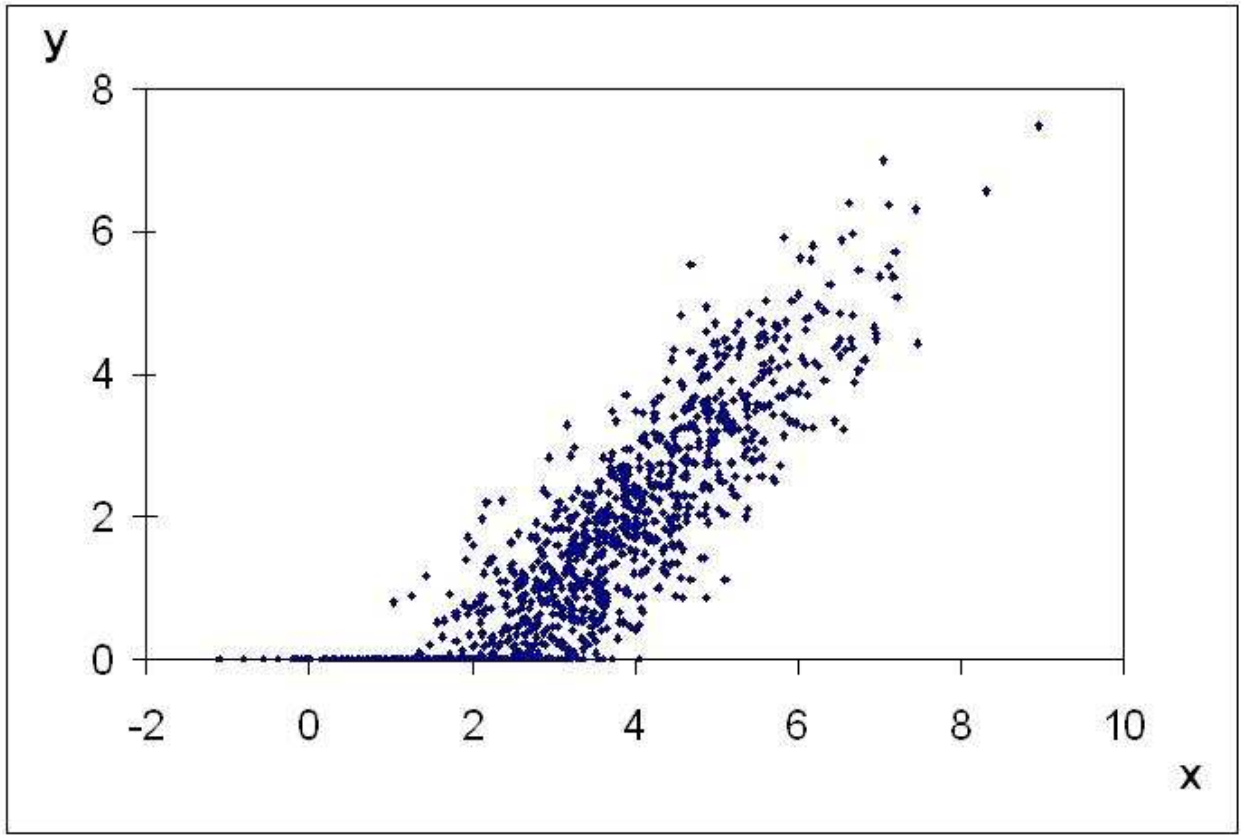

Fig. 1. Population

$\mathbf{x}_k = (1, x_k)', \quad x_k \sim \mathcal{N}(3.49, 1.56), \quad \varepsilon_k \sim \mathcal{N}(0, 0.8), \quad \boldsymbol{\beta} = (-2.42, 1.1), \quad 20\%$ zeroes

$$Y_k^* = -2.42 + 1.1 x_k + \varepsilon_k \quad \Rightarrow \quad Y_k$$

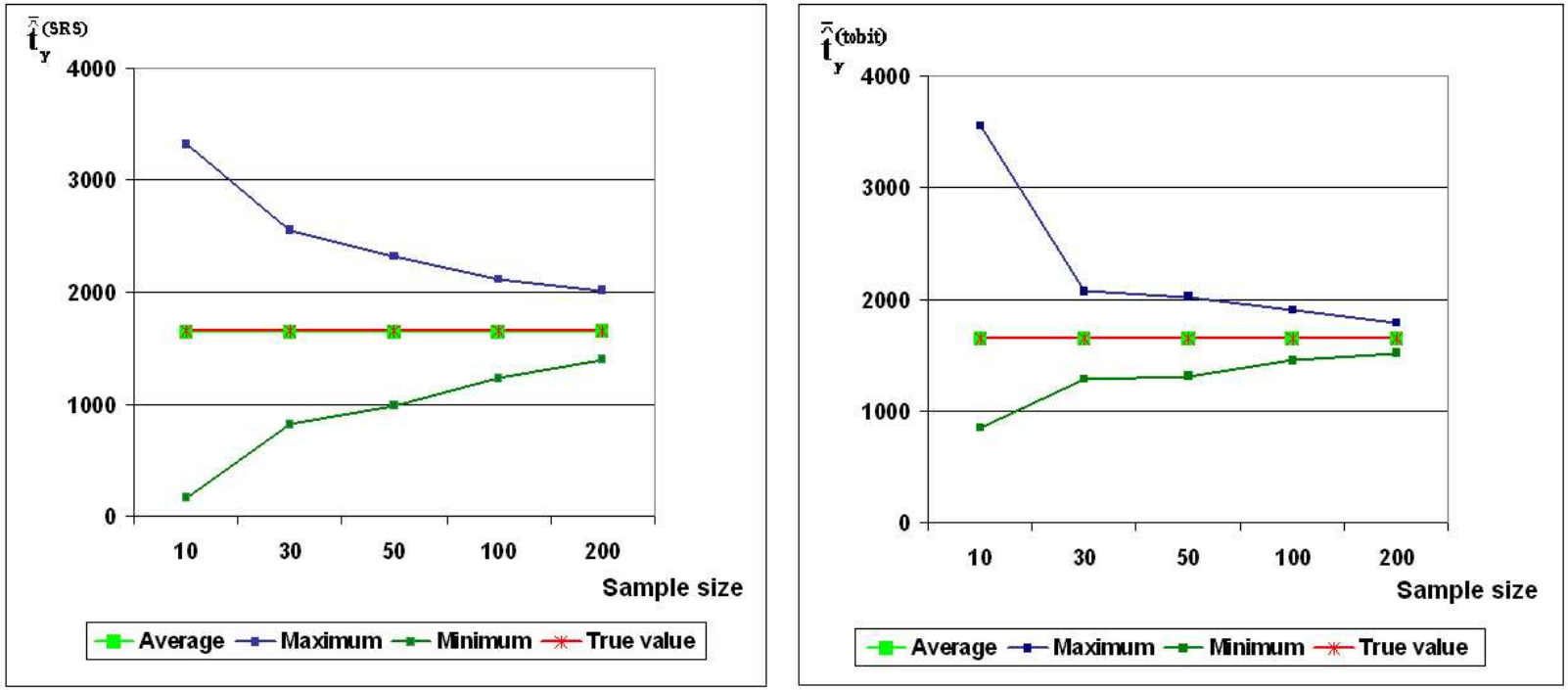

Fig. 2. Dependence oft $\bar{\hat{t}}_y^{(SRS)}$ (left) and $\bar{\hat{t}}_y^{(tobit)}$ (right) on the sample size n

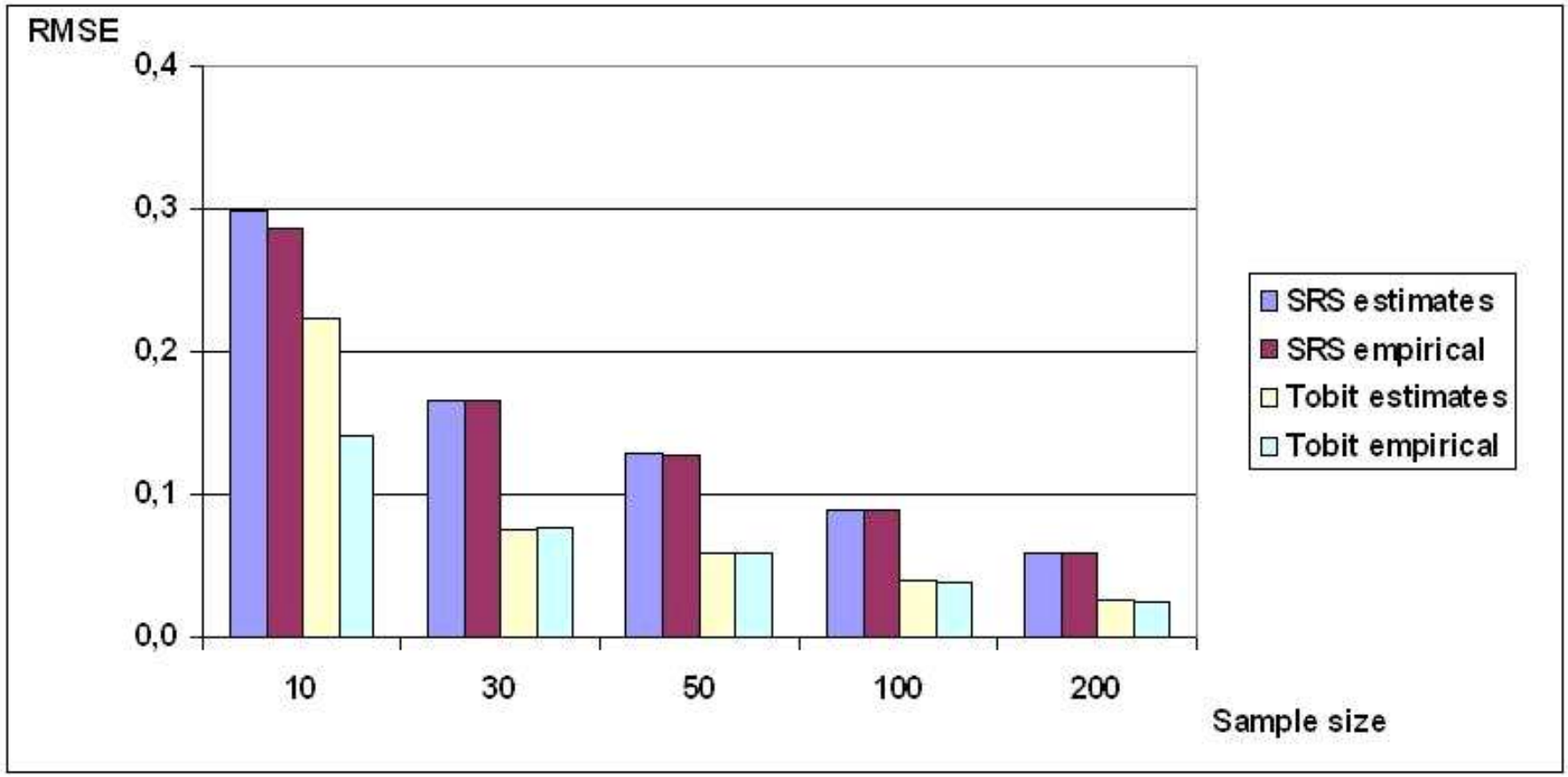

Fig. 3. Dependence of $RMSE(\widehat{t_y})$ on the sample size n

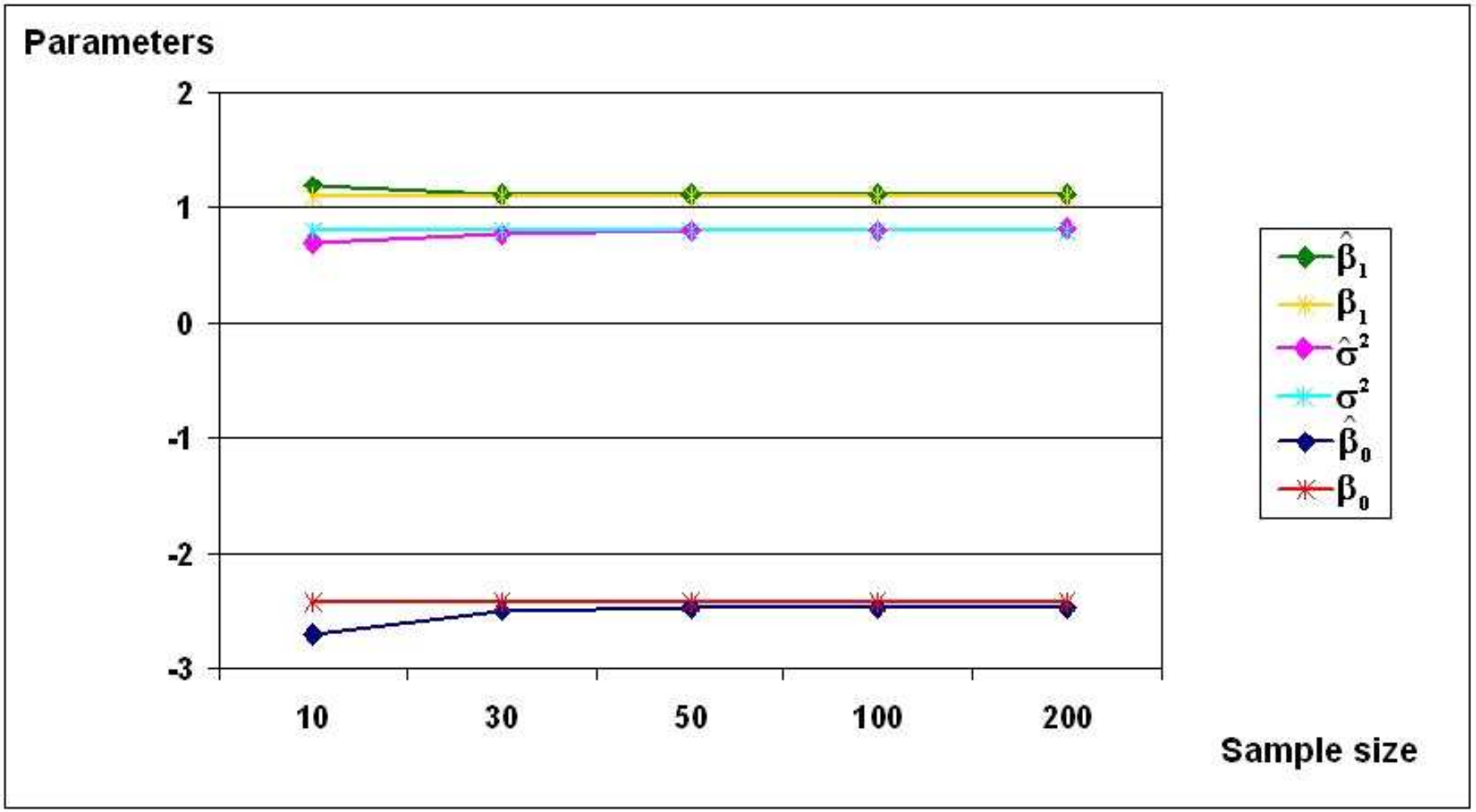

Fig. 4. Dependence of $\widehat{\sigma}^2$, $\widehat{\beta}_0$, $\widehat{\beta}_1$ on the sample size n

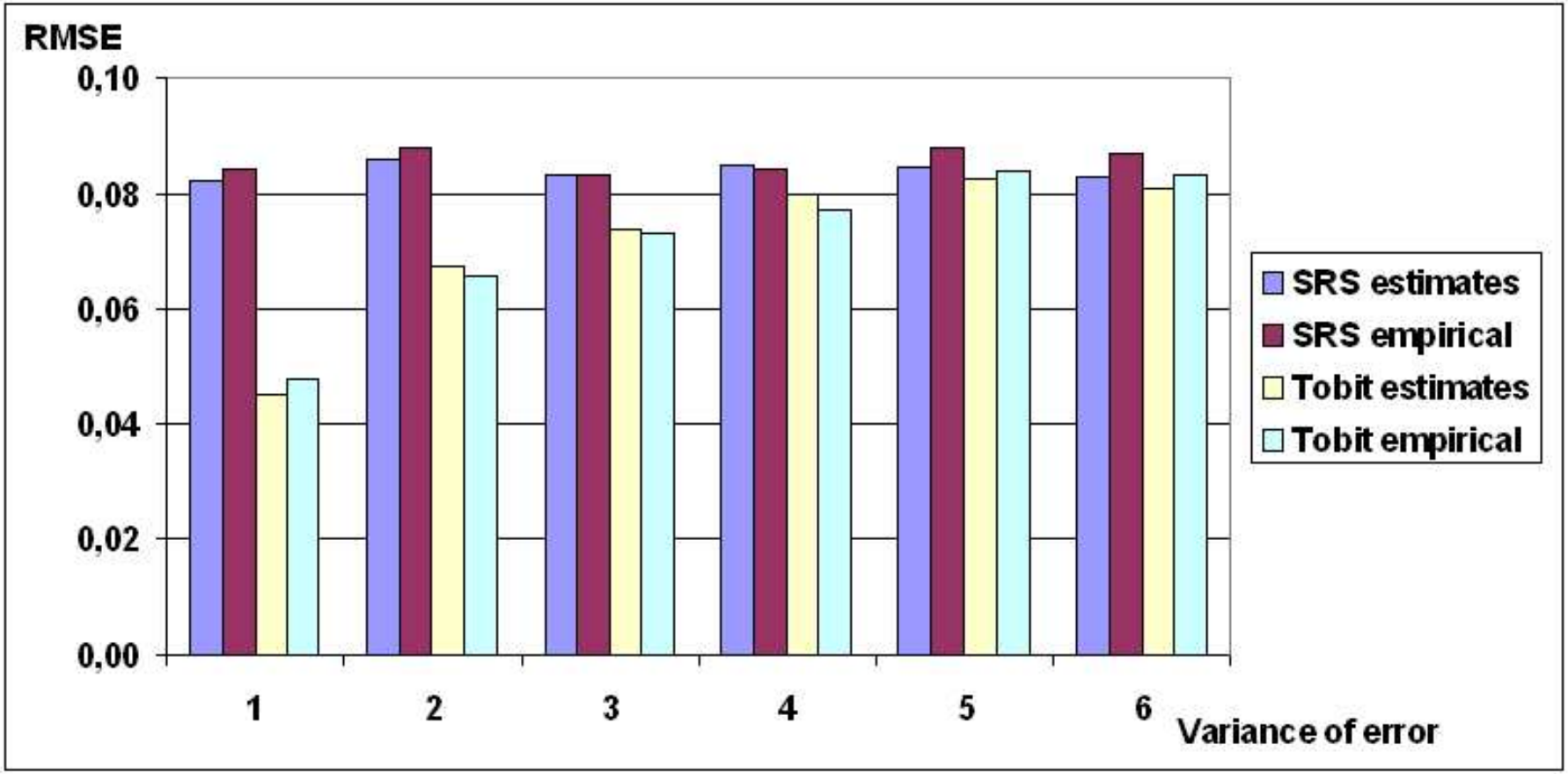

Fig. 5. Dependence of $RMSE(\widehat{t_y})$ on the variance of error σ ($\sigma \uparrow \Rightarrow \beta_0 \uparrow$)

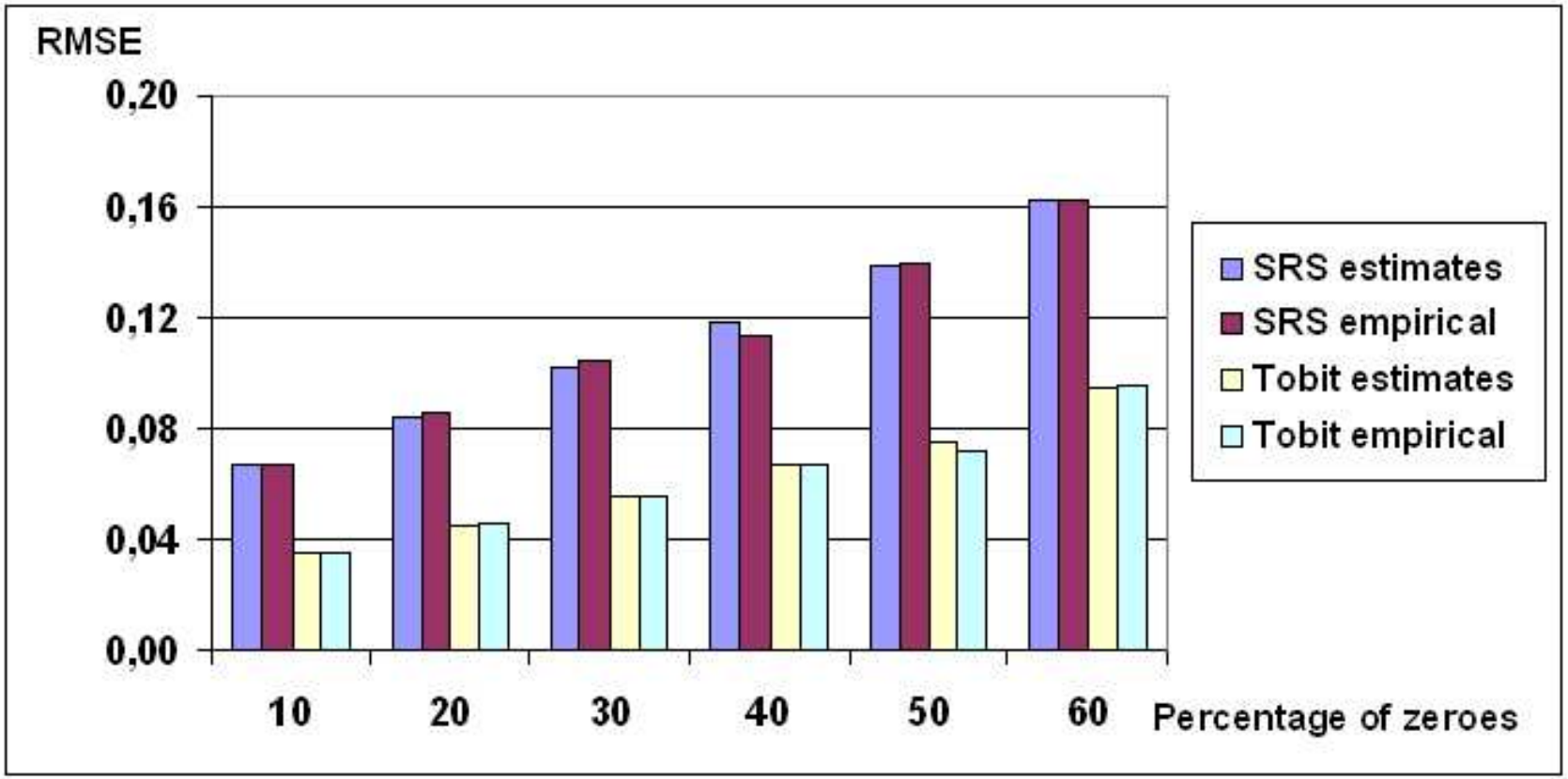

Fig. 6. Dependence of $RMSE(\widehat{t_y})$ on the percentage of zeros in the population

(zeroes $\uparrow \Rightarrow \mu \downarrow$)

Fig. 7. The case when $RMSE(\hat{t}_y^{(tobit)}) > RMSE(\hat{t}_y^{SRS})$

$\mathbf{x}_k = (1, x_k)', \quad x_k \sim \mathcal{N}(1,2), \quad \varepsilon_k \sim \mathcal{N}(0,6), \quad \boldsymbol{\beta} = (0.5, 0.3), \quad 45\%$ zeroes

$$y_k^* = 0.5 + 0.3x_k + \varepsilon_k \quad \Rightarrow \quad y_k$$

Conclusions

For small sample size ($n = 10$) average $RMSE$ and empirical $RMSE$ of model-based estimator can differ much.

The efficiency of the model-based estimator does not show here any dependence on the sample size or percentage of zeroes in the population.

The efficiency of the model-based estimator depends on the variance of the error σ.

Acknowledgements

Support: Nordic Council of Ministers

Library facilities: Umeå university

Thank you for your attention!