RESAMPLING MEDIAN ESTIMATORS FOR LINEAR REGRESSION MODEL

Helen Afanasyeva

Transport and Telecommunication Institute,
Lomonosov Str.1, LV-1019, Riga, Latvia
E-mail: jelena_a@tsi.lv

1. INTRODUCTION

As it is known, that linear regression model is one of the most popular statistical models. It has the following form:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\,, \tag{1}$$

where $\mathbf{X}$ is the $n\times m$ matrix of independent variables, n is the number of observations and m is the number of independent variables, $\mathbf{Y}$ is the $n\times 1$ vector of dependent variables, $\mathbf{Z}$ is the $n\times 1$ vector whose components Z_1, Z_2, ..., Z_n are independent, identically distributed random variables with mean zero and variance σ^2, so $Cov(\mathbf{Z}) = \sigma^2\mathbf{I}_n$, and $\boldsymbol{\beta}$ is the $m\times 1$ vector of parameters of the regression model to be estimated.

The classical least square estimator (LSE) of parameter $\boldsymbol{\beta}$ is well known**:**

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}\,. \tag{2}$$

With this estimator (2) in hand we can predict the value of dependent variable Y for some selected previous or future observation $\mathbf{x_d}$ by the following formula:

$$\hat{Y} = \mathbf{x}_d \cdot \hat{\boldsymbol{\beta}}\,.$$

where $\mathbf{x}_d$ is the $1\times m$ vector of the independent variables values for selected observation d.

The estimator (2) can be successfully used if there are no nuisance observations in the sample, otherwise the quality of obtained estimator is not very good, it may be biased. To improve the quality of obtained estimator, we use resampling median approach instead of classical one. In our paper we investigate suggested method efficiency, taking the bias of the estimators as efficiency criterion.

As it is supposed by intensive computer methods [3,4], we use the given data a lot of times in various combinations. This procedure can give us estimators with less bias because the observed sample contains all the available information about the underlying population, and the observed sample can be considered to be the general population. Hence, we can simulate the distribution of any relevant test statistic by using random samples from the "population" consisting of the original sample [4]. We will describe this approach in details in the second section of the paper. The application of the resampling procedure gives us the sequence of resampling estimators. Some combinations can be used as the estimators of the regression model. We will describe this technique in the third section of the paper. As it was mentioned above, the classical approach gives bad, biased estimators of parameters in the case of disturbed model. We will give a formal definition of the disturbed model in the forth section of the paper. The numerical example concludes the paper.

2. RESAMPLING METHOD

Resampling methods were widely used to solve various kinds of problems for example in Efron and Tibshirani [3] or Andronov and Merkuryev [1], where they showed considerably good results. We would like to implement resampling approach to estimate disturbed regression model.

In order to calculate the estimator $\boldsymbol{\beta}^*$ of the model parameter $\boldsymbol{\beta}$, we propose the following resampling procedure. At the step number l, l=1,…,r, to form each resample, we extract, without replacement, $k \geq m$ observations from n, and calculate the least square estimator $\boldsymbol{\beta}^*(l)$ by (2) using only the extracted elements. We repeat this procedure r times, and obtain the sequence of the resample-estimators

$$\boldsymbol{\beta}^*(1), \boldsymbol{\beta}^*(2), \ldots, \boldsymbol{\beta}^*(r). \quad (3)$$

We can use various combinations of $\boldsymbol{\beta}^*(i)$ as resampling estimators of parameter of interest $\boldsymbol{\beta}$. It can be, for example, an average of $\boldsymbol{\beta}^*(i)$[7]:

$$\boldsymbol{\beta}^* = \frac{1}{r}\sum_{i=1}^{r}\boldsymbol{\beta}^*(i). \quad (4)$$

In some situations it is better to use other estimators instead of average ones (4), for example, the median of $\boldsymbol{\beta}^*(i)$. This estimator is described in the following section.

3. MEDIAN ESTIMATORS

In general case we can describe the method of obtaining the expectation θ=E(T) of random variable T in the following way. Let T_1, T_2, … , T_r be a sample of the random variable T realizations. Well-known unbiased estimator of θ is an average of the sample elements. Alternatively, we can use another estimator of θ, for example, median of the sample elements T_i. We order the sample elements T_i in increasing order of magnitude $T_{(1)}$, $T_{(2)}$, … ,$T_{(s+1)}$, ... , $T_{(r)}$, where r=2s+1 and take the middle value $T_{(s+1)}$ of this sequence as the median estimator of θ.

Let us discuss the procedure of obtaining the resampling median estimator for the regression model. We cannot use the just described median approach directly for the sequence (3), because its elements are vectors and it is not possible to order them naturally. That's why we will employ this approach to the estimation of the predictors. We consider the situation, when we wish to get predictor of dependent variable for some existing or future time moment. Using each of the resample-estimators (3) we obtain the sequence of resample-estimators for predicted value of Y: $Y^*(l)=\mathbf{x_d}\cdot\boldsymbol{\beta}^*(l)$, l=1,…,r, where $\mathbf{x_d}$ – the row corresponding to the selected previous or future observation with number d**.** So we obtain the sequence of resample-estimators $Y^*(1), Y^*(2), \ldots, Y^*(r)$ for predicted value of Y**.** Then we order them in increasing order of magnitude $Y_{d(1)}^*$, $Y_{d(2)}^*$, ... , $Y_{d(s+1)}^*$, …, $Y_{d(r)}^*$, where r=2s+1 and the resample-estimator, that caused the middle value $Y_{d(s+1)}^*$ in the ordered sequence, will have the name *resampling median estimator* of regression parameters.

Despite the fact that the average is the unbiased estimator of expectation, and the median estimator is biased, there are some advantages of using median estimator of $\boldsymbol{\beta}$ instead of the average estimator (4). The first aspect is that unbiasness as criterion has some disadvantages; so unbiased estimators may have "incorrect" sign, big variance and bad results in case of "noisy" data (our case). The second one is that if we take $E(\theta^*-\theta)^2$ as efficiency criterion, the biased estimator may be better. The last one is that, median estimators have the following advantage: they are median unbiased robust estimators, when we deflect from statistical hypothesis.

4. DISTURBED LINEAR REGRESSION MODEL

We will apply our method to the regression model with nuisance observations, called disturbed model. In this model, the nuisance observations among all observations can be present. So the disturbed model can be described in the following way. Let us have n observations that correspond to the rows of matrix $\mathbf{X}$ in our model. Let n-h observations from them be "true", which means that we can describe them by our regression model (1), but h observations be "false", which means that they belong to other regression model. Without loss of generality we suppose that the true observations correspond to the first n – h rows of $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ and will be denoted as $\mathbf{X_t}$, $\mathbf{Y_t}$ and $\mathbf{Z_t}$ correspondingly, but false observations correspond to the last h rows of $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ and will be denoted as $\mathbf{X_f}$, $\mathbf{Y_f}$ and $\mathbf{Z_f}$ correspondingly. Then the regression model (1) can be represented in the following form:

$$\mathbf{Y}=\begin{bmatrix}\mathbf{Y}_t\\ \mathbf{Y}_f\end{bmatrix},\quad \mathbf{X}=\begin{bmatrix}\mathbf{X}_t\\ \mathbf{X}_f\end{bmatrix},\quad \mathbf{Z}=\begin{bmatrix}\mathbf{Z}_t\\ \mathbf{Z}_f\end{bmatrix},$$

where $\mathbf{Y}_t=(Y_1\ Y_2\ \dots\ Y_{n-h})^T$, $\mathbf{Y}_f=(Y_{n-h+1}\ Y_{n-h+2}\ \dots\ Y_n)^T$, $\mathbf{X}_t=(\mathbf{x}_1^T\ \mathbf{x}_2^T\ \dots\ \mathbf{x}_{n-h}^T)^T$,
$\mathbf{X}_f=(\mathbf{x}_{n-h+1}^T\ \mathbf{x}_{n-h+2}^T\ \dots\ \mathbf{x}_n^T)^T$, $\mathbf{Z}_t=(Z_1\ Z_2\ \dots\ Z_{n-h})^T$, $\mathbf{Z}_f=(Z_{n-h+1}\ Z_{n-h+2}\ \dots\ Z_n)^T$,

$$\mathbf{Y}_t=\mathbf{X}_t\boldsymbol{\beta}+\mathbf{Z}_t,\qquad \mathbf{Y}_f=\mathbf{X}_f\boldsymbol{\beta}_f+\mathbf{Z}_f,$$

$$E(\mathbf{Z}_t)=\mathbf{0},\quad Cov(\mathbf{Z}_t)=\sigma^2\mathbf{I}_{n-h},\quad E(\mathbf{Z}_f)=\mathbf{0},\quad Cov(Z_f)=\sigma^2\mathbf{I}_h,\quad Cov(\mathbf{Z}_t,\mathbf{Z}_f)=\mathbf{0}$$

Here $\boldsymbol{\beta}_f$ – is the regression parameter of the model with only false observations. The model of this kind will be considered in the numerical example. It will be presented in the next section.

Usually we don't know if we there are false observations in our sample. We will implement two described methods to investigate witch gives more robust estimators of the regression model.

5. NUMERICAL EXAMPLE

In order to illustrate the efficiency of the proposed approach, let us consider a numerical example. The data are taken from the book of Draper and Smith [2]. The model contains n=9 observations and m=3 independent variables. The data are illustrated at the figure 1, where 3 axes correspond to the independent variables x_1, x_2 and x_3. The points at the histogram identify the observations. The observations with numbers 2, 6 and 9 are the most possible candidates to be nuisance observations, because they are located too far from the center. But this distance from the center, comparing with others, is not so significant to call them nuisance observations. We also calculated the Mahalonobis square distance [5] that shows the distance between each observation and the center:

$$D_i=(\mathbf{x}_i-\bar{\mathbf{x}})\mathbf{S}^{-1}(\mathbf{x}_i-\bar{\mathbf{x}})', \tag{5}$$

where $\mathbf{x}_1,\dots,\mathbf{x}_n$ is the sample of m-dimensional vectors (observations), and it corresponds to the rows of matrix **X,** $\bar{\mathbf{x}}$ **-** the vector of means of each column of matrix **X, S -** the sample covariance matrix.

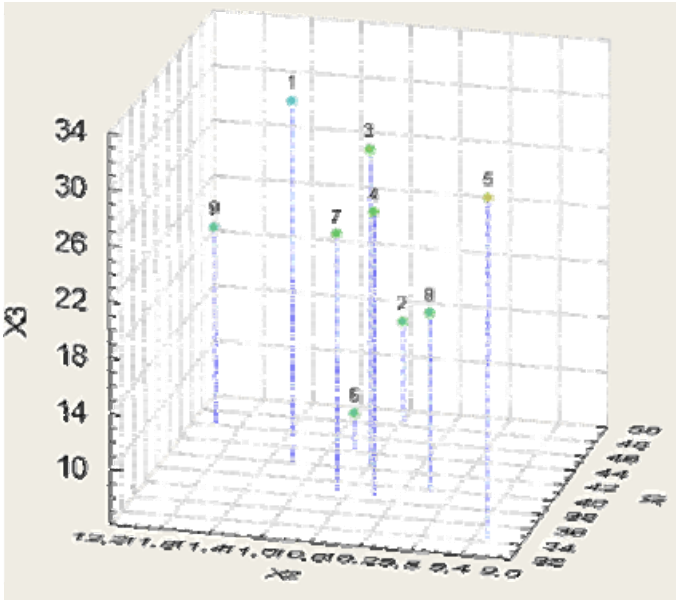

Figure 1. *The histogram of the observations*

The Mahalonobis distance is superior to Euclidean distance, because it takes distribution of the points (correlation) into account. For our example we obtained the following vector of Mahalonobis distances for each observation, where element's index in the vector corresponds to the number of observation: **D**=(2.62 5.281 1.95 0.524 3.686 3.695 1.379 1.205 3.659). Analyzing those distances we can come to the same conclusion, that observations with numbers 2, 6, 5 and 9 are the most probable candidates to be nuisance observations, because of the big distance. We will suppose that the number of false observations in our model is equal to one, i.e. h=1.

Table 1. The bias of resampling median estimators for the example

Param. title	**The bias of LSE**	**The bias of resampling median estimators of predictors**					
		k=3	k=4	k=5	k=6	k=7	k=8
Y_1	0.115	0.617	0.43	0.332	0.139	**0.034**	0.333
Y_2	5.793	7.57	7.441	5.802	**4.85**	**5.037**	**5.631**
Y_3	1.957	5.131	3.165	2.301	**1.702**	**1.682**	1.967
Y_4	0.497	2.571	**0.108**	**0.243**	0.545	0.739	0.772
Y_5	0.129	1.012	0.636	0.282	**0.024**	**0.042**	**0.066**
Y_6	1.481	4.385	3.53	**1.242**	**0.523**	**0.517**	**1.246**
Y_7	1.519	**0.416**	**0,681**	**0.959**	**1.204**	1.572	1.935
Y_8	1.199	1.557	1.523	**0.111**	**0.073**	**0.539**	**0.749**
Y_9	0.499	2.857	0.956	**0.261**	**0.493**	0.543	0.778

For considering model with one false observation we take the observation number 2, because it is the most possible candidate in accordance with its distance from the center. We want to obtain the predicted value for selected future or past observation. Then according to the resampling approach described in the section 2 we calculate the predicted value for each resample. The next procedure is obtaining the resampling median estimator of predicted value according to above description of the section 3. These results we compare consequently with the results of another method taking the bias as the efficiency criterion.

To investigate the resample size k influence on the estimators' properties we vary it $m \le k < n$. Note that in our example the number of resamples r for obtaining each resampling estimator is equal to all possible combinations k from n. In general case the number of resamples is considerably smaller than all possible combinations.

We tried to obtain the resampling median estimators of the predicted values of dependent variable for all existing observations and compare the results with classical estimators, and clean estimators, calculated without false observation. It means that we remove the false observation from our model and use classical LSE estimator (2) for our purposes.

The results are presented in the Table 1. The first column contains the names of parameters of interest the predictors for all observations. The second column contains the bias - the difference between the classical estimator and the clean estimator for all observations. In the next columns we can see the results of implementing resampling median estimators approach for different resample sizes k. It contains the bias – the difference between the resampling median estimator and clean estimators for all observations. Analyzing the obtained results we can conclude that for all observations there are the resampling estimators with the smaller value of bias, than the classical one, choosing the right resample size. Especially good results were obtained for the resample size equal to 6. For example, for the 7-th observation classical approach gives the bias 1.519, but the resampling median estimators are in the interval of (0.416-1.572) depending on the resample size.

6. CONCLUSIONS

The application of the proposed approach to the regression model with nuisance observations leads to obtaining good results. The analysis of the numerical results shows that the considered approach gives the better estimators than the classical methods, if we take the bias of expectation $E(Y_d)$ as the estimators' efficiency criterion.

7. ACKNOWLEDGEMENTS

I would like to thank my scientific supervisor professor Alexander Andronov for his useful and valuable ideas and comments during this paper preparation.

References

[1] Andronov A., Merkuryev Yu. Optimization of Statistical Sample Sizes in Simulation, *Journal of Statistical Planning and Inference,* Vol. 85, 2000, pp. 93-102.

[2] Draper N.R., Smith H. *Applied Regression Analysis*. New York: John Wiley & Sons, 1986.

[3] Efron B., Tibshirani R.Y. *Introduction to the Bootstrap*. London: Chapman & Hall, 1993.

[4] Gentle E. James. *Elements of Computational Statistics*. New York: Springer-Verlag, 2002.

[5] Srivastava Muni S. *Methods of Multivariate Statistics*. New York: Wiley-Interscience, 2002.

[6] Strelen J.Ch. Median Confidence Intervals. In: *Proc. of the ESM2001*. Erlangen: The SCS Publishing House, 2001, pp. 771-775.

[7] Wu C.F.J. Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis, *The Annals of Statistics*, Vol. 14, No 3, 1986, pp. 1261-1295.