Please cite this article as:
Andrzej Z. Grzybowski, On some new method of incorporation prior information along with its uncertainty in regression estimation, Scientific Research of the Institute of Mathematics and Computer Science, 2006, Volume 5, Issue 1, pages 23-31.
The website: http://www.amcm.pcz.pl/

ON SOME NEW METHOD OF INCORPORATION PRIOR INFORMATION ALONG WITH ITS UNCERTAINTY IN REGRESSION ESTIMATION

Andrzej Z. Grzybowski

Institute of Mathematics and Computer Science, Czestochowa University of Technology, Poland

Abstract. The paper is devoted to the problem of incorporating prior information in the regression estimation. In series of papers, see [3-6], we have proposed and analyzed some model of uncertainty which allow incorporating prior information along with its uncertainty via some Bayes estimators. We also introduced the notion of an Index of Uncertainty (IU) which indicate how useful the information and consequently the proposed estimators are. The results and methodology are summarized in [7]. Here, assuming different than in the mentioned papers prior knowledge about the regression problems, we propose a new description of uncertainty along with an index of uncertainty which was developed on the base of computer simulation.

Introduction

Consider the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}$, where $\mathbf{Y}$ is a vector of observations of the dependent variable, $\mathbf{X}$ is a nonstochastic (n x k) matrix of the observations of explanatory variables, $\boldsymbol{\beta}$ is a k-dimensional regression parameter (i.e. vector of unknown regression coefficients) and $\mathbf{Z}$ is an n-dimensional vector of random disturbances. Assume $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$, $\mathrm{Cov}(\mathbf{Y}) = \Sigma$. The paper is devoted to the problem of incorporating prior information when estimating $\boldsymbol{\beta}$.

Assume the prior information $\boldsymbol{\beta} = \boldsymbol{\beta}_p$ is derived from regression analysis applied (perhaps by someone else) to some phenomenon described by the same regression equation. However, we cannot be sure that the two phenomena are described by exactly the same regression equation and we do not know how reliable the previous results are - the prior information is uncertain. So, we must decide whether to use the information. If yes, we must choose a proper estimator. The usual least-squares estimator $\mathbf{b}^{LS}$ does not incorporate prior information and so, to use this information we need some alternative - the statistical theory help us here. We are presented with various Bayes, robust Bayes and minimax estimators, see e.g. [1, 7]. However, the optimal performance of the estimators depends on the problem formulation and the description of the prior information. In actual usage it is often difficult to decide what description of prior knowledge is most suitable - the knowledge may have different nature and various origins. In papers

[3, 4] we use computer simulations to compare various methods for choosing parameters of robust estimators incorporating the prior information $\boldsymbol{\beta} = \boldsymbol{\beta}_p$ and we introduce the notion of *Indices of Uncertainty* (*IU*) which role is to express the uncertainty connected with the information. The indices show how useful (or misleading) is such information in terms of risk reduction.

Assuming different prior information about the regression problem in this paper we propose another description of its uncertainty and then focus on simulation based methodology of determining indices of uncertainty. As a result we obtain some improved version of IU which appear to be very well correlated with the relative risk reduction gained by the estimators based on the information. In our simulations the prior information is generated along with the observations for regression analysis. All other characteristics of examined models are randomly changed as well. Consequently we study the performance of considered estimators for thousands data sets.

1. Problem statement and notation

In what follows the model used to obtain the prior information is called the *previous* model. The model to be examined is called the *current* model. In various symbols lower indices *p* and *c* point out what model are a given quantities from. For instance symbols $\mathbf{b}_p$, and $\mathbf{b}_c$ denote the least-squares estimates of the true parameters $\boldsymbol{\beta}_p$, $\boldsymbol{\beta}_c$ of the previous and current models, respectively, S_p and S_c denote the estimates of the standard deviations of random disturbances for each model.
Now, let us consider the following class of linear estimators:

$$\mathbf{b}_{(\vartheta,\Delta,\Sigma)}(\mathbf{Y}) = C(\Delta,\Sigma)\mathbf{X}^T\Sigma^{-1}\mathbf{Y} + C(\Delta,\Sigma)\Delta^{-1}\vartheta \tag{1}$$

where $C(\Delta,\Sigma) = (\mathbf{X}^T\Sigma^{-1}\mathbf{X}+\Delta^{-1})^{-1}$.

Estimators having such a structure arise as solutions to some problems of Bayes estimation. The value of ϑ may be thought of as a prior guess on $\boldsymbol{\beta}$, while a matrix Δ reflects our uncertainty connected with the guess. To make use of the estimators given by (1) we must specify the parameters $(\vartheta,\Delta,\Sigma)$ and usually it is not clear how to do it. Most easy case is connected with the matrix Σ. Theory of so called empirical (or feasible) generalized least squares estimation provide us with methods of estimating the covariance matrix Σ. The computer simulations also show that the intuitive method of determining the parameter $\boldsymbol{\beta}$ as $\mathbf{b}_p$ is quite satisfactory, see [3]. However, the most confusing point is how to determine the matrix Δ describing our uncertainty connected with the prior information $\boldsymbol{\beta} = \mathbf{b}_p$. And that is the problem we consider in the paper. Our second aim is to answer the question whether or not the obtained estimator $\mathbf{b}_{(\vartheta,\Delta,\Sigma)}$ is in a given situation better than

the usual *LS* - estimator. How can we know it? This question leads to the notion of an index of uncertainty.

2. Description of the uncertainty

As we have mentioned above, the matrix Δ in (1) reflects our uncertainty connected with the information $\boldsymbol{\beta} = \mathbf{b}_p$. For any diagonal positively definite ($k \times k$) matrix Δ the greater are the diagonal elements, the greater is the region in a *k-dimensional* parameter space in which the estimator $\mathbf{b}_{(\vartheta,\Delta,\Sigma)}$ has smaller risk function than the estimator $\mathbf{b}^{LS}$. The region is called an *improvement region.* On the other hand it is well known that the greater is the *improvement region* the smaller is the risk reduction, for more details see e.g. [6, 7]. Thus it is very important to determine the matrix Δ properly.

For a given loss function $L(.,.)$ an improvement gained by any given estimator $\mathbf{b}$ with respect to the estimator $\mathbf{b}^{LS}$ can be measured by a *symmetric relative loss reduction* index given by

$$LR(\mathbf{b}) = \frac{L(\boldsymbol{\beta},\mathbf{b}^{LS}) - L(\boldsymbol{\beta},\mathbf{b})}{L(\boldsymbol{\beta},\mathbf{b}^{LS}) + L(\boldsymbol{\beta},\mathbf{b})}$$

A justification for such a formula can be found in [5, 7].

In this paper we examine the case where the matrix Δ is defined as diagonal one with the elements $\Delta_{ii} = t_i^2$, where $t_i = \dfrac{\mathbf{b}_{ci} - \mathbf{b}_{pi}}{S_{ci}}$. Here $\mathbf{b}_{pi}$ is the *i*-th component of $\mathbf{b}_p$ and S_{ci} is the standard error of $\mathbf{b}_{ci}$. We denote this matrix by Δ^*. For convenience, the estimators $\mathbf{b}_{(\vartheta,\Delta,\Sigma)}$ with $\vartheta = \mathbf{b}_p$, $\Delta = \Delta^*$ and Σ estimated as usually in the empirical generalized LS method will be denoted as $\mathbf{b}^*$. It is obvious that sometimes the estimator is better than $\mathbf{b}^{LS}$, sometimes not. So it would be desirable to obtain a quantity which would show us whether or not the usage of the estimator $\mathbf{b}^*$ is profitable or, in other words, whether the prior information is useful or misleading. Such an indicator is called an index of uncertainty. More precisely, an *index of uncertainty* is an arbitrary quantity which has high negative correlation with the value of a risk reduction gained by the estimator $\mathbf{b}^*$.

Now we are to choose quantities which would possibly reflect the uncertainty understood as described above. It is quite clear that the information is the more profitable the less trustful are our current estimates on one hand and, on the other hand, the more trustful are the previous ones.

Given the data, the tool of least squares can be employed. However how trustful the results are depends on the data at hand and thus as quantities which potentially

reflects our uncertainty of prior information we consider the following well known characteristics of both the data and the model:

- R_c^2, R_p^2 - multiple coefficients of determination for the current and previous model, respectively,
- Statistics $T = \frac{1}{k}\sum_i t_i$,
- CN_c, CN_p - condition numbers of the matrices of observations of explanatory variables for the current and previous model. Let us remember that the condition number of any $(n \times k)$ matrix $\mathbf{X}$ is given by $CN = \frac{\lambda_{\max}}{\lambda_{\min}}$, with $\lambda_{\max}, \lambda_{\min}$ being the maximal and minimal singular value of the matrix $(\mathrm{X}^T\mathrm{X})^{-1}$, see [1],
- df_c, df_p - degrees of freedom for the current and previous model, respectively.

With the help of computer simulations we verify this idea and choose the most useful index IU.

3. Description of simulations

The simulations are based on two procedures: *Single Regression Simulation* and *Main Simulation*. All procedures are programmed using Mathematica 4.0 software.

Single Regression Simulation Procedure (SRSP)

The input for this procedure consists of the matrices $\mathbf{X}_p$, $\mathbf{X}_c$ of the observations of explanatory variables for both models, the true regression parameters $\boldsymbol{\beta}_p$, $\boldsymbol{\beta}_c$ (possibly different), the distributions of $\mathbf{Z}_p$ and $\mathbf{Z}_c$. During *SRSP* the dependent variables, $\mathbf{Y}_p$, $\mathbf{Y}_c$ are generated, each according to the appropriate model. The prior information $\mathbf{b}_p$, S_p, S_{pi}, $i = 1,..., k$ is generated as well. Then the values $\mathbf{b}(\mathbf{Y}_c)$ of all estimators $\mathbf{b}$ under consideration are computed as well as $\mathbf{b}_c$ - the value of $\mathbf{b}^{LS}$. For each considered estimator $\mathbf{b}$ we record a value of relative loss reduction *LR* along with many other characteristics of data and the regression problem, among them the values of the quantities stated previously, i.e. R_c^2, R_p^2, T, CN_c, CN_p, df_c, df_p.

Main Simulation Procedure (MSP)

An input for this procedure consists of the distributions for $\mathbf{Z}_p$, $\mathbf{Z}_c$ (in our research the distributions were normal or uniform). As a first step of this procedure we randomly generate the quantities which form an input for SESP i.e.:

dimensions k, the numbers of observations n_p, n_c, matrices $\mathbf{X}_p$, $\mathbf{X}_c$, vectors $\boldsymbol{\beta}_p$, $\boldsymbol{\beta}_c$. The regression parameter $\boldsymbol{\beta}_c$ is obtained as a random transformation of $\boldsymbol{\beta}_p$, reflecting the fact that the investigated model may be different from the previous one. These generated quantities remain unchanged during a single *MSP*. As a second step of *MSP* we execute single regression simulation procedure *Ns* times and
record average values for all the quantities computed.

With the help of presented above procedures we simulate over a million regression settings. For each case the dimension of regression parameter is drawn from the set [3,...,15] and the degrees of freedom chosen randomly between 3 and 200. The matrices $\mathbf{X}_p$ and $\mathbf{X}_c$ are also randomly chosen. All the matrices have one constant column, what reflect the fact that we perform regression analysis for models with an intercept. Some other characteristics of the generated data are presented in Table 1.

Table 1

Location and dispersion characteristics of generated data

	R_c^2	R_p^2	CN_c	CN_p	k	n_c	n_p
Mean	0.77	0.75	1031	1050	8.9	80.7	81.3
Standard Deviation	0.13	0.11	2311	2345	3.7	62	61
Min	0.25	0.25	1.27	1.2	3	6	6
Max	0.99	0.99	9993	9999	15	200	200
Lower quartile	0.68	0.68	7.11	2.7	6	20	24
Median	0.76	0.75	23.9	6.06	9	62	68
Upper quartile	0.87	0.81	191.9	32.6	12	136	135

4. Index of Uncertainty

Now our aim is to find index *IU* which will show how useful or misleading is the prior information when it is incorporated into regression analysis via the estimator **b***. The index should be well correlated with the performance of the estimator in a sense we described previously.

To develop such a quantity we perform regression parameter estimation for 100 000 regression settings and then adopt standard regression techniques to obtain the model describing the relation between the LR index and the quantities

R_c^2, R_p^2, T, CN_c, CN_p, df_c, df_p. Because we need only *one* quantity explaining the behavior of LR we tried to build log-linear model. Finally, dropping all insignificant variables, we obtain a proposal for an uncertainty index in the following form:

$$IU_1 = 2 - \left(\frac{CN_c}{CN_f}\right)^{0.04} \left(\frac{df_f}{df_c}\right)^{0.03}$$

To simplify the form of the index in next simulation we examine also other form of the index, among them the following:

$$IU_2 = 2 - \left(\frac{CN_c}{CN_f}\right)^{0.04}$$

$$IU_3 = 2 - \left(\frac{CN_c}{CN_f}\frac{df_f}{df_c}\right)^{0.04}$$

$$IU_4 = 2 - \left(\frac{CN_c}{CN_f}\frac{df_f}{df_c}\right)^{0.05}$$

$$IU_5 = 2 - \left(\frac{CN_c}{CN_f}\frac{df_f}{df_c}\right)^{0.1}$$

In our simulation we consider a loss function given by

$$L(\beta, \mathrm{b}) = \frac{1}{k}\sum_{i=1}^{k}\left|\frac{\beta_i - b_i}{\beta_i}\right| \tag{2}$$

In Table 2 we show the values of the Pearson correlation coefficient r between $LR(\mathbf{b}^*)$ calculated for such a loss and the values of given function IU_i, $i = 1,\ldots,5$. The coefficients are computed on the base of whole data gathered during the second part of our research and consisting of 25 500 records. Because the number of loops Ns was equal to 1 each record contains *exact* values of both LR and IU.

Table 2

Pearson correlation coefficients *r* between *exact value* of *LR* and indexes IU_i, $i = 1,\ldots,5$ ($Ns = 1$)

IU_1	IU_2	IU_3	IU_4	IU_5

0.56	0.63	0.64	0.63	0.62

We see that the correlation is very high.
In Table 3 we present the results obtained in the situation where number *Ns* of loops in MSP was equal to 30. Thus the presented numbers *r* can be interpreted as the measure of correlation between *expected value* of *LR* and proposed indexes. The coefficients are computed on the base of whole data gathered during next part of our research and consisting of another 25 500 records and thus they are based on above 750 000 regression settings.

Table 3

Pearson correlation coefficients *r* between *expected value* of *LR* and indexes IU_i, $i = 1,\ldots,5$ ($Ns = 30$)

IU_1	IU_2	IU_3	IU_4	IU_5
0.76	0.81	0.82	0.82	0.79

As we could expect the performance (expressed in terms of an average *LR*) of the estimator **b*** demonstrates even higher correlation with the values of indices IU_i than in the previous case.
Because the performance depends upon the prior information the results suggest the indices could indicate how useful is the information incorporated by the estimator.
In view of Tables 2, 3 and 4 the function IU_3 seems to be best proposal for the index of uncertainty because it demonstrates high correlation with the performance of the estimator **b*** and has simple and intuitive form. To emphasize our choice we denote the index *IU**.
Now, to study how the uncertainty incorporated into the estimate depends on the value of *IU** we compute and compare the average values of both *IU** and *LR* obtained for ten classes of values of *IU**. As limits of the classes we took deciles of the observed values of the index. Table 4 presents the average values of both *LR* and *IU** for these classes.

Table 4

Average values of *LR* and *IU for classes determined by deciles of *IU****

Class of *IU**	Average value of *IU**	Average value *LR*
0.36÷0.70	0.618208	0.551724
0.70÷0.79	0.750889	0.353421
0.79÷0.86	0.824029	0.230933
0.86÷0.93	0.894041	0.129576
0.93÷0.98	0.956151	0.0139668

0.98÷1.00	0.99029	–0.0715046
1.00÷1.03	1.01975	–0.118613
1.03÷1.09	1.06464	–0.227802
1.09÷1.15	1.11694	–0.338404
1.15÷1.35	1.20802	–0.456357

One can notice that the correlation coefficient between the averages equals to 0.997! Similar results obtained for classes determined by percentiles are presented in Figure 1.

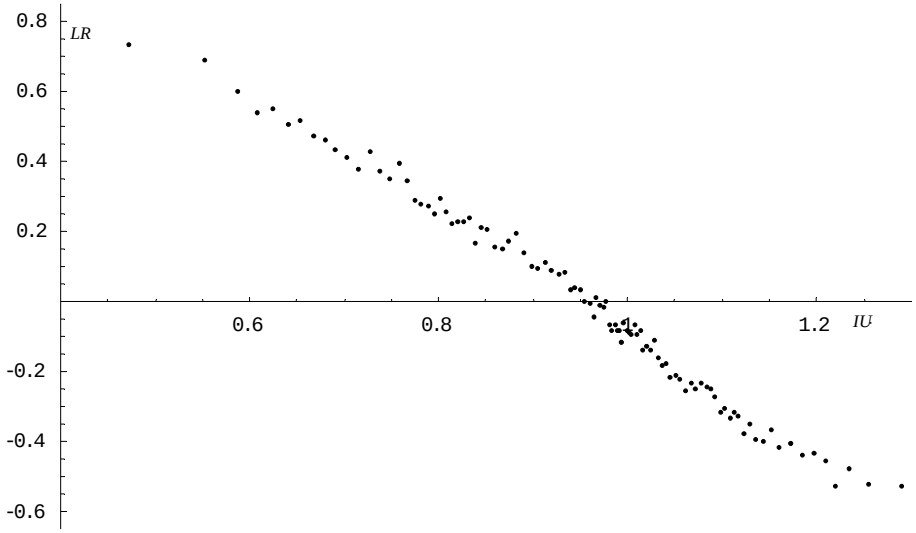

Fig. 1. Average values of *LR* and *IU** for classes determined by percentiles of *IU**

In this case the correlation coefficient between the averages equals to 0.993!

5. Final remarks

Based on computer simulations and assuming different prior knowledge we obtain here new proposal for an uncertainty index *IU**. The index differs from the ones proposed in papers [6, 7]. The main difference is that its value depends on the condition number of the matrix $\mathbf{X}_p$. In many real world problems we are not provided with such a knowledge. However if we have such an information then the proposed index is worth consideration because it exhibits very good features. Its correlation with the value of symmetric relative loss reduction *LR* is amazingly very high. Thus we can be almost sure answering the question whether we should

choose the least square estimator or the estimator incorporating prior information. Moreover, we can also estimate how big loss reduction can be expected.
Is the proposed index the best possible? Perhaps not. We do not neither think that there exists "the best" choice of *IU* and we are aware that computer simulations can prove nothing. On the other hand we think that the index *IU** is really very good and satisfactory proposition for an indicator of the uncertainty of the prior knowledge.

Concluding we should also stress that our results were obtained under the loss given by (2) and when the distributions of disturbances were normal or uniform. We are sure however, that the proposed methodology, can be used to determine the form of the coefficient of uncertainty when the criterion of performance is given by other loss functions (e.g. quadratic or Euclidean norm) as well as for other than considered here distributions.

References

[1] Belslay D., Kuh E., Welsch R., Regression Diagnostics, Wiley & Sons, New York 1980.

[2] Berger J.O., Statistical Decision Theory and Bayesian Analysis, Springer Verlag, New York 1985.

[3] Birkes D., Dodge Y., Alternative methods of regression, Wiley & Sons, New York 1993.

[4] Grzybowski A., Simulation analysis of some regression estimators incorporating prior information, Proceedings of International Workshop on Statistical Modelling, Graz 1999, 548-551.

[5] Grzybowski A., Simulation analysis of some regression estimators incorporating prior information - performance for different loss functions, Proceedings of 16th IMACS World Congress 2000 on Scientific Computation, Applied Mathematics and Simulation, Lausanne 2000 (CD).

[6] Grzybowski A., Computer Simulations in Constructing a Coefficient of Uncertainty in Regression Estimation - Methodology and Results, Lectures Notes in Computer Science No 2338, Springer Verlag, Berlin 2002, vol. 2328, 671-678.

[7] Grzybowski A., Analiza funkcji ryzyka estymatorów opartych na pewnym opisie niepewności, Modelowanie Preferencji a Ryzyko'02, Praca zbiorowa pod redakcją naukową T. Trzaskalika, Wydawnictwo WE w Katowicach, Katowice 2002, 139-149.

[8] Grzybowski A., Metody wykorzystania informacji a priori w estymacji parametrów regresji, Seria Monografie No 89, Wydawnictwo Politechniki Częstochowskiej, Częstochowa 2002.