Please cite this article as:
Andrzej Grzybowski, On a coefficient of uncertainty based on the condition number of the matrix of observations, Scientific Research of the Institute of Mathematics and Computer Science, 2007, Volume 6, Issue 1, pages 69-74.
The website: http://www.amcm.pcz.pl/



# ON A COEFFICIENT OF UNCERTAINTY BASED ON THE CONDITION NUMBER OF THE MATRIX OF OBSERVATIONS

***Andrzej Grzybowski***

*Institute of Mathematics and Computer Science, Czestochowa University of Technology, Poland*
*email: azgrzybowski@gmail.com*

**Abstract.** The paper is devoted to the problem of incorporating prior information into regression model estimation. We assume the prior information about regression parameter is derived from regression analysis applied to some phenomenon described by the same regression equation. However, usually the prior information is uncertain. On the base of computer simulation we construct a coefficient which allows incorporating the prior information along with its uncertainty. The coefficient is based upon the index number of the matrix of observations of the explanatory variables. Performance of estimators based upon the coefficient of uncertainty is examined through computer simulations.

## 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 and $\mathbf{Z}$ is an $n$-dimensional vector of random disturbances. Assume $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$, $\mathrm{Cov}(\mathbf{Y}) = \Sigma$.

Assume now the prior information $\boldsymbol{\beta} = \mathbf{b}_p$ 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. In this paper we propose such an alternative - on the base of computer simulation we also construct an index and a coefficient of uncertainty which allow us incorporating the prior information along with its uncertainty. The coefficient is based upon the condition number of the matrix of

observations of the explanatory variables. Performance of the estimators based upon the coefficient of uncertainty is examined through computer simulations. Another indexes and coefficients of uncertainty were proposed in [4].

## 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 estimated is called the *current* model. In various symbols lower indices *p* and *c* point out what model are the given quantities connected with.

Now, let us consider the following class of linear estimators:

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

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 $\vartheta$ may be thought of as a prior guess on $\mathbf{b}$, while a matrix $\Delta$ reflects our uncertainty connected with the guess. To make use of the estimators 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 $\Sigma$. The theory of so called empirical (or feasible) generalized least squares estimation provides us with methods of estimating the covariance matrix $\Sigma$. The computer simulations also show that the intuitive method of determining the parameter $\vartheta$ as $\mathbf{b}_p$ is quite satisfactory (here $\mathbf{b}_p$ is the *LS*-estimate for $\boldsymbol{\beta}$ in the previous model).

However, the most confusing point is *how to determine* the matrix $\Delta$ describing our *uncertainty* connected with the prior information $\mathbf{b} = \mathbf{b}_p$. And that is the problem we consider here.

Our second aim is to answer the question *whether or not* the 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. The index of uncertainty

In our research we examine the case where the matrix $\boldsymbol{\Delta}$, which reflects our uncertainty connected with the information $\mathbf{b} = \mathbf{b}_p$, is defined as diagonal one with the elements $\Delta_{ii} = t_i^2$ where

$$t_i = \frac{\mathbf{b}_{ci} - \mathbf{b}_{pi}}{S_{ci}}$$

Here $\mathbf{b}_{pi}$ is the *i*-th component of $\mathbf{b}_p$, Here $\mathbf{b}_{ci}$ is the *i*-th component of $\mathbf{b}_c$ (the *LS*-estimate for $\boldsymbol{\beta}$ in the current model) and $S_{ci}$ is the standard error of $\mathbf{b}_{ci}$. We

denote this matrix by $\boldsymbol{\Delta}^*$. The estimators $\mathbf{b}^{(\vartheta,\Delta,\Sigma)}$ with $\vartheta = \mathbf{b}_p$, $\boldsymbol{\Delta} = \boldsymbol{\Delta}^*$ and $\boldsymbol{\Sigma}$ estimated as 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 loss reduction gained by the estimator $\mathbf{b}^*$.

For a given loss function $L(\cdot,\cdot)$ 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

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

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 and, on the other hand, the more trustful are the previous ones. 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: multiple coefficients of determination for the current and previous model, respectively, $CN_c$, $CN_p$ - condition numbers of the matrices of observations of explanatory variables for the current and previous model, $df_c$, $df_p$ - degrees of freedom for the current and previous model, respectively.

Let us remember, see e.g. Belsley [1], that the condition number *CN* of any $(n\times k)$ matrix $\mathbf{X}$ is given by

$$CN = \frac{\lambda_{\max}}{\lambda_{\min}}$$

with $\lambda_{\max}$ and $\lambda_{\min}$ being the maximal and minimal singular value of the matrix $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$.

With the help of computer simulations we generate over 20 000 regression settings. The simulations are described in Grzybowski [2]. The main statistical characteristics of the obtained data are presented in Table 1.

Given the data and employing the tool of least squares we have found several proposals for the uncertainty index. Among them the best was an index *IU** given by the following formula:

$$IU^* = 2 - \left( \frac{CN_c}{CN_p} \frac{df_p}{df_c} \right)^{0.04}$$

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 |

The Pearson correlation coefficient $r$ between $LR(\mathbf{b}^*)$ calculated for loss function given by

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

and the values of $IU^*$ equals −0.66. The value of $r$ is similar for various data sets (representing various regression settings) generated independently and consisting of thousands of records.

## 3. Coefficient of uncertainty

For any given estimator $\mathbf{b}$ the region in a $k$-dimensional parameter space in which the estimator $\mathbf{b}$ has smaller values of the risk function than the estimator $\mathbf{b}^{LS}$ we call an *improvement region* (connected with **b)**.

It is known that for any positively definite ($k\times k$) matrix $\boldsymbol{\Delta}$ and any given number $K > 0$, the improvement region for the estimator $\mathbf{b}_{(\vartheta,K\Delta,\Sigma)}$ is an ellipsoid in the parameter space. The ellipsoid is the greater, the greater is $K$. On the other hand it is also well known that the value of risk reduction tends to 0 when $K$ tends to infinity, see Grzybowski [3].

Let $\mathbf{b}^K$ denote the estimator $\mathbf{b}_{(\vartheta,\Delta,\Sigma)}$ with the parameters defined as follows: $\boldsymbol{\vartheta} = \mathbf{b}_p$, $\boldsymbol{\Delta} = K\boldsymbol{\Delta}^*$ and $\boldsymbol{\Sigma}$ is estimated as in the empirical generalized *LS* method. We see that $\mathbf{b}^1 = \mathbf{b}^*$. It follows from the above remarks that a proper choice of the constant $K$ is a crucial point when applying the estimators $\mathbf{b}^K$.

To study how the uncertainty incorporated into the estimate depends on the value of $IU^*$ we compare performance of the estimators $\mathbf{b}^K$ for various values of $K$ (the bigger $K$ the larger amount of uncertainty is incorporated). Table 2, providing us with the results of the comparison, is based on next 25 500 records.

The results presented in Table 2 suggest that the uncertainty incorporated into regression should be described by a matrix

$$\Delta = CU(IU^*)\,\Delta^*$$

where *CU* is some positive and increasing function of its argument - the function will be called *Coefficient of Uncertainty*. The estimator $\mathbf{b}_{(\vartheta,\Delta,\Sigma)}$ with the above given matrix $\Delta$ we denote as $\mathbf{b}^{CU}$.

Table 2

**Average *LR* gained by the estimators $\mathbf{b}^K$ for deferent value classes of $IU^*$**

| ***K*= 1000** | ***K*= 100** | ***K*=10** | ***K*= 1** | ***K*=0,1** | ***K*= 0,01** | ***K*= 0,001** | **IU** |
|---|---|---|---|---|---|---|---|
| 44,9% | 49,0% | 52,8% | 56,0% | 59,1% | 61,1% | **62,0%** | 0,787 |
| 42,5% | 45,6% | 49,0% | 52,4% | 55,3% | 57,1% | **57,8%** | 0,839 |
| 41,5% | 44,6% | 46,8% | 49,0% | 51,1% | 52,1% | **52,3%** | 0,881 |
| 32,4% | 34,6% | 36,5% | 37,8% | 39,4% | 40,1% | **40,2%** | 0,921 |
| 23,1% | 25,2% | 27,0% | 28,4% | 29,8% | 31,1% | **31,6%** | 0,955 |
| 20,2% | 21,4% | 23,3% | 24,5% | 25,4% | 25,8% | **25,9%** | 0,977 |
| 12,1% | 13,7% | 14,9% | 16,2% | 17,3% | **18,1%** | 17,8% | 0,992 |
| 11,5% | 12,5% | 13,6% | 14,3% | **14,8%** | **14,8%** | 14,5% | 1,004 |
| 8,6% | 9,8% | 10,4% | **10,9%** | **10,9%** | 10,6% | 10,4% | 1,019 |
| 4,1% | **5,4%** | **5,4%** | **5,4%** | 5,1% | 4,6% | 4,1% | 1,038 |
| **-3,1%** | -3,4% | -3,5% | -3,8% | -4,3% | -5,1% | -6,0% | 1,064 |
| **-10,7%** | -11,9% | -12,9% | -13,3% | -14,4% | -15,7% | -17,1% | 1,093 |
| **-15,7%** | -16,8% | -18,6% | -19,9% | -21,3% | -22,7% | -24,8% | 1,120 |
| **-19,7%** | -21,1% | -22,5% | -25,1% | -27,0% | -29,7% | -32,2% | 1,158 |

To construct a satisfactory proposal for coefficient *CU* we perform another simulations. We record the relative improvement gained by the estimators $\mathbf{b}^K$ for 50 different values of the constant *K*. So, for each record we have the value of $IU^*$ and - approximately - the best value of the constant *K*. We obtained data consisting of 27 000 records. On the base of analysis of the data we propose the following formula for the coefficient *CU*.

$$\mathrm{CU}(x)=\begin{cases}(-14000x+13500)^{-1}, & x\in[0.0,0.9634)\\ (-145x+152)^{-1}, & x\in[0.9634,1.048)\\ 10000, & x\geq 1.048\end{cases}$$

In Table 3 we compare the performance of the estimator $\mathbf{b}^{CU}$ and the estimators $\mathbf{b}^K$, for different values of *K*. As we can see the performance of the estimator $\mathbf{b}^{CU}$ is quite satisfactory - it works better than any given estimator $\mathbf{b}^K$.

Table 3

**Average *LR* gained by the estimators $b^{CU}$ and $b^{K}$ for deferent value classes of $IU^{*}$ - comparison**

| *CU* | *K*= 1000 | *K*= 100 | *K*= 1 | *K* =0,1 | *K*= 0,01 | *K*= 0,001 | IU |
|---|---|---|---|---|---|---|---|
| **48,3%** | 35,6% | 43,3% | 46,1% | 47,5% | 48,1% | **48,3%** | 0,786 |
| **43,1%** | 32,4% | 38,1% | 40,7% | 42,5% | **43,4%** | 43,2% | 0,839 |
| **34,5%** | 28,4% | 32,6% | 33,9% | 34,9% | 35,0% | 34,5% | 0,881 |
| **23,9%** | 21,2% | 23,4% | **24,2%** | **24,2%** | 24,1% | 23,9% | 0,921 |
| **15,6%** | 12,8% | 14,3% | 15,0% | 15,4% | 15,5% | 15,1% | 0,955 |
| **6,8%** | 5,8% | 6,5% | **6,8%** | 6,8% | 6,2% | 5,5% | 0,978 |
| **1,5%** | 2,9% | **2,9%** | 2,1% | 1,4% | 0,6% | -0,4% | 0,994 |
| **-1,8%** | **0,5%** | -0,1% | -1,3% | -2,0% | -2,7% | -3,9% | 1,008 |

## Concluding remarks

We should stress that we did not prove that the proposed index *IU** and coefficient *CU* are best possible. Moreover, we do not think that there exists "the best" choice of *IU* and *CU*. However, based on computer simulations we find that our proposal for *IU* is a very good indicator of the uncertainty of the prior information and the incorporation of prior information via proposed *CU* leads to significant risk reduction and so the proposals seem to be very satisfactory.

We should also stress, that our results were obtained in the case where the loss function is given by the formula (1). For any different loss function the appropriate indexes *IU* and coefficients CU may be given by different formulae.

## References

[1] Belsley D.A., Conditioning diagnostics: collinearity and weak data in regression, Wiley, New York 1991.

[2] Grzybowski A., On some New Method of Incorporation Prior Information Along with its Uncertainty in Regression Estimation, Scientific Research of the Inst. of Math. and Comp. Sci. of TUC, 2006, 1(5), 23-31.

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

[4] 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.