Simulation Analysis Of Regression Estimators Based On Coefficients of Uncertainty

Andrzej Grzybowski*

Abstract

The paper is devoted to the problem of incorporating prior information in the regression analysis. Some indices of uncertainty of the prior knowledge are proposed and their usefulness is studied. To incorporate prior information together with its uncertainty into regression estimation some coefficients of uncertainty are introduced as well. Performance of estimators based upon proposed descriptions of uncertainty is examined via computer simulations.

1 Introduction

Let us consider the ordinary linear model $Y = X\beta + Z$, where Y is a vector of observations of the dependent variable, X is a matrix of the observations of explanatory variables, β is a vector of unknown regression coefficients and Z is a vector of random disturbances (all quantities of appropriate dimensions). Let us assume $E(Y) = X\beta$ and $\mathrm{Cov}(Y) = \Sigma$ and the matrix X has a full rank. The paper is devoted to the problem of incorporating prior information in the estimation of the model coefficients. We assume the prior information $\beta = b_p$ was derived from previous regression analysis performed (perhaps by someone else) for some phenomenon which was described by the same regression equation as the one we investigate. However, we cannot be sure that the phenomenon was described by *exactly* the same regression equation and we do not know how reliable are the obtained results - the prior information is *uncertain*. In such a case first we have to decide whether to make use of the information or not. If yes, we should choose a proper estimator. The usual least-squares estimator d_{LS} does not incorporate any prior information about the regression equation. To make use of the information we need some alternative and the statistical theory help us. We are presented with various Bayes, robust Bayes and minimax estimators, see e.g. [1, 2, 4]. However, their optimal performance depends on the problem formulation (e.g on the performance criterion) as well as on the description of the prior information. In practical applications, however, it is usually difficult to decide what description of our prior knowledge would be the most suitable -

*Institute of Mathematics & Computer Science, Technical University of Czestochowa, 42-200 Czestochowa, Poland, email: agrzyb@matinf.pcz.czest.pl

the knowledge may have different nature and various origins. In the paper [3] we compare various methods for choosing parameters of robust linear estimators incorporating the prior information in the situation described above. In our paper we introduce some indices to "measure" the uncertainty connected with such information. Next we propose a method of incorporating the uncertainty into regression estimation. The usefulness of various uncertainty indices as well as the performance of the introduced estimators are examined via computer simulations. During the simulations we generate the prior information as well as the observations for regression analysis (changing at random all characteristics of examined models). Consequently we study the performance of considered estimators for thousands data sets.

2 Problem Statement and Notation

In the sequel the model our prior information is obtained from will be called the *previous* model. The model we are to examine will be called the *current* one. Symbols $b_p,\ b_c$ denote the least-squares estimates of the *true* parameters β_p, β_c of the previous and current models while S_p, S_c denote the estimates of the standard deviations of random disturbances for each model, respectively.

We examine the following class of linear estimators:

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

where $C(\Delta,\Sigma) = (X^T\Sigma^{-1}X + \Delta^{-1})^{-1}$. Such estimators arise as solutions to some problems of Bayes estimation. The value of ϑ may be thought of as a prior guess for β, while Δ reflects our uncertainty connected with the guess. The estimator $d_{(\vartheta,\Delta,\Sigma)}$ is also minimax linear for some problems with unknown matrices Δ, Σ and given ϑ, see [2]. Similar in structure estimators were also obtained as solutions to the problem of minimax linear estimation when the set of the states of nature was given in the form of restricted parameter space, see [4] for references.

To make use of the estimators given by (1) one has to set up the parameters $\Delta, \Sigma, \vartheta$ and usually it is not very clear how to do it. The computer simulations show that in our case, when we know b_p and some other fundamental quantities obtained during previous regression analysis, the intuitive methods of determining the parameters ϑ as b_p and Σ as diagonal matrix with the elements S_c on principal diagonal are quite satisfactory, see [3]. The most confusing point is how to determine the matrix Δ describing our uncertainty connected with the prior information. In the paper we deal with the problem. We examine the case where the matrix is defined as diagonal one with the elements Δ_{ii} equal to $\mathrm{Max}(b_{p_i}, s_{p_i})$. Here b_{p_i} is the i-th component of b_p and s_{p_i} stands for the standard error of estimation of b_{p_i}. We denote this matrix by Δ^*. As potentially good uncertainty indices we consider the following functions:

$$IU_{ijl} = T^i RD^j (\frac{R_c^2}{R_p^2})^l\ ,\ i = 0, ..., 3\ ,\ j = 0, ..., 3\ ,\ l = 0, ..., 3 \tag{2}$$

where a statistic T is given by $T =\parallel (\frac{b_{p_1} - b_{c_1}}{s_{c_1}}, ..., \frac{b_{p_k} - b_{c_k}}{s_{c_k}}) \parallel$, a relative distance RD between estimates is given by $RD = \frac{\|b_p - b_c\|}{\|b_c\|}$ and R_c^2, R_p^2 are multiple coefficients of determination for the current and previous model, respectively. With the help of computer simulations we verify this idea and choose the most useful index IU^*. Next we propose some method of determining the matrix Δ based upon the chosen index IU^*.

3 Description of Simulations

Simulations performed in our studies were based on two described below procedures - *Single Estimation Simulation Procedure* and *Main Simulation Procedure* . All procedures were programmed with the help of the programming tools of Mathematica 4.0, the product of Wolfram Research, Inc.

Single Estimation Simulation Procedure (*SESP*).

An input for this procedure consists of the matrices of the observations of explanatory variables for both models i.e. X_p, X_c, the true regression parameters β_p, β_c (maybe different), the distributions π_p, π_c, i.e. their shapes and moments. During SESP we generate random vectors of observations of the dependent variables, Y_p, Y_c, each according to an appropriate model. We also obtain and write down the prior information $b_p, S_p, s_{p_i}, i = 1, ..., k$. Next we compute the values $d(Y_c)$ of all estimators d under consideration as well as b_c - the value of d_{LS}. For each estimator d we write down $L(d(Y_c), \beta_c) =\parallel d(Y_c) - \beta_c \parallel$ and *Relative Improvements* (w.r.t d_{LS}) $RI(d) = \frac{L(d_{LS}(Y_c),\beta_c) - L(d(Y_c),\beta_c)}{L(d_{LS}(Y_c),\beta_c)}$ For each examined estimator d we additionally write down a variable called $Better(d)$ which is equal to 1 if $RI(d) > 0$ or equals 0 otherwise. An average value of $Better(d)$ is an estimated probability that given estimator is better than d_{LS} in terms of considered loss function. It will be denoted by $PB(d)$. Apart from the above quantities we remember as well many other characteristics, among them the values of the indices IU_{ijl}.

Main Simulation Procedure (*MSP*).

An input for this procedure consists of the distributions π_p, π_c (in our research the distributions were normal). As a first step of this procedure we randomly generate the quantities which form an input for SESP i.e.: dimensions k, n_p, n_c , matrices $X_p, X_c,$, vectors β_p, β_c. The regression parameter β_c is obtained by random transformation of β_p, what reflects the fact that the investigated model may be different from the previous one. These generated quantities do not change during a single MSP. As a second step of MSP we execute SESP over a hundred times and write down average values of all quantities computed during these SESPs.

With the help of presented above procedures we simulated over a million problems of regression estimation. For each of the problem the dimension of regression parameter was drawn from the set [3,...,15], the degree of freedom was a random number between 3 and 150. The matrices of observations of explanatory variables were random as well.

4 Results

Some characteristics of generated data are presented in Table 1.

Tab.1 Some characteristics of generated data

	Mean	Median	Max	Min
S_p	462	314	1874	3.13
S_c	514	331	2001	3.59
S_c/S_p	1,33	1,21	3,8	0,45
T	10.97	6.02	102	1.65
RD	1.41	0.57	25.8	0.04
R_p^2	0,74	0,79	0,98	0,29
R_c^2	0,82	0,87	0,98	0,30

Let for a given value K the symbol d_K denote an estimator $d_{(b_p,K\cdot\Delta^*,S_cI)}$, see Section 2.

In the next table we show the values of the Pearson contingency coefficient C between both RI and PB gained by the estimator d_1 and the values of a given function IU_{ijl} (indicated at the first row by the value of indices ijl). The coefficients are computed on the base of whole data gathered during the first part of our research and consisting of 6 901 records. Each record contains average values of the above mentioned quantities - see description of simulation procedures - computed for a hundred of SESP. Thus it is based on 690 100 simulations of regression problems. In Table 2 we present the results for only few functions IU_{ijl}- the most promising ones and some other to compare.

Tab. 2. Pearson contingency coefficients C for the whole data.

	110	120	130	111	121	131	112	122	132
$\overline{RI}(d_1)$	0.37	0.25	0.22	0.48	0.32	0.30	0.41	0.36	0.25
$PB(d_1)$	0.42	0.20	0.14	0.49	0.27	0.19	0.44	0.29	0.19

We see that the performance (in terms of RI as well as PB) of the estimator d_1 is clearly dependent on the values of indices $IU_{111}, IU_{110}, IU_{112}$. Because the performance depends upon the prior information it suggests the indices could indicate how useful is the information incorporated by the estimator. This fact is confirmed by Pearson correlation coefficients r presented in Table 3.

Tab. 3. Pearson correlation coefficients r for the whole data.

	110	120	130	111	121	131	112	122	132
$\overline{RI}(d_1)$	-0.21	-0.12	-0.08	-0.25	-0.17	-0.10	-0.24	-0.19	-0.12
$PB(d_1)$	-0.28	-0.09	-0.05	-0.32	-0.12	-0.06	-0.30	-0.13	-0.07

One can see that the correlation between $IU_{111}, IU_{110}, IU_{112}$ and both RI and PB is negative so, the greater is the value of any of the indices the worse is the performance of the estimator. It means that the information the estimator d_1 is based on is the more uncertain (and misleading) the greater are the indices.The correlation is even stronger when we ignore small changes in the values of the indices. To verify this we sorted all data according values of each of the indices and then we divided it into 50 Classes of Values (CoV). Next we compute the correlation coefficients r between the average value of the index for a given CoV and average values of both RI and PB in this group of data. The results are presented in Table 4.

Tab. 4. Pearson correlation coefficients r for 50 CoV of IU_{ijl}

	110	120	130	111	121	131	112	122	132
$\overline{RI}(d_1)$	-0.72	-0.48	-0.41	-0.74	-0.49	-0.42	-0.70	-0.48	-0.41
$PB(d_1)$	-0.60	-0.22	-0.15	-0.62	-0.22	-0.15	-0.59	-0.23	-0.16

In view of Tables 2, 3 and 4 the function IU_{111} seems to be most correlated with the performance of d_1. Thus we choose it as uncertainty index and denote IU^*. To study how the amount of uncertainty which should be incorporated into estimation depends on the value of IU^* we compare performance of the estimators d_K for various values of K (the bigger K the larger amount of uncertainty is incorporated). Table 5, providing us with the results of the comparison, is based on next 418 400 simulations.

Tab. 5. Average RI of d_K for different CoV of IU^*

CoV of IU^*	$d_{0.5}$	d_1	d_9	d_{49}	d_{64}	d_{100}
0-1	10,0%	9,0%	3,0%	0,8%	0,6%	0,4%
1-2	10,0%	9,0%	3,0%	0,6%	0,5%	0,3%
2-4	4,0%	7,0%	5,0%	2,0%	2,0%	1,5%
4-8	-39,0%	-24,0%	0,5%	1,3%	1,2%	1,1%
8-20	-95,0%	-61,0%	-8,0%	-0,5%	-0,2%	0,01%
20-50	-143,0%	-85,0%	-9,0%	-0,9%	-0,6%	-0,3%
over 50	-224,0%	-131,0%	-15,0%	-2,0%	-1,3%	-0,7%

We see that the estimator d_1 can be used when $IU^* < 4$ (RI is positive). However, when IU^* is smaller than 2 more profitable is estimator $d_{0.5}$ what means that we can be more trustful. When the index has value greater than 4 we lose using d_1 - our information is misleading. We cannot trust in it. The uncertainty can be reflected by greater value of K, compare the performance of the remaining estimators presented in Table 5. The results suggest that the uncertainty incorporated into regression can be described by a matrix $\Delta = CU(IU^*) \cdot \Delta^*$ for some increasing function CU - the function will be called Coefficient of Uncertainty. In our studies we have examined various proposals for the coefficient and we obtained good estimators for CU given by:

$$\begin{aligned} CU(x) \quad &= \quad (0.07x^2 + 0.3)\mathbf{1}_{[0,2)}(x) + (0.1x^2 + 0.1x)\mathbf{1}_{[2,20)}(x) \\ &\quad\quad + (15x^2 - 100x - 3958)\mathbf{1}_{[20,\infty)}(x) \end{aligned} \tag{3}$$

where $\mathbf{1}_A(\cdot)$ is a characteristic function of the indicated set A.

The comparison of the estimator d_{CU} based upon the coefficient with estimators d_K is provided in Table 6. The results are based on another 376 500 simulations of regression problems.

Tab. 6. Average RI of d_K and d_{CU} for different CoV of IU^*

CoV IU^*	d_{CU}	$d_{0.5}$	d_1	d_9	d_{49}	d_{100}	d_{200}
0-1	7,8%	6,8%	6,2%	1,9%	0,4%	0,2%	0,1%
1-2	7,2%	6,5%	6,7%	1,7%	0,4%	0,2%	0,1%
2-4	6,9%	2,6%	6,7%	2,6%	0,7%	0,4%	0,2%
4-8	11,3%	-19,4%	-6,9%	2,3%	1,4%	0,9%	0,5%
8-20	7,2%	-70,1%	-38,3%	-3,6%	0,0%	0,2%	0,2%
20-50	2,2%	-123,8%	-63,7%	-5,9%	-0,6%	-0,2%	0,0%
over 50	0,3%	-234,3%	-118,2%	-14,7%	-2,0%	-0,8%	-0,3%

Note that the estimator d_{CU} has positive average relative improvement for all classes of values of the uncertainty index.

5 Concluding Remarks

On the base of the performed computer simulations we can judge that the function IU_{111} is a good indicator of the uncertainty of the prior information. In a case where we do not know the value of R_p^2 the index can be replaced by IU_{110}, see Tables 2,3, and 4. The amount of uncertainty introduced into regression estimation should be an increasing function of the uncertainty index. A good proposal for the function (called coefficient of uncertainty) is the function given by (3). We should stress however, that our results were obtained under the loss given by Euclidean norm: $L(d, \beta) = \parallel d - \beta \parallel$ and when the distributions of disturbances were normal. The studies should be carried on to determine the form of the coefficient of uncertainty when the criterion of performance is given by other loss functions (eg. quadratic) or for other than normal distributions. It should be also verified whether in such cases the index IU^* is still correlated well enough with the results of estimation.

References

[1] J. Berger, *A robust generalized bayes estimator and confidence region for a multivariate normal mean* Ann. Statist. 8 (1980), s. 716-761,

[2] A. Grzybowski *On uncertainty classes and minimax estimators in the linear regression models with heteroscedasticity and correlated errors*, Folia Oeconomica Acta Universitatis Lodziensis, Lodz, to appear,

[3] A. Grzybowski *Simulation analysis of some regression estimators incorporating prior information*, Proceedings of the 14^{th} International Workshop on Statistical Modelling, Graz (Austria), July 19-23, 1999, pp 547-550,

[4] P. Stahlecker, G. Trankler, (ed.)., Acta Applicandae Mathematicae, vol. 43 No 1, 1996, Special Issue on Minimax Estimation,