A C T A U N I V E R S I T A T I S L O D Z I E N S I S

FOLIA OECONOMICA 68, 1987

Halina Klepacz*

THE EFFICIENCY OF ESTIMATION METHODS FOR MODELS WITH ERRORS IN EXPLANATORY VARIABLES

1. Introduction

Although the estimation problem for models with observation errors appeared already in the thirties it has been reconsidered very rarely in econometrics. The first significant works in this field were those by R. F r i s c h [3], T. K o o p m a n s [6], and D. L i n d l e y [7]. Unfortunately, the methods proposed by these authors required information that was usually not available and computations that were complicated and laborious.

Models with errors in variables gained importance when economists started to deal with certain economic phenomena, e.g. consumption, investments, production etc. Relations describing these phenomena very often contain non-observable variables; their non-observability comes either from the character of these variables, or from measurement errors which can appear. In both cases a model with errors in variables can be used.

2. Estimation methods for models with observation errors

The problem of deriving and applying estimation methods for models with observation errors has been explored by M. Bartlett, J. Durbin, J. Johnston, I. Kmenta, E. Malinvaud, A. Wald, A. Zellner and others (e.g. cf. [5]). These methods can be divided into two groups; one of them comprises:

* Lecturer Institute of Econometrics and Statistics University of Łódź.

- methods based on the maximum likelihood principle, under the assumption of knowing the variance and covariance matrix of measurement errors and independence of these errors from the random component of the model;

- the revised method of least squares that includes a relation between the measurement error covariance and the non-observable explanatory variables variance.

The other group includes methods based on the instrumental variables method. These methods eventually lead to a specific grouping of variables.

We shall briefly characterize the mentioned estimation methods (e.g. cf. [1] and [3]). Taking into account the brevity of our presentation we shall not derive any formulae; only in Appendix I we shall present analytical forms of estimators of structural parameters and their statistical characteristics.

1. The revised method of least squares (RLS), i.e. the modified version of LSM, accounting for the magnitude of structural parameters underestimation in dependence on the variance of measurement errors.

2. The maximum likelihood method (MLM) is derived under the assumption that non-observable variables have multi-dimensional normal distribution known moments. They are determined by using parameters of the distribution of variables with observation errors and parameters of the distribution of errors themselves.

3. The instrumental variables method (IVM) is an estimation method constructed for models in which explanatory (random) variables are correlated with the model's random component. A basic difficulty in the application of this method is a choice of proper instrumental variables that are uncorelated with non-observable random components.

The possibility of choosing different variables as the so called "instruments" underlies the formulation of some other estimation methods based on IVM. They are:

a) Wald grouping method in which the observation set for the variable with a measurement error is divided into two subsets; we determine proper averages for each of them and we construct a straight line passing through them;

b) Bartlett method, which is a generalization of the Wald method; the observation set is divided into three subsets, and only the last two of them (the first and the last one?) are used further on.

4. The method proposed by M. F e l d s t e i n [2] is a combination of two estimation methods: the method of least squares and the instrumental variables method. The estimator of the parameter standing at the variable with measurement error is determined as a convey linear combination of the LSM and IVM estimators. Properties of methods presented here will be studied by means of a Monte Carlo experiment for properly constructed sample spaces.

3. Construction of sample data for our numerical experiment

For a given set XT_i, $i = 1, \ldots, n$ and for given values of the parameters α_0 and α_1 we determine such theoretical values YT_i of the variable Y that

$$(1) \qquad YT_i = \alpha_0 + \alpha_1 XT_i.$$

The sample values Y_i are the sum of the values of YT_i and random disturbance i generated form the normal distribution $N(0, \sqrt{(\frac{1}{R^2} - 1) \cdot S^2(YT)})$, where R^2 is the square of the correlation coefficient between Y and YT, and $S^2(YT)$ is the variance of Y from the sample. Eventually, for the sample size n we repeat the generation IP times in order to obtain IP replications of the sample $\{(y_i^{(s)}, xT_i;\ i = 1, \ldots, n;\ s = 1, \ldots, IP\}$.

By assumption, the variable XT is non-observable, so - instead of the value of XT - we observe the values of the variable X, as the sum of XT and the disturbance term

$$X_i = XT_i + V_i,$$

where V_i is a random variable with normal distribution $N(0, \sigma_V)$ and $\sigma_V = \sqrt{RB \cdot S^2(YT)}$ which is equivalent to the assumption that

the share of the variance of error in the second central moment of the variable XT is equal to RB·100% ($RB = \frac{\sigma_V^2}{S^2(XT)}$).

Then, just like in the case of Y, we make IP samplings of n realizations of errors V. We obtain the following realizations of the samples:

$$\left\{(x_i^{(s)}, xT_i) \quad i = 1, \ldots, n; \quad s = 1, \ldots, IP\right\}.$$

Finally we obtain the sample space with the levels of R^2 and RB determined in a given experiment:

$$\left\{(y_i^{(s)}, x_i^{(2)}), \quad i = 1, \ldots, n; \quad s = 1, \ldots, IP\right\}.$$

Using this space we determine IP-element sequences of estimates of parameters of the model

$$Y_i = \alpha_o + \alpha_1 X_i + (\varepsilon_i - \alpha_1 V_i),$$

by means of the methods: LSM, RLS, MLM, Wald IVM, Bartlett IVM, Durbin IVM.

The sequences of estimates $a_o^{(s)}$, $a_1^{(s)}$ will be used in determining the following characteristics: average values of estimates from IP replications, standard deviations for the estimates from the sample, variability coefficients, magnitudes of bias of average values of estimates, the spread of estimates with regard to the actual value of a parameter, and the measures of skewness and kurtosis. The respective characteristics obtained from different methods are compared to one another and to the results obtained by means of a standard method, which will be the ordinary method of least squares calculated for the sample values

$$\left\{(y_i^{(s)}, xT_i), \quad i = 1, \ldots, n; \quad s = 1, \ldots, IP\right\},$$

satisfying the ordinary assumptions that the explanatory is observed without errors and is not random.

4. Numerical realization of Monte Carlo experiments

Programme #HAS 1, according to the diagram 1, has been done for the numerical realization of Monte Carlo experiments. This programme is now available from the Programmes Library of the Institute of Econometrics and Statistics of the University of Łódź.

In all experiments we have assumed that $\alpha_0 = 1000$, $\alpha_1 = 2$; the values of XT have been selected from 4-digit tables of random numbers; we have taken sample sizes n = 20, 30, 40, 50, the number of replications being IP = 5, 10, 15, ..., 495, 500. The following levels have been assumed: R^2 = 0,99; 0.95; 0.90; 0.85; RB = 0.01; 0.05; 0.10; 0.15.

5. An analysis of the results of the Monte Carlo experiments

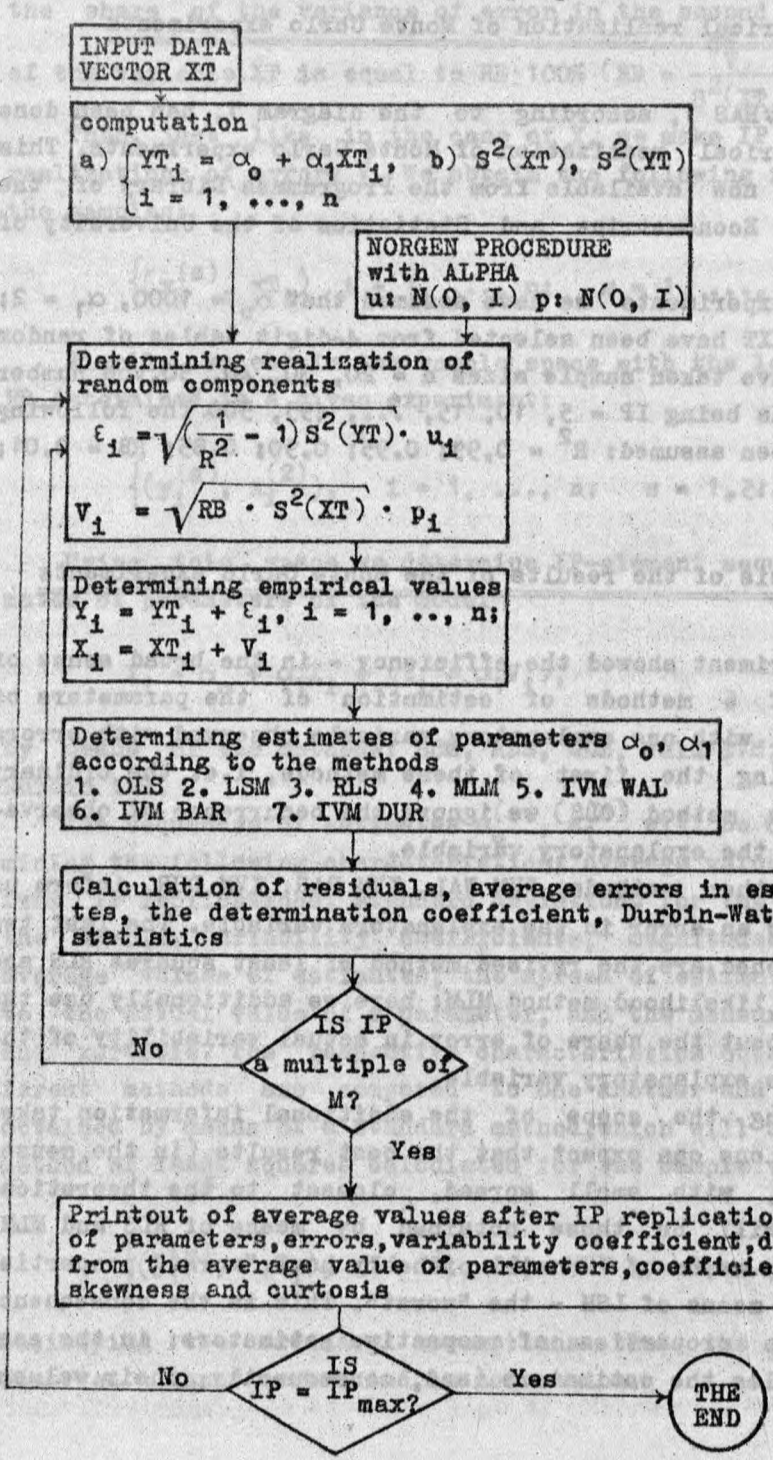
Our experiment showed the efficiency - in the broad sense of the word - of 6 methods of estimation of the parameters of a linear model with one explanatory variable observed with errors.

While using the first of these methods, i.e. the ordinary least squares method (OLS) we ignore the occurrence of observation error in the explanatory variable.

Three further methods, IVM WAL, IVM BAR, IVM DUR, inform us that there is an error in the explanatora variable. The last two to be mentioned are the revised method of least squares RLS and the maximum likelihood method MLM; here we additionally use the information about the share of error in actual variability of the non-observable explanatory variable.

Considering the scope of the additional information taken into account one can expect that the best results (in the sense: least biased, with small spread, closest to the theoretical parameters) will be those obtained by means of RLS and MLM; the results by means of IVM will probably have "worse" properties, and those by means of LSM - the "worst". This is the consequence of asymptotic properties of respective estimators; in the case of small samples the estimators (and, consequently, their values) are biased.

Block diagram of the programme # HAS1



Hence, we are interested in finding answers to the following questions:

1. How much a researcher using a given estimation method can gain compared to LSM?

2. Whether the methods allowing for the magnitude of observation error are, in the case of small samples, better that the others and how much?

3. What the sign of the bias is?

4. How the magnitude of bias changes with regard to the increase of sample size?

The comparison of methods will be done with regard to:

1. The level of the determination coefficient.

2. The sample size.

3. The level of observation error.

No significant differences between the estimate of the parameter α_1 obtained for different levels of R^2 for a given method have been found[1]. The differences in average estimates of a parameter (the magnitude of bias with relation to the actual value of the parameter) are the consequence of the properties of the generated samples, which can be best observed for this parameter's average estimates obtained by the standard method. For instance for $R^2 = 0.90$ we obtain an overestimated estimate of α_1 by the standard method, and BAR consequently in the methods LSM, IVM WAL, IVM BAR, IVM DUR underestimations are smaller, and is RLS and MLM overestimations are smaller than in the case of $R^2 = 0.99$ where $\bar{a}_1 = 1.999$ obtained by the standard was underestimated. In Tab. 1 we present the results for RB = 0.10 and n = 20, as an illustration of the interrelations among the estimates of α_1 obtained by different methods, with the changing levels of R^2.

For the other studied levels of RB these relations are similar.

We can observe small differences in the values of average estimates of the parameter α_1 with relation to the sample size.

We can clearly see that the bias of average estimates for RLS and MLM decreases with the increase of the sample size. For

---

[1] Usually the intercept in a linear model is economically well interpreted, so its analysis is omitted here.

Table 1

Average estimates of the parameter α_1 obtained for RB = 0.10, n = 20 from IP = 500 repetitions

Method \ R^2	0.99	0.95	0.90
LSM	1.825	1.824	1.843
RLS	2.027	2.026	2.030
MLM	2.036	2.034	2.034
IVM WAL	1.853	1.851	1.870
IVM BAR	1.852	1.854	1.875
IVM DUR	1.847	1.845	1.864
Standard method	1.999	1.997	2.002

samples of 20 elements when RB = 0.10 and R^2 = 0.99, the parameter α_1 is overestimated by 1.5% in the average, whereas for samples of 30 or 40 elements the overestimation is about 0.5%, which is the consequence of the consistency of these estimators.

Estimates obtained by means of LSM, regardless of the sample size, are usually underestimated (for n = 20, 30, 40, 50) by 8.5% with RB = 10% (their asymptotic bias, corresponding to the value of RB, is greater). We can conclude that the "correction" for the estimate of the parameter in MLM and RLS should depend not only on the quantity of RB, but also on the sample size.

Estimates obtained by means of instrumental variables methods are underestimated for all studied sample sizes and levels of RB and R^2. This underestimation is relatively smaller than that from LSM, but the differences are not statistically significant as with relation compared to the estimates obtained by means of the standard method. It is worth noting that the "worst" estimates, in the sense of their bias, are those from IVM DUR, and no significant differences in the scatter or average estimates of parameters have been noticed. An example of the results is given in Tab. 2.

Interesting dependences can be observed for the estimates obtained by means of the suggested presented methods in depen-

Table 2

Average estimates of the parameter α_1 with IP = 500, for $R^2 = 0.99$ and RB = 0.10

Method \ n	20	30	40	50
LSM	1.825	1.832	1.828	1.830
RLS	2.027	2.015	2.011	2.013
MLM	2.035	2.018	2.013	2.017
IVM WAL	1.853	1.867	1.872	1.868
IVM BAR	1.852	1.848	1.848	1.858
IVM DUR	1.847	1.849	1.840	1.839
Standard method	1.999	2.000	2.000	1.999

dence on pre-determined levels of RB, i.e. the share of the variance of measurement error of the non-observable variable in its variance.

Generally speaking, without taking the changes in R^2 and n into consideration, we have found out that the average estimates of the parameter α_1 corresponding to the subsequent levels of RB = 1%, 5%, 10%, 15% obtained by means of

- LSM are underestimated by: 1%, 4%, 8%, 12%, respectively;
- RLS and MLM are overestimated by: less than 0.5%, 1%, 3-4% and 4-5%, respectively;
- IVM WAL and IVM BAR are underestimated by: 0.5%, 3-3.5%, 7-7.5%, 10%, respectively;
- in the case of IVM DUR the underestimation is in all cases 0.5% greater than the values obtained by means of IVM WAL and IVM BAR.

Average estimates for the case when $R^2 = 0.99$ and n = 20 are presented in Tab. 3.

The analysis presented here concerns the results of IP = 500 replications. In the statistical sense such a sample is large, but in Monte Carlo experiments we cannot consider it as large (in the sense that the results of these experiments depend on quasi-random numbers generators (cf. [8]) and are conducted for very

Table 3

Average estimates of the parameter $\alpha_1 = 2$ obtained for IP = 500, $R^2 = 0.99$ and n = 20

Method \ RB	0%	1%	5%	10%	15%
LSM	1.999	1.981	1.912	1.825	1.772
RLS	1.999	2.001	2.012	2.027	2.046
MLM	1.999	2.001	2.014	2.036	2.054
IVM WAL	1.999	1.997	1.928	1.853	1.801
IVM BAR	1.998	1.992	1.938	1.853	1.797
IVM DUR	1.999	1.986	1.926	1.846	1.795
Standard method	1.999	1.999	1.999	1.999	1.999

large numbers). However, regarding the costs of calculations not more than 500 repetitions were made.

In order to ilustrate the fact that 500 repetitions is a sufficient quantity, we shall present 4 diagrams of the average estimates of the parameter α_1 after IP replications (IP = 5, 10, ..., 495, 500). All the studied measures and statistics stabilize their behaviour already after 100 replications i.e. there are no differences between them in significat decimal points.

6. Conclusions

We can conclude from the results just presented that the application of IVM methods has little advantage over the LSM in the sense of the decrease of bias. The advantage does not increase with the increase of sample size.

In the case of RLS and MLM, however, the bias is much smaller in small samples, and probably it could be decreased by means of introducing a correction connected with the sample size to proper estimators. These methods assume that we know the share of the variance of measurement error oi the non-observable explanatory variable in its variance. In practice, however, this information

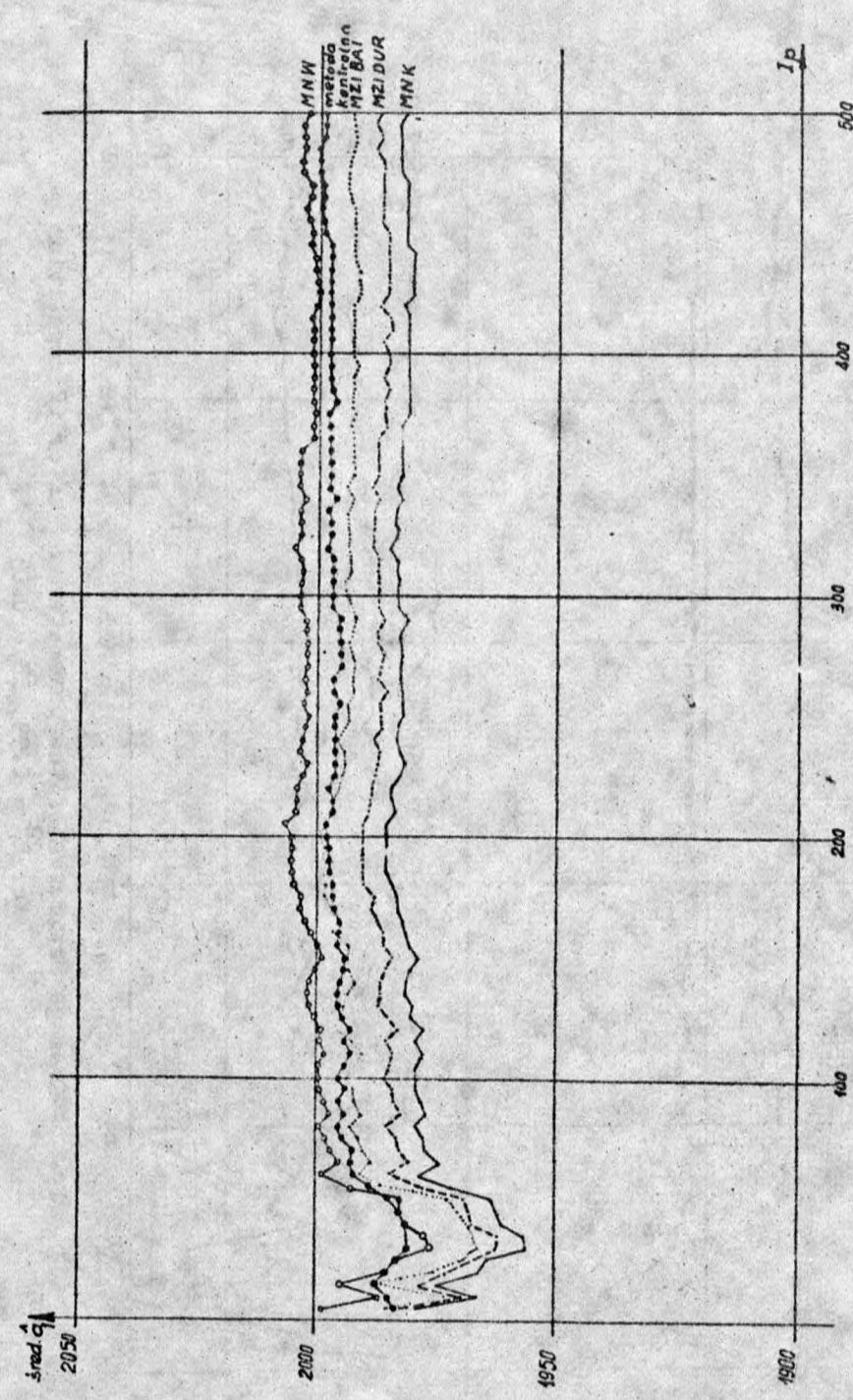


Fig. 1. Average estimates of the parameter α_1 after IP repetitions for $R^2 = 0.99$, RB = 0.01

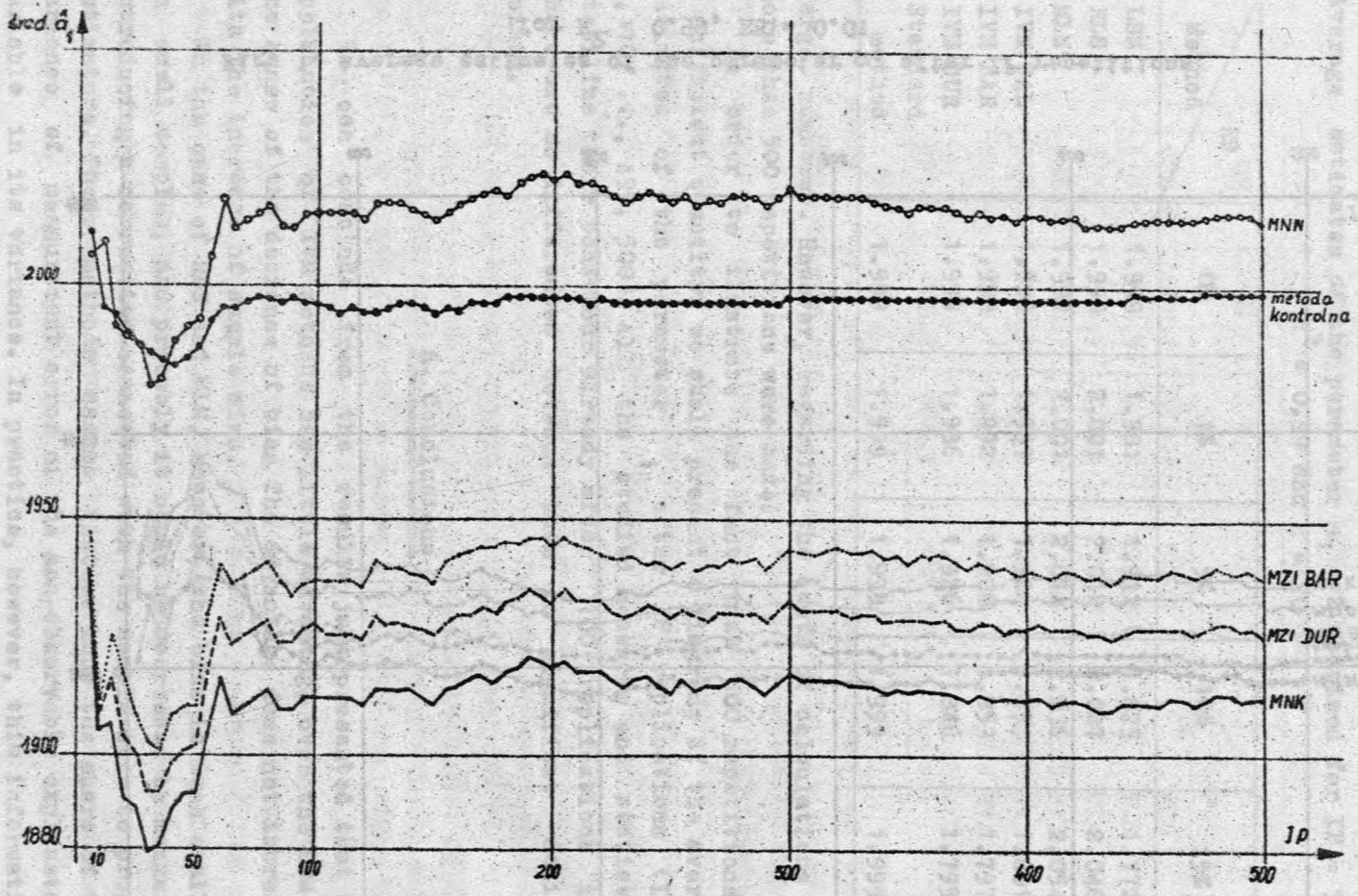


Fig. 2. Average estimates of the parameter α_1 after IP repetition for $R^2 = 0.99$, RB = 0.05

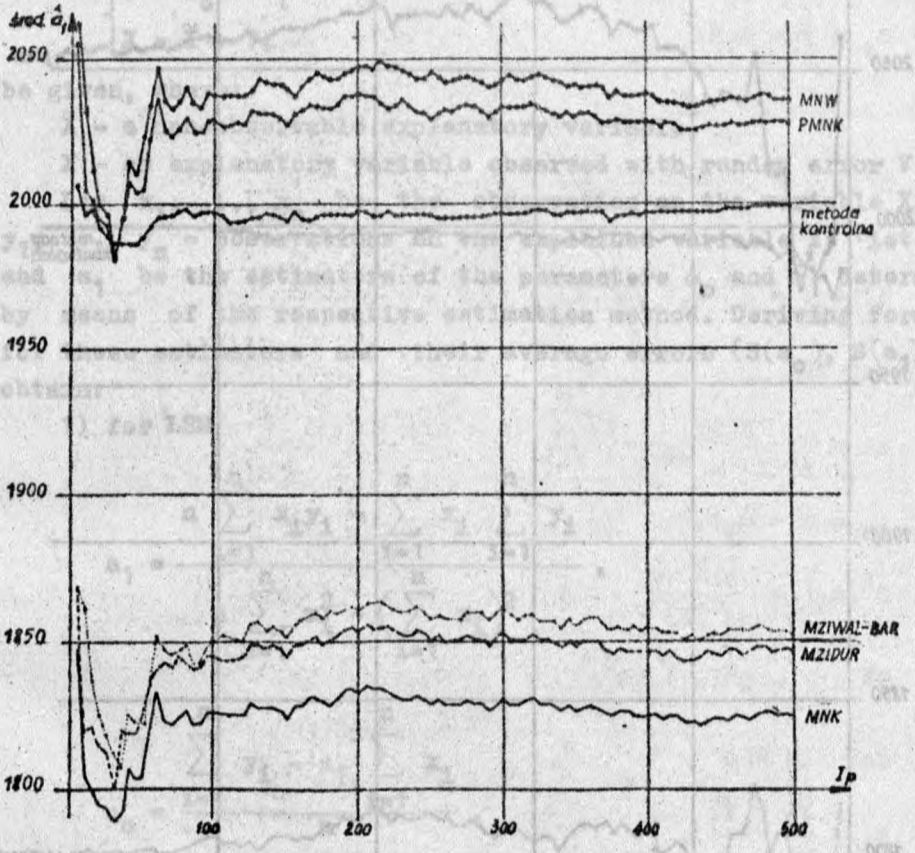


Fig. 3. Average estimates of the parameter α_1 after IP repetitions for $R^2 = 0.99$, RB = 0.10

need not be accurate. Hence it might be a good idea to evaluate these methods when the information concerning RB is not discrete but continuous, which means equal up to this discrete value some intervale.

From all methods given in our bibliography, the Feldstein method has not been studied. In this method the estimator of the parameters of the model is the average weighted estimator LSM and IVM. As both of these estimators are biased in the same direction, its properties can be inferred from the separate analyses of these two estimators.

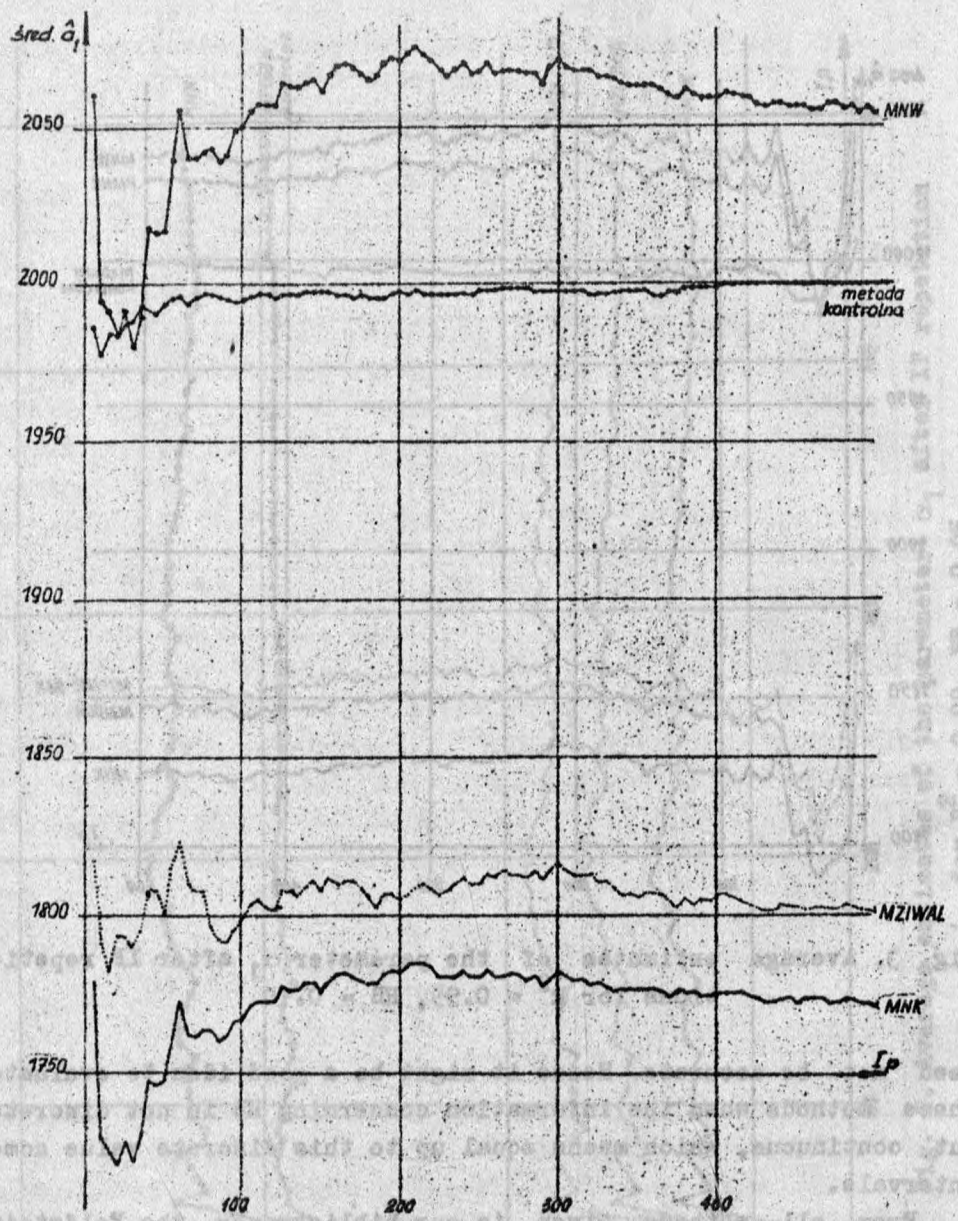


Fig.4. Average estimates of the parameter α_1 after IP repetitions for $R^2 = 0.99$, RB = 0.15

Appendix. Let a model

$$\begin{cases} Y = \alpha_0 + \alpha_1 X + \varepsilon, \\ X = \tilde{X} + V, \end{cases}$$

be given, where:

$\tilde{X}$ - a non-observable explanatory variable,

X - an explanatory variable observed with random error V.

Let $x_1, \ldots, x_n$ be the observation on the variable X, and $y_1, \ldots, y_n$ - observations on the explained variable Y; let a_0 and a_1 be the estimators of the parameters α_0 and α_1 determined by means of the respective estimation method. Deriving formulae for these estimators and their average errors ($S(a_0)$, $S(a_1)$) we obtain:

1) for LSM

$$a_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2},$$

$$a_0 = \frac{\sum_{i=1}^{n} y_i - a_1 \sum_{i=1}^{n} x_i}{n},$$

$$S^2(a_1) = \frac{Se^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2},$$

$$S^2(a_0) = \frac{Se^2 \sum_{i=1}^{n} x_i^2}{n\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)};$$

2) for RLS

$$a_1 = a_1(\text{MNK})\left(1 + \frac{S^2(V)}{S^2(\tilde{X})}\right),$$

$$a_o = \bar{y} - a_1\bar{x},$$

$$S^2(a_1) = \frac{\sum_{i=1}^{n} e_1^2}{n(n-2)\,S^2(X)},$$

$$S^2(a_o) = S^2(a_1)\,(S^2(X) + (\bar{x})^2);$$

3) for MLM

$$a_1 = \frac{S(XY)}{S^2(X) - S^2(V)},$$

$$a_o = \bar{y} - a_1\bar{x},$$

$$S^2(a_1) = \frac{Se^2}{nS^2(X)},$$

$$S^2(a_o) = S^2(a_1)\,(S^2(X) + (\bar{x})^2);$$

4) for IVM WAL

$$a_1 = \frac{\bar{y}_2 - \bar{y}_1}{\bar{x}_2 - \bar{x}_1},$$

$$a_o = \bar{y} - a_1\bar{x},$$

$$S^2(a_1) = \frac{4\,S_e^2}{n(\bar{x}_2 - \bar{x}_1)^2},$$

$$S^2(a_o) = \frac{Se^2}{n}\left(1 + \left(\frac{\bar{x}}{\bar{x}_3 - \bar{x}_1}\right)^2\right);$$

5) for IVM BAR

$$a_1 = \frac{\bar{y}_3 - \bar{y}_1}{\bar{x}_3 - \bar{x}_1},$$

$$a_0 = \bar{y} - a_1\bar{x},$$

$$S^2(a_1) = \frac{4S_e^2}{n\,\bar{x}_3 - \bar{x}_1^{\,2}},$$

$$S^2(a_0) = \frac{Se^2}{n}\left(1 + \frac{\bar{x}}{\bar{x}_3 - \bar{x}_1}\right)^2;$$

6) for IVM DUR

$$a_1 = \frac{2n\sum_{i=1}^{n} iy_i - n(n+1)\sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} ix_i - n(n+1)\sum_{i=1}^{n} x_i},$$

$$a_0 = \frac{\left(\sum_{i=1}^{n} ix_i \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right)}{2n\sum_{i=1}^{n} ix_i - n(n+1)\sum_{i=1}^{n} x_i},$$

$$S^2(a_1) = \frac{n^2 - 1}{12\,n} \cdot \frac{Se^2\bar{x}}{\left(\frac{\sum_{i=1}^{n} ix_i}{n} - \frac{n+1}{2}\bar{x}\right)^2},$$

$$S^2(a_0) = \left(\frac{1}{n} + S^2(a_1)\right) S_e^2.$$

In all these formulae S_e^2 denotes a residual variance, $\bar{x}$, $\bar{y}$ denote arithmetic means of the sample values of X and Y respectively, and $\bar{x}_1$, $\bar{x}_2$, $\bar{x}_3$; $\bar{y}_1$, $\bar{y}_2$, $\bar{y}_3$ are respective group averages.

Bibliography

[1] W e l f e W., (ed.) (1977): Ekonometryczne modele rynku, Vol. 1, Warszawa.

[2] F e l d s t e i n M. (1974): Errors in Variables: A Consistent Estimator with Smaller MSE in Finite Simples, "Journal of the American Statistical Association", 69(343), p. 990-996.

[3] F r i s c h R. (1934): Statistical Confluence Analysis by Means of Complete Regression Systems, University Institute of Economics, Oslo.

[4] G o l d b e r g e r A. S. (1975): Teoria ekonometrii, Warszawa.

[5] K l e p a c z H. (1984): Przegląd metod estymacji modeli jednorównaniowych z błędami w zmiennych, "Zeszyty Naukowe Akademii Ekonomicznej", 181, Kraków, p. 81-103.

[6] K o o p m a n s T. C. (1936): Linear regression analysis of economic time series, Haarlem.

[7] L i n d l e y D. V. (1947): Regression Lines and the Linear Functional Relationship, "Journal of the Royal Statistical Society", B.

[8] Z i e l i ń s k i R. (1979): Generatory liczb losowych. Programowanie i testowanie na maszynach cyfrowych, WNT, Warszawa.

Halina Klepacz

EFEKTYWNOŚĆ METOD ESTYMACJI MODELI Z BŁĘDAMI W ZMIENNYCH OBJAŚNIAJĄCYCH

W artykule przeanalizowano wielkości średnich obciążeń ocen parametru kierunkowego modelu z jedną zmienną objaśniającą w zależności od liczebności próby, poziomów: współczynnika determinacji, wariancji błędu pomiaru, ilości powtórzeń itp. Parametr kierunkowy estymowano sześcioma metodami: najmniejszych kwadratów, "poprawioną" metodą najmniejszych kwadratów, metodą największej wiarygodności oraz trzema metodami zmiennych instrumentalnych: Walda, Bartletta i Durbina. Ogólnie stwierdzono, bez uwzględniania zmian R^2 i n,, że oceny średnie parametru α_1 odpowiadające kolejnym poziomom RB dla MNK i metod zmiennych instrumentalnych są niedoszacowane, zaś dla "poprawionej" metody najmniejszych kwadratów i MNW są przeszacowane.