Toutenburg, Shalabh:

Improving the Estimation of Incomplete Regression Models through Pilot Investigations and Repeated Studies

Sonderforschungsbereich 386, Paper 154 (1999)

Online unter: http://epub.ub.uni-muenchen.de/

Improving the Estimation of Incomplete Regression Models through Pilot Investigations and Repeated Studies

H. Toutenburg [*] Shalabh [**]

18th May 1999

Abstract

Pilot investigations and repeated studies often provide some useful information which can be utilized for the estimation of coefficients in a linear regression model when some observations on the study variable are missing. A suitable framework for this purpose is described and several unbiased estimators for the coefficient vectors are presented. Their efficiency properties are analyzed and a comparison is made.

1 Introduction

Quite often some pilot investigations are carried out to gather some preliminary information before launching the main study. Such pilot investigations may not be required when the same or similar studies are conducted repeatedly and regularly at various points of time. In both the cases, the statistical analyses may provide some potential and useful information about the parameters which can be fruitfully employed in the statistical analysis of current data. Use of such prior information, it is well documented, yields generally more efficient inferences under Bayesian as well non-Bayesian frameworks.

In the context of regression analysis, the pilot investigations may provide unbiased estimates of some or all the regression coefficients along with their standard errors. Same experiments conducted simultaneously at different stations under the same protocol may also provide reliable information of this kind. Similarly, estimates of some coefficients and/or few ratios of some coefficients and/or some linear combinations of coefficients may exhibit considerable stability in repeated studies. Similar investigations by other researchers and the knowledge acquired through experience and long association may also serve as a potential source for this kind of prior information in the form of a set of stochastic linear constraints binding the regression coefficients.

When the prior information specifies unbiased estimates of some linear combinations of regression coefficients, the technique of mixed regression estimation introduced by Theil and Goldberger (1961) provides improved estimators of

[*] Institut für Statistik, Universität München, 80799 München, Germany

[**] Department of Statistics, Panjab University, Chandigarh, India

the regression coefficients; see, e.g.,Srivastava (1980) for annotated bibliography of earlier work and Judge, Griffiths, Hill, Lütkepohl and Lee (1985), Rao and Toutenburg (1995) and Toutenburg (1982) for an interesting exposition, extensions and other developments. If we screen the literature dealing with the technique of mixed regression estimation for the utilization of linear stochastic constraints, it may reveal that all the investigations are limited to the situations where there are no missing values in the data. It may, however, be pertinent to mention the reference of Toutenburg, Heumann, Fieger and Park (1995) who have employed the mixed regression framework for the estimation of regression parameters when some observations on an explanatory variable are missing but no prior information related to coefficients is available. This has motivated us to study the role of prior information in the improved estimation of coefficients when some observations on the study variable are missing.

The plan of this paper is as follows. In Section 2, we describe the model and discuss the estimation of regression coefficients. Their efficiency properties are analyzed in Section 3 while the effect of missing observations is studied in Section 4. Finally, some concluding remarks are placed in Section 5.

2 The Model and Estimators

Let us consider a linear regression model in which there are n_c complete and n_m incomplete observations.

If y_c is a $n_c \times 1$ vector of n_c observations on the study variable, X_c is a $n_c \times K$ full column rank matrix of n_c observations on K explanatory variables, β is the column vector of regression coefficients and ϵ_c is a $n_c \times 1$ vector of disturbances, we can write

$$y_c = X_c\beta + \epsilon_c. \tag{2.1}$$

Similarly, if y_{mis} denotes a $n_m \times 1$ vector of missing observations on the study variable, X_m is a $n_m \times K$ matrix (not necessarily of full column rank) of n_m observations on the explanatory variables and ϵ_m is a $n_m \times 1$ vector of n_m disturbances, we have

$$y_{mis} = X_m\beta + \epsilon_m. \tag{2.2}$$

It is assumed that the elements of ϵ_c and ϵ_m are independently and identically distributed with mean zero and variance σ^2.

In addition, we are given unbiased estimates of a set of linear combinations of regression coefficients. As these are assumed to have been obtained from pilot studies and/or repeated studies, we can express the prior information as follows:

$$r = R\beta + \epsilon \tag{2.3}$$

where the $J \times 1$ vector r and $J \times K$ matrix R contain known elements and ϵ is a $J \times 1$ random vector with null mean vector and $\sigma^2\Sigma^{-1}$ variance covariance matrix in which the elements of Σ are known.

As prior information is independent of the sample observations, we assume that ϵ is stochastically independent of ϵ_c and ϵ_m.

When we ignore the prior information and use only the complete observations, the least squares estimator of β is given by

$$b_c = (X_c'X_c)^{-1}X_c'y_c. \tag{2.4}$$

If we incorporate the prior information and discard the incomplete observations, the technique of mixed regression estimation provides the following estimator of β:

$$b_{MR} = (X_c'X_c + R'\Sigma R)^{-1}(X_c'y_c + R'\Sigma r). \tag{2.5}$$

On the other hand, if we ignore the prior information and utilize the entire set of observations, the estimator of β is given by

$$\tilde{\beta}^* = (X_c'X_c + X_m'X_m)^{-1}(X_c'y_c + X_m'y_{mis}). \tag{2.6}$$

Such an estimator has no utility owing to lack of knowledge of y_{mis}. A popular practice is to replace the missing observations by their predicted values such as $X_m b_c$ and $X_m b_{MR}$; see, e.g., Toutenburg and Shalabh (1996) for the predictive performance. This proposition yields the following two estimators of β:

$$\begin{aligned} b_1 &= (X_c'X_c + X_m'X_m)^{-1}(X_c'y_c + X_m'X_m b_c) && (2.7)\\ &= b_c \\ b_2 &= (X_c'X_c + X_m'X_m)^{-1}(X_c'y_c + X_m'X_m b_{MR}) && (2.8)\\ &= (X_c'X_c + X_m'X_m)^{-1}(X_c'X_c b_c + X_m'X_m b_{MR}) \end{aligned}$$

We thus observe that b_2 is a matrix weighted average of the estimators b_c and b_{MR}.

Finally, if we write (2.1), (2.2) and (2.3) compactly and apply the method of generalized least squares, we find the following etimator of β:

$$\hat{\beta}^* = (X_c'X_c + X_m'X_m + R'\Sigma R)^{-1}(X_c'y_c + X_m'y_{mis} + R'\Sigma r) \tag{2.9}$$

which again does not serve any useful purpose due to involvement of missing observations.

Replacing the missing observations by their predicted values, we obtain the following feasible versions of (2.9):

$$\begin{aligned} b_3 &= (X_c'X_c + X_m'X_m + R'\Sigma R)^{-1}(X_c'y_c + X_m'X_m b_c + R'\Sigma r) && (2.10)\\ &= (X_c'X_c + X_m'X_m + R'\Sigma R)^{-1}\left[X_m'X_m b_c + (X_c'X_c + R'\Sigma R)b_{MR}\right] \\ b_4 &= (X_c'X_c + X_m'X_m + R'\Sigma R)^{-1}(X_c'y_c + X_m'X_m b_{MR} + R'\Sigma r) && (2.11)\\ &= (X_c'X_c + X_m'X_m + R'\Sigma R)^{-1}\left[X_m'X_m b_{MR} + (X_c'X_c + R'\Sigma R)b_c\right] \end{aligned}$$

From the above expressions, we observe that both the estimators are matrix weighted averages of b_c and b_{MR}. Further, the weighting matrices of b_c and b_{MR} in one estimator are interchanged in the other estimator.

Thus we observe that the estimator b_c utilizes neither the incomplete observations nor the prior information. When incomplete observations are used but the prior information is not incorporated, no improvement is achieved and the estimator remains b_c. Such is, however, not the case when incomplete observations are discarded and prior information is incorporated. Then we get the estimator b_{MR} which is different from b_c. Finally, when both the incomplete observations and the prior information are utilized simultaneously, we get three estimators b_2, b_3 and b_4 which are incidentally found to be matrix weighted averages of b_c and b_{MR}.

3 Comparison of Estimators

It is easy to see from (2.1) and (2.3) that all the five estimators, viz., b_c, b_{MR}, b_2, b_3 and b_4 are unbiased for β.

The variance covariance matrices of b_c and b_{MR} are given by

$$\begin{aligned} \mathrm{V}(b_c) &= \mathrm{E}(b_c-\beta)(b_c-\beta)' \qquad (3.1)\\ &= \sigma^2(X_c'X_c)^{-1}\\ \mathrm{V}(b_{MR}) &= \mathrm{E}(b_{MR}-\beta)(b_{MR}-\beta)' \qquad (3.2)\\ &= \sigma^2(X_c'X_c+R'\Sigma R)^{-1}\,. \end{aligned}$$

Using the result in Appendix and writing

$$\begin{aligned} \Delta &= (X_c'X_c)^{-1}-(X_c'X_c+R'\Sigma R)^{-1} \qquad (3.3)\\ &= (X_c'X_c)^{-1}R'\Sigma R(X_c'X_c+R'\Sigma R)^{-1}\\ &= (X_c'X_c+R'\Sigma R)^{-1}R'\Sigma R(X_c'X_c)^{-1} \end{aligned}$$

it can be easily seen that

$$\begin{aligned} \mathrm{V}(b_2) &= \mathrm{E}(b_2-\beta)(b_2-\beta)' \qquad (3.4)\\ &= \sigma^2(X_c'X_c+R'\Sigma R)^{-1}+\sigma^2G\Delta G'\\ \mathrm{V}(b_3) &= \mathrm{E}(b_3-\beta)(b_3-\beta)' \qquad (3.5)\\ &= \sigma^2(X_c'X_c+R'\Sigma R)^{-1}+\sigma^2(I_K-H)\Delta(I_K-H')\\ \mathrm{V}(b_4) &= \mathrm{E}(b_4-\beta)(b_4-\beta)' \qquad (3.6)\\ &= \sigma^2(X_c'X_c+R'\Sigma R)^{-1}+\sigma^2H\Delta H' \end{aligned}$$

where

$$\begin{aligned} G &= (X_c'X_c+X_m'X_m)^{-1}X_c'X_c \qquad (3.7)\\ H &= (X_c'X_c+X_m'X_m+R'\Sigma R)^{-1}(X_c'X_c+R'\Sigma R). \qquad (3.8) \end{aligned}$$

Comparing b_c with the remaining four estimators, we observe that

$$\begin{aligned} \mathrm{D}(b_c;b_{MR}) &= \mathrm{V}(b_c)-\mathrm{V}(b_{MR}) \qquad (3.9)\\ &= \sigma^2\Delta\\ \mathrm{D}(b_c;b_2) &= \mathrm{V}(b_c)-\mathrm{V}(b_2) \qquad (3.10)\\ &= \sigma^2(\Delta-G\Delta G')\\ \mathrm{D}(b_c;b_3) &= \mathrm{V}(b_c)-\mathrm{V}(b_3) \qquad (3.11)\\ &= \sigma^2\left[\Delta-(I_K-H)\Delta(I_K-H')\right]\\ \mathrm{D}(b_c;b_4) &= \mathrm{V}(b_c)-\mathrm{V}(b_4) \qquad (3.12)\\ &= \sigma^2(\Delta-H\Delta H'). \end{aligned}$$

As Δ is a nonnegative definite matrix and the characteristic roots of the matrices G and H are nonnegative and cannot exceed 1, the matrix expressions (3.9)–(3.12) are nonnegative definite implying the superiority of b_{MR}, b_2, b_3 and b_4 over b_c.

Similarly, if we compare b_{MR} with b_2, b_3 and b_4, it clearly follows from (3.2), (3.4), (3.5) and (3.6) that b_{MR} is superior to all the three estimators b_2, b_3 and b_4.

Next, let us compare b_2 with b_3 and b_4.

From (3.4) and (3.5) we observe that

$$\begin{aligned} \mathrm{D}(b_3; b_2) &= \mathrm{V}(b_3) - \mathrm{V}(b_2) \qquad (3.13) \\ &= \sigma^2 \left[(I_K - H)\Delta(I_K - H') - G\Delta G' \right]. \end{aligned}$$

Suppose that the minimum and maximum characteristic roots are g_{min} and g_{max} for the matrix G and h_{min} and h_{max} for the matrix H. It is then seen that the matrix expression on the right hand side of (3.13) is nonnegative definite as long as

$$(g_{max} + h_{max}) < 1 \qquad (3.14)$$

which is a sufficient condition for the superiority of b_2 over b_3.

On the other hand, the estimator b_3 is better than b_2 so long as the following condition is satisfied

$$(g_{min} + h_{min}) > 1 \qquad (3.15)$$

Similarly, from (3.4) and (3.6), we have

$$\begin{aligned} \mathrm{D}(b_4; b_2) &= \mathrm{V}(b_4) - \mathrm{V}(b_2) \qquad (3.16) \\ &= \sigma^2 (H\Delta H' - G\Delta G'). \end{aligned}$$

(3.17)

As $(G^{-1} - H^{-1}) = \Delta X_m' X_m$ and hence $(H - G)$ are nonnegative definite, the matrix expression (3.16) is also nonnegative definite implying the superiority of b_2 over b_4.

Finally, comparing (3.5) and (3.6), we see that

$$\begin{aligned} \mathrm{D}(b_3; b_4) &= \mathrm{V}(b_3) - \mathrm{V}(b_4) \qquad (3.18) \\ &= \sigma^2 \left[(I_K - H)\Delta(I_K - H') - H\Delta H' \right]. \end{aligned}$$

(3.19)

which is nonnegative definite when all the characteristic roots of H are less than 0.5. This holds true so long as h_{max} is smaller than 0.5 which is a sufficient condition for the superiority of b_4 over b_3.

The reverse is true, i.e., the estimator b_3 is superior to b_4 when all the characteristic roots of H are greater than 0.5. Such a condition is satisfied as long as h_{min} is larger than 0.5.

4 Effect of Missing Observations

Let us now study the effect of the missing observations on the efficiency of estimating β.

Assuming for a moment that no observation is missing, we can interpret the estimator b_c as obtained from a sub-model (2.1). Similarly, b_{MR} is the estimator found from sub-model (2.1) by using the prior information while the estimator $\tilde{\beta}^*$ given by (2.6) uses the whole model (2.1) and (2.2) but ignores the prior information. Simultaneous utilization of whole model and prior information is achieved in the estimator $\hat{\beta}^*$ defined by (2.9).

It is easy to see that $\tilde{\beta}^*$ and $\hat{\beta}^*$ are unbiased with variance covariance matrices as

$$\mathrm{V}(\tilde{\beta}^*) = \sigma^2(X_c'X_c + X_m'X_m)^{-1} \tag{4.1}$$

$$\mathrm{V}(\hat{\beta}^*) = \sigma^2(X_c'X_c + X_m'X_m + R'\Sigma R)^{-1} \tag{4.2}$$

Comparing (4.1) with (3.1) and (4.2) with (3.2), one can clearly appreciate the loss of efficiency in the estimation of β. These losses arise when we have to discard the sub-model (2.2) due to missing observations.

The strategy of repairing the data set through substitution of imputed values in place of missing observations yields the estimators $b_1 = b_c$ and b_2 from $\tilde{\beta}^*$ and the estimators b_3 and b_4 from $\hat{\beta}^*$.

Comparing (3.1) and (3.4) with (4.1) and (3.5) and (3.6) with (4.2), one can get an idea of the losses in efficiency due to repairing of data in order to take into account the sub-model (2.2) whether the prior information is ignored or incorporated.

These comparisons thus highlight the effect of some missing observations and clearly reveal the reduction in the efficiency, which could be substantial at times, of estimating the regression coefficients.

5 Some Concluding Remarks

Assuming the missingness of some observations on the study variable and the availability of some prior information in the form of unbiased estimates of a set of linear combinations of regression coefficients in a linear regression model, we have discussed the estimation of the vector of regression coefficients and have presented six estimators. The first estimator is the traditional least squares estimator b_c that discards the incomplete observations as well as the prior information. The second estimator is the mixed regression estimator b_{MR} which incorporates the prior information but ignores the incomplete observations. In order to take the incomplete observations into account, the data set is repaired by substituting imputed values in place of missing observations. These imputed values are nothing but the predicted values derived from an analysis of complete observations using and not using the prior information. This proposition has provided four estimators b_1, b_2, b_3 and b_4. Incidentally, the estimator b_1 turns out to be identically equal to b_c while the remaining three are found to be the matrix weighted averages of the least squares and mixed regression estimators. Thus we have five distinct estimators in all.

Analyzing the efficiency properties, it is seen that all the five estimators are unbiased. Comparing them with respect to the criterion of variance covariance matrix, it is observed that the least squares estimator is beaten by all the remaining four estimators while the mixed regression estimator beats all the other estimators and emerges as the best choice. For the remaining three estimators, conditions for the superiority of one estimator over the other are obtained. An attractive feature of these conditions is that they are easy to verify in practice.

If we group the estimators as (b_c, b_{MR}) and (b_2, b_3, b_4), then the first group can be regarded as the outcome of discarding the incomplete observations outrightly while the second group can be treated as reflecting the strategy of utilizing the entire set of available observation. From this viewpoint, the repairing of data through the given imputation procedure and then estimating the regression coefficients in the given manner do not seem to bring any gain in efficiency.

Examining the impact of missingness of some observations on the efficiency of estimating the regression coefficients, it is observed that it always leads to loss in efficiency whether one discards the incomplete observations or employs an imputation procedure for repairing the data set.

A general conclusion emerging from our investigations is thus that discarding the incomplete observations and incorporating the prior information is the most successful strategy so far as the estimation of regression coefficients is concerned. Further, our investigations supply the expressions which can be utilized to evaluate the loss in efficiency arising from the use of some alternative strategy in any given application and to judge whether this loss is substantial or not.

Appendix

If β is estimated by

$$\hat{\beta} = Wb_c + (I_K - W)b_{MR}$$

then the variance covariance matrix of $\hat{\beta}$ is given by

$$\mathrm{V}(\hat{\beta}) = \sigma^2(X_c'X_c + R'\Sigma R)^{-1} + \sigma^2 W\left[(X_c'X_c)^{-1} - (X_c'X_c + R'\Sigma R)^{-1}\right]W'$$

where W denotes a $K \times K$ matrix with nonstochastic elements.

Proof: From (2.1) and (2.3), we observe that

$$\begin{aligned}(b_c - \beta) &= (X_c'X_c)^{-1}X_c'\epsilon_c \\ (b_{MR} - \beta) &= (X_c'X_c + R'\Sigma R)^{-1}(X_c\epsilon_c + R'\Sigma\epsilon)\end{aligned}$$

whence we can write

$$(\hat{\beta} - \beta) = W(b_c - \beta) + (I - W)(b_{MR} - \beta)$$

so that

$$\begin{aligned}\mathrm{V}(\hat{\beta}) &= \mathrm{E}(\hat{\beta} - \beta)(\hat{\beta} - \beta)' \\ &= W\,\mathrm{E}(b_c - \beta)(b_c - \beta)'W' \\ &\quad + W\,\mathrm{E}(b_c - \beta)(b_{MR} - \beta)'(I_K - W') \\ &\quad + (I_K - W)\,\mathrm{E}(b_{MR} - \beta)(b_c - \beta)'W' \\ &\quad + (I_K - W)\,\mathrm{E}(b_{MR} - \beta)(b_{MR} - \beta)'(I_K - W').\end{aligned}$$

Observing that

$$\mathrm{E}(b_c - \beta)(b_{MR} - \beta)' = \sigma^2 (X_c' X_c + R' \Sigma R)^{-1}$$

and using (3.1) and (3.2), we obtain the desired expression for the variance covariance matrix of $\hat{\beta}$.

References

Judge, G. G., Griffiths, W. E., Hill, R. C., Lütkepohl, H. and Lee, T.-C. (1985). *The theory and practice of econometrics*, 2 edn, Wiley, New York.

Rao, C. R. and Toutenburg, H. (1995). *Linear Models: Least Squares and Alternatives (corrected second printing, 1997)*, Springer, New York.

Srivastava, V. K. (1980). Estimation of linear single-equation and simultaneous-equation models under stochastic linear constraints: An annotated bibliography, *International Statistical Review* **48**: 79–82.

Theil, H. and Goldberger, A. S. (1961). On pure and mixed estimation in econometrics, *International Economic Review* **2**: 65–78.

Toutenburg, H. (1982). *Prior Information in Linear Models*, Wiley, New York.

Toutenburg, H., Heumann, C., Fieger, A. and Park, S. H. (1995). Missing values in regression: Mixed and weighted mixed estimation, *in* V. Mammitzsch and H. Schneeweiß (eds), *Gauss Symposium*, de Gruyter, Berlin, pp. 289–301.

Toutenburg, H. and Shalabh (1996). Predictive performance of the methods of restricted and mixed regression estimators, *Biometrical Journal* **38**: 951–959.