Tilburg University

Efficiency gains due to using missing data procedures in regression models

Nijman, T.E.; Palm, F.C.

Publication date:
1986

[Link to publication in Tilburg University Research Portal](#)

Citation for published version (APA):

Nijman, T. E., & Palm, F. C. (1986). *Efficiency gains due to using missing data procedures in regression models*. (Research memorandum / Tilburg University, Department of Economics; Vol. FEW 240). Unknown Publisher.

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

70
CBM
R

986-240
7626
1986
240

STY

UNIVERSITEIT
BRABANT

POSTBOX 90153
5000 LE TILBURG
THE NETHERLANDS

* C I N O O 3 4 8 *

DEPARTMENT OF ECONOMICS
RESEARCH MEMORANDUM

FEW 240

EFFICIENCY GAINS DUE TO USING
MISSING DATA PROCEDURES
IN REGRESSION MODELS

Th.E. Nijman
F.C. Palm

November 1986

EFFICIENCY GAINS DUE TO USING MISSING DATA PROCEDURES IN REGRESSION MODELS

Th.E. Nijman *
F.C. Palm **

november 1986

In the chapter on "Economic Data Issues" of the Handbook of Econometrics, Griliches (1986) analyzes the asymptotic variance of an estimator of a regression coefficient using imputations for the missing regressor values and he compares it with that of an estimation procedure based on the complete observations only. His derivation of an expression for the relative efficiency is incorrect. In this note, we give the correct result and show that the relative efficiency of three estimators designed to handle incomplete samples depends on parameters that have a straightforward statistical interpretation. In terms of a gain of asymptotic efficiency, the use of these estimators is equivalent to the observation of a percentage of the values which are actually missing. This percentage depends on three R^2 -measures only, which can be straightforwardly computed in applied work. Therefore it should be easy in practice to check whether it is worthwhile to use a more elaborate estimator.

The authors thank Professor Z. Griliches for his comments on an earlier version of this note.

* Department of Econometrics, Tilburg University, P.O.B. 90153,
5000 LE Tilburg, The Netherlands.

** Department of Economics, University of Limburg, P.O.B. 616,
6200 MD Maastricht, The Netherlands.

Griliches (1986) considers the following regression model

$$y_i = \beta x_i + \gamma z_i + e_i, \quad e_i \sim \text{IN}(0, \sigma^2), \quad (1)$$

and

$$x_i = \delta z_i + v_i, \quad v_i \sim \text{IN}(0, \sigma_v^2), \quad (2)$$

where the regressors x_i and z_i are assumed to be independent of the corresponding disturbances e_i and v_i . Actually, the normality assumption is not made by Griliches, but it will be required below for maximum likelihood (ML) estimation and it does not affect the results for the other estimators. The variables y_i and z_i are observed for $i = 1, \dots, N_1 + N_2$, whereas x_i is observed for $i = 1, \dots, N_1$ only.

Besides the OLS estimator of the regression of y_i on x_i and z_i for $i = 1, \dots, N_1$ only, denoted by $\hat{\beta}_a$ and $\hat{\gamma}_a$, Griliches (1986) considers an estimation procedure in which the missing x_i 's are replaced by $\hat{\delta}_a z_i$, where $\hat{\delta}_a$ is the OLS estimate of δ in (2) using the first N_1 observations. The estimate $\hat{\gamma}_{a+b}$ is subsequently computed by OLS on

$$y_i - \hat{\beta}_a \hat{x}_i = \gamma z_i + w_i, \quad (3)$$

where $w_i = e_i + \theta_i \beta v_i + \theta_i (\beta \delta - \hat{\delta}_a \hat{\beta}_a) z_i$, with $\hat{x}_i = x_i$ and $\theta_i = 0$ if $i \leq N_1$ and $\hat{x}_i = \hat{\delta}_a z_i$ and $\theta_i = 1$ otherwise.

It is straightforward to show that $\hat{\gamma}_{a+b}$ can be alternatively computed by OLS of y_i on $\hat{x}_i$ and z_i

$$y_i = \hat{\beta}_a \hat{x}_i + \gamma z_i + \{e_i + \theta_i \beta v_i + \beta \theta_i (\delta - \hat{\delta}_a) z_i\}. \quad (4)$$

Contrary to what is stated by Griliches (1986), the contribution of $\beta \theta_i (\delta - \hat{\delta}_a) z_i$ to the asymptotic variance of $\hat{\gamma}_{a+b}$ is not negligible if $\text{plim } N_2 N_1^{-1} = \lambda \neq 0$ for $N \rightarrow \infty$, with $N = N_1 + N_2$.

As the three components of the disturbance w_i in (4) are independent, the large sample distribution of $\hat{\gamma}_{a+b}$ is given by

$$\sqrt{N} (\hat{\gamma}_{a+b} - \gamma) \underset{a}{\sim} N(0, V)$$

$$\text{with } V = \text{plim } N (X'X)^{-1} \{X' \Omega X + X' W \beta^2 \Sigma W' X\} (X'X)^{-1}, \quad (5)$$

where X is the matrix of regressors in (4), W is a vector with typical element $\theta_i z_i$, Ω is a diagonal matrix with typical element $\sigma^2 + \theta_i^2 \sigma_v^2$, and Σ is the asymptotic variance of $\hat{\delta}_a$.

After some algebra, we get

$$V = \frac{\sigma^2}{\lambda \sigma_z^2} \{ (1 - r_{xz}^2)^{-1} + \lambda (\mu^{-1} - 2) \}, \quad (6)$$

where $\sigma_z^2 = \text{plim } N^{-1} \sum_{i=1}^N z_i^2$ (assumed to exist), $\mu = \sigma^2 (\beta^2 \sigma_v^2 + \sigma^2)^{-1}$ and r_{xz}^2 is the theoretical R^2 of the regression (2) of x on z . The result in (6) has been obtained by Gouriéroux and Monfort [1981, expression (11) on p. 583].

The relative efficiency of $\hat{\gamma}_{a+b}$ with respect to $\hat{\gamma}_a$ is

$$\text{Eff}(\hat{\gamma}_{a+b}) = \frac{\text{Avar}(\sqrt{N} \hat{\gamma}_{a+b})}{\text{Avar}(\sqrt{N} \hat{\gamma}_a)} = 1 + \lambda (\mu^{-1} - 2) (1 - r_{xz}^2). \quad (7)$$

According to (7), using imputed values as in (3) leads to a gain of efficiency compared with using complete observations only if $\mu > \frac{1}{2}$, which is more stringent than the condition $\mu > \frac{1 - \lambda}{2 - \lambda}$ given by Griliches (1986). Both conditions require that the unpredictable part of x from z is not too important relative to σ^2 , the overall noise level of (1).

$$\text{As } \mu = \frac{1 - r_{yxz}^2}{1 - r_{yz}^2}, \quad (8)$$

where r_{yxz}^2 and r_{yz}^2 denote the theoretical R^2 's of a regression of y on respectively x and z and on z only, it is obvious that a sufficient condition for an efficiency gain is $r_{yxz}^2 < \frac{1}{2}$, i.e. the predictable part of y is small.

As noted by Griliches (1986) and others, an efficiency gain is assured if (4) is estimated by a generalized least squares (GLS) method which

takes the correlation structure of the disturbance in (4) into account. Again, the term $\Theta_i \beta (\delta - \hat{\delta}_a) z_i$ cannot be neglected (see Palm and Nijman (1982) and Nijman and Palm (1985)). Alternatively, the fully efficient ML estimator can be computed, e.g. using the convenient reparametrisation suggested by Gouriéroux and Monfort (1981). From their results, the relative efficiency of the GLS and ML estimators with respect to that of $\hat{\gamma}_a$ can be obtained

$$\text{Eff}(\hat{\gamma}_{\text{GLS}}) = 1 - \lambda\mu(1 - r_{xz}^2) \quad (9)$$

$$\text{and } \text{Eff}(\hat{\gamma}_{\text{ML}}) = 1 - \lambda\mu(1 - r_{xz}^2) - 2\lambda\mu(1 - \mu)r_{xz}^2. \quad (10)$$

The relative efficiency in (7), (9) and (10) only depends on the three magnitudes λ , μ and r_{xz}^2 . Equation (9) indicates that in terms of a gain of asymptotic efficiency, the use of GLS is equivalent to the observation of $100 \mu(1 - r_{xz}^2) \%$ of the values of x_i that are actually missing. Similar expressions can be obtained from (7) and (10) for $\hat{\gamma}_{a+b}$ and $\hat{\gamma}_{\text{ML}}$ respectively. The values in Table 1 illustrate this result.

TABLE 1 Percentage of missing observations that are regained by the use of missing data procedures instead of the complete data only.

$\mu = \frac{1 - r_{yxz}^2}{1 - r_{yz}^2}$	r_{xz}^2	Gain in percentage points for		
		$\hat{\gamma}_{a+b}$	$\hat{\gamma}_{\text{GLS}}$	$\hat{\gamma}_{\text{ML}}$
.3	.2	-106	24	32
.3	.8	- 27	6	40
.6	.2	27	48	58
.6	.8	7	12	50
.9	.2	71	72	76
.9	.8	18	18	32

Note that a good fit in (2) yielding a "good proxy" for the missing values of x_i does not imply that a large part of the missing information on x_i can be recovered, because of the induced multicollinearity between $\hat{x}_i$ and z_i in (4). Especially, when r_{xz}^2 is small, the efficiency gain obtained by using the appropriate estimators can be substantial. The value of μ is crucial for the efficiency of $\hat{\gamma}_{a+b}$. The loss of efficiency

can be important when $\mu < \frac{1}{2}$. This loss increases as r_{xz}^2 decreases. Finally, if μ is close to one, i.e. x_1 is not very important in explaining y in equation (1), all three approaches which take into account the incomplete data, yield about equally efficient estimators.

References

- Gouriéroux, C., and A. Monfort (1981), "On the problem of missing data in linear models", Review of Economic Studies, 48, 579-586.
- Griliches, Z. (1986), "Economic data issues", in Z. Griliches and M.D. Intriligator, eds, Handbook of Econometrics, North Holland, Amsterdam, 1466-1514.
- Nijman, Th.E., and F.C. Palm (1985), "Consistent estimation of a regression model with incompletely observed exogenous variable", Netherlands Central Bureau of Statistics, unpublished paper.
- Palm, F.C., and Th.E. Nijman (1982), "Linear regression using both temporally aggregated and temporally disaggregated data", Journal of Econometrics, 19, 333-343.

IN 1985 REEDS VERSCHENEN

- 168 T.M. Doup, A.J.J. Talman
A continuous deformation algorithm on the product space of unit simplices
- 169 P.A. Bekker
A note on the identification of restricted factor loading matrices
- 170 J.H.M. Donders, A.M. van Nunen
Economische politiek in een twee-sectoren-model
- 171 L.H.M. Bosch, W.A.M. de Lange
Shift work in health care
- 172 B.B. van der Genugten
Asymptotic Normality of Least Squares Estimators in Autoregressive Linear Regression Models
- 173 R.J. de Groof
Geïsoleerde versus gecoördineerde economische politiek in een twee-regiomodel
- 174 G. van der Laan, A.J.J. Talman
Adjustment processes for finding economic equilibria
- 175 B.R. Meijboom
Horizontal mixed decomposition
- 176 F. van der Ploeg, A.J. de Zeeuw
Non-cooperative strategies for dynamic policy games and the problem of time inconsistency: a comment
- 177 B.R. Meijboom
A two-level planning procedure with respect to make-or-buy decisions, including cost allocations
- 178 N.J. de Beer
Voorspelprestaties van het Centraal Planbureau in de periode 1953 t/m 1980
- 178a N.J. de Beer
BIJLAGEN bij Voorspelprestaties van het Centraal Planbureau in de periode 1953 t/m 1980
- 179 R.J.M. Alessie, A. Kapteyn, W.H.J. de Freytas
De invloed van demografische factoren en inkomen op consumptieve uitgaven
- 180 P. Kooreman, A. Kapteyn
Estimation of a game theoretic model of household labor supply
- 181 A.J. de Zeeuw, A.C. Meijdam
On Expectations, Information and Dynamic Game Equilibria

- 182 Cristina Pennavaja
Periodization approaches of capitalist development.
A critical survey
- 183 J.P.C. Kleijnen, G.L.J. Kloppenburg and F.L. Meeuwssen
Testing the mean of an asymmetric population: Johnson's modified T
test revisited
- 184 M.O. Nijkamp, A.M. van Nunen
Freia versus Vintaf, een analyse
- 185 A.H.M. Gerards
Homomorphisms of graphs to odd cycles
- 186 P. Bekker, A. Kapteyn, T. Wansbeek
Consistent sets of estimates for regressions with correlated or
uncorrelated measurement errors in arbitrary subsets of all
variables
- 187 P. Bekker, J. de Leeuw
The rank of reduced dispersion matrices
- 188 A.J. de Zeeuw, F. van der Ploeg
Consistency of conjectures and reactions: a critique
- 189 E.N. Kertzman
Belastingstructuur en privatisering
- 190 J.P.C. Kleijnen
Simulation with too many factors: review of random and group-
screening designs
- 191 J.P.C. Kleijnen
A Scenario for Sequential Experimentation
- 192 A. Dortmans
De loonvergelijking
Afwenteling van collectieve lasten door loontrekkers?
- 193 R. Heuts, J. van Lieshout, K. Baken
The quality of some approximation formulas in a continuous review
inventory model
- 194 J.P.C. Kleijnen
Analyzing simulation experiments with common random numbers
- 195 P.M. Kort
Optimal dynamic investment policy under financial restrictions and
adjustment costs
- 196 A.H. van den Elzen, G. van der Laan, A.J.J. Talman
Adjustment processes for finding equilibria on the simplotope

- 197 J.P.C. Kleijnen
Variance heterogeneity in experimental design
- 198 J.P.C. Kleijnen
Selecting random number seeds in practice
- 199 J.P.C. Kleijnen
Regression analysis of simulation experiments: functional software specification
- 200 G. van der Laan and A.J.J. Talman
An algorithm for the linear complementarity problem with upper and lower bounds
- 201 P. Kooreman
Alternative specification tests for Tobit and related models

IN 1986 REEDS VERSCHENEN

- 202 J.H.F. Schilderinck
Interregional Structure of the European Community. Part III
- 203 Antoon van den Elzen and Dolf Talman
A new strategy-adjustment process for computing a Nash equilibrium
in a noncooperative more-person game
- 204 Jan Vingerhoets
Fabrication of copper and copper semis in developing countries.
A review of evidence and opportunities.
- 205 R. Heuts, J. v. Lieshout, K. Baken
An inventory model: what is the influence of the shape of the lead
time demand distribution?
- 206 A. v. Soest, P. Kooreman
A Microeconomic Analysis of Vacation Behavior
- 207 F. Boekema, A. Nagelkerke
Labour Relations, Networks, Job-creation and Regional Development
A view to the consequences of technological change
- 208 R. Alessie, A. Kapteyn
Habit Formation and Interdependent Preferences in the Almost Ideal
Demand System
- 209 T. Wansbeek, A. Kapteyn
Estimation of the error components model with incomplete panels
- 210 A.L. Hempenius
The relation between dividends and profits
- 211 J. Kriens, J.Th. van Lieshout
A generalisation and some properties of Markowitz' portfolio
selection method
- 212 Jack P.C. Kleijnen and Charles R. Standridge
Experimental design and regression analysis in simulation: an FMS
case study
- 213 T.M. Doup, A.H. van den Elzen and A.J.J. Talman
Simplicial algorithms for solving the non-linear complementarity
problem on the simplotope
- 214 A.J.W. van de Gevel
The theory of wage differentials: a correction
- 215 J.P.C. Kleijnen, W. van Groenendaal
Regression analysis of factorial designs with sequential replica-
tion

- 216 T.E. Nijman and F.C. Palm
Consistent estimation of rational expectations models
- 217 P.M. Kort
The firm's investment policy under a concave adjustment cost function
- 218 J.P.C. Kleijnen
Decision Support Systems (DSS), en de kleren van de keizer ...
- 219 T.M. Doup and A.J.J. Talman
A continuous deformation algorithm on the product space of unit simplices
- 220 T.M. Doup and A.J.J. Talman
The 2-ray algorithm for solving equilibrium problems on the unit simplex
- 221 Th. van de Klundert, P. Peters
Price Inertia in a Macroeconomic Model of Monopolistic Competition
- 222 Christian Mulder
Testing Korteweg's rational expectations model for a small open economy
- 223 A.C. Meijdam, J.E.J. Plasmans
Maximum Likelihood Estimation of Econometric Models with Rational Expectations of Current Endogenous Variables
- 224 Arie Kapteyn, Peter Kooreman, Arthur van Soest
Non-convex budget sets, institutional constraints and imposition of concavity in a flexibele household labor supply model.
- 225 R.J. de Groof
Internationale coördinatie van economische politiek in een twee-regio-twee-sectoren model.
- 226 Arthur van Soest, Peter Kooreman
Comment on 'Microeconometric Demand Systems with Binding Non-Negativity Constraints: The Dual Approach'
- 227 A.J.J. Talman and Y. Yamamoto
A globally convergent simplicial algorithm for stationary point problems on polytopes
- 228 Jack P.C. Kleijnen, Peter C.A. Karremans, Wim K. Oortwijn, Willem J.H. van Groenendaal
Jackknifing estimated weighted least squares
- 229 A.H. van den Elzen and G. van der Laan
A price adjustment for an economy with a block-diagonal pattern
- 230 M.H.C. Paardekooper
Jacobi-type algorithms for eigenvalues on vector- and parallel computer

- 231 J.P.C. Kleijnen
Analyzing simulation experiments with common random numbers
- 232 A.B.T.M. van Schaik, R.J. Mulder
On Superimposed Recurrent Cycles
- 233 M.H.C. Paardekooper
Sameh's parallel eigenvalue algorithm revisited
- 234 Pieter H.M. Ruys and Ton J.A. Storcken
Preferences revealed by the choice of friends
- 235 C.J.J. Huys en E.N. Kertzman
Effectieve belastingtarieven en kapitaalkosten
- 236 A.M.H. Gerards
An extension of König's theorem to graphs with no odd- K_4
- 237 A.M.H. Gerards and A. Schrijver
Signed Graphs - Regular Matroids - Grafts
- 238 Rob J.M. Alessie and Arie Kapteyn
Consumption, Savings and Demography
- 239 A.J. van Reeken
Begrippen rondom "kwaliteit"

Bibliotheek K. U. Brabant

17 000 01059702 0