STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 26
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Data Editing
(Cardiff, United Kingdom, 18-20 October 2000)

Topic III: New techniques and tools for editing imputation

DATA IMPUTATION BASED ON REGRESSION MODELS WITH VARIATIONS OF ENTROPY

Submitted by the State Committee of the Russian Federation on Statistics[1]

Contributed paper

I. INTRODUCTION

1. The majority of methods used for statistical data imputation may be divided into two broad groups: imputation by estimation and donor imputation. Imputation by estimation in contrast with donor imputation is used when there are reasons for using a certain data model. Regression models are used rather often for this purpose, including autoregression. However, unlikely regression models for parameters estimation and prediction, applying regression methods to imputation poses some specific problems.

2. While donor imputation assures that an imputed record meets all the edits, imputation by regression estimators and estimators upon the whole, does not provide such guarantees. That is why after regression imputation, a repeated checking of edited data is usually needed, and if the result is negative, it may require another stage of imputation [1]. However, even a repeated imputation does not guarantee that a record will pass all the edits.

3. Another shortcoming of applying estimation methods is that variability in the imputed values is underestimated in comparison with variability of non-imputed values. To solve this problem, it was suggested to add random error to the value fitted by regression estimator [2]. Multiple imputation is another way to overcome this problem [3].

4. This paper proposes methods of imputing missing or inconsistent data items based on linear regression models, which allow obtaining a set of estimates for each imputed value. All of these estimates belong to the linear regression model with different criteria of modeling. Entropy is used as a criterion of the regression modeling. From this point of view, usual regression estimates used for prediction and parameter estimation are the maximum entropy estimates. In this paper, the minimum entropy estimates for time series will be obtained, as well as a whole set of estimates between the minimum and maximum entropy. For cross-sectional data, the analogues of these estimates will be obtained.

[1] Prepared by Sergey S. Kuzin

5. An opportunity to select estimates from a certain range increases the probability that an imputed record will satisfy all the edits. An alternative method of using this set of regression estimates is multiple imputation. Actually, here we propose another way of solving the above-mentioned problems, which arise when data are imputed by estimation methods.

6. Section II deals with imputation of longitudinal data by autoregression where, in addition to the known estimate of the maximum entropy, estimates of the minimum entropy are derived together with a whole range of intermediate estimates. Section III deals with the analogues of the estimates of maximum and minimum entropy for cross-sectional data. In Section IV the estimates of multichannel autoregression are proposed for longitudinal surveys, which take into account both the correlations between variables with different lags and autocorrelations. In Section V, some examples of imputation of households budget survey data by the derived regression methods are given.

II. LONGITUDINAL DATA AUTOREGRESSION

7. A problem of imputing a missing or inconsistent value on the basis of previous observations may be formulated as a problem of a one-step-forward prediction of a time series. This prediction, or extending a time series beyond the observed period, is always based on a certain model of data and a specified criterion. Maximizing entropy is an example of such criterion. It means that an extension of a time series is chosen, which autocorrelation function is the most random of all autocorrelation function conforming to the available data. In the case of Gaussian statistics it corresponds to the known solution that minimizes prediction error for a stationary time series. [4].

8. Let us consider a stationary time series *x(t)* with a correlation function (CF), where *N* values are known {r[0], r[1], ..., r[n-1]}. It is necessary to predict the next value *r*[*n*] of the correlation function on the basis of known values. Linear prediction filters for CF and time series are determined as follows:

$$r[n] = -\sum_{k=1}^{p} a[i]\, r[n-k]\,, \tag{2.1}$$

$$x[n] = -\sum_{ki=1}^{p} a[i]\, x[n-k]\,, \tag{2.2}$$

where $a = [a[1], a[2], \ldots, a[p]]^T$ is a vector of a linear prediction coefficients; *p* - order of autoregression; *T* — matrix transposition.

9. Let ζ be the unknown element r_n of the CF (to distinguish it from the known elements of CF). Thus, we search for the extension $\boldsymbol{R}_n$ of a Toeplitz covariance matrix $\boldsymbol{R}_{n\text{-}1}$:

$$\boldsymbol{R}_n = \left[\begin{array}{c:cccc} r[0] & r[1] & \cdots & r[n-1] & \zeta \\ \hdashline r[1] & r[0] & \cdots & r[n-2] & r[n-1] \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r[n-1] & r[n-2] & \cdots & r[0] & r[1] \\ \zeta & r[n-1] & \cdots & r[1] & r[0] \end{array}\right] = \left[\begin{array}{c:c} r[0] & \boldsymbol{r}_n^T(\text{æ}) \\ \hdashline \boldsymbol{r}_n(\text{æ}) & \boldsymbol{R}_{n-1} \end{array}\right], \tag{2.3}$$

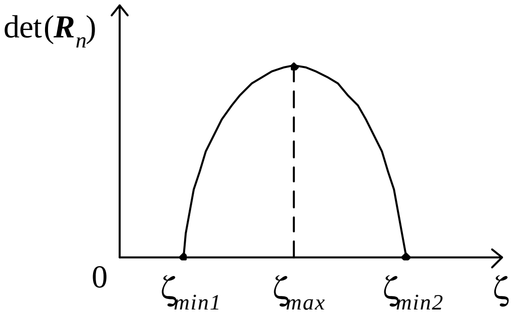
with the constraint of a positive semi definite extended covariance matrix: $\det(\boldsymbol{R}_n) \geq 0$.

10. The entropy of a zero mean the Gaussian process can be specified using the determinant of the covariance matrix of the process [5]

(2.4) $$H = \frac{1}{2}\ln[\det(\boldsymbol{R})].$$

11. Thus, the problem of maximizing or minimizing the entropy of the extended time series is reduced to a problem of maximizing or minimizing the determinant of the extended covariance matrix (2.2).

12. As $\det(\boldsymbol{R}_n)$ is a quadratic function of the predicted element ζ, there are two values of ζ where the matrix determinant is equal to zero. These two values define the boundaries of an interval, which ζ should belong to.



13. The known estimate of a vector of the prediction parameters [5]

(2.5) $$\begin{bmatrix} 1 \\ -- \\ \boldsymbol{a} \end{bmatrix} = \frac{1}{\left(\boldsymbol{R}_{n-1}^{-1}\right)_{11}} \boldsymbol{R}_{n-1}^{-1} \boldsymbol{e}_{1,} \qquad \boldsymbol{e}_1 = [1, 0, \ldots, 0]^T,$$

is an estimate of the maximum entropy, that is a predicted element of the correlation function $r[n] = \zeta_{\max}$.

14. To find the solutions for the minimum entropy and the solutions for intermediate values of entropy, let us define the determinant of the Toeplitz covariance matrix $\boldsymbol{R}_n$ as a function of the predicted value ζ [6]

(2.6) $$\det(\boldsymbol{R}_n) \equiv D_n(\zeta) = \frac{D_{n-1}^2 - \left|D_{n-1,(1)}(\zeta)\right|^2}{D_{n-2}},$$

where $D_{n-i} = \det(\boldsymbol{R}_{n-i})$, $$D_{n\text{-}1,(1)}(\zeta) = \det\left[\begin{array}{ccc:c} r[1] & \cdots & r[n\text{-}1] & \zeta \\ \hdashline r[0] & \cdots & r[n-2] & r[n\text{-}1] \\ \cdots & \cdots & \cdots & \cdots \\ r[n-2] & \cdots & r[0] & r[1] \end{array}\right].$$

15. By applying the Sylvester identity to the determinant $D_{n\text{-}1,(1)}(\zeta)$, we transform (2.6) to

(2.7) $$D_n(\zeta) = \frac{D^2_{n-1} - \left|\text{æ}D_{n-2} + D_{n-1,(1)}(0)\right|^2}{D_{n-2}},$$

where, assuming $D_{n\text{-}2} > 0$ and $D_n(\zeta) \geq 0$, we find a single solution of the maximum entropy and two solutions of the minimum entropy ($D_n(\zeta) = 0$):

(2.8) $$\zeta_{max} = -\frac{D_{n-1,(1)}(0)}{D_{n-2}},$$

(2.9) $$\zeta_{min1} = -\frac{D_{n-1,(1)}(0) + D_{n-1}}{D_{n-2}},$$

(2.10) $$\zeta_{min2} = -\frac{D_{n-1,(1)}(0) - D_{n-1}}{D_{n-2}}.$$

16. Estimates of the linear prediction coefficients $a[i]$ may be obtained from the following matrix equation:

(2.11) $$\boldsymbol{R}_n(\zeta)\begin{bmatrix} 1 \\ \text{--} \\ \boldsymbol{a} \end{bmatrix} = \begin{bmatrix} \text{ñ}_n \\ \text{---} \\ \boldsymbol{0} \end{bmatrix} \quad \text{or} \quad \left[\begin{array}{c:c} r[0] & \boldsymbol{r}_n^T(\text{æ}) \\ \hdashline \boldsymbol{r}_n(\text{æ}) & \boldsymbol{R}_{n-1} \end{array}\right]\begin{bmatrix} 1 \\ \text{-} \\ \boldsymbol{a} \end{bmatrix} = \begin{bmatrix} \text{ñ}_n \\ \text{---} \\ \boldsymbol{0} \end{bmatrix},$$

where ρ_n is a prediction error variance, which, in the case of minimum entropy solution, is assumed to be zero.

17. Independently of ζ value, it follows from (2.11) that

(2.12) $$\left[\boldsymbol{r}_n^T(\text{æ}) \;\vdots\; \boldsymbol{R}_{n-1}\right]\begin{bmatrix} 1 \\ \text{-} \\ \boldsymbol{a} \end{bmatrix} = \boldsymbol{0},$$

hence

(2.13) $$\boldsymbol{a} = -\boldsymbol{R}_{n-1}^{-1}\boldsymbol{r}_n(\zeta).$$

18. Substitution of a respective values of ζ from the range determined by (2.8), (2.9), (2.10) into vector $\boldsymbol{r}_n(\zeta)$ gives correspondent prediction vectors $\boldsymbol{a}$ for the minimum and maximum entropy together with all intermediate solutions. The imputed value is determined by (2.2), where the obtained regression coefficients are substituted to.

19. Thus, we have not a single regression estimate for each imputed value, but a whole set of regression estimates with entropy in the range from zero to maximum. Two ways of using these estimates seem to be the most reasonable:

a) One estimate is chosen for imputation from the set of available solutions, which meets all the edits. On the one hand, the availability of a set of estimates increases the probability for an imputed value to hit the feasible interval, and, on the other hand, all these estimates belong to one class of autoregression estimates.

b) The obtained set of estimates may be used as initial estimates for a method of multiple imputation, which provides more realistic variances for imputed values.

III. CROSS-SECTIONAL DATA MULTIPLE REGRESSION

20. In the case of multiple regression, we are not speaking strictly of minimizing or maximizing the entropy of the process, however, some analogues could be suggested for the estimates proposed above. Here, the value of the determinant of the covariance matrix will change as well, but, not due to changing of the predicted element but due to subtracting a noise component from the matrix.

21. Let y_1 is a dependent variable and y_2, ..., y_{n} are independent variables of a given survey. If $\boldsymbol{R}$ is a covariance matrix for these variables, then the vector of the regression coefficients which minimizes the standard error, is determined as follows:

(2.14) $$\boldsymbol{a} = -\boldsymbol{R}_{n-1}^{-1}\boldsymbol{r}_n,$$

where $$\boldsymbol{R} = \left[\begin{array}{c:c} r_{11} & \boldsymbol{r}_n^T \\ \hdashline \boldsymbol{r}_n & \boldsymbol{R}_{n-1} \end{array}\right].$$

22. The solution (2.14) is equal to the solution of the maximum entropy for a stationary process. We get a certain analogue for the solution of the minimum entropy if we consider the following matrix:

(2.15) $$\boldsymbol{R}_\sigma = \boldsymbol{R} - \sigma^2\boldsymbol{I},$$

where $0 \le \sigma^2 \le \lambda_{min}$; $\boldsymbol{I}$ is an identity matrix and λ_{min} is the minimum eigenvalue of the matrix $\boldsymbol{R}$.

23. Physically, operation (2.15) may be thought of as subtracting a part of uncorrelated noise component, which is always presented in statistical measurements, from a covariance matrix. The minimum eigenvalue λ_{min} of the matrix $\boldsymbol{R}$ corresponds to the variance of this uncorrelated component.

24. Let $\boldsymbol{u}$ be the first column of the matrix $\boldsymbol{R}_\sigma^{-1}$, if $\sigma^2 \neq \lambda_{min}$, or an eigenvector corresponding to a zero eigenvalue of $\boldsymbol{R}_\sigma$, if $\sigma^2 = \lambda_{min}$, then a vector $\boldsymbol{a}_\sigma$ of regression coefficients will be defined as:

(2.16) $$\boldsymbol{a}_\sigma = \left[\frac{u_2}{u_1}, \frac{u_3}{u_1}, \ldots, \frac{u_n}{u_1}\right]^T.$$

25. By specifying various values to σ^2, we can obtain a set of regression estimates for a value to be imputed. The ways of using these estimates are the same as described above for autoregression with various values of entropy.

IV. MULTICHANNEL AUTOREGRESSION

26. The models of auto- and multiple regression considered above take into account respectively either time dependences, or relationships between variables. However, there is a model of multichannel autoregression, which allows us to take into account both these relations at the same time. Of course, there should be reasons for applying a more complicated model, and if there are no serious "pros" for using this model, then less complicated models should be applied. But there are situations, when it is advisable to take into account both types of relations and to apply a multichannel autoregression model.

27. In case of multichannel autoregression, we can also obtain a set of estimates for each value to be imputed. Multichannel regression can be viewed in this context as a regression of current values of several variables to the available historic values of the same variables. Different variables here are understood as channels of measurement, and an entire covariance matrix is defined in a following way:

(2.17)
$$\boldsymbol{R} = \left[\begin{array}{c|c|c|c} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} & \cdots & \boldsymbol{R}_{1n} \\ \hline \boldsymbol{R}_{21} & \boldsymbol{R}_{22} & \cdots & \boldsymbol{R}_{2n} \\ \hline \cdots & \cdots & \cdots & \cdots \\ \hline \boldsymbol{R}_{n1} & \boldsymbol{R}_{n2} & \cdots & \boldsymbol{R}_{nn} \end{array}\right],$$

where $\boldsymbol{R}_{ik}$ is a *p ´ p* covariance matrix of the same variables in the *i-th* and *k-th* surveys.

28. Simultaneously, *p* variables are predicted. We will define them as $\boldsymbol{x}_k$ vector:

(2.18)
$$\boldsymbol{x}_k = -\sum_{i=1}^{n-1} \boldsymbol{A}_i^T \, \boldsymbol{x}_{k-i},$$

where $\boldsymbol{A}_i^T$ are the matrix of the autoregression coefficients, which are determined by the following equation:

(2.19)
$$\left[\begin{array}{c|c|c|c} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} & \cdots & \boldsymbol{R}_{1n} \\ \hline \boldsymbol{R}_{21} & \boldsymbol{R}_{22} & \cdots & \boldsymbol{R}_{2n} \\ \hline \cdots & \cdots & \cdots & \cdots \\ \hline \boldsymbol{R}_{n1} & \boldsymbol{R}_{n2} & \cdots & \boldsymbol{R}_{nn} \end{array}\right] \cdot \left[\begin{array}{c} \boldsymbol{I} \\ \hline \boldsymbol{A}_1 \\ \hline \cdots \\ \hline \boldsymbol{A}_{n-1} \end{array}\right] = \left[\begin{array}{c} \boldsymbol{P}_R \\ \hline \boldsymbol{0} \\ \hline \cdots \\ \hline \boldsymbol{0} \end{array}\right],$$

where $\boldsymbol{P}_R$ is a covariance matrix for residual errors.

29. A maximum entropy solution (2.19) for the matrix coefficients $\boldsymbol{A}_i$ may be written as follows:

(2.20)
$$\left[\begin{array}{c} \boldsymbol{A}_1 \\ \hline \boldsymbol{A}_2 \\ \hline \cdots \\ \hline \boldsymbol{A}_{n-1} \end{array}\right] = -\left[\begin{array}{c|c|c|c} \boldsymbol{R}_{22} & \boldsymbol{R}_{23} & \cdots & \boldsymbol{R}_{2n} \\ \hline \boldsymbol{R}_{32} & \boldsymbol{R}_{33} & \cdots & \boldsymbol{R}_{3n} \\ \hline \cdots & \cdots & \cdots & \cdots \\ \hline \boldsymbol{R}_{n2} & \boldsymbol{R}_{n3} & \cdots & \boldsymbol{R}_{nn} \end{array}\right]^{-1} \cdot \left[\begin{array}{c} \boldsymbol{R}_{21} \\ \hline \boldsymbol{R}_{31} \\ \hline \cdots \\ \hline \boldsymbol{R}_{n1} \end{array}\right].$$

30. Solutions corresponding to other values of entropy may be obtained by the following equation:

$$(2.21) \qquad \left[\begin{array}{c} \boldsymbol{A}_{(\sigma)1} \\ \hdashline \boldsymbol{A}_{(\sigma)2} \\ \hdashline \cdots \\ \hdashline \boldsymbol{A}_{(\sigma)n-1} \end{array}\right] = -\left[\begin{array}{c:c:c:c} \boldsymbol{R}_{(\sigma)22} & \boldsymbol{R}_{23} & \cdots & \boldsymbol{R}_{2n} \\ \hdashline \boldsymbol{R}_{32} & \boldsymbol{R}_{(\sigma)33} & \cdots & \boldsymbol{R}_{3n} \\ \hdashline \cdots & \cdots & \cdots & \cdots \\ \hdashline \boldsymbol{R}_{n2} & \boldsymbol{R}_{n3} & \cdots & \boldsymbol{R}_{(\sigma)nn} \end{array}\right]^{-1} \cdot \left[\begin{array}{c} \boldsymbol{R}_{21} \\ \hdashline \boldsymbol{R}_{31} \\ \hdashline \cdots \\ \hdashline \boldsymbol{R}_{n1} \end{array}\right],$$

where $\boldsymbol{R}_{(\sigma)kk} = \boldsymbol{R}_{kk} - \sigma^2\boldsymbol{I}$; $0 \le \sigma^2 \le \lambda_{min}$ and λ_{min} is the minimum eigenvalue of the matrix $\boldsymbol{R}$.

V. AN APPLICATION TO THE SURVEY OF HOUSEHOLDS

31. Regression models with variations of entropy discussed above provide a set of estimates for each value to be imputed, of which the values satisfied all edits could be chosen. Now, we consider some examples of these estimates in view of quarterly household's expenses survey data.

32. Table 1 shows the estimates with different rates of entropy and the reported values of the expenses of households on food in the 4-th quarter 1999. The estimates on the basis of autoregression with maximum, minimum and medium levels of entropy have been obtained from 11 previous surveys.

Table 1. Autoregression estimates

Estimate type	Household #				
	1	2	3	4	5
Reported value	*1567*	*3070*	*1757*	*1698*	*2301*
Minimum entropy 1	2320	3541	1884	906	2018
Medium entropy 1	2064	3581	1727	956	1741
Maximum entropy	1807	3622	1570	1004	1464
Medium entropy 2	1551	3663	1414	1054	1188
Minimum entropy 2	1294	3704	1257	1104	911

33. Related estimates for values to be imputed have been obtained on the basis of multiple regressions with variations of entropy. As independent variables were chosen expenses of households in the 4-th quarter 1999 on: (1) bread and bakery products; (2) meat and meat products; (3) milk and dairy products. The dependent variable is expenses of households on butter, margarine and fat. The results of the estimation of imputed values are presented in Table 2.

Table 2. Multiple regression estimates

Estimate type	Household #				
	1	2	3	4	5
Reported value	*416*	*123*	*370*	*169*	*130*
Maximum entropy	283	186	298	206	91
Medium entropy	279	202	284	185	105
Minimum entropy	249	285	208	106	206

43. Finally, we should note that the suggested approach for obtaining a set of linear regression estimates for each data item to be imputed is applicable both to longitudinal and to cross-sectional surveys. In either case, the results of using these techniques may be as follows:

i) The opportunity to choosing the imputed value from a set of regression estimates, resulting in a higher probability to pass all the edits by an imputed record.

ii) The provision of more realistic standard errors of imputed values in comparison with the reported values by using a set of regression estimates with different levels of entropy in accordance with the multiple imputation method.

REFERENCES

1. "Functional Description of the Generalized Edit and Imputation System." Business Survey Methods Division, Statistics Canada. July 25, 1991. Revised May 31, 1999.

2. West S.A., Butani S., Witt M Alternative Imputation Methods for Wage Data. U.S. Bureau of Labor Statistics.

3. Rubin D. (1987). Multiple Imputation for Nonresponse in Surveys. Wiley, New York.

4. Box G. E. P., Jenkins G. M. (1970). Time Series Analysis Forecasting and Control. Holden-day, San Francisco.

5. Marple S. L. (1987) Digital Spectral Analysis with Applications. Prentice-Hall, Englewood Cliffs.

6. Iohvidov I. S. (1974). Hankel and Toeplitz Matrices and Forms. Nauka, Moscow.