Communications of the Korean Statistical Society
2009, Vol. 16, No. 2, 317–323

Imputation Using Factor Score Regression

Sang Eun Lee[a], Hee-Jin Hwang[b], Key-Il Shin[1,b]

[a]Dept. of Applied Statistics, Kyonggi Univ., [b]Dept. of Statistics, Hankuk Univ. of Foreign Studies

Abstract

Recently not even government polices but small town decisions are based on the survey data/information, so the most of government agencies/organizations demand various sample surveys in each fields for more detail information. However in conducting the sample survey, nonresponse problem rises very often and it becomes a major issue on judging the accuracy of survey. For that matters, one solution can be using the administration data. However unfortunately most of administration data are restricted to the common users. The other solution can be the imputation. Therefore several methods of imputation are studied in various fields. In this study, in stead of the simple regression imputation method which is commonly used, factor score regression method is applied specially to the incomplete data which have the unit and item missing values in survey data. Here for simulation study, Consumer Expenditure Surveys in Korea are used.

Keywords: Imputation, factor analysis, factor score regression, mean squared percentage error, mean absolute percentage error.

1. Introduction

Recently in most of surveys, filling the questionnaires completely is getting uneasy. Therefore imputation techniques have been developed in many different aspects. One of common methods is regression imputation in recent. Regression imputation, where the missing variables for a unit are estimated by predicted values from the regression on the known variables for that unit, is the most popular in practice. The classical and standard approach to missing data in regression is due in general to Yates, which was filling in the least squares estimates of all missing values, $\hat{y}_i = x_i\hat{\beta}$ where $\hat{\beta}$ is defined by ordinary least squares estimates on observed y_i and complete data x_i. After Yates method in Alexander (1993), Bartlett's ANOVA method in Little and Rubin (1987, 2002) was useful procedures. However, both Yates and Bartlett's procedures need to set the best model with some auxiliary variables. In practice, finding the proper auxiliary variables and also selecting best variables among the huge number of available variables are not quite simple. Therefore in this study, factor score regression imputation which can be improved the efficiency is investigated and compared with the standard regression imputation using the data from Consumer Expenditure Survey in Korea.

2. Regression Imputation

Regression imputation replaces missing values by predicted values from a regression model and also is one of the most handful and commonly used approaches. Basically regression imputation is a modeling technique.

This research was supported by the research fund of Hankuk University of Foreign Studies 2009.

[1] Corresponding author: Professor, Department of Statistics, Hankuk University of Foreign Studies, Kyonggi 449-791, Korea. E-mail: keyshin@hufs.ac.kr

Suppose that simple random sample of size N, $(y_i, x_{i1}, \ldots, x_{iK})$, where the K auxiliary variables $x_1, \ldots, x_K$ are observed for all in the sample, then model for Y is as following:

$$y_i = \beta_0 + \sum_{k=1}^{K} \beta_k x_{ik} + \epsilon_i, \quad i = 1, \ldots, N,$$

where $\epsilon_i \sim N(0, \sigma^2)$ and let the parameter, $\theta = (\beta_0, \ldots, \beta_K, \sigma^2)$. Now $\hat{\bar{Y}}^{RI}$ is the estimated population mean obtained by observed and imputed values using regression imputation(RI) and expressed by:

$$\hat{\bar{Y}}^{RI} = \frac{m\hat{\bar{y}}_{RI} + (N-m)\bar{y}_R}{N}. \tag{2.1}$$

Here $\bar{y}_R$ is the mean of observed values, $\hat{\bar{y}}_{RI}$ is the mean of imputed values using RI, and $\hat{\bar{y}}_{RI}$ can be calculated by

$$\hat{\bar{y}}_{RI} = \frac{1}{m} \sum_{i=1}^{m} \hat{y}_{iRI},$$

where $\hat{y}_{iRI} = E[(y_i | x_{i1}, \ldots, x_{iK}, \theta)| x]$ and x'_{ik}s are the auxiliary values corresponding missing value on y_i.

To investigate the properties of $\hat{\bar{y}}_{RI}$, let $y = [y'_R \;\; y'_{NR}]'$, $X = [X'_R \;\; X'_{NR}]'$ and $n = N - m$. Here y_R is the vector of the observed sample and y_{NR} is the vector of missing sample. X_R is the matrix of the auxiliary variables corresponding to the observed sample, y_R, and X_{NR} is the matrix corresponding to y_{NR}.

Therefore y_R and y_{NR} are $n \times 1$ and $m \times 1$ vectors and X_R and X_{NR} are $n \times (K+1)$ and $m \times (K+1)$ matrices, respectively. Then using regression analysis, we obtain $\hat{\beta}_{RI} = (\hat{\beta}_{0RI}, \hat{\beta}_{1RI}, \ldots, \hat{\beta}_{KRI})$, the estimate of $\beta_{RI} = (\beta_0, \beta_1, \ldots, \beta_K)$ as follows:

$$\hat{\beta}_{RI} = (X'_R X_R)^{-1} X'_R y_R. \tag{2.2}$$

So we have $\hat{y}_{RI} = X_{NR}\hat{\beta}_{RI}$ and hence we have

$$\hat{\bar{y}}_{RI} = \frac{1}{m} \bar{1}' X_{NR} \hat{\beta}_{RI}, \tag{2.3}$$

where $\bar{1} = [1, 1, \ldots, 1]'$. Since there are no missing values in the auxiliary variables, $\hat{\bar{y}}_{RI}$ is an unbiased estimator of $\bar{Y}^{RI}$. Therefore, to obtain the MSE of RI, we only need the $\mathrm{Var}(\hat{\bar{Y}}^{RI})$.

Theorem 1. *Let $\bar{Y}$ be the population mean and $\hat{\bar{Y}}^{RI}$ be defined in Equation (2.1). Then*

$$MSE_R = Var_R\left(\bar{Y} - \hat{\bar{Y}}^{RI}\right) = n^{-2}\left(m + \bar{1}' X_{NR} (X'_R X_R)^{-1} X'_{NR} \bar{1}\right)\sigma^2. \tag{2.4}$$

Proof: Since $\mathrm{Cov}(\sum_{i=1}^{m} y_{iNR}, \sum_{i=1}^{m} \hat{y}_{iRI}) = 0$, we have

$$\mathrm{Var}\left(\bar{Y} - \hat{\bar{Y}}^{RI}\right) = n^{-2}\mathrm{Var}\left(\sum_{i=1}^{m} y_{iNR} - \sum_{i=1}^{m} \hat{y}_{iRI}\right) = mn^{-2}\sigma^2 + n^{-2}\mathrm{Var}\left(\sum_{i=1}^{m} \hat{y}_{iRI}\right).$$

Also we have

$$\mathrm{Var}\left(\sum_{i=1}^{m} \hat{y}_{iRI}\right) = \mathrm{Var}\left(\bar{1}\hat{\bar{y}}_{RI}\right) = \mathrm{Var}\left(\bar{1}' X_{NR}\hat{\beta}_{RI}\right) = \bar{1}' X_{NR} \mathrm{Var}\left(\hat{\beta}_{RI}\right) X'_{NR} \bar{1}.$$

Therefore using $\mathrm{Var}(\hat{\beta}_{RI}) = (X'_R X_R)^{-1}\sigma^2$, we have the result. □

3. Factor Score Regression Imputation

Define p-variate observation vectors, $(x_1, \ldots, x_N) = X' = [X'_R \;\; X'_{NR}]'$ on N objects. The means of x'_is are assumed to have been subtracted out, so that $E(X') = 0$. The Factor Analysis model is

$$x_{i(p\times 1)} = \Lambda_{(p\times k)} f_{i(k\times 1)} + \epsilon_{i(p\times 1)}, \quad i = 1, \ldots, N,\ k < p,$$

where Λ denotes factor loading matrix, f_i denotes the factor score vector for subject i and $F' = (f_1, \ldots, f_N)$, $\epsilon'_i s$ are assumed to be mutually uncorrelated and normally distributed as $N(0, \Psi)$, where Ψ is a symmetric positive definite matrix, *i.e*, $\Psi > 0$. See more details on factor analysis in Johnson and Wichern (2002).

From the model, estimates of Λ and Ψ can be obtained by M.L.E. And for factor score, given any vector of observations x'_is and taking $\hat{\Lambda}$ and $\hat{\Psi}$ as estimated values using M.L.E, we can have the i^{th} factor score vector as following:

$$\hat{f}_i = \hat{\Lambda}' S^{-1} x_i, \quad i = 1, \ldots, N,$$

where S is the sample covariance matrix.

Once we have all the estimates, $\hat{\Lambda}, \hat{\Psi}$ and $\hat{f}_i$, we set up for the factor score regression imputation:

$$y = \hat{F}\beta_{FI} + \eta,$$

here $y = [y'_R \;\; y'_{NR}]'$ and η are $N \times 1$ vectors and $\eta \sim N(0, I\dot{\sigma}^2)$.

Let $\hat{F} = [\hat{F}'_R \;\; \hat{F}'_{NR}]'$ where $\hat{F}_R = X'_R S^{-1}\hat{\Lambda}$ and $\hat{F}_{NR} = X'_{NR} S^{-1}\hat{\Lambda}$. Then imputing values by factor score regression, $\hat{y}^F_{NR} = \hat{y}_{FI}$ becomes following:

$$\hat{y}_{FI} = \hat{F}_R \hat{\beta}_{FI},$$

here $\hat{\beta}_{FI} = (\hat{F}'_R \hat{F}_R)^{-1} \hat{F}'_R y_R$.

Now let $\hat{\bar{Y}}^{FI}$ be the estimated population mean obtained by observed and imputed values using factor score regression imputation(FI). Then we have

$$\hat{\bar{Y}}^{FI} = \frac{m\hat{\bar{y}}_{FI} + (N-m)\bar{y}_R}{N},$$

where $\bar{y}_R$ is the mean of observed values and $\hat{\bar{y}}_{FI}$ is the mean of estimated imputed values.

Now we assume that the number of common factors is $k < p$ and $X_R, X_{NR}, \hat{F}_R$ and $\hat{F}_{NR}$ are defined as above.

Theorem 2. *If we have X, and we only use X_R for calculating $\hat{\Lambda}$ and S^{-1} denoted by $\hat{\Lambda}_R$ and S_R^{-1}, then*

$$\begin{aligned} Var_F\left(\bar{Y} - \hat{\bar{Y}}^{FI}\right) &= n^{-2}\left(m + \bar{1}'\hat{F}_{NR}\left(\hat{F}'_R\hat{F}_R\right)^{-1}\hat{F}'_{NR}\bar{1}\right)\sigma^2 \\ &\approx n^{-2}\left(m + \bar{1}' X_{NR}\left(X'_R X_R\right)^{-1} X'_{NR}\bar{1}\right)\sigma^2. \end{aligned} \tag{3.1}$$

Proof: By the same argument as Theorem 1, we can easily obtain

$$\mathrm{Var}_F\left(\bar{Y} - \hat{\bar{Y}}^{FI}\right) = n^{-2}\left(m + \bar{1}'\hat{F}_{NR}\left(\hat{F}'_R\hat{F}_R\right)^{-1}\hat{F}'_{NR}\bar{1}\right)\sigma^2.$$

Also, $\mathrm{Var}(F) = I$ by the assumption of factor score regression, which means $\hat{F}_R'\hat{F}_R \approx nI$. Therefore we have $\hat{F}_{NR}(\hat{F}_R'\hat{F}_R)^{-1}\hat{F}_{NR}' \approx 1/n\hat{F}_{NR}\hat{F}_{NR}'$, where $\hat{F}_{NR}\hat{F}_{NR}' = X_{NR}'S_R^{-1}\hat{\Lambda}_R\hat{\Lambda}_R'S_R^{-1}X_{NR}$. The fundamental representation of factor analysis says $S_R = \hat{\Lambda}_R\hat{\Lambda}_R' + \Psi_R$ or $S_R \approx \hat{\Lambda}_R\hat{\Lambda}_R'$, we have $\hat{F}_{NR}\hat{F}_{NR}' \approx X_{NR}'S_R^{-1}X_{NR}$. Therefore

$$\hat{F}_{NR}\left(\hat{F}_R'\hat{F}_R\right)^{-1}\hat{F}_{NR}' \approx X_{NR}\left(X_R'X_R\right)^{-1}X_{NR}'.$$

This completes the proof. □

Theorem 2 says that the methods of regression imputation and factor regression imputation give the approximately same results. However when researchers have a large number of variables, they usually select a portion of variables to avoid multilinearity or to simplify the analysis. In that case, we can see that the factor regression imputation is superior to the regression imputation.

4. Example

In this example section, we use the usual imputation procedure which is as followings;

Let y_i be the i^{th} response variable, Y, $i = 1, \ldots, N$. Then the population mean, $\bar{Y}$, is estimated by the weighted average of observed and imputed values and expressed by:

$$\bar{Y} = \frac{m\hat{\bar{y}}_{NR} + (N-m)\bar{y}_R}{N},$$

where $\bar{y}_R$ is the mean of observed values, $\hat{\bar{y}}_{NR}$ is the mean of imputed values as estimates, m is the number of missing values and N is the sample size. And we assume that these auxiliary variables are complete.

4.1. Data description

In consumer's expenditure survey, we select the household income as dependent variable, y and choose 10 variables as independent variables, X: 1) number of members 2) number of employees 3) gender of the main person 4) age of main person 5) level of education 6) working industrial 7) working job 8) price of living house 9) house expense 10) total expense. We use complete data set on April, 2001 and the size of sample is 2386. For incomplete data set, we make the missing values randomly only in dependent variable and use complete data for independent variables.

4.2. Simulation results

For the simulation study, we decide the number of missing values, m, from 300 to 800 and for each case we replicate 1000 times. After imputing the missing values, using only the observed data, we obtain the models for regression imputation and factor score regression imputation.

First, for regression imputation we need to find the best regression model with observed data. Using Stepwise procedure four variables are selected: mem = number of employees, edu = level of education, $h_$exp = house expense, and $e_$exp =total expense. And the selected equation is as following:

$$y_i = \beta_0 + \beta_1 \mathrm{mem}_{i1} + \beta_2 \mathrm{edu}_{i2} + \beta_3 h\_\mathrm{exp}_{i3} + \beta_4 e\_\mathrm{exp}_{i4} + \epsilon_i, \quad i = 1, \ldots, n.$$

With observed data, we calculate the $\hat{\beta}$'s first and then impute the missing values by predicted value, $\hat{y}_{jRI} = \hat{\beta}_0 + \hat{\beta}_1 \mathrm{mem}_{j1} + \hat{\beta}_2 \mathrm{edu}_{j2} + \hat{\beta}_3 h\_\mathrm{exp}_{j3} + \hat{\beta}_4 e\_\mathrm{exp}_{j4}$, $j = 1, \ldots, m$. Finally, we have the estimated

mean of population, $\hat{\bar{Y}}^{RI}$, as following:

$$\hat{\bar{Y}}^{RI} = \frac{m\hat{\bar{y}}_{RI} + (2386 - m)\bar{y}_R}{2386},$$

where m can be varied from 300 to 800.

Now for the factor score regression imputation, let's consider the factor analysis procedure. First, number of common factors, k is decided. In this case, we postulate the number of common factors as 4 which is obtained by principle component analysis. Then the factor analysis model is

$$x_{i(10\times1)} = \Lambda_{(10\times4)} f_{i(4\times1)} + \epsilon_{i(10\times1)}, \quad i = 1, \ldots, 2386.$$

Before getting into factor score regression, from SAS PROC FACTOR, we obtain M.L.E of Λ and Ψ are obtained with only X_R, the corresponding values for the observed dependent data, y_R. And importantly $\hat{F}$ are obtained as following:

$$\hat{F} = XS_R^{-1}\hat{\Lambda}_R', \quad i = 1, \ldots, 2386, \tag{4.1}$$

where $X = [X_R' \; X_{NR}']'$ and S_R is the sample covariance matrix of X_R.

Finally we have the factor score regression equation:

$$y = \hat{F}\beta_{FI} + \eta,$$

where $\hat{F}$ is defined in Equation (4.1) and $\beta_{FI} = (\beta_1, \beta_2, \beta_3, \beta_4)'$. Now without loss of generality, $\hat{y}_{FI} = \hat{F}_{NR}\hat{\beta}_{FI}$ can be computed using factor score regression with imputed values. Now let $\hat{\bar{y}}_R$ be the mean of the observed values then we have the estimated mean of population, $\hat{\bar{Y}}^{FI}$:

$$\hat{\bar{Y}}^{FI} = \frac{m\hat{\bar{y}}_{FI} + (2386 - m)\bar{y}_R}{2386},$$

where m can be varied from 300 to 800. Regarding to the comparison of two methods, we use the followings:

$$\text{MSE} = \frac{1}{R}\sum_{r=1}^{R}\left(\bar{Y}_r - \hat{\bar{Y}}_r\right)^2,$$

$$\text{MAB} = \frac{1}{R}\sum_{r=1}^{R}\left|\bar{Y}_r - \hat{\bar{Y}}_r\right|,$$

$$\text{MSPE} = \frac{1}{R}\sum_{r=1}^{R}\left(\frac{\bar{Y}_r - \hat{\bar{Y}}_r}{\bar{Y}_r}\right)^2,$$

$$\text{MAPE} = \frac{1}{R}\sum_{r=1}^{R}\left|\frac{\bar{Y}_r - \hat{\bar{Y}}_r}{\bar{Y}_r}\right|.$$

Here $\bar{Y}_r$ is the true population mean and $\hat{\bar{Y}}_r$ is the estimated population mean with imputed values and we use the 1000 replications ($R = 1000$). Table 1 shows the result of two methods which is calculated the ratio of two MSE values, for example, RMSE = $\text{MSE}_R/\text{MSE}_F$, where MSE_R and MSE_F denote that MSE of regression and factor score regression imputation results respectively. And m is the

Table 1: Ratio of regression and factor score regression imputation

m	300	400	500	600	700	800
RMSE	5.39	4.37	3.19	3.00	2.26	1.54
RMAB	2.32	2.13	1.80	1.75	1.50	1.24
RT_MSE	11.47	8.41	4.21	2.31	1.63	1.16
RT_MSPE	11.39	8.26	4.15	2.31	1.66	1.17
RT_MAB	3.58	2.97	2.05	1.49	1.25	1.05
RT_MAPE	3.56	2.94	2.04	1.49	1.25	1.06

Table 2: Ratio of standard deviation regression and factor score regression imputation

m	300	400	500	600	700	800
$RMSE_s$	5.61	3.80	3.09	2.89	2.24	1.57
$RMAB_s$	2.32	2.02	1.76	1.69	1.52	1.25
RT_MSE_s	9.08	7.03	4.10	2.79	2.16	1.59
RT_MSPE_s	8.95	6.87	4.01	2.76	2.13	1.56
RT_MAB_s	2.76	2.58	2.07	1.85	1.73	1.51
RT_MAPE_s	2.74	2.55	2.05	1.84	1.71	1.49

number of missing dependent variables and since we use the log-transformation, the re-transformed results RT_(.)'s are calculated.

Similar to Table 1, 2 also shows ratio of standard deviation of two values, for example, $RMSE_s = SMSE_R/SMSE_F$, where $SMSE_R$ and $SMSE_F$ denote that standard deviation of regression and factor score regression imputation results respectively.

5. Summary

From Table 1 and 2 in Section 4.2, all the cases(m = 300 through 800) of results, factor score regression method(FI) is better than regression method(RI) and the standard deviations of all cases are smaller in FI case. This result was expected before the simulation because, in any regression model, more information will give the better explanation about the data through the model. In this study, instead of using the selected variables in regression method(RI), FI uses all 10 variables in analysis. In practice, for finding the best model, we usually spend quite some time to choose the proper variables among available variables. However FI method doesn't have to choose some variables, only grouping the variables and use them all, then may gives not only improve the efficiency but also can save the effort to find proper variables. And another notice is that in Table 1 and 2, as getting increased the number of missing values in dependent variable, the values of ratio become close to "1". That means this if we do have the missing values in surveyed data itself it can not be improved as much as we do have the decent surveyed data. Aa a further study, applying the Bayesian factor technique in Press (1982, 2003) to the factor score regression may fill the missing values in surveyed data.

References

Alexander, B. (1993). *Statistical Factor Analysis and Related Methods: Theory and Applications*, John Wiley & Sons, New York.

Johnson, R. A. and Wichern, D. W. (2002). *Applied Multivariate Statistical Analysis*, Prentice-Hall International, New York.

Little, R. J. A. and Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, John Wiley & Sons, New York.

Press, S. J. (1982). *Applied Multivariate Analysis: Using Bayesian Frequentist Methods of Inference*, Robert E. Krieger Publishing Company, Florida.

Press, S. J. (2003). *Subjective and Objective Bayesian Statistics -Principles, Models, and Applications*, John Wiley & Sons, New York.

Received February 2009; Accepted February 2009