

Robust bootstrap multiple imputation with missing data and outliers

Sunil Sapra
Department of Economics and Statistics,
California State University
5151 State University Dr.
Los Angeles, CA 90032

E-mail: ssapra@calstatela.edu

Keywords: Missing at Random (MAR) data; Multiple Imputation; Nonparametric Bootstrap; Huber's M estimator; Quantile Regression Estimator.

ABSTRACT

Given that survey data used in economics and other social sciences is plagued by missing data and outliers, this paper develops two bootstrap-based robust multiple imputation approaches for dealing with incomplete data and outliers on response variable: bootstrapping cases and bootstrapping residuals. Our methods are based on nonparametric bootstrap of cases and residuals and use robust regression techniques to generate forecasts and residuals for use in imputations. In contrast with the EM algorithm and Bayesian multiple imputation (Rubin (1987)), our approaches make minimal distributional assumptions with respect to the imputation model and use smooth bootstrap to smooth over the discreteness of robust regression estimators.

1. Introduction

Missing data and outliers are pervasive in survey data used in economics and other social sciences. Several approaches to dealing with missing data have been suggested in the literature. These approaches include ad hoc methods such as complete case analysis, likelihood –based methods such as EM algorithm, and imputation methods such as hot-deck, mean, regression-based imputation, and multiple imputation (MI) (Little and Rubin (2002)). Ad hoc methods and single random regression-based imputation (Buck (1960)) result in underestimation of standard errors, while likelihood-based methods are too sensitive to departures from distributional assumptions about the imputation and analysis models. The Bayesian MI approach of Rubin (1987) and bootstrap-based MI approach of Efron (1994) are motivated by the need to correct underestimation of variability in parameters under single imputation. Over the past two decades, interest has focused on multiple imputation methods for filling in missing data and estimating efficiently a parameter of interest and its variability in a missing data situation. The EM algorithm based on complete-data likelihood function and the Bayesian MI approach based on posterior distribution of the parameters under the model for observed data make strong assumptions about the imputation and analysis models, which can fail in common real-world settings. The bootstrap MI approach first suggested by Efron (1994) involves resampling the complete cases and using the bootstrap samples to fit a linear regression model of y on x in order to impute y -values from a normal distribution with estimated linear conditional mean function and estimated constant variance. Nevertheless, these approaches do not perform well in the presence of outliers on response or explanatory variables.

This paper extends Efron's nonparametric bootstrap for data on response variable missing at random (MAR) to such situations involving outliers on the response variable. These methods are likely to be useful in applications involving relationships between financial series, which often involve long-tailed distributions and missing data. Our methods are based on nonparametric bootstrap, which uses empirical distribution function of the population instead of a known distribution function and a robust regression method for imputation. The first approach adapts Efron's (1994) nonparametric bootstrap multiple imputation approach to missing data with outliers on the response variable and uses Huber's M estimator (Huber (1981)) or quantile regression estimator (see Koenker (2005)) instead of OLS estimator and smooth bootstrap to generate imputations. The second approach constructs imputations by applying smooth bootstrap to residuals generated by Huber's M or quantile regression methods applied to complete cases instead of OLS estimator and adding these residuals to forecasts from a robust regression on complete cases. The main advantages of our bootstrap-based approaches over the existing approaches are that they guard against misspecification of imputation models by making minimal assumptions about the imputation model and are resistant to outliers in the data on response variable. Section 2 presents robust bootstrap MI algorithms, section 3 presents simulation results and section 4 concludes.

2. Robust Regression Bootstrap-based Multiple Imputation Algorithms

Consider a data set $z_i = (x_i, y_i)$, $i = 1, 2, \dots, n$, where $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$. Assume that data on y is missing at random (MAR) for some values of x , but data on x is complete. We discuss two bootstrap-based multiple imputation approaches, one based on bootstrapping (x, y) pairs and the other based on bootstrapping robust regression residuals. The first

approach assumes that the x variables are random and allows for heteroscedasticity while the second approach assumes that the x variables are fixed and does not allow for heteroscedasticity.

Robust Multiple Imputation via Bootstrapping (x,y) Pairs

Algorithm 1

1. Select B independent bootstrap (x,y) samples using nonparametric bootstrap from the original incomplete sample with missing data on response variable y.
2. For each bootstrap sample generated in step 1, drop the cases with missing data on the response variable and retain the complete cases.
3. Regress y on x for cases with complete data using a robust estimation procedure such as Huber's M or quantile regression. Compute the imputations for missing values of y as predicted values of y for the corresponding x values. Replace missing values of y with their imputed values to generate complete data on y and a complete (x,y) sample.
4. *Smooth Bootstrap for Huber's M and Quantile Regression Estimators:* In order to smooth over the discreteness of these estimators, add a $N(0, 1/\sqrt{n})$ random noise to each bootstrap sample, where n is the sample size.
5. Repeat steps 2 through 4 for each of B bootstrap samples to generate B multiple imputations.

Step 1 is needed to ensure that multiple imputations are proper in the sense that it allows for variation in parameters and residual variability (Efron (1994), p. 464).

Robust Multiple Imputation via Bootstrapping Residuals

Algorithm 2

1. Drop the cases with missing data on the response variable and retain the remaining cases.
2. Regress y on x for cases with complete data using a robust estimation procedure such as Huber's M or quantile regression estimator and compute the predicted values of y for values of x for which y is missing.
3. Compute the residuals in the regression in step 2. Select B independent bootstrap samples from the sample of residuals, each of size k , the number of incomplete cases.
4. *Smooth Bootstrap for Huber's M and Quantile Regression Estimators*: In order to smooth over the discreteness of these estimators, add a $N(0, 1/\sqrt{k})$ random noise to each bootstrap sample of residuals.
5. Compute the imputations for missing values of y by adding predicted values of y for the corresponding x values from step 2 and the residuals in step 4.
6. Repeat steps 1 through 5 for each of B bootstrap samples of residuals generated in step 4 generating B imputations.

As under Algorithm 1, Step 3 is needed to ensure that multiple imputations are proper in the sense that it allows for variation in parameters and residual variability (Efron (1994), p. 464).

Let B be the number of bootstrap samples, r_k be the estimate of the parameter of interest (e.g., a regression coefficient or correlation coefficient), and s_k be its estimated

standard error in the k -th bootstrap sample. The estimate of the standard error of \bar{r} , the mean of estimates of the parameter of interest across B repetitions is given by

$$s.e.(\bar{r}) = \sqrt{1/B \sum_k s_k^2 + (1+1/B)(1/(B-1)) \sum_k (r_k - \bar{r})^2}. \quad (1)$$

(Rubin (1987), p. 76).

3. Monte Carlo Comparison of Robust and Non-robust Imputation Methods

1000 samples of size 100 were generated by mixing 90 observations generated according to a multivariate normal distribution with mean $= \mu = (10,20)'$ and covariance matrix =

$\Sigma = \begin{pmatrix} 4 & 2 \\ 2 & 3 \end{pmatrix}$ with 10 observations generated according to a multivariate t distribution with mean $(10,10)'$, covariance matrix $\Sigma = \text{diag}(2,2)$ and 1 degree of freedom resulting in

samples of size 100 with 10% outliers. The percentage of observations drawn from the multivariate normal distribution was subsequently reduced to 70% of the data and the percentage of observations generated according to multivariate t-distribution was increased to 30% resulting in 30% outliers. Furthermore, samples with 10% missing observations were generated by dropping every tenth observation in the original samples and samples with 20% missing observations were generated by dropping every fifth observation in original samples. Algorithms 1 and 2 were applied for generating 5 imputations under each approach. Results of these experiments are presented in tables 1 through 6.

The performance of each imputation approach is measured in terms of mean bias, overall standard error of r given by the formula in equation (1) and mean squared error (MSE). Our results can be summarized as follows.

1. *Comparing the two bootstrap MI approaches under equal missing data and outlier percentages:* The two imputation approaches are compared in tables 1 and 2. For the same sample size, percentage of outliers and percentage of missing data, bootstrapping residuals approach generally leads to lower standard errors than bootstrapping (x,y) pairs approach for each MI approach and the former approach leads to upward bias while the latter approach leads to downward bias.
2. *Comparing MI approaches under a common bootstrap approach:* In the presence of even a small percentage of outliers and missing data, robust MI based on Huber's M and quantile regression outperforms imputation based on listwise deletion and OLS as reflected in smaller mean squared errors of r in tables 1, 3, and 5 for the bootstrapping (x,y) pairs approach and in tables 2, 4, and 6 for bootstrapping residuals approach.
3. An inspection of tables 1 through 6 suggests that in all of the scenarios, the differences between mean squared errors of sample correlation coefficient r based on OLS-MI, Huber-MI and Quantile-MI under bootstrapping (x,y) pairs approach are significant. Nevertheless, differences among these MI methods based on mean squared errors of r under bootstrapping residuals approach are insignificant and no single method seems to dominate the other methods. Under both approaches, MI based on Huber's M and quantile regression methods lead to lower standard errors than MI based on OLS in the presence of outliers and missing data.
4. *MI approaches under increasing percentage of missing data:* As the percentage of missing data increases, performance of each MI approach deteriorates as reflected in higher mean squared errors of r in tables 5 and 6 relative to tables 1

and 2 respectively. However, performance of OLS-based MI worsens faster relative to performances of MI based on robust methods: Huber's M and quantile regression. Performances of various MI approaches are compared for bootstrapping pairs approach in tables 1 and 5 and for bootstrapping residuals approach in tables 2 and 6. A comparison of tables 1 and 5 suggests that the performances of MI based on listwise deletion and OLS using bootstrapping (x,y) approach worsen significantly as reflected in much higher mean squared errors of r in table 5 than in table 1 while those of Huber's M and quantile regression worsens only slightly as reflected in the small increase in mean squared errors of r . The situation is less clear under bootstrapping residuals approach as revealed by a comparison of tables 2 and 6: with increasing percentage of missing data, MI based on robust estimators Huber's M and quantile regression estimators performs better than MI based on OLS as reflected in lower standard errors of r under the former than the latter but worse as reflected in higher biases and mean squared errors.

5. *MI approaches under increasing percentage of outliers:* A comparison of tables 1 and 3 suggests that as the percentage of outliers increases from 10% to 30% of data, performance of each MI approach deteriorates as reflected in higher MSEs under bootstrapping (x,y) pairs approach. A comparison of tables 2 and 4 seems to suggest that the performance of each method under bootstrapping residuals approach improves with increasing percentage of outliers as reflected in lower MSE for all methods except listwise deletion. However, a closer inspection reveals that this is caused by a faster decrease in mean absolute bias than the

increase in standard errors for each MI approach as the percentage of outliers increases.

4. Conclusion

This article has extended Efron's bootstrap MI for missing data to missing data with outliers on response variable. Imputations are based on robust estimation methods such as Huber's M or quantile regression estimators instead of OLS estimators and smooth bootstrap is applied to either cases or residuals. Evidence from simulation experiments suggests that these bootstrap MI approaches based on robust estimators generally outperform bootstrap MI approach based on listwise deletion as well as OLS in the presence of outliers in the response variable in terms of mean squared errors.

Comparison among various Bootstrap MI approaches

Table 1. Bootstrapping (x,y) pairs, n = 100, 10% outliers, and 10% missing observations with 1000 replications

Method	Mean Bias	Standard Error	MSE
Listwise Deletion	-0.2422837	0.2828748	0.138719544
OLS	-0.2546361	0.2848616	0.145985675
Huber's M	-0.01279331	0.2419862	0.05872099
Quantile	-0.01279355	0.2420035	0.058729369

Table 2. Bootstrapping Residuals, n = 100, 10% outliers, and 10% missing observations with 1000 replications

Method	Mean Bias	Standard Error	MSE
Listwise Deletion	-0.2422837	0.2828748	0.138719544
OLS	0.532881	0.08214618	0.290710155
Huber's M	0.5331839	0.081154	0.290871043
Quantile	0.5330151	0.07984261	0.290479939

Table 3. Bootstrapping (x,y) pairs, n = 100, 30% outliers, and 10% missing observations with 1000 replications

Method	Mean Bias	Standard Error	MSE
Listwise Deletion	-0.1506254	0.4153728	0.195222574
OLS	-0.1609974	0.4193898	0.201807967
Huber's M	-0.0509343	0.3808517	0.14764232
Quantile	-0.05095848	0.2046435	0.044475729

Table 4. Bootstrapping Residuals, n = 100, 30% outliers, and 10% missing observations with 1000 replications

Method	Mean Bias	Standard Error	MSE
Listwise Deletion	-0.1506254	0.4153728	0.195222574
OLS	0.1434735	0.2046435	0.062463607
Huber's M	0.1506375	0.2083522	0.066102296
Quantile	0.1491985	0.2055609	0.064515476

Table 5. Bootstrapping (x,y) pairs, n = 100, 10% outliers, and 20% missing observations with 1000 replications

Method	Mean Bias	Standard Error	MSE
Listwise Deletion	-0.3549752	0.3227382	0.230167338
OLS	-0.38018	0.3264436	0.251102256
Huber's M	-0.01116601	0.2547076	0.065000641
Quantile	-0.01118069	0.2546997	0.064996945

Table 6. Bootstrapping residuals, n = 100, 10% outliers, and 20% missing observations with 1000 replications

Method	Mean Bias	Standard Error	MSE
Listwise Deletion	-0.3549752	0.3227382	0.230167338
OLS	0.3587174	0.08683382	0.136218285
Huber's M	0.3620657	0.084283	0.138195195
Quantile	0.3679083	0.08293019	0.142233934

References

Buck, S. F. (1980) A method for estimation of missing values in multivariate data models for use with an electronic computer, *Journal of Royal Statistical Society, Series B* **22**, 302-306.

Efron, B. (1994) Missing Data, imputation, and the bootstrap, *Journal of the American Statistical Association* **89**, 463-475.

Huber, P. J. (1981) *Robust Statistics*, Wiley, New York.

Koenker, R. (2005). *Quantile Regression*, Cambridge University Press, New York.

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

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