Quantile Regression with Univariate Non-Response

Valentina Zangirolami*, Marco Zanotti*, and Muhammad Amir Saeed

University of Milano-Bicocca, Milan, Italy

Abstract. Our study examines the impact of Missing Completely At Random missing values on quantile regression estimators. In doing so, we conduct a comparative analysis between standard estimators and bootstrap estimators to assess the estimates, considering the extra variability induced by imputation methods for univariate non-response. Overall, our empirical study reveals that listwise deletion maintains similar statistical properties compared to the complete dataset while the sample imputation method exhibits certain issues, suggesting potential limitations in its application.

Keywords: missing data, quantile regression, bootstrap estimators

1 Introduction

The challenge of missing data is a common occurrence in practice, often hiding meaningful values for the statistical analysis. When covariates are affected by missing values, it becomes essential to analyze the non-response mechanism to discern its potential impact on inference. Typically, addressing Missing Completely At Random (MCAR) should yield comparable results in inference when analyzing the subset of data without missing values, as there is no systematic relationship across the population with or without missing values [1]. However, it can lead to consequences on sample size. Therefore, we delve into a scenario of incomplete data with a MCAR mechanism for missing values, aiming to compare complete-case analysis and single-imputation methods. With this aim, we study the impact on estimates when employing quantile regression. Moreover, we adjust the standard errors by using bootstrap resampling techniques taking into account the extra-variability generated by the imputation methods.

This work is organized as follows. Section 2 provides a theoretical overview of missing data, the main methods used in the analysis and quantile regression. Section 3 describes the simulation studies, the experimental settings and the related results. Section 4 concludes this work with final considerations.

2 Methodology

In this section, we begin with outlining the missing data problem and specifically the case of univariate non-response. Finally, we summarize the quantile regression in our framework and the bootstrap estimators.

2.1 Background: missing data problem

Missing data problem arises when some values of one or more variables are not observed. This issue can affect both the response variable and the covariates, as well as one variable (univariate non-response) or more variables (multivariate non-response) in the same dataset. In our framework, we consider univariate non-response case where missing values are generated by a MCAR mechanism.

Let D be the data matrix with dimension $n \times k$, where n is the total number of observations while k represents the number of variables (both dependent and independent). D is composed by observed and missing values, i.e. $D = (D_{obs}, D_{miss})$. Defining the missing-data indicator, such that

$$M = \begin{cases} 1, & \text{missing values} \\ 0, & \text{otherwise} \end{cases}, \tag{1}$$

the MCAR statement guarantees $p(M|D, \phi) = p(M|\phi)$, $\forall D, \phi$ [2].

In this work, we consider two kind of strategies for non-response data: listwise deletion and single imputation methods. The listwise deletion lead to remove all the records which contain a missing value, while the single imputation methods replace the missing value with a single value (e.g. the mean of that variable). Although the listwise deletion might bring about a loss of information, imputation methods may cause a reduction of data variability.

2.2 Quantile regression with univariate non-response

In our case, we consider the following model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i, \tag{2}$$

where Y is the response variable, X is a complete covariate, and Z is a covariate with missing values. Following a quantile regression fashion, the coefficients $\beta = (\beta_0, \beta_1, \beta_2)$ can be estimated by minimizing the following quantity [3]

$$\sum_{i=1}^{n} \rho_\tau(Y - (\beta_0 + \beta_1 x_i + \beta_2 z_i)), \tag{3}$$

where $\tau \in (0, 1)$ and $\rho_\tau(u) = (\tau - I(u < 0))u$. Hence, the estimated model corresponds to $\hat{Q}_\tau(Y|x, z; \hat{\beta}) = \hat{\beta}_{0,\tau} + \hat{\beta}_{1,\tau} x + \hat{\beta}_{2,\tau} z$.

2.3 Bootstrap estimators

Since the use of imputation methods to handle the non-response covariate, we employ bootstrap estimators to involve the extra-variability.

Considering B bootstrap sample D^* from D (i.e. the incomplete dataset), each value of D^*_{miss} should be replaced by a single value following the chosen imputation method. Hence, the bootstrap estimators for the quantile regression coefficients and related standard errors are

$$\hat{\beta}^*_\tau = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}^*_{\tau;b}, \text{ and} \tag{4}$$

$$se^*(\hat{\beta}^*_\tau) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\beta}^*_{\tau;b}-\hat{\beta}^*_\tau)(\hat{\beta}^*_{\tau;b}-\hat{\beta}^*_\tau)^T}. \tag{5}$$

Moreover, if the bootstrap distribution is approximately normal, the coverage for the scalar β_τ considering a $(1-\alpha)$ bootstrap confidence interval corresponds to

$$p(\hat{\beta}^*_\tau - z_{\alpha/2}se^*(\hat{\beta}^*_\tau) < \beta_\tau < \hat{\beta}^*_\tau + z_{\alpha/2}se^*(\hat{\beta}^*_\tau)), \tag{6}$$

where $z_{\alpha/2}$ is the $\alpha/2$-quantile of a normal distribution.

3 Simulation studies

In the context of missing data, we investigate its influence on inference when quantile regression (Section 2.2) is used. Specifically, we contrast the performance of standard estimators with bootstrap estimators (Section 2.3).

In this simulation, we assume $p = 2$ covariates (X, Z) with $X = (x_1, \dots, x_n) \sim U(3, 8)$ and $Z = (z_1, \dots, z_n) \sim U(-1, 5)$ and gaussian errors $\epsilon \sim N(0, 1)$, such that

$$y_i^{(j)} = 3x_i - 0.5z_i^{(j)} + \epsilon_i \quad \forall i = 1, \dots, n. \tag{7}$$

We consider a sample size $n = 500$ and Z to be a variable which contains missing data. In our analysis, we study two regression models which include different $Z^{(j)}$ $(j = 1, 2)$ depending on its percentage of missing values. We assume $Z^{(1)}$ contains 10% of missing value while $Z^{(2)}$ contains 40%, and both satisfy the MCAR mechanism.

3.1 Experimental settings

We compare several methods for handling missing data. Initially, we conduct a complete-case analysis (listwise deletion). Subsequently, we explore imputation techniques, including mean imputation, median imputation, and random imputation, where missing values are replaced with the respective statistic or randomly drawn from observed values. Additionally, we assess these imputation methods using both the original simulated dataset and bootstrap samples. Follow to this, we implement bootstrapping on incomplete datasets. For each bootstrap sample, we apply the previously mentioned imputation methods to replace missing values, resulting in bootstrap estimators along with their associated standard errors.

Furthermore, we employ the quantile regression model to estimate regression coefficients for each non-response method by considering the quantiles of order 0.25, 0.5 and 0.75. Finally, we compare the regression coefficient estimates and standard errors across all models. In the case of bootstrapping on incomplete data, we also evaluate the coverage of confidence intervals by fixing a number of repetitions equal to $B = 200$.

3.2 Results

In Table 1 and Table 2, we provide the regression coefficient estimates for each covariate along with the standard errors.Our comparison involves evaluating these estimates using both the completed simulated dataset (i.e., no missing values) and the incomplete dataset.

Looking the tables, we can observe that the coefficient estimates of both covariates in the case of listwise deletion are closed to the coefficient estimates of the completed dataset. Notably, when the percentage of missing values is low, the estimates of regression coefficients and standard errors exhibit similarity to those derived from the complete dataset. Conversely, a greater percentage of missing significantly affect the estimates. Furthermore, we observe that the 'sample' imputation method tends to yield worse estimates compared to other imputation methods.

	Variable Z: $\tau = 0.25$			Variable Z: $\tau = 0.5$			Variable Z: $\tau = 0.75$		
Method	**0%**	**10%**	**40%**	**0%**	**10%**	**40%**	**0%**	**10%**	**40%**
no-missing	-0.481			-0.429			-0.486		
	(0.036)			(0.037)			(0.04)		
deletion		-0.477	-0.476		-0.426	-0.451		-0.489	-0.455
		(0.037)	(0.043)		(0.038)	(0.053)		(0.042)	(0.059)
sample		-0.42	-0.343		-0.398	-0.289		-0.445	-0.21
		(0.038)	(0.046)		(0.039)	(0.047)		(0.043)	(0.05)
mean		-0.458	-0.449		-0.429	-0.452		-0.479	-0.476
		(0.038)	(0.045)		(0.039)	(0.051)		(0.042)	(0.058)
median		-0.458	-0.449		-0.429	-0.451		-0.475	-0.463
		(0.038)	(0.046)		(0.039)	(0.051)		(0.042)	(0.057)

Table 1. Estimates of regression coefficients related to Z and their standard errors (within the brackets) for each percentage of missing values and non-response method

The presence of missing values on Z significantly affect the estimates of the variable X which does not contain missing values. In Table 3 and Table 4, we provide the bootstrap estimates for each covariate in terms of regression coefficients and standard errors. Moreover, we show the coefficients' coverage considering a bootstrap confidence interval with 95% confidence level. From the tables, we can observe that generally the coverage is around 95% and differences in bootstrap estimates compared to standard estimates.

	Variable X: $\tau = 0.25$			Variable X: $\tau = 0.5$			Variable X: $\tau = 0.75$		
Method	**0%**	**10%**	**40%**	**0%**	**10%**	**40%**	**0%**	**10%**	**40%**
no-missing	3.012			3.02			2.938		
	(0.045)			(0.044)			(0.046)		
deletion		3.015	3.029		3.019	3.023		2.93	2.918
		(0.037)	(0.06)		(0.046)	(0.063)		(0.048)	(0.064)
sample		2.976	2.907		3.01	3.014		2.959	3.047
		(0.048)	(0.055)		(0.045)	(0.052)		(0.047)	(0.054)
mean		2.995	2.927		3.016	3.022		2.942	3.028
		(0.046)	(0.054)		(0.045)	(0.048)		(0.047)	(0.05)
median		2.995	2.929		3.016	3.027		2.946	3.012
		(0.046)	(0.054)		(0.045)	(0.048)		(0.047)	(0.049)

Table 2. Estimates of regression coefficients related to X and their standard errors (within the brackets) for each percentage of missing values and non-response method

	Variable Z: $\tau = 0.25$		Variable Z: $\tau = 0.5$		Variable Z: $\tau = 0.75$	
Method	**10%**	**40%**	**10%**	**40%**	**10%**	**40%**
sample	-0.428	-0.308	-0.408	-0.308	-0.45	-0.264
	(0.041 -0.955)	(0.042-0.965)	(0.036-0.945)	(0.038-0.945)	(0.042-0.955)	(0.057-0.96)
mean	-0.457	-0.453	-0.433	-0.457	-0.482	-0.48
	(0.04 -0.965)	(0.04-0.94)	(0.038-0.95)	(0.047-0.96)	(0.043-0.96)	(0.065-0.975)
median	-0.459	-0.45	-0.434	-0.455	-0.483	-0.473
	(0.039 -0.96)	(0.037-0.95)	(0.038-0.955)	(0.046-0.955)	(0.044-0.96)	(0.064-0.975)

Table 3. Bootstrap estimates of regression coefficients (standard errors - coverage) related to Z for each percentage of missing values and imputation method

4 Conclusion

This work investigates the impact of univariate non-response in quantile regression. Missing data often complicates statistical analysis requiring specific tools for inference. We focus on scenarios of univariate non-response with Missing Completely At Random (MCAR) mechanisms, where missing values occur in only one covariate. To address missing data, we explore various techniques, including imputation methods. Additionally, we leverage bootstrap estimators to account for the extra variability introduced by these methods in quantile regression estimates. From the experimental study, we observe that the listwise deletion method yields estimates closely to those obtained with the complete dataset, whereas imputation methods provide less accurate estimates. Notably, the sample imputation method exhibits really different estimates. Furthermore, our analysis of bootstrap estimators indicates that some standard errors increase due to the additional variability introduced by the imputation method. Since the empirical distribution of estimated coefficients highlight a gaussian behavior, we involve bootstrap confidence intervals assuming gaussian distribution of estimators which were useful to compute the coverage. The latter shows results around

	Variable X: $\tau = 0.25$		Variable X: $\tau = 0.5$		Variable X: $\tau = 0.75$	
Method	**10%**	**40%**	**10%**	**40%**	**10%**	**40%**
sample	2.974	2.916	3.009	3.029	2.959	3.03
	(0.048 -0.98)	(0.05-0.95)	(0.033-0.95)	(0.04-0.955)	(0.049-0.95)	(0.05-0.955)
mean	2.979	2.933	3.011	3.027	2.963	3.014
	(0.05 -0.97)	(0.047-0.94)	(0.032-0.935)	(0.033-0.96)	(0.045-0.945)	(0.052-0.96)
median	2.98	2.93	3.011	3.03	2.961	3.021
	(0.049 -0.975)	(0.05-0.95)	(0.031-0.94)	(0.033-0.94)	(0.044-0.945)	(0.049-0.945)

Table 4. Bootstrap estimates of regression coefficients (standard errors - coverage) related to X for each percentage of missing values and imputation method

0.95. Future enhancements could involve incorporating empirical bootstrap confidence intervals for comparative purposes.

References

1. Nicolini G., Marasini D., Montanari G.E., Pratesi M., Ranalli M.G., and Rocco E.: Metodi inferenziali in presenza di mancate risposte parziali. In: Metodi di stima in presenza di errori non campionari (2013). UNITEXT. Springer, Milano. 1
2. Little J.A., and Rubin D.: Statistical Analysis With Missing Data. 2nd Edition. Wiley Series in Probability and Statistics Book (2002) 2
3. Koenker, R.: Quantile Regression. Econometric Society Monographs. Cambridge University Press (2005) 2