Quantile Regression with Univariate Non-response

Valentina Zangirolami, Marco Zanotti, Muhammad Amir Saeed

University Milano-Bicocca

Contents

1. Missing Data Problem

We considered univariate non-response case where missing values are generated by a MCAR mechanism.

Let D be the data matrix with dimension $n \times k$. D is composed by observed and missing values, i.e. $D = (D_{obs}, D_{miss})$.

The missing-data indicator can be defined such that

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

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

We considered two kind of strategies for non-response data: listwise deletion and single imputation methods.

2. Quantile Regression

We considered the following model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i$$

The coefficients $\beta = (\beta_0, \beta_1, \beta_2)$ can be estimated by minimizing

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

where $\tau \in (0,1)$ and $\rho_\tau(u) = (\tau - I(u < 0))u$.

The estimated model corresponds to

$$\hat{Q}_\tau(Y|x,z;\hat{\beta}) = \hat{\beta}_{0,\tau} + \hat{\beta}_{1,\tau} x + \hat{\beta}_{2,\tau} z$$

3. Bootstrap Estimation

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.

The bootstrap estimators are

$$\hat{\beta}^*_\tau = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}^*_{\tau;b}$$

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

Moreover, the coverage for β_τ corresponds to

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

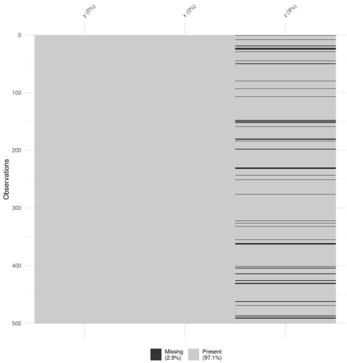
4. Simulation & Results

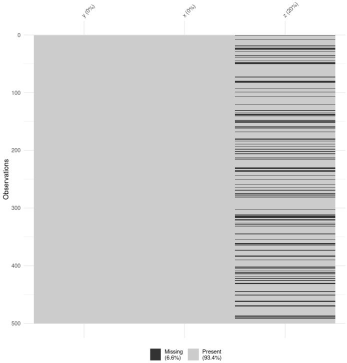
In this simulation, we assumed:

- a sample size $n = 500$
- $p = 2$ covariates (X, Z) with $X = (x_1, \ldots, x_n) \sim U(3, 8)$ and $Z = (z_1, \ldots, z_n) \sim U(-1, 5)$
- gaussian errors $\epsilon \sim N(0, 1)$, such that

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

Z is the variable containing missing values, with two scenarios





Both scenarios satisfy the MCAR assumption.

We compared several methods for handling missing data:

- complete-case analysis
- random imputation
- mean imputation
- median imputation

We estimated the quantile regression model considering the quantiles of order 0.25, 0.5 and 0.75.

We assessed the imputation methods using the original simulated dataset and bootstrap samples.

	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

	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

Method	Variable Z: $\tau = 0.25$ 10%	40%	Variable Z: $\tau = 0.5$ 10%	40%	Variable Z: $\tau = 0.75$ 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

Method	Variable X: $\tau = 0.25$ 10%	40%	Variable X: $\tau = 0.5$ 10%	40%	Variable X: $\tau = 0.75$ 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

5. Conclusions

We investigated the impact of univariate non-response in quantile regression and we concluded that:

- the listwise deletion method yields estimates closely to those obtained with the complete dataset
- the other imputation methods provide less accurate estimates
- the sample imputation method exhibits really different estimates
- the analysis of bootstrap estimators indicates that some standard errors increased due to the additional variability introduced by the imputation method.

Bibliografy

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.

Little J.A., and Rubin D.: Statistical Analysis With Missing Data. 2nd Edition. Wiley Series in Probability and Statistics Book (2002)

Koenker, R.: Quantile Regression. Econometric Society Monographs. Cambridge University Press (2005)

Thank you!