Scholars Journal of Physics, Mathematics and Statistics
Abbreviated Key Title: Sch J Phys Math Stat
ISSN 2393-8056 (Print) | ISSN 2393-8064 (Online)
Journal homepage: https://saspublishers.com

OPEN ACCESS

Quantile Regression-based Multiple Imputation of Skewed Data with Different Percentages of Missingness

Nwakuya, M. T[1*] and Onyegbuchulam B. O.[2]

[1]Department of Mathematics & Statistics, University of Port Harcourt
[2]Department of Mathematics & Statistics, Imo State Polythenic, Umuagwo

DOI: 10.36347/sjpms.2022.v09i04.002 | **Received:** 29.03.2022 | **Accepted:** 05.05.2022 | **Published:** 10.05.2022

***Corresponding author:** Nwakuya, M. T
Department of Mathematics & Statistics, University of Port Harcourt

Abstract

Original Research Article

This study investigates the Quantile Regression-Based Multiple Imputation (QR-based MI) on a simulated right skewed data with 5% and 25% missing data points. Quantile regression analysis on three data sets that comprises of the complete skewed data without missing values, data set with 5% missing values and data set with 25% missing values was performed at 0.25, 0.5, 0.75 and 0.95 quantiles. The data sets with 5% and 25% missing values were imputed using QR-based MI technique, giving rise to two complete data sets. This analysis was performed using both transformed and untransformed version of the three data sets. The transformation was carried out by applying the Yeo-Johnson transformation technique and comparison of results was based on the Mean Square Error (MSE), Akiake Information Criteria (AIC) and Bayesian Information Criteria (BIC). The result from the original complete right skewed data shows that the untransformed data presented better results at 0.25 and 0.50 quantiles compared to the transformed data while results at 0.75 and 0.95 quantiles of the transformed data showed a better result compared to the untransformed. This result is attributed to the fact that the data was right skewed, so that the transformation will benefit the heavy tails on the right while the lighter tail on the left needs not to be transformed hence the 0.25 and 0.50 quantile better result with untransformed data and the 0.75 and 0.95 better result with transformed data. Considering the imputed complete data sets from the 5% and 25% missingness, it was seen that for both data sets at all quantiles considered, the untransformed data produced better results than the transformed data. This led us to conclude that the QR-based MI is not distribution dependent hence it is not sensitive to skewness. Therefore it can be stated based on the results that QR-based MI is robust to skewness, thus can be applied to skewed data sets.

Keywords: Missing data, Quantile Regression-Based Multiple Imputation, Yeo-Johnson Transformation, Quantile regression and Skewed data.

1. INTRODUCTION

Missing data in research has been a topic of discussion for a long time now. Considering the fact that almost all statistical approaches were structured to work with complete data, the repercussions of missingness is a gaping concern. Data points can miss in the response variables and/or in predictor variables. Researcher's aim of obtaining valid and efficient inferences about the population of interest remains the same notwithstanding that data points may or may not be missing. Due to this, the quest to bring solutions have quaked the research world for many years and this has yielded many breakthroughs in that area. The inverse probability weighting (IPW) method is a method to deal with the missing data problem it is also called propensity scoring method but it is sensitive to influential weight and suffers loss of efficiency, Seaman & White (2011). Little and Rubin (1987) came up with a nomenclature for different types of missingness based on the process that generated the missing values. These missingness mechanism as expounded by Rubin and Little (1987) are; missing completely at random (MCAR), this is a mechanism were the missing data are assumed not to be related to the missing values and the observed values; missing at random (MAR) this is where the missing data is said to depend on the observed values but do not depend on the missing values and finally not missing at random (NMAR) this is when the missing data depend on certain missing values. They hinted that MCAR and MAR can be ignored but MNAR cannot be ignored. Hence these mechanisms help as a guide in dealing with missing data during analysis. Multiple Imputation, Maximum Likelihood and Fully Bayesian methods are the three most commonly used model-based approaches in missing data problems, Chen and Ibrahim (2013). There exists two generic pathways for imputing missing values in a multivariate data known as; Joint Modeling (JM) and Fully Conditional Specification (FCS), which is also called multivariate imputation by chained equations (MICE), Van Buuren S. and G-oudshoorn K. (2011). MICE stipulates the multivariate imputation model based on individual conditional density distributions of the variables with missing values, that is each incomplete variable is imputed by a separate model. This paper considers MICE using quantile regression imputation models, known as Quantile Regression-based Multiple Imputation. Parametric analysis is hinged on many assumptions that aid valid inferences but in most cases these assumptions are violated leading to invalid inferences. Researchers have also come up with solutions to such matters, such as generalized methods, quantile regression methods amongst other. When conventional parametric modelling assumptions are not met in the presence of missing data, Quantile regression (QR) is an effective technique for the multiple imputation of the missing data points and also for the data analysis. In Quantile Regression-based Multiple Imputation (QR-based MI), modeling the likelihood is not really essential and it also has some fetching attributes that may be competent in an empirical situation. QR-based MI could be applied in the imputation of the dependent data, censored, bounded and count data. Matteo and Huiling (2013), in their simulation studies, noted that QR-based MI exhibits an edge over other methods in respect to the mean squared error in all frameworks and also in non-normally distributed data but for normally distributed data all methods investigated accomplished satisfactory results. Kleinke et al (2021) noted that irrespective of the glaring upper hand of QR-based MI over normal model-based multiple imputations, formal assessments of QR and Generalized Additive Models for Location Scale and Shape (GAMLSS) based MI are still sparse. This paper presents an implementation of QR-based MI at 5% and 25% missingness in the response variable of a simulated right skewed data with transformation and without the transformation technique. It is assumed that the predictor variables are fully observed. The transformation technique adopted is the Yeo-Johnson transformation that was proposed but Yeo and Johnson (2000). The analysis of this paper is in bi-fold: Firstly simulation of skewed data, injecting of missingness and application of QR-based MI and the second part involves the Yeo-Johnson transformation of the data sets and analysis using both transformed and untransformed data. The aim of this work is probe the effectiveness of QR-based MI in skewed data. This paper assumes that the missingness is ignorable.

2. MULTIPLE IMPUTATION

Multiple imputation (MI) was formally introduced by Rubin (1987). The key idea of the multiple imputation procedure is to replace each missing value with a set of m plausible values, i.e., values “drawn” from the distribution of one’s data that represent the uncertainty about the right value to impute. The conventional multiple imputation adopts the Bayesian technique to create pseudo values from the posterior predictive distribution. This method is hinged on assumptions of the parametric modelling and the prior distributions of the model parameters. The outline of this multiple imputation is given as; Firstly for each missing point in the data a value is imputed from the posterior predictive distribution. Thus forming a complete data, then the required parameters are estimated. This method is repeated m times resulting to m different data sets with m different estimators and variances produced. The average of these estimators and their variances are computed to get a single estimator with its variance. Even though multiple imputation is commonly used, below are some of its limitations.

- It relies on a parametric model and thus it is not robust to model misspecification.
- The checking of the convergence of posterior predictive distribution is back-breaking, Gelman et al. (1996).
- The estimator of the multiple imputation variances usually fails the congeniality condition that is overestimating the variance, Meng (1994).
- Finally misspecification of prior distribution could lead to biased results (Nielsen 2003).

Multiple imputation MI consists of three stages: (1) imputation, (2) analysis, and (3) pooling. Some literatures of imputation includes the work of Hargarten and Wheeler, (2020) were they applied the WQS (weighted quantile sum) regression in the multiple imputation (MI) framework in order to fully account for the uncertainty due to censoring.

3. QUANTILE REGRESSION-BASED MULTIPLE IMPUTATION

Quantile regression over the years has proved to represent a holistic regression method in Econometrics, applied statistics and in many other fields of research contrary to the least square regression which is a centered regression and yields optimal results in normally distributed data. Quantile regression has the advantage of working well with skewed or non-normally distributed data and it also identifies the varying effects of the predictors on different segments of the distribution and it is equally robust to outliers. The estimates of the model parameters of the least square regression are obtained through the minimization of the loss function of the mean square error.

Given a linear model; $y_i = X\beta + e$………………. (1)

Where y is the response variable, x is the regressor variable, β is the regression parameter and e is the error term. The estimated response is given as have $\hat{y}_i = X\hat{\beta}\ hence, the\ error\ term\ e = y_i - X\hat{\beta}.$ Given the conditional mean function,

$E(Y/ X = x) = \; X\beta$ ……………………… (2)
β is estimated thus,
$\hat{\beta} = \text{argmin}_{\beta \in R} \sum (y - X\beta)^2$ ……………………. (3)

Given the linear conditional quantile function
$Q_\tau(Y \,/\, X) = \; X\beta_\tau$ ………………… (4)

Koenker and Bassett (1978) proposed the estimation of the quantile regression model parameters as given by;
$\hat{\beta}_\tau = \text{argmin}_{\beta \in R} \sum \varrho_\tau(Y - X\beta)$ ………………… (5)

Where $\varrho_\tau(e)$ the loss function is defined as $\varrho_\tau(e) = e(\tau - 1(e \leq 0))$, given that 0<τ<1.

The Quantile regression-based multiple imputation method was proposed and implemented in R by Geraci (2016) and also by Geraci and McLain (2018). The quantile regression-based MI method is implemented in the Qtools package in R using the function mice.impute.rq(). It utilizes the universality of the uniform distribution known as the probability integral transform theorem, were given that U ~ Unif(0, 1), then F^{-1} (U) ~ F. To obtain the multiple imputation, let's assume $f(Y \,/X\,)$ to be the conditional density function for which y is the response variable that is not fully observed while the predictor variable x is fully observed, and $Q_\tau(Y \,/\, X)$ is the τ^{th} conditional quantile function and it is equivalent to the inverse conditional distribution function that is;
$Q_\tau(Y \,/\, X) = F_\tau^{-1}(Y \,/\, X)$ ……………. (6)

The imputation entails estimating of $Q_\tau(Y \,/\, X)$ using observed data under the ignorable missingness assumption. Then multiple imputed values $y_i^*(i = 1, \dots I)$ are obtained $y_i^* = \hat{Q}_\tau(Y \,/\, X)$, τ is simulated independently from a uniform distribution.

4. METHODOLOGY

A right skewed data set of n=500 with no missingness was simulated using the mnonr package in R and quantile regression analysis was carried out on the data and the transformed version of the data at 0.25, 0.50, 0.75 and 0.95 quantiles. Later the simulated data was injected with 5% missingness producing an incomplete data set with 5% of the values missing and 25% missingness was also injected producing an incomplete data set with 25% of the values missing. These two data sets with missing values were then imputed independently using QR-based MI technique, producing two complete data sets. Quantile regression analysis was carried out on these two data sets, then the data sets were transformed and quantile regression analysis was equally applied again on the transformed data sets, producing results from untransformed and transformed data sets. The results of the analysis were compared based on the Mean Square Error (MSE), Aikake Information Criteria (AIC) and Bayesain Information Criteria (BIC). The transformation of the data sets was applied using the Yeo-Johnson transformation method; this was used because the simulated data has both negative and positive values.

Missingness Mechanism: The missingness mechanism adopted was ignorable missingness. In other to show the missing mechanism, a random sample $\{x_i, y_i\}_{i=1}^n$ where only y_i is missing and x_i is fully observed was considered. Let ϑ_i be the missingness indicator, where $\vartheta_i = 1$ means that y_i is not missing and $\vartheta_i = 0$ means y_i was not observed. It is assumed that the missingness probability doesn't depend on either X or Y, hence ignorable and it is denoted as,
$P(\vartheta_i /\, x_i, y_i) = P(\vartheta_i)$ ……………… (6)

QR-based MI Procedure: Quantile regression-Based multiple imputations as given by Chen (2014) proceeds as follows;

- Draw τ_i independently from a uniform (0,1) ifor i = 1, 2, · · · , I.
- $\hat{\beta}_{\tau_i}$are estimated for each $i = 1, \dots, I$ using the formula $\hat{\beta}_\tau = \text{argmin}_{\beta \in R} \sum \vartheta_i \varrho_\tau(Y - X\beta)$
- I independent values are imputed for every missing point as $y_i^* = \hat{Q}_\tau(Y \,/\, X) = X\hat{\beta}_{\tau_i}$, this procedure is repeated for every missing point in the data set and this eventually forms a complete data set.

Yeo-Johnson Transformation Procedure: Yeo and Johnson (2000) came up an alternative family of transformations that addresses the limitation of transforming response variable with only positive values and extends it to accommodate negative response values. These transformations are described by the function below:

$$G_\lambda(y^*) = \begin{cases} \dfrac{\left((1+y)^\lambda - 1\right)}{\lambda} \; if \; \lambda \neq 0 \; and \; y \geq 0 \\ \log(y+1) \;\; if \; \lambda = 0 \; and \; y \geq 0 \\ \dfrac{-[(-y+1)^{2-\lambda} - 1)]}{(2-\lambda)} \; if \; \lambda \neq 2 \; and \; y < 0 \\ -\log(-y+1) \; if \; \lambda = 2 \; and \; y < 0 \end{cases}$$

Where y* is the transformed response and λ is the transformation parameter.

5. RESULTS

Table 1: Results from analysis without missingness values

Comparison Criteria	Tau	Untransformed	Transformed
MSE		**1.130385**	1.145189
AIC	0.25	**67.27912**	73.78467
BIC		**19.77421**	19.78901
MSE		**0.7168231**	0.755267
AIC	0.50	**-160.4631**	-134.3422
BIC		**19.36065**	19.39909
MSE		1.155202	**1.154916**
AIC	0.75	78.13771	**78.01395**
BIC		19.79903	**19.79874**
MSE		3.934031	**2.457656**
AIC	0.95	690.8323	**455.6041**
BIC		22.57786	**21.10148**

The result from table 1 above reveals that for the complete right skewed data, 0.25 and 0.50 quantiles of the untransformed data presented better results compared to the transformed data while 0.75 and 0.95 quantiles of the transformed data showed a better result. This result is attributed to the fact that the data is right skewed.

Table 1: Results from analysis 5% missingness values

Comparison Criteria	Tau	Untransformed	Transformed
MSE		**1.005203**	1.955456
AIC	0.25	**8.594631**	341.3116
BIC		**19.64903**	20.59928
MSE		**0.5446861**	1.16777
AIC	0.50	**-297.7728**	83.54815
BIC		**19.18851**	19.81159
MSE		**0.7675921**	2.013601
AIC	0.75	**-126.2451**	355.9624
BIC		**19.41142**	20.65743
MSE		**1.83434**	4.485849
AIC	0.95	**309.3423**	756.4639
BIC		**20.47816**	23.12967

The result in table 2 above shows that at all quantiles considered the untransformed data produced better results. This led us to conclude that the QR-based MI is not distribution dependent hence it is not sensitive to skewness.

Table 3: Results from analysis with 25% missingness values

Comparison Criteria	TAU	Untransformed	Transformed
Mean Square Error	0.25	**1.3842637**	1.621586
AIC		**168.584**	247.7025
BIC		**20.02809**	20.2654
Mean Square Error	0.50	**0.9883006**	1.136581
AIC		**0.1158274**	70.01246
BIC		**19.63212**	19.78041
Mean Square Error	0.75	**1.074175**	1.456702
AIC		**41.77627**	194.0875
BIC		**19.71800**	20.10073
Mean Square Error	0.95	**4.785968**	5.021932
AIC		**788.8437**	812.9074
BIC		**23.42979**	23.66576

Table 3 above also shows that the untransformed data produced better results at all quantiles considered. This we also attribute to the fact that QR-base MI is not distribution dependent therefore it is not sensitive to skewness.

6. DISCUSSION AND CONCLUSSIONS

The paper studies the QR-based MI on both transformed and untransformed right skewed data with missing values and without missing values. The transformation technique adopted was the Yeo-Johnson transformation, the technique was considered because the simulated data sets have both negative and positive values. In the analysis three different data sets were considered; the original complete skewed data, the data with 5% missing values in the response variable and data with 25% missing values in the response variable. The two data sets with missing values were both imputed independently using the QR-based MI method. Analysis was carried out on the original complete skewed data with transformation and without transformation. Then analysis was also carried out on the two data sets that were imputed without transforming the data sets. These imputed data sets were later transformed and analysis also carried out on the transformed data sets. The quantile regression analysis at 0.25, 0.5, 0.75 and 0.95 quantiles was employed. The result from table 1 discloses that for the complete right skewed data, results from the untransformed data presented better results at 0.25 and 0.50 quantiles compared to the transformed data while results at 0.75 and 0.95 quantiles of the transformed data showed a better result. The result from table 1 is attributed to the fact that the data was right skewed, so that the transformation will benefit the heavy tails on the right while the lighter tail on the left needs not to be transformed; hence the 0.25 and 0.50 quantile better result with untransformed data and the 0.75 and 0.95 better result with transformed data. The result in Table 2 discloses that at all quantiles considered the untransformed data produced better results. This led us to conclude that the QR-based MI is not distribution dependent hence it is not sensitive to skewness. Likewise the table 3 also discloses that at all the quantiles considered the untransformed data produced better results, agreeing with the fact that QR-based MI is distribution free. Therefore it can be stated based on the results that QR-based MI is robust to skewness, thus can be applied to skewed data sets.

REFERENCES

- Bottai, M., & Zhen, H. (2013). Multiple Imputation based on conditional quantile estimation. *Epidemiol Biostat Pub Health*, 19(1). DOI: 10.2427/8758.
- Chen, S. (2014). Imputation of missing values using quantile regression. Graduate Theses and Dissertations. 13924. https://lib.dr.iastate.edu/etd/13924
- Chen Q., & Ibrahim J. G., (2013). A note on the relationships between multiple imputation, maximum likelihood and fully Bayesian methods for missing responses in linear regression models. *Statistics and Its Interface*, 6(3), pp 315–324. DOI: https://dx.doi.org/10.4310/SII.2013.v6.n3.a2
- Gelman, A., Hill, J., Su, Y. S., Ya, M., & Pittau, M. G. (2013). Missing data imputation and model checking. http://www.R-project.org/
- Geraci, M. (2016). Estimation of regression quantiles in complex surveys with data missing at random: An application to birthweight determinants. *Statistical Methods in Medical Research*, 25(4), 1393-1421.
- Geraci, M., and McLain, A. (2018). Multiple imputation for bounded variables. *Psychometrika*, 83(4), 919-940.
- Hargarten, P. M., & Wheeler, D. C. (2020). Accounting for the Uncertainty Due to Chemicals Below the Detection Limit in Mixture Analysis. *Environmental Research*, 186(109466). https://doi.org/10.1016/j.envres.2020.109466. [p226, 227, 228, 230, 234, 244]
- Kleinke K., Fritsch M., Stemmler M., Reinecke J., & Lösel, F. (2021). Missing Values — An Evaluation and Application to Corporal Punishment Data. M*ethodology*, 17(3), 205–230. https://doi.org/10.5964/meth.2317
- Meng, X. L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statist Sci*, 9, 538–558.
- Nielsen, S. F. (2003). Proper and Improper Multiple Imputation. International statistical review = Revue internationale de statistique, 71, 593–607.
- Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. Wiley, New York
- Little, R. J. A., & Rubin, D. B. (1987). Statistical Analysis with Missing Data. Wiley, New York, 1987, IV+278 pp.
- Seaman, S. R., & White, I. R. (2011). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, 22, 278–295.
- Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1–67.
- Quantile Regression for Longitudinal Data. *Journal of Statistical Research*, 55(1), pp 43-58. https://doi.org/10.47302/jsr.2021550105
- Yeo, I., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4).