Conditional Logistic Quantile Imputation for Continuous Bounded Outcomes with Moderate to Heavy Skew

Daniel Jang[1,2], Robert Thiesmeier[2], Nicola Orsini[2]

[1]Amherst College

[2]Karolinska Institutet

May 4, 2025

Abstract

The recording of biomarkers, which are variables that are often continuous and bounded, in studies is difficult, expensive, and time-consuming. Except in ideal scenarios, a study may not have a complete and accurate set of biomarker data for statistical analysis, leading to missing data and biased estimates of parameters of interest. Existing Missing Data methods - such as predictive mean matching (PMM) - can produce unbiased estimates for this type of data, but has limitations. We propose Conditional Logistic Quantile Imputation (CLQI) as another method to impute continuous and bounded variables which 1) does not produce implausible values, 2) can handle a higher proportion of data missing without introducing bias to the estimates, and 3) is simple in its implementation. We perform a simulation study to compare the two methods and found that CLQI is a valid MI method that is statistically more efficient method for this type of missing data compared to PMM.

1 Introduction

The recording of biomarkers - defined by the World Health Organization as "any substance, structure, or process that can be measured in the body or its products and influence or predict the incidence of outcome or disease" - is difficult, expensive, and time-consuming. Thus, except in ideal scenarios, studies that involve biomarkers can run into problems with a lack of funding, time restraints, or technical failures, leading to missing data. This leads to potentially biased estimates of parameters of interest, even when performing a complete-case analysis. To remedy this problem, certain Multiple Imputation (MI) procedures can be implemented on the Biomarker data to maintain statistical power, reduce bias, and increase the efficiency of estimates. However, Biomarkers are often continuous, bounded, and heavily skewed, which adds various methodological hurdles.

Predictive Mean Matching (PMM), first introduced by Rubin [1987], is a standard semi-parametric MI technique that imputes an actual observed value from a set of $k > 1$ values closest to the algorithm's prediction for the missing value. This immediately distinguishes the method in two ways: (a) it successfully avoids imputing implausible values for the data, and (b) PMM is better able to preserve the distribution of the empirical data compared to parametric MI methods. Furthermore, PMM is well-integrated to many MI packages, most notably the `mice` package in R. However, several studies have identified key limitations for PMM that are often ignored. [Kleinke, 2017]. 1) The algorithm is based on means and not the shape of the distribution. This creates a problem when attempting to impute values when the true distribution is quite skewed. Imputed values can cluster in a narrow range, resulting in biased estimates. 2) Imputed values are equivalent to existing data. Although this avoids imputing any implausible values - i.e. negative values in the context of biomarkers - this introduces problems such as a higher standard error in the estimate and a failure to capture extreme values that are in the population but are not present in the data. This can lead to bias in the estimates, especially when the true max of the population is known. Furthermore, when the effect size between the true effect and no effect is small, PMM has much lower power. 3) Since the method is based on linear regression, there is an assumption of homoscedasticity and errors being normally distributed and centered around zero, which are often violated in real-world applications. Transformations may help, but this is not ideal. 4) The method only works best when the proportion of missing data does not exceed 30% and a decently large sample size.

What we propose is a new Multiple Imputation method called Conditional Logistic

Quantile Imputation (CLQI) based on Logistic Quantile Regression (LQR) - first described by Bottai et al. [2010] - and Conditional Quantile Imputation (CQI) - first described by Bottai and Zhen [2013]. CLQI addresses the aforementioned problems while maintaining PMM's greatest advantage - avoiding imputing implausible values. For one, since CLQI relies on quantiles instead of means, this method is more flexible with a wider range of non-standard distributions. Furthermore, there is no assumption of homoscedasticity and symmetric results with LQR. Second, not only are the imputed values plausible by design, they are not necessarily equivalent to any observed data, which can potentially decrease standard error in our final estimate. These two advantages alone make CLQI an appropriate alternative to PMM with heavily-skewed continuous bounded outcomes.

The paper is structured as follows. First, we describe the PMM and CLQI methods in detail, along with important assumptions and conditions. Next, we introduce simulations as a means to compare both methods. We then discuss the potential use cases and limitations CLQI has, both compared with PMM and in general.

2 Methods

The following section describes how the PMM and CLQI algorithms work, as well as define important assumptions and conditions. Before describing them, however, we must first establish notation we will be using. Y represents the continuous and bounded variable for imputation. $I \subseteq \boldsymbol{N}$ represents the number of observations in our study, and i represents its index. Let $\boldsymbol{X} = [X_1, X_2, \ldots, X_p] \in \mathbb{R}^{i \times p}$ denote the matrix of predictors for Y, where each column $X_k = (x_{k1}, x_{k2}, \ldots, x_{ki})^\top$ represents the values of predictor k across all i observations. We assume that there is no missingness in $\boldsymbol{X}$; that is, $\forall k \in \{1, 2, \ldots, p\}$, the variable X_k is fully observed. Let $I_C \subset I$ be the subset of data with y_i observed, and let its complement $I_M \subset I$ be the subset of data with y_i missing. Thus, define $\boldsymbol{X_C} = \{\boldsymbol{x}_i | i \in I_C\}$ to be the matrix of predictor values corresponding to complete cases and $\boldsymbol{X_M} = \{\boldsymbol{x}_i | i \in I_M\}$ to be the matrix of predictor values corresponding to incomplete cases. Correspondingly, y_m represents the mth missing observation and $y_m^{(imp)}$ represents the imputed value. Lastly, let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_{p+1})$ be the vector of estimated regression coefficients for our model.

2.1 Predictive Mean Matching

Predictive Mean Matching (PMM) works in the following steps:

1. Develop a predictive model using complete-case data with our set of predictors - $\boldsymbol{X}_C$. In our case, we will be using linear regression.

$$\hat{Y} = \boldsymbol{X}_C^T \hat{\boldsymbol{\beta}} \tag{1}$$

2. Use the model to generate predicted observations $\hat{y}_i$ for both complete and incomplete cases.

3. For each missing observation y_m, where $m \in I_M$, and its corresponding predicted observation $\hat{y}_m$, identify n closest values from the predicted observations of the fully observed data $\hat{y}_c$, where $c \in I_C$. We notate the set of these values C of cardinality k, which is our so-called "donor pool". Common donor pool sizes are 3, 5, and 10. With smaller samples (i.e. $N < 100$), a smaller donor pool has been shown to give more accurate results.

$$C = \min_n \{|\hat{y}_m - \hat{y}_c| \mid c \in I_C\} \tag{2}$$

4. Randomly choose with equal probability from the set C and impute:

$$y_m^{(imp)} = y_c, \quad \text{where } y_c \sim \text{Unif}(C) \tag{3}$$

2.2 Conditional Logistic Quantile Imputation

Conditional Logistic Quantile Imputation (CLQI) works in the following steps:

1. Log-transform all observed values of Y using $y_{min} = 0$ and y_{max}. The value of y_{max} that a study chooses can vary depending on context. We will discuss further in Section 3 what an appropriate value for y_{max} should be, and if the different choices make substantial differences in the final analysis.

$$h(y_i) = log(\frac{y_i - y_{min}}{y_{max} - y_i}) = logit(y_i) \tag{4}$$

2. For all missing observations y_m where $m \in I_M$:

   - Draw a random value from $u \sim Unif(0, 99)$
   - Fit an LQR model on the uth quantile using complete-case data with our set of predictors - $\boldsymbol{X}_C$

$$Q_u(p \mid x_i) = \boldsymbol{X}_C^T \hat{\boldsymbol{\beta}}_u \tag{5}$$

- Generate a predicted value of y_m based on the data for the observation's set of predictors. This will be our imputed value. Call this value $\hat{z}_m = logit(y_m^{(imp)})$

3. Untransform $logit(Y)$ after all missing data is imputed using the same y_{min} and y_{max} values as in step 1

$$Q_y(p) = \frac{exp(z_i)y_{max} + y_{min}}{1 + exp(z_i)} \tag{6}$$

For both PMM and CLQI, according to MI procedures, we repeat the algorithm M times, saving both $\hat{\theta}_m$ and $\hat{V}(\hat{\theta}_m)$. We then combine the results using Rubin Rules to obtain our final estimates $\hat{\theta}$ and $\hat{V}(\hat{\theta})$.

2.3 Assumptions and Conditions

Both PMM and CLQI are semi-parametric methods, but CLQI has fewer assumptions due to relying on Logistic Quantile Regression and not linear regression. That being said, there are still a common set of assumptions and conditions that we follow throughout the simulation sections

1. Variable of interest must be Missing Completely at Random (MCAR) or Missing at Random (MAR)

2. We must have a sufficient sample size for our donors for regression. This is probably around the range of $n = 100$ actual observed observations.

3 Simulation Study

The following section introduces the simulation study to compare PMM and CLQI in imputing continuous, bounded, and skewed variables. We first describe the Data Generating Mechanism (DGM) to generate our skewed explanatory variable for imputation. We then describe two distance metrics to assess the quality of one imputed distribution compared to the theoretical distribution. After, we describe XX simulation scenarios we want to explore in comparing PMM and CLQI, with the max value informed by the Sensitivity Analysis done in Section 3. The aim of the simulation study is to show CLQI's performance under settings where PMM would be inappropriate, i.e. with higher proportion of missing data and higher levels of skew. All simulations were done in R version 4.4.1, and PMM was done with the `mice` package.

3.1 Data Generating Mechanism (DGM)

For our simulation study, we chose a simple setting with a binary confounder, a continuous biomarker, and a binary outcome.

- **Binary Confounder:** C

$$C \sim \text{Bern}(\pi_c), \quad \pi_c = 0.4$$

- **Continuous Predictor:** V

$$V \sim \text{Norm}(\mu = 0, \sigma = 2.5)$$

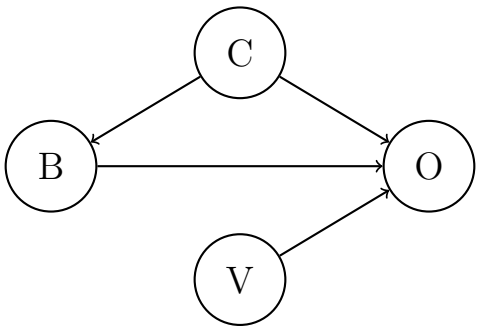
- **Continuous Biomarker:** B

$$B \sim \chi^2(5 + 3C)$$

- **Binary Outcome:** O

$$O \sim \text{Bern}(\pi_d), \quad \pi_d = \text{expit}(\beta_0 + \beta_1 B + \beta_2 C + \beta_3 V)$$

We will be using the following parameter values: $\beta_0 = \text{logit}(0.1)$, $\beta_1 = \ln(1.1)$, $\beta_2 = \ln(0.7)$, $\beta_3 = \ln(0.85)$. Our target parameter is β_1. Below is a Causal Diagram of the DGM, summarizing the relationships we defined above.



3.2 Sensitivity Analysis for CLQI

This subsection goes through a brief Sensitivity Analysis to determine if the choice of y_{max} in CLQI affects the estimate of the parameter of interest or not. We have four potential candidates: **1)** $y_{max} = \max(B_{obs})$ from the data **after** missingness is induced, **2)** $y_{max} = \max(B)$ from the data **before** missingness is induced, **3)** the 99.99^{th} quantile of the theoretical mixture distribution for C, and **4)** the 99.9999^{th} quantile of the theoretical mixture distribution for C. 1) is the naive choice for a max value if no other assumptions of the data can be made, which can prevent CLQI from covering more extreme values. 2) is the ideal scenario where the true max of the observed data is captured every time. 3) and 4) are meant to imitate scenarios where a max value for B is pre-specified in the analysis, potentially from previous knowledge of the population distribution, biological maximum limit, etc.

We chose a 30% missing data and $n = 1000$ scenario to test the sensitivity of our choice of y_{max}, with $m = 10$ imputations and $N = 1000$ iterations of the simulation. To maintain consistency, for each iteration, the exact same data was generated, the same data was induced missing, and the random u value was the same for each row.

From Figure 1, what we found is that the choice for y_{max} does not seem to greatly affect the parameter estimates, coverage, and power. However, it is worth noting that choosing a y_{max} value that cuts off potential extreme values from being imputed will perform worse overall than a higher y_{max} value, even if the higher y_{max} value is extremely unlikely to occur. Therefore, we will be approaching simulations for CLQI with y_{max} being the $99.9999th$ quantile of the theoretical mixture distribution of B.

Table 1: Sensitivity Analysis of y_{max} choice in CLQI

Metric	$\mathbf{max}(B_{obs})$	$\mathbf{max}(B)$	$99.99^{th}Q$	$99.9999^{th}Q$
Bias	0.0012	0.0010	0.0010	0.0010
RMSE	0.0391	0.0389	0.0385	0.0385
Coverage	94.8	92.9	94.4	94.3
Power	83.1	85.1	85.2	85.7

3.3 Distance Metrics

We will use the Kolmorogrov-Smirnov (KS) distance metric to compare the quality of the imputation for PMM and CLQI compared with the distributions they were

generated from. The KS distance reports the largest distance observed. Let F_T be our theoretical distribution and let F_I be our distribution with imputation. For the future, we will also report the test statistic and p-value.

$$D_{KS}(F_T, F_I) = sup_x|F_T(x) - F_I(x)| \tag{7}$$

3.4 Simulation Scenario

We will simulate one (for now) scenario, with $N = 1000$ simulations: 30% of B missing completely at random (MCAR) with $n = 1000$. We set the number of imputations to be $m = 10$ for both PMM and CLQI for computational efficiency. Lastly, we also perform Complete-Case (CC) analysis on the data as a benchmark comparison. We will be focusing on four performance metrics: 1) Relative Bias, 2) Root Mean Square Error (RMSE), 3) Coverage, and 4) Power. For PMM, we will use a donor pool size of $n = 5$, and for CLQI, we will use $y_{max} = 99.9999^{th}$ quantile of the theoretical mixture distribution B. Note that different simulation scenarios may require adjustments to these choices.

3.5 Results

3.5.1 CLQI Quality for One Imputation

One imputation using CLQI shows that our individual imputations are approximately following the conditional theoretical distributions ($\chi^2(5)$ and $\chi^2(8)$) in the absence (blue) and presence (orange) of the confounder shown in Figure 1. The KS distances are also small enough, as shown in Table 2

Table 2: KS Distance for Imputed Distributions by Confounder

Confounder	Metric	Distance
0	Kolmogorov–Smirnov	0.041
1	Kolmogorov–Smirnov	0.028

3.5.2 Repeated Sampling: Comparison with PMM, CLQI, and CC

The results for our simulation scenario are given by Figure 2 and Table 3. Under many replications, although there is a slight increase in relative bias (3.34%) for

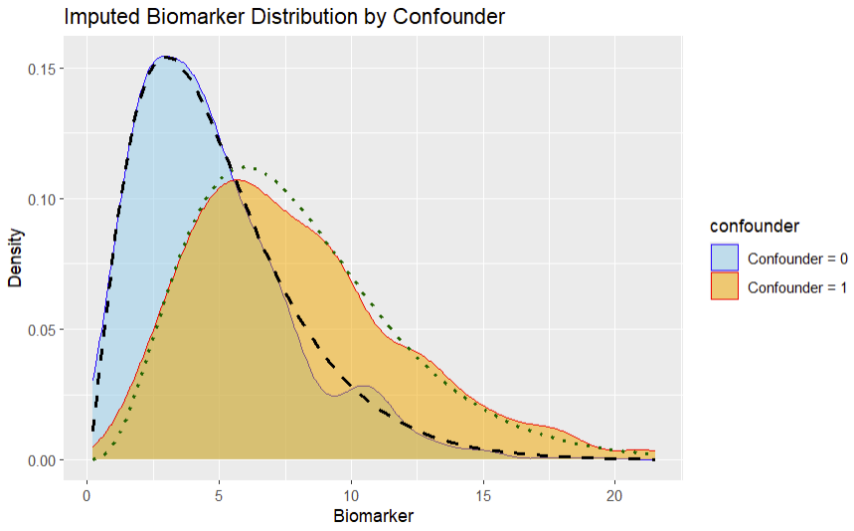


Figure 1: CLQI: Theoretical distributions (dotted lines) and imputed distributions (blue and yellow) faceted on the binary confounder variable with 30% of data missing

CLQI, it is a much more efficient method compared to traditional PMM. This is noted by the much lower RMSE (0.039) and higher power (85.7%) (Table 1), and in the narrower distribution (Figure 3). We also have a good coverage (approximately 95%) for CLQI, so we do not have a negative tradeoff with a smaller standard error.

Table 3: Performance Comparison of Methods: 30% Missing and $n = 1000$

Method	**Rel Bias (%)**	**RMSE**	**Coverage (%)**	**Power (%)**
PMM	0.97	0.055	99.2	32.9
CLQI	3.34	0.039	94.5	85.7
CC	1.00	0.099	98.5	90.6

4 Discussion

CLQI is a valid and promising alternative MI method to PMM that can be more efficient when dealing with continuous and bounded variables with moderate to heavy skew. More testing must be done to confirm our initial results, especially under different scenarios to see if performance metrics include. Furthermore, we are looking

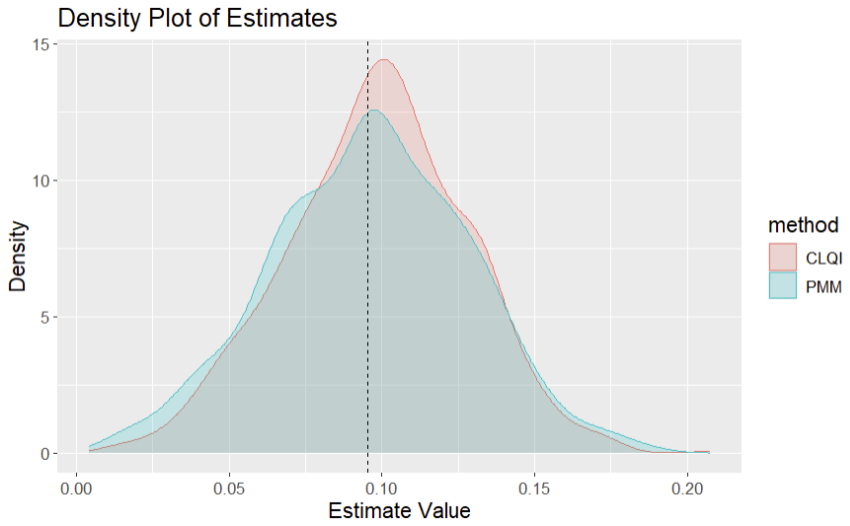


Figure 2: Distribution of Estimates for the Biomarker given by PMM and CLQI respectively over 1000 simulations for 30% missing data, with a line at the set parameter value of $ln(1.1) = 0.095$

to improve runtime, as CLQI is a very resource-intensive method at its current state. R code for the simulation study can be found on Github, which will be linked in the future.

We are planning to develop an R package for easier implementation of our method, which will be called `CLQI`, as well as expand this method for limit of detection data with aggregate data sharing between studies.

References

Matteo Bottai and Huiling Zhen. Multiple imputation based on conditional quantile estimation. *Epidemiology, Biostatistics, and Public Health*, 10(1), 2013. doi: 10.2427/8758. URL https://riviste.unimi.it/index.php/ebph/article/view/18261.

Matteo Bottai, Bo Cai, and Robert E. McKeown. Logistic quantile regression for bounded outcomes. *Statistics in Medicine*, 29(2):309–317, 2010. doi: https://doi.org/10.1002/sim.3781. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.3781.

Kristian Kleinke. Multiple imputation under violated distributional assumptions: A systematic evaluation of the assumed robustness of predictive mean matching. *Journal of Educational and Behavioral Statistics*, 42(4):371–404, 2017. doi: 10.3102/1076998616687084.

Donald B. Rubin. Multiple imputation for nonresponse in surveys. *John Wiley & Sons*, 1987.