Journal of Statistical Research *https://doi.org/10.47302/jsr.2021550105*
2021, Vol. 55, No. 1, pp. 43-58 *ISSN 0256 - 422 X*

THE IMPACT OF INCOMPLETE DATA ON QUANTILE REGRESSION FOR LONGITUDINAL DATA

ANNELEEN VERHASSELT

Data Science Institute, I-Biostat, Universiteit Hasselt, Diepenbeek, Belgium.

Email: anneleen.verhasselt@uhasselt.be

ALVARO J. FLÓREZ

Data Science Institute, I-Biostat, Universiteit Hasselt, Diepenbeek, Belgium.
School of Statistics, Universidad del Valle, Cali, Colombia.

Email: alvaro.florez@uhasselt.be

INGRID VAN KEILEGOM

Research Centre for Operations Research and Statistics, KU Leuven, Leuven, Belgium.

Email: ingrid.vankeilegom@kuleuven.be

GEERT MOLENBERGHS*

Data Science Institute, I-Biostat Universiteit Hasselt, Diepenbeek, Belgium.
I-Biostat, KU Leuven, Leuven, Belgium.

Email: geert.molenberghs@uhasselt.be

SUMMARY

We investigate the performance of methods for estimating the conditional quantile of a response based on longitudinal data, when outcomes are incomplete and when the correlation between repeated responses is ignored. In a simulation study, we compare the performance of the quantile regression estimator based on the complete cases, the available cases, quantile-based multiple imputation, and quantile-based inverse probability weighting. In the data setting considered, quantile-based multiple imputation is the most promising method with the best bias-efficiency trade-off. A potential drawback, however, is its computation time.

Keywords and phrases: Dropout; Inverse probability weighting; Missing data; Multiple imputation; Quantile regression

* Corresponding author

1 Introduction

In a setting where longitudinal data are collected, often not all scheduled measurements of a subject's outcome are observed (see Fitzmaurice et al. (2008) for an overview of longitudinal data analysis). In this paper, we focus on dropout, where a subject drops out from the study at a certain occasion, after which time there are no recordings. Inference based on only the complete cases in a longitudinal study generally leads to biased and inefficient estimators (as shown in the simulation study, see Section 3). Therefore, there is a need for techniques that properly handle missing data (see among others Little and Rubin (2014) and Molenberghs et al. (2014)).

The nature of the missingness mechanism highly influences the performance of statistical techniques that deal with missing data. According to Rubin (1976), there are three main missing data mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Under MCAR, missingness does not depend on either the observed or unobserved variables, apart from perhaps covariates. When missingness is independent of the unobserved measurements conditional on the observed ones, the process is called MAR. MNAR occurs when neither MAR nor MCAR hold. The latter is a very general and therefore realistic assumption, of course, but leads to challenging inferential problems. The most commonly used assumption, at least for a primary analysis in a variety of settings, such as clinical trials, is MAR.

Two commonly used methods to deal with missing data are inverse probability weighting (IPW) methods (Horvitz and Thompson, 1952; Robins et al., 1994, 1995; Lipsitz et al., 1998) and multiple imputation (MI) (Rubin, 1987; Lipsitz et al., 1998; Aerts et al., 2002; Wei et al., 2012; van Buuren, 2012; Carpenter and Kenward, 2013). MI uses the observed data to generate values for the missing data from the predictive distribution of what is unobserved given what is observed. Once the data are imputed, standard inference techniques for complete data are used together with relatively straightforward combination rules. IPW methods, on the other hand, change the optimization problem, by re-weighting observations as a a function of the underlying missingness mechanism. Several authors have compared IPW and MI, predominantly in fully parametric or semi-parametric settings, such as generalized estimating equations (GEE). For example, Beunckens et al. (2008) concluded that MI combined with GEE performed better in terms of mean squared error than weighted GEE. This is not totally surprising. Indeed, as is known from the survey literature as well, a weighted analysis tends to decrease precision relative to the corresponding unweighted analysis. The effect is exacerbated when the variability in observation-driven weights, rather than design-based weights is taken into account, but the decreased precision manifests itself, even when this additional variability is ignored.

The focus of this paper is placed on techniques dealing with missing data when the aim is to estimate the effect of covariates on a quantile of the response, i.e., quantile regression (Koenker and Bassett, 1978). Quantile regression allows to examine the effect of covariates on different quantiles of a response (and not only the center of the distribution). The method is therefore also useful for asymmetrically distributed responses and models with heteroscedastic errors.

The model and methods are described in Section 2. Section 3 contains a simulation study. A motivating study on ophthalmology is analyzed in Section 4. Finally, we end the paper with concluding remarks in Section 5.

2 Model and Methodology

We focus on estimating the τ-th quantile (with $0 < \tau < 1$) of a response given the covariates. We assume that we have observations $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$ (for $i = 1, \ldots, n$), where $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$ is a n_i-dimensional response vector for individual $i = 1, \ldots, N$. Consider the multivariate linear quantile regression model

$$Q_{\boldsymbol{Y}_i}(\tau|\boldsymbol{X}_i) = \boldsymbol{X}_i'\boldsymbol{\beta}^\tau, \tag{2.1}$$

where $\boldsymbol{X}_i'$ is a $(n_i \times p)-$matrix of covariates, $\boldsymbol{\beta}^\tau = (\beta_1^\tau, \ldots, \beta_p^\tau)'$ is a vector of regression coefficients and $Q_{\boldsymbol{Y}_i}(\tau|\boldsymbol{X}_i)$ is the elementwise τ-th conditional quantile of $\boldsymbol{Y}_i$.

2.1 Quantile regression

The conditional quantile $Q_{\boldsymbol{Y}_i}(\tau|\boldsymbol{X}_i)$ can easily be estimated, given an estimator for $\boldsymbol{\beta}^\tau$. Koenker and Bassett (1978) proposed the following quantile regression estimator for $\boldsymbol{\beta}^\tau$:

$$\hat{\boldsymbol{\beta}}^\tau = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \rho_\tau(Y_{ij} - \boldsymbol{X}_{ij}'\boldsymbol{\beta}), \tag{2.2}$$

where $\rho_\tau(u) = u[\tau - I(u < 0)]$ is the check-loss function used in quantile regression. The minimization in (2.2) can be done very fast, as the minimization problem can be written as a linear programming problem. However, explicit expressions for $\hat{\boldsymbol{\beta}}^\tau$ are not available.

When several quantiles are estimated separately, the estimated quantiles could cross and therefore the estimated quantile function is not monotone in τ. Monotonicity can be enforced by using restricted quantile regression (He, 1997). This is out of the scope of the current paper, as we focus on the performance when a single quantile of the response is estimated.

The estimation procedure (2.2) thus far considered completely observed data. We add a dropout mechanism, where the first time point is never missing Denote the dropout indicator $D_i = j$ when subject i drops out at occasion j and $D_i = n_i + 1$ if the profile for subject i is complete. The response vector $\boldsymbol{Y}_i$ for subject i is divided into its observed $\boldsymbol{Y}_i^{\text{obs}}$ and missing $\boldsymbol{Y}_i^{\text{mis}}$ components. In the next subsections we describe two methods (MI and IPW) to deal with this missingness.

Clearly, the estimator employed does not make use of the correlation between the repeated measures. This is intentionally the case, to examine this choice's effect. It is known that in linear models, as well as in generalized estimating equations for repeated measures, ignoring the correlation leads to a consistent and asymptotically normal estimator, perhaps at the expense of efficiency loss. It is well known that linear regression as well as generalized linear regression are easy to fit, either in closed form or at least by way of stable and fast iterative maximization of the log-likelihood function. This is in contrast to their hierarchical versions, linear mixed models and generalized linear models (cf. Molenberghs and Verbeke 2005); especially the latter are ridden with issues of numerical inaccuracy, slow convergence, and strong dependence of starting values. Their advantage is that these latter methods take correlation between repeated measures into account. Multivariate versions of quantile regression definitely require more research but already at this stage, we can examine the behavior of estimators when ignoring correlation between repetitions, i.e., with cross-sectional

quantile methods. This behavior will be studied both in terms of statistical as well as numerical properties.

2.2 Quantile regression-based multiple imputation

In MI, the missing $\boldsymbol{Y}_i^{\text{mis}}$ are replaced stochastically with a set of M possible estimates $\hat{\boldsymbol{Y}}_i^{(m)}$ (for $m = 1, \ldots, M$), drawn from the predictive distribution of the missing given the observed measurements. Once the data set is complete, estimation procedures (such as (2.2)) for complete data are applied and lead to M estimates $\hat{\boldsymbol{\beta}}^{\tau,1}, \ldots, \hat{\boldsymbol{\beta}}^{\tau,M}$ for $\boldsymbol{\beta}^\tau$. Using Rubin's rule, the parameter estimates from these M analyses are combined into a final parameter estimator

$$\hat{\boldsymbol{\beta}}^\tau = \frac{1}{M} \sum_{m=1}^{M} \hat{\boldsymbol{\beta}}^{\tau,m}.$$

The quantile regression-based MI method is implemented in the Qtools package in R (Geraci, 2016). It exploits the probability integral transform theorem, i.e., if $U \sim \text{Unif}(0, 1)$, then $F^{-1}(U) \sim F$. In the first step M copies of U are sampled from a $\text{Unif}(0, 1)$. In practice, the U's are sampled from $\text{Unif}(\omega, 1 - \omega)$ (with ω sufficiently small), to avoid computational problems in the next step. In the second step, a quantile regression model of the form $Q_{\boldsymbol{Y}_i}(U_m|\boldsymbol{X}_i) = \boldsymbol{X}_i'\boldsymbol{\beta}^{U_m}$ is fitted for each U_m (for $m = 1, \ldots, M$) using a consistent estimator, such as (2.2). Finally, the missing $\boldsymbol{Y}_i^{\text{mis}}$ are imputed: $\hat{\boldsymbol{Y}}_i^{(m)} = \boldsymbol{X}_i\hat{\boldsymbol{\beta}}^{U_m}$.

For longitudinal data with dropouts, the MI procedure imputes progressively each variable Y_{ij} conditional on its covariates and history, using both observed and latest-imputed values, i.e., $\hat{Y}_{ij}^{(m)} = \hat{Q}_{Y_{ij}}(U_m|\boldsymbol{Y}_{i\bar{j}}, \boldsymbol{X}_i)$, with $\boldsymbol{Y}_{i\bar{j}} = (Y_{i1}, \ldots, Y_{i(j-1)})'$.

2.3 Quantile regression-based inverse probability weighting

In IPW, the contribution of each subject to the optimization problem in (2.2) is weighted. The IPW estimator for $\boldsymbol{\beta}^\tau$ is

$$\hat{\boldsymbol{\beta}}^\tau = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \frac{R_{ij}}{\pi_{ij}} \rho_\tau(Y_{ij} - \boldsymbol{X}_{ij}'\boldsymbol{\beta}),$$

where R_{ij} is the missing data indicator and π_{ij} is the probability of being observed up to and including occasion j. The former is defined as $R_{ij} = 1$ if Y_{ij} is observed and $R_{ij} = 0$ otherwise. Furthermore, the probabilities π_{ij} $(j = 2, \ldots, n_i)$ are obtained as follows (recall that the first time point is always observed):

- if the subject drops out at occasion j:

$$\pi_{ij} = p_{ij} \prod_{l=2}^{j-1} (1 - p_{il})$$

- if the subject does not drop out at occasion j:

$$\pi_{ij} = \prod_{l=2}^{j} (1 - p_{il}),$$

with $p_{il} = P(D_i = l | D_i \geq l, \boldsymbol{Y}_{i\bar{l}}, \boldsymbol{X}_i)$ (the probability of dropping out at occasion l given the subject is still in the study).

In practice, the probabilities p_{il} are unknown and need to be estimated, for example by assuming a logistic regression model using the outcome $\boldsymbol{Y}_{i\bar{l}}$ and covariates $\boldsymbol{X}_i$ as regressors.

3 Simulation Study

In a simulation study, we investigate the performance of the MI and IPW method in a quantile regression setting. We compare these two methods with estimators based on the complete cases (i.e., only using subjects with all measurements available) and the available cases (i.e., using all available measurements from all subjects).

We generate data from the following heteroscedastic linear model:

$$Y_{ij} = \beta_0 + t_j\beta_1 + T_i\, t_j\beta_2 + (\gamma_0 + t_j\gamma_1 + T_i\, t_j\gamma_2)\, \varepsilon_{ij}, \tag{3.1}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, 6$, where $t_j = j - 1$ is the measurement time (note that $n_i = 6$ for $i = 1, \ldots, n$), T_i represents a treatment indicator (0 for control and 1 for treatment), and ε_{ij} is the error. Assume that $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n$, with $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{i6})'$, are i.i.d. multivariate normally $N_6(\mathbf{0}_6, \boldsymbol{\Sigma})$ distributed, and $\boldsymbol{\varepsilon}_i$ is independent of T_i.

For the simulation, we set

$$\boldsymbol{\beta} = \begin{pmatrix} 4 \\ 0.5 \\ 0.2 \end{pmatrix}, \boldsymbol{\gamma} = \begin{pmatrix} 0.5 \\ 0.15 \\ 0 \end{pmatrix}, \text{ and } \boldsymbol{\Sigma} = (\sigma_{jk})_{j,k=1}^{6},$$

with $\sigma_{jk} = 0.75^{|j-k|}$, $1 \leq j, k \leq 6$. Furthermore, T_i is generated from a Bernoulli distribution with $P(T_i = 1) = 0.5$. We consider three different sample sizes: $n = 50, 200$, and 1000.

Note that model (3.1) fits in the general model (2.1) of Section 2, with $\boldsymbol{X}_i^{(1)} = (t_1, \ldots, t_6)'$ and $\boldsymbol{X}_i^{(2)} = (T_i\, t_1, \ldots, T_i\, t_6)'$. Therefore, the (elementwise) τ-th conditional quantile of $\boldsymbol{Y}_i$ is given by:

$$\begin{aligned} Q_{\boldsymbol{Y}_i}(\tau|\boldsymbol{X}_i) =& \beta_0 + \gamma_0\Phi^{-1}(\tau) + t_j\left[\beta_1 + \gamma_1\Phi^{-1}(\tau)\right] + T_i t_j\left[\beta_1 + \gamma_2\Phi^{-1}(\tau)\right] \\ =& \beta_0^\tau + t_j\beta_1^\tau + T_i t_j\beta_2^\tau, \end{aligned}$$

where $\Phi^{-1}(\cdot)$ is the quantile function of the standard normal distribution. Since $\gamma_2 = 0$, the coefficient of $T_i t_j$ does not depend on τ ($\beta_2^\tau = \beta_2$). The estimation of $\boldsymbol{\beta}^\tau = (\beta_0^\tau, \beta_1^\tau, \beta_2^\tau)$ is done by (2.2).

3.1 Missing data mechanism

We consider dropouts at two occasions ($t_3 = 2$ and $t_5 = 4$), where the probability of withdrawal at each point is determined by

$$P\left(D_i = 3 | D_i \geq 3, \boldsymbol{Y}_{i\bar{3}}, \boldsymbol{X}_i\right) = \left[1 + e^{-\alpha_0 - \alpha_1 (Y_{i2} - \mu_{i2})}\right]^{-1},$$

and

$$P\left(D_i = 5 | D_i \geq 5, \boldsymbol{Y}_{i\bar{5}}, \boldsymbol{X}_i\right) = \left[1 + e^{-\alpha_0 - \alpha_1 (Y_{i4} - \mu_{i4})}\right]^{-1},$$

respectively. $\mu_{i2} = E\left(Y_{i2}|T_i\right)$ and $\mu_{i4} = E\left(Y_{i4}|T_i\right)$ are computed by (3.1), i.e., $\mu_{i2} = 4.5$ and $\mu_{i4} = 5.5$ if $T_i = 0$; $\mu_{i2} = 4.7$ and $\mu_{i4} = 6.1$ otherwise.

Setting $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)$ we consider four different missing data mechanisms: Firstly, a missing completely at random (MCAR) mechanism by fixing $\boldsymbol{\alpha} = (-0.66, 0)$; two stochastic missing at random mechanisms, named MAR1 and MAR2, by setting $\boldsymbol{\alpha} = (-0.68, 0.8)$ and $\boldsymbol{\alpha} = (-0.68, 5)$, respectively; finally, a deterministic missing at random mechanism, called MAR3, where a dropout occurs at t_3 if $Y_{i2} > 4.8$ and a withdrawal happens at t_5 if $Y_{i4} > 6$. In all cases, we reach around 30% of missing observations.

3.2 Results and conclusion

A total of 1000 data sets were generated for each scenario. Furthermore, we analyze the methods for quantile regression with τ equal to 0.25, 0.5, and 0.75.

Although the data-generating model in equation (3.1) is relatively simple, we evaluate the robustness of the MI procedure by considering proper and improper MI. In the former, the imputation model includes the history (both observed and latest-imputed value) of the outcomes $(Y_{i\bar{j}})$ and the effect of treatment (T_i). On the other hand, in the latter, the imputation model leaves out treatment.

On each regression coefficient separately, we compute the relative bias and relative efficiency. The relative bias of $\hat{\beta}_j^\tau$ is defined as

$$\frac{\hat{\beta}_j^\tau - \beta_j^\tau}{\beta_j^\tau}.$$

The relative efficiency of $\hat{\beta}_j^\tau$ is defined as the ratio of the median absolute deviation (MAD) of the estimator $\hat{\beta}_j^\tau$ based on the observed data over the MAD of the (infeasible) estimator that uses the full data. Additionally, we evaluate the coverage of the 95% confidence interval for the treatment effect (β_2), i.e., the proportion of samples for which the parameter is within the confidence interval. The confidence intervals are computed by the xy-pairs bootstrapping (Koenker, 1994).

Tables 1, 2, 3 and 4 display the relative bias and the relative efficiency of each technique using sample size of $N = 50$, $N = 200$ and $N = 1000$, respectively. As is to be expected, apart from improper MI, all methods work fine under MCAR, given that the data used for analysis are a random subsample of all data. Since the improper MI does not include treatment in the imputation model, this effect is highly biased. That said, some efficiency is lost under complete cases analysis, relative to the other methods that use all data. For the MAR settings, whereas in a likelihood or Bayesian

estimation framework available case analysis would be unbiased, this is not the case here, especially not for β_1 and β_2. This is not surprising given the frequentist nature of our estimation method, and in line with work on, for example, generalized estimating equations (Robins et al., 1995; Bang and Robins, 2005), and on pseudo-likelihood (Molenberghs et al., 2011; Hermans et al., 2020). Pseudo-likelihood in their work is understood as the replacement of the likelihood function by a simpler one that still leads to consistent and asymptotically normal estimators under broad regularity conditions. In our simulations, bias is indeed present for all three sample sizes considered, an issue further aggravated when n increases in terms of increasing bias.

Table 1: Relative bias of the coefficient estimators for $n = 50$. CC: complete cases, AC: available cases, MI_p: proper multiple imputation, MI_i: improper multiple imputation, IPW: inverse probability weighting.

		Relative bias (%)											
		MCAR			MAR1			MAR2			MAR3		
Method	τ	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
CC	0.25	0.09	1.29	1.08	-3.21	-4.27	-6.12	-7.00	-7.49	-32.92	-6.03	-7.90	-33.1
	0.50	-0.06	0.07	-0.16	-3.44	-4.77	-7.42	-8.46	-8.20	-38.76	-7.67	-8.09	-41.18
	0.75	-0.24	-0.90	-0.80	-3.68	-5.13	-8.68	-9.90	-9.14	-44.48	-9.50	-9.32	-48.68
AC	0.25	0.05	1.10	0.97	0.52	-8.91	-3.75	0.84	-22.14	-17.38	0.78	-20.45	-19.67
	0.50	-0.02	-0.04	0.00	0.37	-8.80	-3.85	0.78	-22.19	-19.69	0.69	-21.33	-22.07
	0.75	-0.10	-0.83	-0.79	0.31	-8.60	-4.17	0.77	-22.21	-22.04	0.64	-21.98	-24.49
MI_p	0.25	0.02	0.71	1.98	0.09	0.38	0.13	0.76	-7.39	-9.46	0.75	-8.39	-7.99
	0.50	0.02	-0.76	1.21	0.07	-0.95	0.04	0.26	-4.04	-3.67	0.22	-4.59	-2.17
	0.75	0.03	-1.74	0.49	0.11	-1.72	-0.17	0.13	-0.73	2.54	0.11	-0.91	4.19
MI_i	0.25	0.17	7.15	-32.36	0.31	5.56	-31.79	1.05	-3.60	-37.00	1.08	-6.05	-30.77
	0.50	0.01	6.19	-32.90	0.10	5.47	-34.18	0.38	1.63	-37.82	0.32	0.06	-31.85
	0.75	-0.16	5.46	-33.27	-0.03	5.51	-36.43	0.16	6.46	-38.34	0.11	5.75	-32.71
IPW	0.25	0.09	0.22	0.48	0.10	0.79	-0.35	0.81	-9.05	-10.54	-	-	-
	0.50	0.00	-0.58	1.42	0.00	0.33	-1.44	0.89	-11.4	-12.14	-	-	-
	0.75	-0.01	-0.42	-0.92	0.18	-0.36	-6.40	1.12	-13.29	-18.86	-	-	-

Both proper MI and IPW lead to bias reduction, although it is clear that a decent sample size is needed: when $n = 50$ the bias is still noticeable, although much smaller than for complete case and available case analysis. The bias reduces more quickly with sample size when proper MI is used, as compared to IPW. IPW, though, suffers from two further problems. First, it is less efficient than MI, a finding in line with Beunckens et al. (2008) and references therein. Second, it does not work under the deterministic MAR3. This is well-known because with such a mechanism, the probability of generating observations in certain regions of the sample space is zero, rendering it impossible to use proper weights.

The coverage of the 95% confidence interval for the treatment effect β_2 is exhibited in Table 5. Except for proper MI, the coverage of all methods is considerably lower than 95%. Furthermore,

Table 2: Relative efficiency of the coefficient estimators for $n = 50$. CC: complete cases, AC: available cases, MI_p: proper multiple imputation, MI_i: improper multiple imputation, IPW: inverse probability weighting.

		Relative efficiency											
		MCAR			MAR1			MAR2			MAR3		
Method	τ	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
CC	0.25	1.58	1.66	1.60	1.92	1.49	1.48	3.62	1.34	1.52	3.08	1.35	1.44
	0.50	1.56	1.61	1.47	2.32	1.51	1.45	5.15	1.36	1.51	4.60	1.37	1.50
	0.75	1.71	1.68	1.62	2.45	1.67	1.62	6.31	1.63	1.69	6.09	1.53	1.72
AC	0.25	1.09	1.46	1.39	1.16	1.46	1.32	1.12	2.15	1.40	1.13	1.91	1.26
	0.50	0.98	1.41	1.27	1.08	1.54	1.19	1.04	2.84	1.20	1.05	2.72	1.12
	0.75	1.08	1.49	1.34	1.12	1.71	1.35	1.10	3.22	1.20	1.06	3.28	1.26
MI_p	0.25	1.13	1.34	1.20	1.15	1.23	1.19	1.14	1.33	1.25	1.17	1.30	1.22
	0.50	1.07	1.31	1.12	1.07	1.18	1.20	1.09	1.48	1.17	1.07	1.51	1.16
	0.75	1.14	1.37	1.15	1.14	1.34	1.28	1.25	1.83	1.35	1.18	1.77	1.30
MI_i	0.25	1.12	1.26	1.28	1.15	1.25	1.33	1.19	1.22	1.40	1.19	1.17	1.32
	0.50	1.04	1.27	1.32	1.07	1.22	1.38	1.06	1.31	1.51	1.03	1.35	1.40
	0.75	1.14	1.31	1.34	1.11	1.30	1.47	1.25	1.71	1.59	1.20	1.67	1.52
IPW	0.25	1.09	1.38	1.38	1.11	1.28	1.34	1.18	1.67	1.59	-	-	-
	0.50	1.11	1.44	1.31	1.17	1.39	1.51	1.25	2.14	1.72	-	-	-
	0.75	1.08	1.44	1.34	1.16	1.51	1.65	1.29	2.71	1.82	-	-	-

it gets smaller as the sample size increases. Two reasons can explain this. Firstly, these methods provide biased estimates for these effects. Secondly, the bootstrapping method is not considering that the measurements are correlated, and consequently, the standard errors are underestimated. On the other hand, the proper MI provides intervals with coverage near to the confidence level. However, these are slightly affected by ignoring the correlation.

We conclude that, as expected, proper MI is by far the most promising method among the ones considered. One drawback of the approach though is that it requires more computation time, which is more pronounced when the number of imputations increases. For this reason, it may be worthwhile to explore doubly robust approaches of the type described in Molenberghs et al. (2011) and Hermans et al. (2020) for the context of pseudo-likelihood.

Table 3: Relative bias and efficiency of the coefficient estimators for $n = 200$. CC: complete cases, AC: available cases, MI_p: proper multiple imputation, MI_i: improper multiple imputation, IPW: inverse probability weighting.

(a) Relative bias (%)

		MCAR			MAR1			MAR2			MAR3		
Method	τ	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
CC	0.25	0.08	0.89	-0.98	-3.38	-5.04	-6.86	-7.24	-7.35	-32.43	-6.29	-7.63	-32.25
	0.50	0.02	0.16	-1.07	-3.48	-5.03	-7.06	-8.53	-7.67	-37.02	-7.79	-7.31	-40.18
	0.75	0.02	-0.22	-1.16	-3.53	-4.97	-7.26	-9.90	-8.62	-41.38	-9.52	-8.38	-47.48
AC	0.25	0.08	0.76	-1.09	0.52	-10.57	-2.95	0.78	-22.43	-18.84	0.73	-20.54	-21.78
	0.50	0.04	0.17	-1.05	0.49	-9.62	-2.72	0.86	-22.28	-20.29	0.77	-21.17	-23.98
	0.75	0.03	-0.20	-1.04	0.48	-8.97	-2.50	0.89	-22.28	-21.76	0.75	-21.7	-26.17
MI_p	0.25	0.06	0.63	-0.97	0.05	-0.12	0.61	0.17	-1.83	-0.72	0.17	-1.79	-1.85
	0.50	0.06	-0.16	-0.89	0.07	-0.63	0.69	0.14	-1.49	-0.22	0.10	-1.17	-0.95
	0.75	0.07	-0.67	-0.81	0.05	-1.00	0.79	0.08	-1.26	0.41	0.01	-0.72	0.06
MI_i	0.25	0.22	6.30	-31.94	0.26	4.59	-28.16	0.46	1.82	-28.57	0.48	0.34	-23.06
	0.50	0.06	6.04	-31.9	0.09	5.31	-30.86	0.18	4.87	-36.14	0.16	3.84	-31.55
	0.75	-0.05	5.91	-31.82	-0.03	5.87	-33.48	0.03	7.24	-43.62	0.03	6.67	-40.04
IPW	0.25	0.04	0.42	-0.20	0.07	0.08	0.46	0.60	-5.38	-10.21	-	-	-
	0.50	0.06	-0.04	-0.63	0.03	-0.11	1.00	1.03	-7.85	-13.45	-	-	-
	0.75	0.02	-0.27	0.22	0.04	-0.33	0.27	1.33	-10.49	-15.54	-	-	-

(b) Relative efficiency

		MCAR			MAR1			MAR2			MAR3		
Method	τ	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
CC	0.25	1.47	1.52	1.60	3.25	1.63	1.54	6.84	1.69	2.28	5.88	1.62	2.25
	0.50	1.46	1.54	1.48	4.30	1.73	1.51	10.54	2.08	2.76	9.64	1.94	3.07
	0.75	1.61	1.56	1.51	3.99	1.80	1.40	11.26	2.41	2.67	10.78	2.34	3.10
AC	0.25	0.98	1.32	1.39	1.08	2.05	1.45	1.11	4.12	1.61	1.06	3.74	1.68
	0.5	1.02	1.37	1.32	1.13	2.61	1.35	1.35	5.76	1.72	1.23	5.53	1.84
	0.75	1.04	1.38	1.26	1.02	2.54	1.24	1.26	6.14	1.56	1.14	6.01	1.75
MI_p	0.25	0.99	1.24	1.16	1.00	1.21	1.20	1.05	1.36	1.30	1.10	1.24	1.19
	0.50	1.04	1.27	1.18	1.05	1.29	1.25	1.20	1.63	1.33	1.12	1.47	1.27
	0.75	1.01	1.25	1.13	0.98	1.36	1.20	1.14	1.73	1.17	1.13	1.62	1.15
MI_i	0.25	0.98	1.40	2.20	1.01	1.26	2.00	1.13	1.22	2.06	1.07	1.19	1.71
	0.50	1.03	1.70	2.43	1.05	1.57	2.37	1.22	1.59	2.80	1.12	1.58	2.48
	0.75	1.03	1.69	2.12	1.00	1.71	2.26	1.13	2.05	2.95	1.11	2.09	2.67
IPW	0.25	0.98	1.36	1.36	1.02	1.30	1.36	1.23	2.09	2.06	1.07	3.81	1.60
	0.5	1.08	1.40	1.32	1.08	1.39	1.48	1.68	3.23	2.58	1.23	5.51	1.85
	0.75	0.98	1.36	1.25	1.00	1.52	1.45	1.58	3.80	2.48	1.17	6.02	1.73

Table 4: Relative bias and efficiency of the coefficient estimators for $n = 1000$. CC: complete cases, AC: available cases, MI_p: proper multiple imputation, MI_i: improper multiple imputation, IPW: inverse probability weighting.

(a) Relative bias (%)

		MCAR			MAR1			MAR2			MAR3		
Method	τ	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
CC	0.25	-0.02	0.25	-0.19	-3.51	-5.48	-7.11	-7.31	-8.00	-33.44	-6.39	-8.34	-32.2
	0.50	0.00	0.11	-0.33	-3.53	-4.80	-7.88	-8.59	-7.72	-38.07	-7.83	-7.34	-40.56
	0.75	0.01	0.02	-0.47	-3.59	-4.41	-8.64	-9.98	-8.33	-42.47	-9.57	-8.10	-48.18
AC	0.25	-0.01	0.25	-0.32	0.46	-10.9	-3.41	0.76	-23.2	-19.71	0.69	-21.28	-22.02
	0.50	-0.01	0.13	-0.33	0.45	-9.37	-3.88	0.83	-22.46	-21.28	0.76	-21.32	-24.56
	0.75	-0.01	0.05	-0.34	0.45	-8.37	-4.34	0.88	-22.01	-22.85	0.75	-21.47	-27.07
MI_p	0.25	0.00	-0.13	-0.14	0.01	-0.38	0.03	0.01	-0.59	-0.50	0.03	-0.71	-0.59
	0.50	0.01	-0.14	-0.26	0.01	-0.22	-0.15	0.02	-0.32	-0.48	0.03	-0.35	-0.57
	0.75	0.00	-0.18	-0.37	0.00	-0.16	-0.33	-0.01	-0.22	-0.44	0.00	-0.16	-0.51
MI_i	0.25	0.13	5.67	-31.1	0.17	4.53	-28.11	0.26	3.17	-27.37	0.29	1.72	-21.37
	0.50	0.01	6.01	-31.15	0.03	5.78	-31.33	0.06	6.04	-36.3	0.07	4.93	-31.4
	0.75	-0.10	6.20	-31.19	-0.08	6.65	-34.54	-0.03	8.16	-45.29	0.00	7.38	-41.55
IPW	0.25	-0.01	0.24	-0.09	0.00	0.04	0.21	0.17	-2.25	-5.69	-	-	-
	0.50	-0.01	0.13	-0.18	-0.01	0.06	-0.04	0.68	-4.79	-9.43	-	-	-
	0.75	-0.01	0.06	-0.27	-0.01	0.10	-0.29	1.00	-6.58	-13.16	-	-	-

(b) Relative efficiency

		MCAR			MAR1			MAR2			MAR3		
Method	τ	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
CC	0.25	1.59	1.48	1.65	7.59	2.14	1.73	15.89	3.11	5.18	13.87	3.28	4.95
	0.50	1.59	1.47	1.53	10.16	2.73	1.73	24.54	4.39	6.35	22.36	4.19	6.81
	0.75	1.58	1.51	1.52	10.10	2.76	1.69	27.92	5.18	6.06	26.67	5.02	6.91
AC	0.25	1.02	1.32	1.46	1.31	4.14	1.38	1.68	8.98	3.06	1.57	8.19	3.39
	0.50	1.01	1.36	1.35	1.41	5.24	1.32	2.34	12.83	3.54	2.15	12.15	4.09
	0.75	1.09	1.36	1.25	1.39	5.13	1.24	2.40	13.64	3.25	2.02	13.24	3.83
MI_p	0.25	1.04	1.17	1.20	1.00	1.15	1.26	1.05	1.36	1.29	1.07	1.21	1.23
	0.50	1.03	1.25	1.23	1.03	1.28	1.20	1.18	1.62	1.36	1.18	1.55	1.33
	0.75	1.07	1.32	1.12	1.10	1.32	1.15	1.32	1.83	1.20	1.32	1.80	1.19
MI_i	0.25	1.02	2.28	4.79	1.07	1.78	4.35	1.15	1.42	4.33	1.17	1.22	3.29
	0.50	1.03	3.38	5.26	1.06	3.29	5.23	1.18	3.41	6.13	1.19	2.75	5.28
	0.75	1.12	3.85	4.46	1.14	4.07	4.93	1.35	4.95	6.48	1.34	4.54	5.89
IPW	0.25	1.15	1.32	1.47	1.08	1.30	1.52	2.45	3.38	3.58	-	-	-
	0.50	1.05	1.36	1.40	1.06	1.35	1.48	2.69	4.84	4.14	-	-	-
	0.75	1.15	1.37	1.27	1.23	1.45	1.49	3.11	6.10	4.20	-	-	-

Table 5: Coverage of the 95% confidence interval of the treatment effect (β_2). CC: complete cases, AC: available cases, MI_p: proper multiple imputation, MI_i: improper multiple imputation, IPW: inverse probability weighting.

(a) $N = 50$

		MCAR			MAR1			MAR2			MAR3		
Method	τ	$N = 50$	$N = 200$	$N = 1000$	$N = 50$	$N = 200$	$N = 1000$	$N = 50$	$N = 200$	$N = 1000$	$N = 50$	$N = 200$	$N = 1000$
CC	0.25	79.2	79.4	77.6	79.3	78.9	76.5	76.4	58.8	10.5	78.0	55.1	9.3
	0.50	80.6	81.4	79.6	79.4	79.9	72.9	75.8	47.3	1.6	72.3	36.2	0.3
	0.75	80.6	79.6	79.5	77.9	79.1	74.1	74.5	45.1	1.4	67.4	25.3	0.1
AC	0.25	78.6	79.5	78.9	80.9	79.8	78.5	81.6	71.4	40.2	80.3	66.4	29.4
	0.50	81.1	82.1	81.2	81.0	83.5	77.9	83.5	70.5	27.7	80.3	62.8	13.1
	0.75	81.5	81.3	79.0	79.1	80.4	78.5	83.5	73.3	28.7	81.3	63.9	11.9
MI_p	0.25	93.1	91.6	90.2	90.3	91.7	88.9	95.3	91.7	90.6	94.0	90.6	90.5
	0.50	91.0	89.1	89.0	89.5	90.9	88.8	94.6	91.8	90.3	92.9	92.4	90.8
	0.75	93.2	91.2	90.2	93.1	93.7	89.5	97.0	96.0	92.8	95.7	95.1	90.9
MI_i	0.25	87.2	65.6	10.6	85.8	70.9	16.0	88.7	73.4	23.4	89.2	78.6	42.0
	0.50	82.7	58.0	4.4	82.5	58.6	4.2	85.1	54.3	3.1	87.2	62.4	7.5
	0.75	87.0	65.2	8.8	86.2	62.7	5.7	91.8	55.5	1.3	92.8	61.7	3.4
IPW	0.25	84.4	84.1	79.9	83.1	84.2	80.5	78.9	78.0	67.8	80.7	70.0	29.0
	0.50	82.8	84.2	83.5	81.4	83.6	82.6	79.0	78.8	78.0	79.7	62.8	13.5
	0.75	85.8	84.5	81.3	80.4	85.2	83.9	77.2	78.9	76.3	83.8	66.3	12.2

4 Age-related Macular Degeneration Trial

The age-related macular degeneration (ARMD) trial is a randomized multi-centric clinical trial comparing an experimental treatment (interferon-α) to a corresponding placebo in the treatment of patients with ARMD, a medical condition in which individuals progressively lose sight. The full results of this trial have been reported by Pharmacological Therapy for Macular Degeneration Study Group (1997). The data are available in the R package *nlmeU*. The outcome of interest is the visual acuity over time measured as the ability to read lines of letters on standardized vision charts. Here, we focus on the comparison between placebo and the highest dose (6 millions units daily) of interferon-α at three different quantiles (0.25, 0.5 and 0.75). To handle missingness, we implement a quantile regression-based MI and IPW procedure. For the latter, patients with intermittent missing data are ignored.

The ARMD data contains the patients' visual acuity at four different time points (4 weeks, 12 weeks, 24 weeks, and 52 weeks) of the two treatment groups. Although the total number of longitudinal profiles is 240, only 188 (78.33%) of these have the four follow-up measurements been made, 40 (18%) exhibit monotone missingness, and 8 (3.33%) have intermittent missing values. From the dropouts, 6 subjects have no follow-up measurements.

Defining Z_{ij} as the visual acuity loss (difference in the visual acuity and the baseline value) measured at patient i in week t_j, the quantile model takes the form:

$$Q_{Z_{ij}}(\tau|T_i, t_j) = \beta_0^\tau + T_i\ \beta_1^\tau + t_j\ \beta_2^\tau + T_i\ t_j\ \beta_3^\tau, \tag{4.1}$$

where $T_i = 0$ if patient i is in the control group, and $T_i = 1$ if patient i is in the treatment group.

For the MI procedure, the model to impute the logarithm of the visual acuity at time t_j ($\log Y_{ij}$) includes its history ($\log \boldsymbol{Y}_{i\bar{j}}$), treatment effect ($T_i$) and level of lesion at baseline (a four-point categorical variable) as covariates. The logarithmic transformation is applied to ensure that all imputed values are positive. Furthermore, 20 multiply imputed datasets are generated. For the quantile regression-based IPW, the following weight model is assumed:

$$\text{logit}\left[P\left(D_i = j | D_i \geq j, \boldsymbol{Y}_{i\bar{j}}, T_i, t_j\right)\right] = \eta_{ij} \tag{4.2}$$

where:

$$\eta_{ij} = \psi_0 + Y_{i,j-1}\psi_1 + T_i\psi_2 + t_j\psi_3 + L_{1i}\psi_4 + L_{2i}\psi_5 + L_{3i}\psi_6,$$

$Y_{i,j-1}$ is the outcome at previous time t_{j-1}, and L_{1i}, L_{2i} and L_{3i} are the three dummy variables associated to the level of lesion at baseline.

Parameter estimates and corresponding standard errors for the model in equation (4.2) are displayed in Table 6. The standard errors were computed using the $xy-$pair bootstrap method (Parzen et al., 1994). There is a significant effect of treatment, level of lesion, and time on the probability of dropout. On the contrary, the previous outcome does not have a significant impact. This shows that there is no strong evidence for MAR.

Table 7 shows the parameter estimates and their standard errors for the model in equation (4.1) using the available cases (AC), multiple imputation (MI), and inverse probability weighting (IPW). There is no substantial difference in the estimates of AC and IPW-based models. Nevertheless, some

Table 6: ARMD data. Parameter estimates and their standard error for a logistic model for dropouts

Effect	parm.	est.	s.e.
Intercept	ψ_0	-2.34	0.694
Previous outcome	ψ_1	-0.02	0.009
Treatment	ψ_2	0.87	0.345
Time	ψ_3	0.04	0.009
Lesion level 1	ψ_4	-1.37	0.463
Lesion level 2	ψ_5	-1.75	0.498
Lesion level 3	ψ_6	-2.76	0.711

differences are observed in the MI-based model. The estimates associated to treatment at baseline (β_1) and treatment effect (β_3) are different for all quantile levels. For the latter, it is even significant for quantile 0.5. There are no noticeable differences for the rest of estimates based on these three methods.

Table 7: ARMD data. Parameter estimates and their standard error for quantile regression with $\tau = 0.25, 0.5$, and 0.75 using available cases (AC), multiple imputation (MI), and inverse probability weighting (IPW).

		AC		MI		IPW	
Effect	parm.	est.	s.e.	est.	s.e.	est.	s.e.
Intercept	$\beta_0^{0.25}$	-4.75	0.987	-4.63	0.997	-4.75	1.034
Treatment	$\beta_1^{0.25}$	-1.75	1.802	-1.36	1.952	-1.75	1.988
Time	$\beta_2^{0.25}$	-0.31	0.073	-0.33	0.076	-0.31	0.081
Treatment×Time	$\beta_3^{0.25}$	-0.06	0.096	-0.11	0.111	-0.06	0.113
Intercept	$\beta_0^{0.5}$	-0.42	0.522	-0.43	0.533	-0.5	0.552
Treatment	$\beta_1^{0.5}$	-0.67	0.804	-0.39	0.903	-0.58	0.919
Time	$\beta_2^{0.5}$	-0.15	0.034	-0.14	0.036	-0.13	0.035
Treatment×Time	$\beta_3^{0.5}$	-0.08	0.060	-0.13	0.066	-0.11	0.066
Intercept	$\beta_0^{0.75}$	3.90	0.898	3.84	0.911	3.9	0.897
Treatment	$\beta_1^{0.75}$	-1.32	1.246	-1.07	1.251	-1.32	1.226
Time	$\beta_2^{0.75}$	-0.08	0.037	-0.07	0.038	-0.08	0.037
Treatment×Time	$\beta_3^{0.75}$	-0.07	0.054	-0.08	0.054	-0.07	0.057

5 Conclusion

We considered a linear regression model for longitudinal data with missingness in the response. The impact of methods dealing with missing data on the estimation of the regression coefficients, when a conditional quantile of the response is estimated, was investigated in a simulation study. In the simulation study an analysis based on the complete cases, the available cases, a quantile-based

multiple imputation estimation, and a quantile-based inverse probability weighting estimation were compared. Multiple imputation is the most promising method as it reduces the bias and is more efficient. A drawback might be its computation time. In our simulations and data analysis, the datasets were relatively small. However, when several imputations are used (which is often not needed) and with very large datasets, the imputation process can be computationally intensive. Moreover, the efficiency can be reduced by not considering relevant covariates on the imputation model. All results are obtained while ignoring the correlation between repeated measures when analyzing the data, but not when drawing imputations. As expected, this does not jeopardize the validity of the estimators, although properly accommodating for dependence within a repeated measures sequence may lead to increased efficiency.

The missing data mechanisms considered in this paper are of an MAR nature, and sometimes MCAR but never MNAR. As has been shown repeatedly, MNAR cannot be ruled out in practice based on data-analytic considerations alone. It is commonly accepted that, to address this problem, a sensitivity analysis to varying the missing-data assumptions would be needed (Little et al., 2010). This requires further research in the quantile-regression context and clearly falls outside of the scope of the current paper.

Acknowledgements

The authors gratefully acknowledge Fonds Wetenschappelijk Onderzoek [N1518917, G0B2615N], Special Research Fund (Bijzonder Onderzoeksfonds) of Hasselt University [BOF14NI06], and the European Research Council (2016-2021, Horizon 2020 / ERC grant agreement No. 694409).

References

Aerts, M., Claeskens, G., Hens, N., and Molenberghs, G. (2002), "Local multiple imputation," *Biometrika*, 89, 375–388.

Bang, H. and Robins, J. (2005), "Doubly robust estimation in missing data and causal inference models," *Biometrics*, 61, 962–972.

Beunckens, C., Sotto, C., and Molenberghs, G. (2008), "A simulation study comparing weighted estimating equations with multiple imputation based estimating equations for longitudinal binary data," *Computational Statistics and Data Analysis*, 52, 1533–1548.

Carpenter, J. and Kenward, M. (2013), *Multiple imputation and its application*, New York: John Wiley & Sons.

Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G. (2008), *Longitudinal data analysis*, New York: Chapman & Hall/CRC.

Geraci, M. (2016), "Qtools: A Collection of Models and Tools for Quantile Inference," *The R Journal*, 8, 117–138.

He, X. (1997), "Quantile curves without crossing," *The American Statistician*, 51, 186–192.

Hermans, L., Ivanova, A., Sotto, C., Molenberghs, G., Verbeke, G., and Kenward, M. (2020), "Doubly robust pseudo-likelihood for incomplete hierarchical binary data," *Statistical Modeling*, 20, 42–57.

Horvitz, D. and Thompson, D. (1952), "A generalization of sampling without replacement from a finite universe," *Journal of the American Statistical Association*, 47, 663–685.

Koenker, R. (1994), "Confidence Intervals for Regression Quantiles," in *Asymptotic Statistics*, Physica-Verlag HD, pp. 349–359.

Koenker, R. and Bassett, Jr., G. (1978), "Regression quantiles," *Econometrica*, 46, 33–50.

Lipsitz, S., Zhao, L., and Molenberghs, G. (1998), "A semiparametric method of multiple imputation," *Journal of the Royal Statistical Society, Series B*, 60, 127–144.

Little, R., D'Agostino, R., Dickersin, K., Emerson, S., Farrar, J., Frangakis, C., Hogan, J., Molenberghs, G., Murphy, S., Neaton, J., Rotnitzky, A., Scharfstein, D., Shih, W., Siegel, J., and Stern, H. (2010), *he Prevention and Treatment of Missing Data in Clinical Trials. Panel on Handling Missing Data in Clinical Trials*, Washington D.C: National Research Council, Committee on National Statistics, Division of Behavioral and Social Sciences and Education, The National Academies Press.

Little, R. J. and Rubin, D. B. (2014), *Statistical analysis with missing data*, Hoboken, New Jersey: John Wiley & Sons.

Molenberghs, G., Fitzmaurice, G., Michael, K., Tsiatis, A., and Verbeke, G. (2014), *Handbook of missing data methodology*, Boca Raton: Chapman & Hall/CRC.

Molenberghs, G., Kenward, M., Verbeke, G., and Teshome Ayele, B. (2011), "Pseudo-likelihood estimation for incomplete data," *Statistica Sinica*, 21, 187–206.

Parzen, M. I., Wei, L. J., and Ying, Z. (1994), "A resampling method based on pivotal estimating functions," *Biometrika*, 81, 341–350.

Pharmacological Therapy for Macular Degeneration Study Group (1997), "Interferon alfa-2a is ineffective for patients with choroidal neovascularization secondary to age-related macular degeneration: Results of a prospective randomized placebo-controlled clinical trial," *Archives of Ophthalmology*, 115, 865–872.

Robins, J., Rotnitzky, A., and Zhao, L. (1994), "Estimation of regression coefficients when some regressors are not always observed," *Journal of the American Statistical Association*, 89, 846–866.

— (1995), "Analysis of semiparametric regression models for repeated outcomes in the presence of missing data," *Journal of the American Statistical Association*, 90, 106–121.

Rubin, D. (1987), *Multiple imputation for nonresponse in surveys*, New York: John Wiley & Sons.

Rubin, D. B. (1976), "Inference and missing data," *Biometrika*, 63, 581–592.

van Buuren, S. (2012), *Flexible imputation of missing data*, Boca Raton: Chapman & Hall/CRC.

Wei, Y., Ma, Y., and Carroll, R. (2012), "Multiple imputation in quantile regression," *Biometrika*, 99, 423–438.

Received: February 10, 2021

Accepted: April 21, 2021