Recovering income distribution in the presence of interval-censored data

Fernando Rios-Avila[1]
Levy Institute at Bard College, New York - USA, friosavi@levy.org

Gustavo Canavire-Bacarreza
The World Bank, Washington DC - USA, gcanavire@worldbank.org

Flavia Sacco-Capurro
The World Bank, Washington DC - USA, fsaccocapurro@worldbank.org

[1] Corresponding Author

This research received no funding, and there are no conflict of interest.

Abstract

We propose a method to analyze interval-censored data using a multiple imputation based on a Heteroskedastic Interval regression approach. The proposed model aims to obtain a synthetic dataset that can be used for standard analysis, including standard linear regression, quantile regression, or poverty and inequality estimation. We present two applications to show the performance of our method. First, we run a Monte Carlo simulation to show the method's performance under the assumption of multiplicative heteroskedasticity, with and without conditional normality. Second, we use the proposed methodology to analyze labor income data in Grenada for 2013-2020, where the salary data are interval-censored according to the salary intervals prespecified in the survey questionnaire. The results obtained are consistent across both exercises.

Keywords: interval-censored data, Monte Carlo simulation, heteroskedastic interval regression, wages

JEL Codes: C150, C340, J3

1. Introduction

Labor force surveys are a useful data source for understanding employment dynamics in both developing and developed countries. These surveys provide vast information on the labor market status at higher frequency levels than living condition surveys. And, in some cases, they are the only source of information to describe and examine the structure of the labor markets. In fact, in the Latin American and the Caribbean region, countries like Bolivia, Chile, Costa Rica, Ecuador, Jamaica, Mexico, Peru, and Uruguay, collect their labor force surveys quarterly as opposed to a yearly basis, which is the case of most household and living standard surveys.

One of the key features of these labor surveys is that they provide information on wages and salaries of workers. This allows us to estimate job market trends and obtain inequality measures of labor income among workers. However, the full income distribution in many countries cannot be retrieved because labor income is reported in brackets. Because of this, the estimation of inequality or poverty measures, as well as regression-type analysis, is difficult. This is the case of the labor force survey for all countries in the Organization of Eastern Caribbean States (OECS).

This is not unique to the Caribbean region. Countries like Colombia, Germany, Australia, New Zealand, Bosnia and Herzegovina, North Macedonia, and Serbia, among others, have similar data collection protocols for their microcensus (Walter & Weimer, 2018). In the U.S., the current population survey (CPS) collects detailed family income only once a year, in the March supplement, but collects family income in brackets on monthly basis.

One argument in favor of using interval-censored questions to collect information on income is the higher response rate compare to questions asking to report exact amounts (Wang et al., 2013). This happens because income information is considered "sensitive", and people are reluctant to report actual earnings, and may choose not to respond those questions at all (Hagenaars & de Vos, 1988; Moore et al., 2000). Field tests conducted in the past have shown that asking follow-up income questions in a series of unfolding brackets achieves superior results in terms of response rates for income amounts, as was the case of the National Health Interview Survey (NHIS) and the Behavioral Risk Factor Surveillance System Survey (BRFSS), both administered by the Center for Disease Control and Prevention of the United States (Angelov & Ekström, 2019; Yan et al., 2018). However, even though this form of data collection reduces the severity of underreporting or misreporting, it raises a problem for recovering the full wage (income) distribution, which is key to understanding and analyzing inequality.

To better use the information from these types of surveys, we propose an imputation approach to simulate the distribution of the data that is only available in brackets. Our method is an extension of the imputation approach described in Royston (2007), that allows for heteroskedastic errors to model the conditional distribution of the censored data. The estimated conditional distribution is then used to impute the data using draws from the estimated conditional distribution. Once the imputed data is obtained, standard aggregation methods (Rubin, 1987) can be used to analyze the censored data as if it were fully observed. For example, it can be used to calculate poverty or inequality measures, as well as perform regression analysis. To demonstrate the flexibility of this approach, we use a Monte Carlo simulation to analyze the sensitivity of our method. As an empirical example, we use the approach to analyze wage inequality in Grenada utilizing their Labor Force Survey.

Other approaches exist in the literature and have been used for analyzing this kind of data. Royston (2007), which our paper expands upon, proposes and implements a strategy for using interval regression under homoskedasticity in the framework of multiple imputation. In contrast, our implementation is more general, as it considers the case of heteroskedastic errors, allowing for a better approximation of the conditional distribution and imputation of the outcome.

To measure income inequality with right-censored (top-coded) data, Jenkins et al. (2011) propose multiple-imputation methods for estimation and inference where censored observations are imputed using draws from a flexible parametric model fitted to the censored distribution, such as Generalized Beta of the second kind (GB2), Sigh-Maddala or Dagum distributions. Chen (2018) provides a generalized approach for the estimation of parametric income distributions using grouped data, showing its consistency through complementary simulation results. More recently, Walter and Weimer (2018) propose an iterative kernel density algorithm that generates pseudo samples from the interval-censored income variable to estimate poverty and inequality indicators. While the interval regression approach we propose fits with the models described in Chen (2018), Jenkins et al. (2011), and Walter and Weimer (Walter & Weimer, 2018), these papers focus on recovering the unconditional distribution of income, without considering the relationship with explanatory variables. The advantage of using multiple imputed data as we propose is that one can just as easily analyze unconditional statistics, as well as analyze data by subgroups or use regression analysis to capture relationships between controls and the outcome of interest.

Zhou et al. (2017) and Hsu, et al. (2021) propose methodologies for the estimation of conditional quantile regressions using interval censored data, under different distributional assumptions. While these approaches can be used for analyzing interval-censored data, they only focus on estimating conditional quantile regressions, requiring specialized software that is not readily available. In contrast, the method we propose can be applied not only for the estimation of conditional quantile regressions, but also for the estimation of unconditional distribution statistics.

Other studies, like the one proposed by Han et al., (2020), construct new measures of income distribution and estimate poverty in the U.S. using data from the monthly Current Population Survey (CPS). They address the problem of censored income data using draws from the empirical income distribution observed in the last March supplement. A similar method is proposed by Parolin & Wimer (2020), who produce monthly updates of the Supplemental Poverty Measure (SPM) rates with demographic data from the CPS and poverty data from the previous March supplement of the CPS. However, these studies seek to obtain income estimates using the uncensored distribution of previous years, which is not always available with other data sources, like the ones analyzed in this paper.

Büttner & Rässler (2008) proposes a multiple imputation approach, similar to ours, to analyze wages from the German Institute of Employment Research (IAB) employment survey. While their method focuses on the analysis top coded data, we expand the approach to analyze data with a more generalized censoring structure.

The paper is organized as follows. Section 2 introduces the model and the econometric issues associated with the imputation method; Section 3 provides a Monte Carlo simulation exercise to analyze the performance of the methodology; Section 4 discusses further considerations regarding the methodology, modeling, and limitations; Section 5 uses the methodology to analyze labor income distribution changes in Grenada using the 2013-2020 series of the Labor Force Survey. Section 6 concludes.

2. Methodology

To address the problem of interval-censored data, we propose a multiple imputation approach based on a heteroskedastic interval regression model. Allowing for heteroskedastic errors provides better flexibility for the modeling of the conditional distribution of the outcome, which allows for better imputation. An interval-regression model is a generalization of the Tobit model that allows the use of a mixture of censored and completely observed data, even if the censoring thresholds are unique to each individual. The goal of the model is to find a set of parameters that maximizes the probability that, given a set of characteristics, the predicted latent earnings fall within the declared earning threshold. Imputations are obtained using random draws of the estimated conditional distributions. In a framework of heteroskedastic errors, the methodology uses the estimates for the conditional mean and conditional variance to obtain simulated errors and impute the data. To facilitate the description of the methodology, we refer to y as the log of earned income.

2.1. Interval regression model

Assume that (log) earned income (y_i) has a data-generating process (d.g.p.) such that:

$$y_i = \mu(x_i) + v_i\sigma(x_i) \tag{1}$$

Where v_i is a homoskedastic i.i.d. error, with mean 0 and standard deviation 1, that is independent of the characteristics x. $\mu(x_i)$ and $\sigma(x_i)$ are flexible functions of x_i. $\mu(x_i)$ represents the conditional mean of y_i, and $\sigma(x_i)$ is a strictly positive function that represents the conditional standard deviation of y_i. Assuming heteroskedastic standard errors, based on a multiplicative structure, provides a more flexible framework to model the potentially more complex unconditional distribution of y.

Following Machado & Santos Silva (2019), the conditional mean $\mu(x_i)$ captures location shift effects of characteristics on the outcome, whereas $\sigma(x_i)$ capture the scale shits, which relates to how much of the spread is explained by differences in characteristics. Following the standard setup of interval-regression models (Stewart, 1983), we impose the assumption that v_i follows a standard normal distribution, so that $y_i|x_i$ is also normally distributed with mean $\mu(x_i)$ and standard deviation $\sigma(x_i)$. [2]

$$if\ v_i \sim N(0,1) \rightarrow y_i|x_i \sim N\big(\mu(x), \sigma(x)\big) \tag{2}$$

Under this assumption, equation 1 can be estimated via maximum likelihood by maximizing the following function:

$$L_i\big(\mu(x), \sigma(x)\big) = f_{y|x}\big(\mu(x), \sigma(x)\big) = \frac{1}{\sigma(x)}\phi\left(\frac{y_i - \mu(x)}{\sigma(x)}\right) \tag{3a}$$

$$\hat{\mu}(x), \hat{\sigma}(x) = \max_{\mu(x),\sigma(x)} \frac{1}{N}\sum \log\big(L_i(\mu(x), \sigma(x)\big) \tag{3b}$$

Where $\hat{\mu}(x)$ and $\hat{\sigma}(x)$ are the solutions that maximize the log-likelihood function.

Under these conditions, and assuming a flexible enough model specification to capture the conditional mean and conditional variance, estimating equation (1) allows us to recover the whole distribution of the dependent variable y_i.

When y_i is fully observed, this variable can be directly used for estimating any measure of poverty or inequality, or to analyze the relationship between observed characteristics X and the outcome y, using standard statistical methods. Often, however, due to survey design, one may only have access to data reported in brackets. In other words, rather than observing y_i, one may only observe that reported income by individual i is within some lower (ll_i) and upper (uu_i) threshold, which may be different for each individual. In this case, unless $ll_i = uu_i$, the likelihood function defined by Equations 3a and 3b is not defined.

An alternative for estimating a model with this type of data is the use of what is known as interval regression. Interval regression is a generalization of the censored regression estimators like the Tobit model (see Cameron & Trivedi (2005) ch 16 for a discussion of censored regressions), where data can be a mixture of left-censored, right-censored, interval-censored, or fully observed. For simplicity, we refer to the case with interval-censored data.

When the data is interval-censored, rather than modeling the outcome itself, the approach focuses on modeling the probability that an individual i reports income to be within the underlying income brackets:

---

[2] While this assumption is unnecessary for the estimation of standard linear regression models, imposing some distribution assumption on the errors is necessary when estimating models via maximum likelihood. Nevertheless, as described in McDonald et al., (2018), it is possible to relax this assumption using more flexible distributions.

$$P(ll_i \leq y_i < uu_i | x_i) \tag{4}$$

Using the data generating process (d.g.p.) defined by equation 1, and the normality assumption of the error v_i, equation (4) can be rewritten as:

$$P\left(\frac{ll_i - \mu(x_i)}{\sigma(x_i)} \leq v_i < \frac{uu_i - \mu(x_i)}{\sigma(x_i)} | x_i\right) = \mathrm{P}\left(v_i < \frac{uu_i - \mu(x_i)}{\sigma(x_i)}\right) - \mathrm{P}\left(v_i < \frac{ll_i - \mu(x_i)}{\sigma(x_i)}\right) \tag{5a}$$

$$= \Phi\left(\frac{uu_i - \mu(x_i)}{\sigma(x_i)}\right) - \Phi\left(\frac{ll_i - \mu(x_i)}{\sigma(x_i)}\right) \tag{5b}$$

Where $\Phi(.)$ is the cumulative normal density function. Using equation (5b), the loglikelihood function that is maximized to identify the parameters $\mu(x_i)$ and $\sigma(x_i)$ is defined as:

$$L_i\big(\mu(x), \sigma(x)\big) = \Phi\left(\frac{uu_i - \mu(x_i)}{\sigma(x_i)}\right) - \Phi\left(\frac{ll_i - \mu(x_i)}{\sigma(x_i)}\right) \; if \text{ data is interval} - \text{censored} \tag{6a}$$

$$L_i\big(\mu(x), \sigma(x)\big) = \Phi\left(\frac{uu_i - \mu(x_i)}{\sigma(x_i)}\right) \; if \text{ data is left} - \text{ censored} \tag{6b}$$

$$L_i\big(\mu(x), \sigma(x)\big) = 1 - \Phi\left(\frac{ll_i - \mu(x_i)}{\sigma(x_i)}\right) \; if \text{ data is right} - \text{censored} \tag{6c}$$

$$L_i\big(\mu(x), \sigma(x)\big) = \frac{1}{\sigma(x_i)}\phi\left(\frac{y_i - \mu(x_i)}{\sigma(x_i)}\right) \; if \text{ data is fully observed} \tag{6d}$$

Which can be used to obtain estimates for $\mu(x)$ and $\sigma(x)$ using maximum likelihood estimation.

2.2. Model Imputation.

As previously described, when dealing with interval-censored data, we have limited access to the observed distribution of the variable of interest. This is in contrast with standard multiple imputation analysis, where the variable of interest is fully unobserved. This distinction has implications for the imputation strategy because it determines the appropriate draw of the imputed error.

Consider the d.g.p stated in equation 1 and define y_i^* to be the true but unobserved variable of interest. By definition, if the data is interval-censored, the range of values that can be potentially used to impute y_i^* are bounded between the lower and upper threshold of a given interval. In addition, conditional on the observed characteristics x, and the parameters $\mu(x_i)$ and $\sigma(x_i)$, it implies that the unobserved error v_i^* is also bounded:

$$v_i^* \in \left[\frac{ll_i - \mu(x_i)}{\sigma(x_i)}, \frac{uu_\mathrm{i} - \mu(x_i)}{\sigma(x_i)}\right] \tag{7}$$

Furthermore, under the assumption that v_i follows a standard normal distribution, we can impute values for y_i^*, by simply getting random draws for v_i^* from a truncated random normal distribution:

$$\tilde{v}_i = \Phi^{-1}(r_i), \text{where } r_i \sim Uniform\left(\Phi\left(\frac{ll_i - \mu(x_i)}{\sigma(x_i)}\right), \Phi\left(\frac{uu_\mathrm{i} - \mu(x_i)}{\sigma(x_i)}\right)\right) \tag{8}$$

Where $\Phi^{-1}(r_i)$ corresponds to the r_{th} quantile for the standard normal distribution. Finally, the imputed value for the outcome of interest y_i^* is given by:

$$\tilde{y}_i = \mu(x_i) + \tilde{v}_i \sigma(x_i) \tag{9}$$

Because the population parameters $\mu(x_i)$ and $\sigma(x_i)$ are unknown, we use the sample equivalents that are estimated using the interval regression estimator via Maximum likelihood.[3] To account for the uncertainty of the regression estimation, we obtain random draws from the following joint normal distribution:

$$\begin{bmatrix}\tilde{\mu}(x)\\ \tilde{\sigma}(x)\end{bmatrix} \sim N\left(\begin{matrix}\hat{\mu}(x)\\ \hat{\sigma}(x)\end{matrix}, \widetilde{\Omega}\right) \; ; \; \widetilde{\Omega} = \widehat{\Omega} * \frac{n}{\tilde{n}}; \tilde{n} \sim \chi_n^2 \tag{10}$$

Where $\widehat{\Omega}$ is the ML variance-covariance matrix estimate, n is the number of observations in the sample, and $\tilde{n}$ is a random draw from a chi-squared distribution n degrees of freedom (χ_n^2). In Royston (2007), and the current implementation in -Stata-, the imputation algorithms assume $\hat{\sigma}(x)$ is constant. This simplifies the draws we need to obtain in equation 10 but imposes a homoskedastic assumption on the conditional distribution of y. . Finally, the imputation for y_i^* will be given by:

$$\tilde{\tilde{y}}_i = \tilde{\mu}(x_i) + \tilde{\tilde{v}}_i \tilde{\sigma}(x_i) \tag{11a}$$

$$\tilde{\tilde{v}}_i = \Phi^{-1}(\tilde{r}_i), \text{where } r_i \sim Uniform\left(\Phi\left(\frac{ll_i - \tilde{\mu}(x_i)}{\tilde{\sigma}(x_i)}\right), \Phi\left(\frac{uu_i - \tilde{\mu}(x_i)}{\tilde{\sigma}(x_i)}\right)\right) \tag{11b}$$

Where $\tilde{\tilde{v}}_i$ is used in (11a) instead of $\tilde{v}_i$, to account for the role of the estimated parameters in the error $\tilde{v}$.

[3] For numerical purposes, it is also important to emphasize that $\sigma(x_i)$ is not estimated directly, but $\ln \sigma(x_i)$ is estimated instead.

In summary, the imputation algorithm is as follows:

1. Estimate the parameters associated with $\mu(x)$ and $\sigma(x)$ using a heteroskedastic interval regression approach via maximum likelihood, as well as the variance-covariance matrix Ω.
2. Obtain $\tilde{n}$ from a random draw from χ^2_n, and estimate $\widetilde{\Omega}$.
3. Obtain a random draw for $\tilde{\mu}(x)$ and $\tilde{\sigma}(x)$ from $N\left(\begin{matrix}\hat{\mu}(x)\\ \hat{\sigma}(x)\end{matrix}, \widetilde{\Omega}\right)$.
4. Obtain random draws for $\tilde{\tilde{v}}_i$, conditional on $\tilde{\mu}(x)$ and $\tilde{\sigma}(x)$, for each observation i.
5. Get the full sample of imputed data $\tilde{\tilde{y}}_i$.
6. Repeat steps 2-4 M times and obtain M sets of imputed samples.

Steps 2-4 correspond to simulating data from the posterior distribution, similar to what is described in Gelman et al., (2014).

After the M imputations have been obtained, one could use the imputed values $\tilde{\tilde{y}}_i$, or any other monotonic transformation $g(\tilde{\tilde{y}}_i)$, for further analysis. In most cases, we may be more interested in analyzing outcomes in levels but may have to model and impute log of the outcome, because the latter will be more likely to fulfill the conditional normality assumption.

2.3. Model estimation and inference

Once the M imputed datasets have been obtained, statistical analysis can be done by independently implementing the desired model estimation across all M imputed samples. The aggregation and summary from the M estimated models could then be done by applying the combination rules described in Rubin (1987).

Let β be the set of parameters of interest, and $\hat{\beta}_m$ and $\hat{V}_m$ be the set of estimated coefficients and corresponding variance-covariance matrix obtained using simulated sample m. The Multiple imputation estimates $\hat{\beta}_M$ for the parameter of interest is given by:

$$\hat{\beta}_M = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}_m \tag{13}$$

Whereas the variance-covariance estimate $\hat{V}_M$ is given by:

$$\hat{V}_M = \frac{1}{M}\sum_{m=1}^{M}V_m + \left(\frac{M+1}{M}\right)\frac{(\hat{\beta}_m - \hat{\beta}_M)'(\hat{\beta}_m - \hat{\beta}_M)}{M-1} \tag{14}$$

3. Monte Carlo Simulations

3.1. Setup

We examine the performance of our proposed estimator under several simulation scenarios, using data structures with explicit multiplicative heteroskedasticity, similar to the ones proposed in Machado and Santos-Silva (2019), and with a varying coefficient model structure, as in Hsu et al., (2021). In both cases, the goal is to simulate data that would show heterogeneity in the distribution of the outcome. This structure is flexible enough to also allow the estimation of other distribution-based regressions such as unconditional quantile regressions (Firpo et al., 2009) and Recentered Influence function regressions in general (Rios-Avila, 2020).

The first set of simulations is designed to study the performance of the estimator under the assumption of multiplicative heteroskedasticity assuming the following functional form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + v\sigma(x_1, x_2) \tag{15}$$

Where x_1 follows a Bernoulli distribution ($x_1 \sim bernulli(0.5)$)and x_2 follows a rescaled chi-squared distribution with 5 degrees of freedom ($x_2 \sim \chi_5^2/5$). Following Machado and Santos-Silva (2019), we use two different functional forms for $\sigma(x_1, x_2)$:

$$\sigma_1(x_1, x_2) = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 \tag{16a}$$

$$\sigma_2(x_1, x_2) = e^{\gamma_0 + \gamma_1 x_1 + \gamma_2 x_2} \tag{16b}$$

In both cases, we require that $\sigma(x_1, x_2)$ to be strictly positive. The first case, equation (16a), imposes the assumption of linear heteroskedasticity and provides a closed-form solution for the corresponding quantile coefficients. The second option, equation (16b), guarantees standard deviation to be strictly positive but does not have a closed-form solution for the corresponding conditional quantile regression coefficients. As described in Machado and Santos-Silva (2019), this data-generating process of multiplicative heteroskedasticity also guarantees that quantiles will not cross, and thus the corresponding coefficients can be estimated directly using standard conditional quantile regression estimators.

Using this data structure, we consider four different distributions for the error v: Normal distribution, logistic distribution, chi-square distribution with 5 degrees of freedom, and uniform distribution. All of them were adjusted to have a mean 0 and standard deviation 1. Whereas the first two distributions are meant to show how sensitive is the estimator to the normality assumption, the third and fourth aim to show how sensitive the results are to cases where the error has a skewed distribution, or a distribution with a limited range. With these considerations, the data-generating process are defined as:

$$y = x_1 + x_2 + v * (1 - 0.5x_1 + 0.2x_2) \tag{17a}$$

$$y = x_1 + x_2 + v * e^{0.6-0.5+0.2x_2} \tag{17b}$$

The second set of simulations uses a data-generating process following a varying coefficient approach, based on the percentile τ an observation belongs to. In this setup, we assume that τ is defined by a random draw from a uniform distribution and that y is given by:

$$y = \beta_0(\tau) + \beta_1(\tau)x_1 + \beta_2(\tau)x_2 \tag{18}$$

To compare the results to Hsu, et al. (2021), we assume the coefficients $\beta(\tau)'s$ are defined as:

$$\beta_0(\tau) = 1 + 0.5\Phi^{-1}(\tau); \beta_1(\tau) = 0.4 + 1.2\Phi^{-1}(\tau); \beta_2(\tau) = 0.6 + 0.5\Phi^{-1}(\tau) \tag{19a}$$

$$\beta_0(\tau) = \beta_1(\tau) = \beta_2(\tau) = 0.5(1 + \Phi^{-1}(\tau) - \log(1-\tau)) \tag{19b}$$

Equation (19a) imposes a structure that is similar to the multiplicative normality under linear heteroskedasticity (equation 17a), whereas the second equation imposes a skew conditional distribution of the outcome. This d.g.p. allows us to present a more general

In all scenarios, we assume that data is subject to interval censoring, such that $ll_i = \lfloor y_i \rfloor$ & $uu_i = \lceil y_i \rceil$, where $\lfloor . \rfloor$ and $\lceil . \rceil$ represent the nearest integer that is lower or higher than y_i respectively. In addition, we also assume if $y_i < -1$ or $y_i > 10$, the lower and upper thresholds, respectively, will be undefined. These steps are not necessary, but allow us to mimic how would data be accessible when bracket bounds are adjusted and transformed (log).

For the implementation and analysis, we use 2500 replications, with a sample size of 1000 observations for the core results. Replications using sample sizes of 500 and 2000 are provided in the appendix, with qualitatively similar results. We focus on the comparison of conditional quantile regressions for the 10th, 50th, and 90th quantiles, as well as for the 10th, 50th, and 90th unconditional quantiles. Quantile regressions were estimated using the fast algorithm developed by Chernozhukov et al. (2022) and implemented via the Stata command -qrprocess-, whereas the unconditional quantile regressions were estimated following Firpo, Fortin, and Lemieux (2009) and implanted via the Stata command -rifhdreg- (Rios-Avila, 2020). Finally, the simulation was implemented using -parallel- (Vega Yon & Quistorff, 2019). Finally, our imputation method is implanted with a new user-written program -intreg_mi-, which is available upon request.

While population parameters for conditional quantile regressions exist for some of the data-generating processes, there are no close-form solutions for the population parameters corresponding to the RIF regressions. Because of this, our comparisons and evaluations assume the estimates using fully observed data to be the truth, which are compared to coefficients based on simulated data.

3.2. Results

Tables 1 to 3 provide a summary of the results for the Montecarlo simulations using the different data-generating processes. In each table, we present the bias of the estimates, comparing the imputation-based estimated coefficients to the coefficients obtained using fully observed data. We also present the mean squared error (MAE) associated with the bias, and provide the Standard error ratio. The latter shows how much larger the standard error of the estimated coefficients using the imputed data is, compared to the coefficients based on fully observed data.

Based on the results in Table 1 and Table 2, when the error v is assumed to follow a normal or a logistic distribution (upper panel), the bias observed for the conditional and unconditional quantile regressions (column 2 of each subpanel) is negligible. In Table 1, when the error v is normally distributed, the largest bias is observed for the quantile regressions at the bottom of the distribution. Instead, if the errors follow a logistic distribution, the bias is somewhat larger, with the largest bias at 0.03. While this bias does not disappear with larger samples (see appendix), is relatively small compared to the expected coefficient. In Table 2, when the multiplicative heteroskedasticity depends on an exponential function, the bias when v follows a logistic distribution is smaller, but still present.

We also observe that the aggregated imputation-based standard errors are between 5 to 29 percent larger than those based on fully observed data. This is expected given the information loss due to the nature of the interval-censored data. Additional simulations (see appendix) suggest that larger sample sizes have no impact on the precision of using imputation-based estimates.

It is interesting to note that when the d.g.p. follows the linear heteroskedastic form, the coefficients associated with conditional and unconditional quantile regressions have similar levels of precision loss (based on the Standard errors ratio) and are similarly close to the coefficients based on fully observed data (based on MAE). When the d.g.p. assumes an exponential function for heteroskedasticity, the precision loss when estimating unconditional quantile regression is almost double that in the former case.

When the errors v follow a chi2 distribution or uniform distribution (lower panel in tables 1 and 2), we observe nonnegligible bias, especially for the lower quantile coefficients.[4] For example, when considering the 10^{th} conditional quantile coefficient, we see a bias of 0.321, almost 30% of the coefficient magnitude. Although the magnitude of the bias is smaller if the data-generating process imposes a functional form with exponential heteroskedasticity (see Table 2), the magnitude of the bias remains high (up to 0.07 for the unconditional quantile case). Based on further simulations with different sample sizes (see appendix), we observe that the bias magnitude

[4] It may be possible that the location of the bias is explained by features in the data generating process.

does not depend on the sample size, but instead depends strongly on the correct model specification. As we show later in section 4.1 however, further improvements could be obtained with a larger number of brackets.

In Table 3, we show results based on a data-generating process that follows a varying coefficient structure. A varying coefficient model structure is usually considered a more flexible characterization of quantile regressions, compared to models with heteroskedastic errors. For the two specifications we use in our simulations, the Montecarlo simulations suggest there is also a small bias across all coefficients, with similar performance to the case with exponential heteroskedasticity.

A different approach to evaluating the quality of the imputation is analyzing the dispersion of the difference between the estimated coefficients obtained using the fully observed data, and the ones obtained using the imputed data. We do this using the mean absolute error (MAE), where *the error* is defined as the difference between estimated coefficients. In absolute terms, we see that the imputation-based coefficients are better at replicating the fully observed coefficients when considering the middle of the distribution (50th conditional and unconditional quantiles). As can be observed in Table 1, the MAE for the 50th quantile regressions is almost half of that for the 10th quantile, and about 10 to 20% smaller than the 90th percentile. Differences in MAE across quantiles and distribution assumptions of v are much smaller when considering the specification results in Table 2, or when considering the varying coefficient structure of Table 3.

Considering the role of sample sizes, if the assumptions of the imputation model hold, and the bias is small, increasing the sample size improves the overall quality of the imputation-based estimates. Based on the simulations in the appendix, doubling the sample size reduces the MAE between 20% to 30%. When the estimated coefficients are severely biased, we see only minor changes in MAE (compare Table 1 with Appendix A4).

Table 1. Monte Carlo Simulation: N=1000, Linear Heteroskedasticity

$y = x\beta + u * \gamma x$		u~normal				u~logistic			
		E($\hat{\beta}_f$)	Bias	MAE	StErr Ratio	E($\hat{\beta}_f$)	Bias	MAE	StErr Ratio
CQR-Q10	x1	2.011	0.008	0.099	23.540	1.955	-0.030	0.102	15.581
	x2	0.798	0.002	0.061	15.471	0.803	-0.006	0.063	13.418
	cons	-2.381	-0.008	0.125	29.196	-2.247	0.022	0.128	20.698
CQR-Q50	x1	1.000	0.001	0.045	6.955	1.003	-0.001	0.042	8.644
	x2	1.001	-0.002	0.036	7.149	0.997	-0.001	0.033	8.822
	cons	-0.001	0.002	0.049	7.165	-0.002	0.002	0.047	8.790
CQR-Q90	x1	-0.009	0.000	0.064	10.306	0.041	0.005	0.066	9.622
	x2	1.199	0.001	0.051	10.964	1.191	0.000	0.054	10.869
	cons	2.383	-0.001	0.071	9.881	2.252	0.007	0.074	9.530
UQR-Q10	x1	2.097	0.008	0.102	15.906	1.915	-0.021	0.099	11.408
	x2	0.611	0.001	0.046	6.041	0.602	-0.008	0.052	4.672
	cons	-2.537	-0.006	0.095	12.972	-2.342	0.016	0.100	10.820
UQR-Q50	x1	1.006	0.000	0.057	13.020	1.026	-0.007	0.056	15.928
	x2	0.929	0.000	0.041	11.030	0.919	-0.002	0.040	12.792

	cons	0.131	0.001	0.059	12.181	0.120	0.004	0.055	13.987
UQR-Q90	x1	0.052	-0.001	0.067	10.256	0.106	0.004	0.072	10.773
	x2	1.466	0.001	0.074	19.167	1.492	-0.004	0.084	21.324
	cons	2.263	-0.001	0.079	12.912	2.134	0.010	0.088	13.397
$y = x\beta + u * \gamma x$		u~Chi2				u~uniform			
		E($\hat{\beta}_f$)	Bias	MAE	StErr Ratio	E($\hat{\beta}_f$)	Bias	MAE	StErr Ratio
CQR-Q10	x1	1.848	0.321	0.321	91.549	2.094	0.258	0.260	68.957
	x2	0.831	0.123	0.123	37.703	0.786	0.052	0.068	29.649
	cons	-1.989	-0.471	0.471	103.460	-2.572	-0.264	0.270	74.863
CQR-Q50	x1	1.168	0.020	0.044	9.368	0.996	0.004	0.053	4.313
	x2	0.969	0.000	0.032	8.434	0.999	0.001	0.042	4.050
	cons	-0.385	0.023	0.047	9.570	0.005	-0.005	0.058	4.468
CQR-Q90	x1	-0.051	-0.011	0.080	4.794	-0.097	-0.008	0.053	14.553
	x2	1.215	-0.004	0.065	4.700	1.216	-0.001	0.040	14.464
	cons	2.486	-0.003	0.086	4.646	2.573	-0.042	0.066	14.118
UQR-Q10	x1	1.840	0.188	0.194	28.691	2.539	0.236	0.268	45.933
	x2	0.684	0.079	0.088	25.193	0.651	0.027	0.063	19.107
	cons	-2.370	-0.174	0.185	33.232	-3.046	-0.163	0.198	40.143
UQR-Q50	x1	1.165	0.039	0.062	16.087	0.945	0.012	0.060	6.623
	x2	0.945	0.000	0.037	14.438	0.921	0.007	0.045	6.169
	cons	-0.199	0.020	0.052	13.652	0.190	-0.004	0.064	7.339
UQR-Q90	x1	0.014	-0.004	0.068	3.359	-0.003	-0.004	0.059	16.617
	x2	1.455	-0.004	0.089	11.090	1.484	-0.015	0.060	23.134
	cons	2.369	0.000	0.097	6.799	2.314	0.010	0.065	17.632

Note: Monte Carlo Simulation Results. E($\hat{\beta}_f$) represent the average estimated coefficients across all simulations, based on uncensored data. Bias is the average difference of the coefficients using uncensored data and Multiple imputed (MI) data. MAE is the average Mean absolute error (MAE) when comparing MI data and the uncensored data. StErr ratio represents how much larger the Std error of the coefficients is using imputed data, compared to the fully observed data. CQR: Conditional Quantile Regression; UQR: Unconditional Quantile Regression.

Table 2 Monte Carlo Simulation: N=1000, exponential Heteroskedasticity

$y = x\beta + u * e^{\gamma x}$		u~normal				u~logistic			
		E($\hat{\beta}_f$)	Bias	MAE	StErr Ratio	E($\hat{\beta}_f$)	Bias	MAE	StErr Ratio
CQR-Q10	x1	1.639	0.001	0.048	16.840	1.603	0.009	0.050	14.282
	x2	0.743	-0.004	0.040	17.292	0.758	-0.003	0.043	16.214
	cons	-1.280	0.003	0.052	15.957	-1.209	0.013	0.056	12.862
CQR-Q50	x1	1.000	0.000	0.033	11.537	0.999	0.000	0.032	15.846
	x2	0.999	0.000	0.027	11.436	0.998	0.003	0.026	15.839
	cons	0.000	-0.001	0.038	11.918	0.004	-0.003	0.036	16.176
CQR-Q90	x1	0.364	0.002	0.050	17.023	0.392	-0.009	0.050	15.391
	x2	1.255	0.001	0.039	16.242	1.239	0.002	0.040	15.707
	cons	1.279	-0.001	0.055	16.401	1.216	-0.012	0.056	14.733
UQR-Q10	x1	1.613	0.002	0.088	25.565	1.478	-0.018	0.085	25.832
	x2	0.582	-0.001	0.044	10.654	0.565	-0.006	0.039	9.973
	cons	-1.533	0.001	0.098	27.431	-1.390	0.019	0.089	25.553
UQR-Q50	x1	1.003	-0.002	0.047	24.581	1.025	0.015	0.050	27.348
	x2	0.850	-0.001	0.033	18.288	0.853	0.009	0.032	20.023
	cons	0.170	0.000	0.045	20.692	0.152	-0.016	0.047	22.121

UQR-Q90	x1	0.430	0.002	0.040	7.693	0.446	0.006	0.037	5.046
	x2	1.624	0.001	0.087	48.404	1.612	0.024	0.087	43.271
	cons	1.212	-0.002	0.092	32.304	1.190	-0.027	0.094	29.040
$y = x\beta + u * e^{\gamma x}$		u~Chi2				u~uniform			
		E($\hat{\beta}_f$)	Bias	MAE	StErr Ratio	E($\hat{\beta}_f$)	Bias	MAE	StErr Ratio
CQR-Q10	x1	1.536	0.030	0.043	48.403	1.691	0.012	0.044	31.959
	x2	0.788	-0.018	0.033	46.521	0.726	-0.046	0.053	26.788
	cons	-1.072	0.050	0.058	45.635	-1.384	0.008	0.050	34.370
CQR-Q50	x1	1.102	-0.025	0.038	15.236	1.001	0.001	0.038	4.488
	x2	0.959	0.009	0.027	15.400	1.002	-0.007	0.033	4.317
	cons	-0.204	-0.053	0.058	15.355	0.000	0.008	0.045	4.453
CQR-Q90	x1	0.331	0.003	0.060	5.324	0.306	-0.014	0.042	27.005
	x2	1.269	0.007	0.048	3.487	1.274	0.012	0.034	25.265
	cons	1.340	0.010	0.067	4.724	1.386	0.040	0.056	26.647
UQR-Q10	x1	1.273	-0.071	0.083	26.493	1.734	0.125	0.130	8.851
	x2	0.648	0.003	0.036	18.002	0.670	0.084	0.088	4.190
	cons	-1.417	0.042	0.081	34.823	-1.843	-0.200	0.203	16.747
UQR-Q50	x1	1.099	-0.037	0.056	28.393	0.922	-0.072	0.079	18.122
	x2	0.860	0.031	0.041	21.901	0.839	-0.042	0.050	13.941
	cons	0.023	-0.059	0.067	23.301	0.245	0.077	0.085	18.233
UQR-Q90	x1	0.383	0.020	0.049	0.030	0.429	-0.009	0.040	12.949
	x2	1.585	0.064	0.113	29.469	1.630	-0.041	0.088	55.819
	cons	1.305	-0.066	0.117	16.299	1.217	0.043	0.093	37.680

Note: Monte Carlo Simulation Results. E($\hat{\beta}_f$) represent the average estimated coefficients across all simulations, based on uncensored data. Bias is the average difference of the coefficients using uncensored data and Multiple imputed (MI) data. MAE is the average Mean absolute error (MAE) when comparing MI data and the uncensored data. StErr ratio represents how much larger the Std error of the coefficients is using imputed data, compared to the fully observed data. CQR: Conditional Quantile Regression; UQR: Unconditional Quantile Regression.

Table 3 Monte Carlo Simulation: N=1000, Varying coefficient structure

$y = x\beta(t)$		Type 1				Type 2			
		E($\hat{\beta}_f$)	Bias	MAE	StErr Ratio	E($\hat{\beta}_f$)	Bias	MAE	StErr Ratio
CQR-Q10	x1	-1.140	0.000	0.064	10.331	-0.092	-0.011	0.061	15.394
	x2	-0.035	-0.010	0.054	12.024	-0.086	-0.010	0.053	15.588
	cons	0.356	0.010	0.061	18.218	-0.086	0.043	0.066	16.975
CQR-Q50	x1	0.404	0.002	0.043	6.726	0.845	0.009	0.051	5.568
	x2	0.601	0.001	0.033	6.521	0.841	0.004	0.042	4.454
	cons	0.998	-0.003	0.037	8.535	0.853	-0.029	0.053	5.877
CQR-Q90	x1	1.938	-0.001	0.063	9.790	2.282	0.001	0.095	4.465
	x2	1.236	0.005	0.051	9.840	2.280	0.010	0.108	5.448
	cons	1.644	-0.004	0.053	13.852	2.309	0.002	0.105	7.259
UQR-Q10	x1	-1.211	-0.002	0.085	17.009	-0.097	-0.018	0.061	17.379
	x2	-0.044	-0.002	0.042	7.209	-0.078	-0.012	0.044	15.686
	cons	0.482	0.004	0.054	6.569	-0.074	0.056	0.075	16.203
UQR-Q50	x1	0.418	0.003	0.046	11.851	0.900	0.008	0.036	3.104
	x2	0.535	0.003	0.036	11.204	0.737	0.005	0.026	2.576
	cons	0.899	-0.007	0.050	12.403	0.757	-0.017	0.038	2.795
UQR-Q90	x1	1.982	-0.001	0.109	15.015	2.214	0.002	0.112	5.050
	x2	1.296	-0.002	0.080	13.614	2.321	-0.001	0.110	7.709

	cons	1.778	0.003	0.111	14.129	2.499	0.006	0.129	5.711

Note: Monte Carlo Simulation Results. E($\hat{\beta}_f$) represent the average estimated coefficients across all simulations, based on uncensored data. Bias is the average difference of the coefficients using uncensored data and Multiple imputed (MI) data. MAE is the average Mean absolute error (MAE) when comparing MI data and the uncensored data. StErr ratio represents how much larger the Std error of the coefficients is using imputed data, compared to the fully observed data. CQR: Conditional Quantile Regression; UQR: Unconditional Quantile Regression.

4. Further Considerations

4.1. On the Role of brackets

As presented in Section 3, the successful implementation of the methodology we propose depends greatly on the model specification assumptions. If the underlying censored data, or some monotonic transformation, has a conditional distribution that can be modeled as a normal distribution with multiplicative error structure, the imputation procedure would do a good job creating imputed data that resembles the true but unobserved data. Yet, if the assumptions are incorrect, we will have a misspecification problem that would generate biases when analyzing the data.

However, because imputed data is constrained to be within the provided brackets, it is possible to improve the quality of the imputed data by using more brackets with narrower limits, even if the assumptions regarding the conditional distribution of the outcome are incorrect. In other words, the imputation quality will improve if the width of the brackets decreases.

To see this, we use the structure described by equation (17b), assuming the error v follows a normal distribution (case 1), and a chi2 distribution (case 2). The second case will be equivalent to having a misspecification problem regarding the distribution of v. We assume, however, that the conditional mean and conditional variance models are correctly specified. For the bracket's width, we consider two cases, one where there are 5 equidistant brackets and one with 15 equidistant brackets. We report the simulation results in Table 4, considering only the estimates for conditional quantile regressions.

Table 4 Monte Carlo Simulation: N=1000, Role of Brackets

			5 Brackets		15 Brackets	
	v~normal	E($\hat{\beta}_f$)	Bias	MAE	Bias	MAE
CQR-Q10	x1	1.387	0.004	0.083	0.000	0.047
	x2	0.805	-0.001	0.083	0.000	0.037
	cons	-1.925	-0.004	0.083	0.000	0.050
CQR-Q50	x1	0.999	-0.001	0.051	0.000	0.030
	x2	1.000	-0.001	0.051	-0.001	0.025
	cons	0.001	0.002	0.051	0.001	0.033
CQR-Q90	x1	0.620	0.000	0.079	0.000	0.047
	x2	1.195	0.001	0.079	-0.001	0.052

	cons	1.919	-0.002	0.079	0.000	0.058
			5 Brackets		15 Brackets	
	v~Chi2	E($\hat{\beta}_f$)	Bias	MAE	Bias	MAE
CQR-Q10	x1	1.321	0.004	0.055	0.002	0.031
	x2	0.836	0.045	0.055	-0.003	0.026
	cons	-1.603	-0.003	0.055	0.009	0.034
CQR-Q50	x1	1.062	-0.020	0.051	-0.002	0.030
	x2	0.966	0.002	0.051	0.001	0.024
	cons	-0.303	-0.076	0.051	-0.012	0.033
CQR-Q90	x1	0.604	-0.007	0.099	0.000	0.058
	x2	1.197	0.011	0.099	0.005	0.053
	cons	2.015	0.031	0.099	0.001	0.066

Note: Monte Carlo Simulation Results. E($\hat{\beta}_f$) represent the average estimated coefficients across all simulations, based on uncensored data. Bias is the average difference of the coefficients using uncensored data and Multiple imputed (MI) data. MAE is the average Mean absolute error (MAE) when comparing MI data and the uncensored data. CQR: Conditional Quantile Regression.

As can be seen in this table, when the conditional normality assumption holds, the imputation approach produces unbiased estimates for the coefficients across all quantiles (10^{th}, $50^{th,}$ and 90^{th}), regardless of the number of brackets considered. However, we also observe that both the bias and the precision of the imputed estimates (measured using the MAE) improve considerably when 15 brackets are utilized. In contrast, when the error v is assumed to follow a Chi2 distribution, instead of a normal distribution, the estimates based on the imputed data show a larger bias. This is similar to what we saw before in Tables 1 and 2. Using more brackets reduces the bias and improves the precision of the estimates, as seen in the bottom right panel of Table 4.

4.2. Non-response and Missing Data

A second aspect of interest is the treatment of survey non-response. Similar to the treatment of missing data elsewhere in the literature (see Enders (2022), chp 1) it is necessary to consider why the data is missing. Under the assumption of missing at random (MAR), we could use interval regression modeling to correctly identify the conditional distribution of the outcome, and impute the outcomes for the censored and the missing data. Alternatively, we could also use an inverse probability weighting (IPW) approach to account for sample composition bias. We provide an example of applying the full imputation in section 5.

If data is not missing at random, for example by people self-selecting and refusing to answer the survey, we face a problem of misspecification and would be unable to identify the true conditional distribution, instead identifying the endogenous sample conditional distribution of the outcome. Smaller brackets would only improve the imputation of the censored data, not that of the missing data. This is not dissimilar to the assumptions used in other multiple imputation approaches. Addressing problems of missing data missing not at random (MNAR) is beyond the scope of this paper.

In terms of implementation, the command that implements our strategy imputes the outcomes for all observations in the data by default, unless it is requested otherwise. In the example we provide in Section 5, we assume that the non-response items are missing at random, imputing earnings for those who refuse to report income. We also provide in the appendix a robustness check where we address the truly missing data using IPW.

4.3. Choice of Covariates and Model Overfitting

Following the literature on imputation (Enders,2022), covariates should be chosen in terms of what factors better predict the outcome of interest. When using the approach to impute non-response items, one should also include covariates that determine why data was missing. As general advice, the set of covariates used for imputation should be at least as extensive as the set used for data modeling. This would help provide a flexible specification for the identification of the conditional distribution of the outcome.

The fewer the covariates available, the more one relies on the identification based on the bracket's boundaries. In contrast, if the number of covariates used in the modeling increases, it may cause problems of overfitting, reducing the quality of imputed values of non-response items, because of the increased variation (standard errors) of the estimated coefficients. This would result in unbiased estimated coefficients but with potentially larger variation. For the case of imputed censored data, because the imputed values depend on the coefficient variation, error variation, and brackets limits, the final effect on the quality of the imputed values may be smaller.

To see this, we run a Monte Carlo simulation using a data structure with multiplicative heteroskedasticity similar to Equation 17b, with some differences. First, we consider 3 explanatory variables (x_1, x_2 and x_3), all of which follow a standard normal distribution. Second, we assume these variables only affect the conditional variance, not the conditional mean. For the imputation step, we consider two scenarios based on the number of brackets, combined with a scenario where the covariates are excluded from the conditional mean modeling (correct model), and one where they are included (overfitting). Results corresponding to the conditional quantile regressions are presented in Table 5. For completeness, we also consider a complementary setup, where the covariates affect both the conditional mean and variance, but they are not considered for modeling the conditional mean. These results are presented in Table 6

As we observe in Table 5, because the underlying assumption of normality holds, the bias of the coefficients is negligible, regardless of the number of brackets or model specification. Similar to Table 4, we observe that using more brackets improves the imputation quality, based on the smaller MAE. Interestingly, by adding unnecessary controls to the conditional mean (overfitting), there is a small loss in efficiency (larger MAE). In contrast, when considering the problem of underfitting (table 6), we see that ignoring important variables in the model generates a non-negligible bias on the estimated coefficients. Nevertheless, as we have shown before, increasing the number of brackets helps reduce such bias.

Table 5 Monte Carlo Simulation: N=1000, Exact fitting vs Overfitting

			Exact Fitting				Overfitting			
			5 Brackets		15 Brackets		5 Brackets		15 Brackets	
	u~normal	E($\hat{\beta}_f$)	Bias	MAE	Bias	MAE	Bias	MAE	Bias	MAE
CQR-Q10	x1	-0.118	0.000	0.026	0.000	0.016	0.000	0.027	0.000	0.016
	x2	0.353	0.000	0.024	0.000	0.014	0.000	0.024	0.000	0.014
	x3	-0.233	0.001	0.026	0.000	0.015	0.001	0.026	0.001	0.015
	cons	-0.369	0.000	0.029	0.000	0.017	-0.001	0.029	0.000	0.017
CQR-Q50	x1	0.000	0.000	0.021	0.000	0.012	0.000	0.022	0.000	0.012
	x2	0.001	0.000	0.021	0.000	0.012	0.001	0.022	0.000	0.012
	x3	0.000	0.000	0.021	0.000	0.012	0.000	0.022	0.000	0.012
	cons	0.999	0.000	0.024	0.000	0.013	0.000	0.024	0.000	0.013
CQR-Q90	x1	0.116	0.000	0.026	-0.001	0.015	0.000	0.027	-0.001	0.016
	x2	-0.351	0.000	0.024	0.000	0.015	0.000	0.024	0.000	0.015
	x3	0.235	0.000	0.025	-0.001	0.015	-0.001	0.025	-0.001	0.015
	cons	2.368	0.002	0.029	0.000	0.017	0.002	0.029	0.000	0.017

Note: Monte Carlo Simulation Results. E($\hat{\beta}_f$) represent the average estimated coefficients across all simulations, based on uncensored data. Bias is the average difference of the coefficients using uncensored data and Multiple imputed (MI) data. MAE is the average Mean absolute error (MAE) when comparing MI data and the uncensored data. CQR: Conditional Quantile Regression.

Table 6 Monte Carlo Simulation: N=1000, Exact fitting vs Underfitting

			Exact Fitting				Underfitting			
			5 Brackets		15 Brackets		5 Brackets		15 Brackets	
	u~normal	E($\hat{\beta}_f$)	Bias	MAE	Bias	MAE	Bias	MAE	Bias	MAE
CQR-Q10	x1	-1.117	0.001	0.036	0.000	0.022	-0.112	0.112	-0.026	0.031
	x2	1.351	0.000	0.034	0.000	0.021	0.168	0.168	0.036	0.038
	x3	-1.234	-0.001	0.035	0.000	0.021	-0.141	0.141	-0.030	0.034
	cons	-0.368	-0.004	0.039	-0.002	0.022	0.290	0.290	0.033	0.037
CQR-Q50	x1	-1.001	0.000	0.024	-0.001	0.014	-0.088	0.088	-0.014	0.018
	x2	1.000	-0.001	0.024	0.000	0.014	0.089	0.089	0.014	0.018
	x3	-0.999	0.001	0.024	0.000	0.014	-0.088	0.088	-0.014	0.018
	cons	1.000	0.000	0.026	0.000	0.015	-0.003	0.027	-0.001	0.015
CQR-Q90	x1	-0.885	0.000	0.035	0.000	0.021	-0.067	0.070	-0.009	0.023
	x2	0.647	-0.002	0.033	0.000	0.020	0.019	0.038	0.002	0.021
	x3	-0.765	0.001	0.034	0.001	0.021	-0.043	0.051	-0.005	0.022
	cons	2.368	0.003	0.039	0.001	0.022	-0.295	0.295	-0.042	0.044

Note: Monte Carlo Simulation Results. E($\hat{\beta}_f$) represent the average estimated coefficients across all simulations, based on uncensored data. Bias is the average difference of the coefficients using uncensored data and Multiple imputed (MI) data. MAE is the average Mean absolute error (MAE) when comparing MI data and the uncensored data. CQR: Conditional Quantile Regression.

4.4. General Considerations on Implementation

As described earlier, there are general considerations one should keep in mind when applying the methodology, including variable choice, functional form specification, and data transformations. First, following the literature on imputation analysis, the variable choice should consider variables that explain the outcome, allowing for a sufficiently flexible model specification for modeling the conditional mean and variance. This may include the use of interactions and high-order polynomials. In addition, based on a few examples available in our repository,[5] the imputation method provides sensible results if the imputation step considers at least all variables used in the analysis step. Although this practice may lead to model overfitting, the drawbacks of misspecification errors outweigh the loss of precision derived from model overfitting.

Regarding model misspecification, it is important to consider that the main assumption of the model is that the outcome of interest, or some transformation of it, follows a conditionally normal distribution. In the example presented in the next section, and the ones available in the online repository, we have used the log transformation as a simple and common approach to model the dependent variable. However, similar to the work on small area poverty estimations (Corral et al., 2021), one can consider other transformations, including log-shift transformation, Box-Cox transformation, or a hyperbolic sine transformation,[6] among others, to help fulfill the model assumptions. If the conditionally normal distribution assumption is questionable, other methods that deal with interval-censored data as described in McDonald et al., (2018) could be applied, and our methodology extended.

Like most imputation methods in the literature, when there are no response items, our methodology relies on the assumption that data is missing at random (MAR). In general, the application of the imputation method becomes problematic in scenarios when data is missing not at random (MNAR), i.e., endogenous sample selection. If data is MAR, the missing responses could be imputed by combining our methodology with a re-weighting approach as shown in section 5. Otherwise, imputation could be done as is, if the covariates used include factors that relate to the missingness mechanism. In such cases, where data is MNAR, it may be necessary to use Heckman-type selection models or pattern mixture models (Enders, 2011, 2022; C. Hsu et al., 2023; Muñoz et al., 2023), to impute the missing information. With sufficiently small brackets, this may not be a problem, however, this is a topic left for further research.

[5] A set of examples that shows the application and performance of the methodology can be found at https://github.com/friosavila/intreg_mi.

[6] We thank an anonymous referee for the suggestion.

5. Wage Inequality in Grenada

This illustration focuses on an empirical application of our proposed method for the case of Grenada, focusing on the description of wage inequality trends in the country between 2013 and 2020 using the annual Labor Force Survey (LFS). This survey provides is the only source of information that can be used to describe the status of the labor market and the distribution of labor income in the country.

One major limitation of this survey, however, is the collection of labor income data. Compared to standard household surveys or labor force surveys in most developed countries, labor income recorded in the LFS in Grenada is only available in brackets. Furthermore, there is a large proportion of the employed population who do not declare their labor income. Table 7 provides an overview of the labor income distribution across time.

Table 7 Labor Income distribution by year

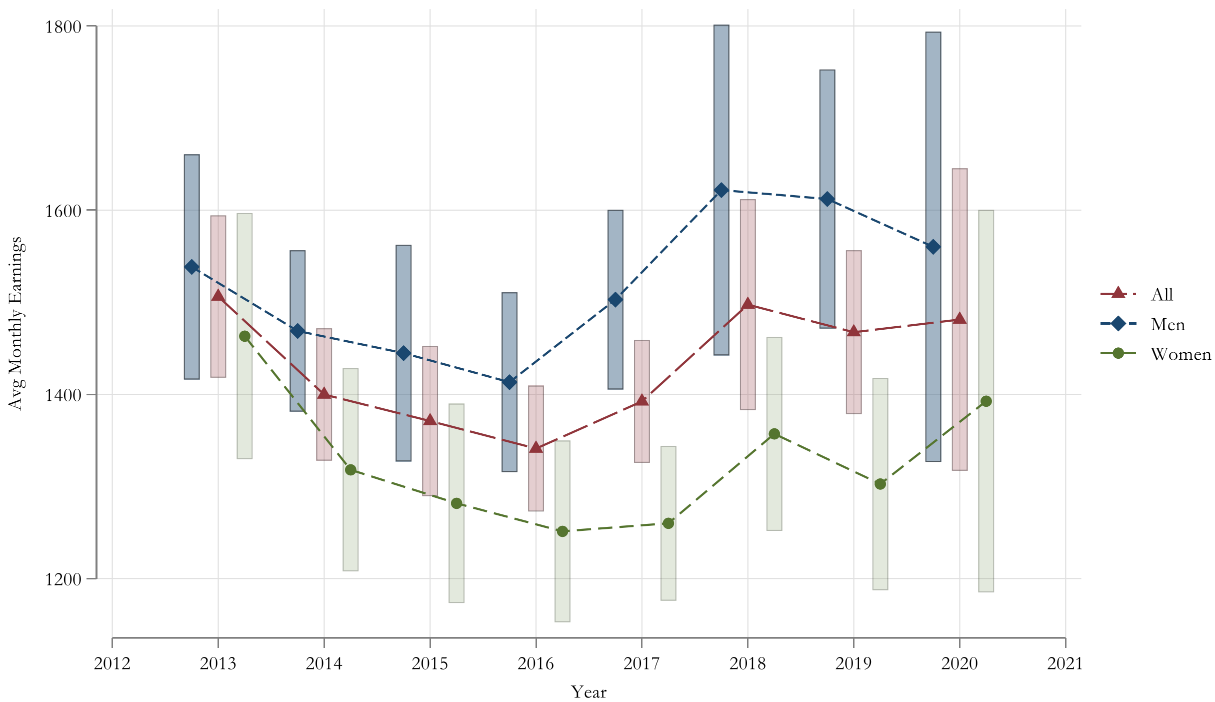
Year	2013	2014	2015	2016	2017	2018	2019	2020
>200	2.9	0.9	3.7	3.5	1.4	0.2	0.0	0.5
200-399	7.1	5.5	6.2	5.4	4.1	1.6	1.2	1.2
400-799	15.1	15.7	12.2	14.2	13.7	9.1	8.3	10.3
800-1199	19.2	20.2	18.3	18.6	21.1	20.5	23.8	23.7
1200-1999	17.5	17.3	13.8	13.1	18.4	14.7	14.9	15.8
2000-3999	15.6	11.4	11.1	11.4	10.5	9.7	12.8	11.1
4000-5999	2.5	2.5	2.4	2.2	2.2	1.5	1.2	2.2
6000+	2.0	1.2	0.6	0.6	0.7	1.0	1.0	0.5
Not stated	18.1	25.3	31.6	31.1	27.9	41.9	36.7	34.7
N	1056	1285	1290	1349	1485	1089	858	460

Note: Censored Data distribution based on Grenada Labour Force Survey,

In this case, we face two types of problems. On the one hand, we only have access to interval-censored data, which is insufficient to analyze changes in the distribution of earnings in the country, and, on the other hand, we have an increasing proportion of individuals who do not declare income. We apply the imputation procedure previously described to address both problems, estimating the interval-censored regression for each year, with a set of household-level characteristics and job type characteristics. The sample of interest includes all adults who declared to be employed, even if they did not state their income. It should be emphasized that the application of this methodology relies on the assumption that nonresponse can be classified as missing at random MAR, and that our modeling accounts both for income-determining factors, as well as factors affecting the likelihood of not declaring income. This is a simplifying assumption that we use for the exercise, but may not be reliable in other settings.

To account for the fact that characteristics may differ across those who did or did not state their incomes, an inverse probability weighting strategy is used to estimate the interval regression model. Finally, the imputation procedure is implemented as discussed in section 3 using the natural logarithm on the bracket limits. Thus, we assume no lower and upper bounds for the imputed log wages. Nevertheless, the maximum imputed wage for those who do not state their income is capped at the maximum predicted among those who declare their income, to avoid extreme outliers.[7] Wages and brackets are measured in Eastern Caribbean dollars (XCD), adjusted by inflation using 2010 as the base year. While we impute log wages, we transform the data back to levels to estimate the different statistics shown in Figures 1 and 2.

Figure 1 Average Monthly Earnings by Year and Gender



Note: Average Monthly earnings by year and Gender, based on full imputed data. 90% CI

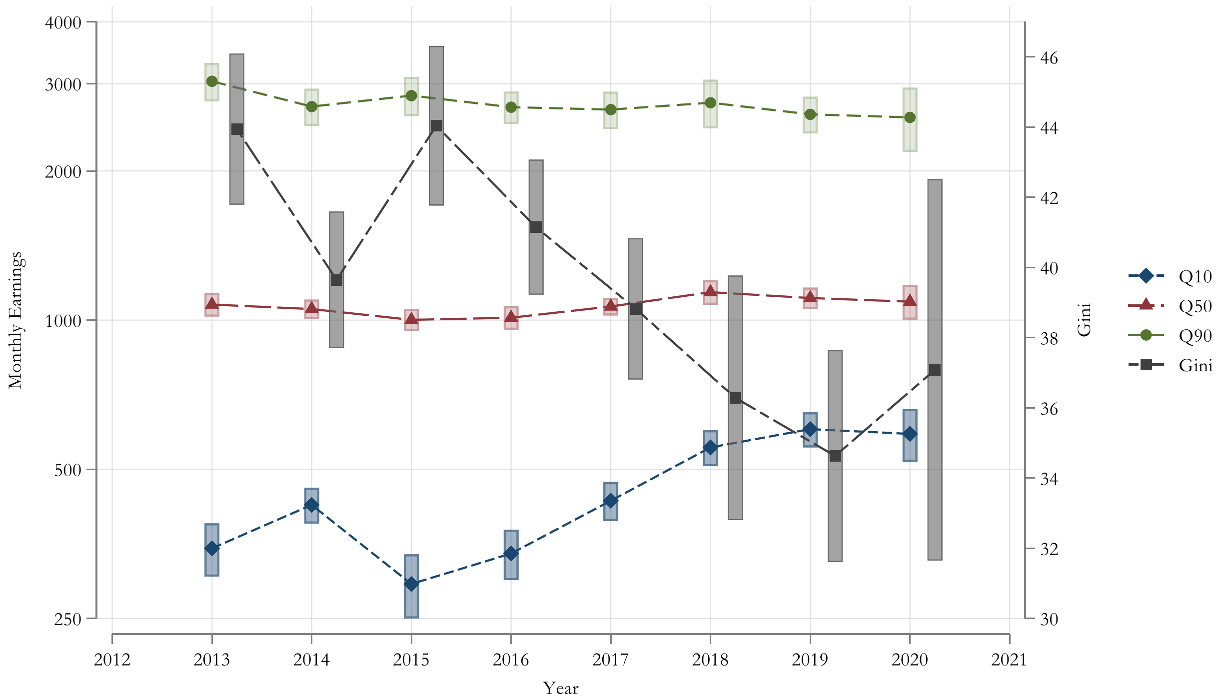
Figure 1 shows the average earned income for the total employed population, as well as for men and women separately, including 90% confidence intervals. The results suggest that after a small decline in average real monthly earnings from 2013 to 2016, there was a slight improvement in the following two years, with a small decline in 2019, with average wages remaining at stable levels in 2020, despite the COVID-19 pandemic.[8] The results also suggest that the gender earnings gap has shown a somewhat increasing trend between 2013 and 2019, although we predict a small

[7] In the simulations, maintaining the assumption of no upperbound limit for the imputed values would create some unusually large imputations among the non-response items. Because of this, we decide to set limits in the data to reduce the possibilities of generating unrealistic imputed datasets.

[8] This estimate does not take into account the decline in labor force participation observed during the pandemic.

decline in 2020. In the appendix, we reproduce a similar plot using excluding non-respondents, but utilizing inverse probability weight, observing similar conclusions.

Figure 2 Selected Quantiles and Gini Coefficient across Years



Note: Selected Quantiles and Gini coefficients, based on full imputed data. 90% CI

Figure 2 provides results using selected inequality statistics. The estimates suggest that inequality has declined substantially across the years. The estimated Gini coefficient fell from 44.2 Gini points in 2015 to 34.1 in 2019, with an increase in 2020. This decline in inequality seems to have been driven by faster growth in the lower and middle sections of the wage distribution and a small decline in the upper section of the distribution.

While such a decline in inequality may seem larger than average, even among other countries in the region, it is unlikely that it is driven by features of the imputation procedure. While less evident, the crosstabulation presented in Table 7 already suggests a concentration of wages, with an increasing proportion of individuals declaring wages in the middle brackets. On the other hand, according to the World Bank Outlook Poverty Report (World Bank, 2020), Grenada experienced a steady growth path before the COVID-19 crisis, driven by an expansion of the tourism and construction sectors. The expansion of these sectors aligns with the estimated wage increases at the bottom of the distribution, as we show in Figure 2.

6. Conclusion

In this paper, we present an imputation strategy that can be used to analyze interval-censored data. Our method proposes that a flexible enough interval regression model can be used to impute censored data, which allows to recover the full distribution of data and can be further analyzed using standard statistical methods.

The main limitation of our strategy is the assumption of conditional normality we impose on the distribution, which is required for the estimation of the interval regression model using standard software. In fact, we have shown that the quality of the imputation depends strongly on the correct model specification of the conditional mean and conditional variance. The principles of the imputation approach, however, could be extended to allow for more flexible moment specifications, as well as error distributions. A second potential limitation is related to the presence of non-response items with endogenous missing data. Following the literature, it may be possible to extend our methodology with other strategies that deal with data missing not at random such as the use of reweighted data, as shown in the empirical example, or combine it with the use of Heckman selection type models.

Nevertheless, the Monte Carlo simulation suggests that as long as the latent error has a symmetric bell-shaped distribution, regression analysis using the imputed data shows small biases, with performance that is comparable to analyzing the uncensored data. Likewise, when the heteroskedasticity structure is given by an exponential function, biases are small even when the latent error follows a skew or a limited distribution. Furthermore, even if the imputation model is misspecified, multiple imputation could still provide a good approximation for analysis if the width of the brackets is narrow. In some cases, it may be the only approach to analyze the data.

For the specific case of Grenada we only had access to interval-censored data, which is insufficient to analyze changes in the distribution of earnings in the country, and, on the other hand, we have an increasing proportion of individuals who do not declare income. We apply the imputation procedure to address both problems, under the assumption that non-response items follow a missing at-random pattern. Interval-censored regressions are estimated for each year, with a set of household-level characteristics and job-type characteristics, and the estimates used for imputation. The results suggest that earned income inequality in this country has declined, which coincides with other economic performance indicators, and the growth of the tourism and construction sector.

While this method aims to provide an imputation approach that facilitates the analysis of interval-censored data, the imputation quality will depend on the identification of the conditional distribution of the outcome, or some monotonic transformation of it, which is unobserved. However, using imputed data may still provide better estimates and insights than not using any imputation at all.

References

Angelov, A. G., & Ekström, M. (2019). Maximum likelihood estimation for survey data with informative interval censoring. *AStA Advances in Statistical Analysis*, *103*(2), 217–236. https://doi.org/10.1007/s10182-018-00329-x

Büttner, T., & Rässler, S. (2008). *Multiple imputation of right-censored wages in the German IAB employment sample considering heteroscedasticity* (Issue 44/2008). Institut für Arbeitsmarkt- und Berufsforschung (IAB). http://hdl.handle.net/10419/32715

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and applications*. Cambridge University Press.

Chen, Y.-T. (2018). A Unified Approach to Estimating and Testing Income Distributions With Grouped Data. *Journal of Business & Economic Statistics*, *36*(3), 438–455. https://doi.org/10.1080/07350015.2016.1194762

Corral, P., Himelein, K., McGee, K., & Molina, I. (2021). A Map of the Poor or a Poor Map? *Mathematics*, *9*(21), 2780. https://doi.org/10.3390/math9212780

Enders, C. K. (2011). Missing not at random models for latent growth curve analyses. *Psychological Methods*, *16*(1), 1–16. https://doi.org/10.1037/a0022640

Enders, C. K. (2022). *Applied missing data analysis* (Second Edition). The Guilford Press.

Firpo, S., Fortin, N. M., & Lemieux, T. (2009). Unconditional Quantile Regressions. *Econometrica*, *77*(3), 953–973.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (Third edition). CRC Press, Taylor and Francis Group.

Hagenaars, A., & de Vos, K. (1988). The Definition and Measurement of Poverty. *The Journal of Human Resources*, *23*(2), 211–221. https://doi.org/10.2307/145776

Han, J., Meyer, B. D., & Sullivan, J. X. (2020). *Income and Poverty in the COVID-19 Pandemic* (Working Paper 27729). National Bureau of Economic Research. https://doi.org/10.3386/w27729

Hsu, C., He, Y., Hu, C., & Zhou, W. (2023). A multiple imputation-based sensitivity analysis approach for regression analysis with a missing not at random covariate. *Statistics in Medicine*, *42*(14), 2275–2292. https://doi.org/10.1002/sim.9723

Hsu, C.-Y., Wen, C.-C., & Chen, Y.-H. (2021). Quantile function regression analysis for interval censored data, with application to salary survey data. *Japanese Journal of Statistics and Data Science*, *4*(2), 999–1018. https://doi.org/10.1007/s42081-021-00113-3

Jenkins, S. P., Burkhauser, R. V., Feng, S., & Larrimore, J. (2011). Measuring Inequality Using Censored Data: A Multiple-Imputation Approach to Estimation and Inference. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *174*(1), 63–81. https://doi.org/10.1111/j.1467-985X.2010.00655.x

Machado, J. A. F., & Santos Silva, J. M. C. (2019). Quantiles via moments. *Journal of Econometrics*, *213*(1), 145–173. https://doi.org/10.1016/j.jeconom.2019.04.009

McDonald, J., Stoddard, O., & Walton, D. (2018). On using interval response data in experimental economics. *Journal of Behavioral and Experimental Economics*, *72*, 9–16. https://doi.org/10.1016/j.socec.2017.10.003

Moore, J. C., Stinson, L. L., & Welniak, E. J. (2000). Income measurement error in surveys: A review. *Journal of Official Statistics-Stockholm-*, *16*(4), 331–362.

Muñoz, J., Efthimiou, O., Audigier, V., De Jong, V. M. T., & Debray, T. P. A. (2023). Multiple imputation of incomplete multilevel data using Heckman selection models. *Statistics in Medicine*, sim.9965. https://doi.org/10.1002/sim.9965

Parolin, Z., & Wimer, C. (2020). Forecasting estimates of poverty during the COVID-19 crisis. *Poverty and Social Policy Brief*, *4*(8), 1–18.

Royston, P. (2007). Multiple Imputation of Missing Values: Further Update of Ice, with an Emphasis on Interval Censoring. *The Stata Journal*, *7*(4), 445–464. https://doi.org/10.1177/1536867X0800700401

Rubin, D. B. (1987). *Multiple Imputation for nonresponse in surveys*. Wiley.

Stewart, M. B. (1983). On Least Squares Estimation when the Dependent Variable is Grouped. *The Review of Economic Studies*, *50*(4), 737–753. JSTOR. https://doi.org/10.2307/2297773

Vega Yon, G. G., & Quistorff, B. (2019). parallel: A command for parallel computing. *The Stata Journal*, *19*(3), 667–684. https://doi.org/10.1177/1536867X19874242

Walter, P., & Weimer, K. (2018). *Estimating poverty and inequality indicators using interval censored income data from the German microcensus* (Discussion Paper 2018/10). Freie Universität Berlin, School of Business & Economics. http://hdl.handle.net/10419/179926

Wang, X., Chen, M.-H., & Yan, J. (2013). Bayesian dynamic regression models for interval censored survival data with application to children dental health. *Lifetime Data Analysis*, *19*(3), 297–316. https://doi.org/10.1007/s10985-013-9246-8

World Bank. (2020). *Macro poverty outlook: Country-by-country analysis and projections for the developing world.* World Bank, Washington, DC.

Yan, T., Qu, L., Li, Z., & Yuan, A. (2018). Conditional kernel density estimation for some incomplete data models. *Electronic Journal of Statistics*, *12*(1), 1299–1329. https://doi.org/10.1214/18-EJS1423

Zhou, X., Feng, Y., & Du, X. (2017). Quantile regression for interval censored data. *Communications in Statistics - Theory and Methods*, *46*(8), 3848–3863. https://doi.org/10.1080/03610926.2015.1073317

Appendix

Table A1. Monte Carlo Simulation: N=500, Linear Heteroskedasticity

$y = x\beta + u * \gamma x$		u~normal				u~logistic			
		E($\hat{\beta}_f$)	Bias	MAE	StErr Ratio	E($\hat{\beta}_f$)	Bias	MAE	StErr Ratio
CQR-Q10	x1	2.005	-0.023	0.141	24.935	1.962	0.028	0.142	17.682
	x2	0.808	-0.006	0.090	16.189	0.802	0.004	0.091	14.581
	cons	-2.385	0.025	0.183	30.300	-2.251	-0.020	0.182	22.701
CQR-Q50	x1	1.003	-0.001	0.063	6.241	0.998	0.001	0.059	7.952
	x2	1.004	0.000	0.050	6.252	0.999	0.000	0.049	7.986
	cons	-0.004	0.000	0.069	6.378	0.005	-0.001	0.068	8.074
CQR-Q90	x1	-0.011	0.000	0.091	10.490	0.040	-0.005	0.095	10.556
	x2	1.205	0.001	0.074	11.240	1.190	0.004	0.077	11.882
	cons	2.376	0.000	0.102	10.343	2.255	-0.011	0.107	10.434
UQR-Q10	x1	2.071	-0.022	0.142	16.525	1.907	0.018	0.129	11.404
	x2	0.613	-0.005	0.066	7.379	0.594	0.006	0.070	4.944
	cons	-2.528	0.016	0.139	14.910	-2.327	-0.012	0.134	11.376
UQR-Q50	x1	1.023	-0.002	0.077	12.916	1.035	0.009	0.080	15.760
	x2	0.941	-0.003	0.057	10.562	0.934	0.001	0.055	12.713
	cons	0.111	0.003	0.079	11.629	0.107	-0.004	0.076	13.627
UQR-Q90	x1	0.042	0.000	0.091	9.811	0.100	-0.006	0.098	10.405
	x2	1.470	0.005	0.101	18.028	1.469	0.004	0.116	20.676
	cons	2.267	-0.001	0.108	11.837	2.158	-0.010	0.118	12.544
$y = x\beta + u * \gamma x$		u~Chi2				u~uniform			
		E($\hat{\beta}_f$)	Bias	MAE	StErr Ratio	E($\hat{\beta}_f$)	Bias	MAE	StErr Ratio
CQR-Q10	x1	1.843	-0.330	0.331	80.730	2.093	-0.265	0.277	58.716
	x2	0.838	-0.123	0.127	33.472	0.783	-0.052	0.086	24.910
	cons	-1.988	0.481	0.481	91.788	-2.563	0.274	0.298	63.838
CQR-Q50	x1	1.160	-0.021	0.062	9.472	1.009	0.000	0.074	5.395
	x2	0.971	0.000	0.045	8.170	0.996	0.000	0.058	5.473
	cons	-0.379	-0.023	0.067	10.375	-0.002	0.002	0.081	5.567
CQR-Q90	x1	-0.047	0.012	0.112	5.447	-0.090	0.010	0.073	11.999
	x2	1.212	0.002	0.089	5.565	1.209	0.000	0.059	11.774
	cons	2.488	0.006	0.124	5.152	2.569	0.039	0.087	11.543
UQR-Q10	x1	1.829	-0.201	0.218	29.506	2.461	-0.264	0.311	43.730
	x2	0.687	-0.087	0.107	25.826	0.631	-0.038	0.091	22.042
	cons	-2.366	0.190	0.220	34.762	-2.983	0.196	0.247	41.332
UQR-Q50	x1	1.169	-0.039	0.079	15.149	0.956	-0.011	0.082	6.445
	x2	0.965	0.004	0.053	13.696	0.921	-0.011	0.065	5.756
	cons	-0.218	-0.023	0.071	13.141	0.183	0.006	0.092	6.837
UQR-Q90	x1	0.018	0.006	0.095	3.929	0.000	0.004	0.081	16.365
	x2	1.430	0.006	0.120	10.946	1.481	0.012	0.081	21.996
	cons	2.389	-0.002	0.128	6.438	2.313	-0.009	0.088	16.681

Note: Monte Carlo Simulation Results. True coefficients represent the average quantile coefficients based on uncensored data. Bias is the average difference of the coefficients using uncensored data and Multiple imputed (MI) data. MAE ratio represents the average Mean absolute error (MAE) ratio between MI data and the uncensored data. StErr ratio represents the average coefficients standard error ratio between MI data and uncensored data. CQR: Conditional Quantile Regression; UQR: Unconditional Quantile Regression.

Table A2 Monte Carlo Simulation: N=500, exponential Heteroskedasticity

$y = x\beta + u * e^{\gamma x}$		u~normal				u~logistic			
		E($\hat{\beta}_f$)	Bias	MAE	StErr Ratio	E($\hat{\beta}_f$)	Bias	MAE	StErr Ratio
CQR-Q10	x1	1.643	0.000	0.065	17.224	1.608	0.008	0.071	15.862
	x2	0.749	-0.004	0.060	17.363	0.762	-0.003	0.059	17.541
	cons	-1.287	0.004	0.078	16.935	-1.219	0.013	0.081	14.726
CQR-Q50	x1	1.001	-0.001	0.048	10.497	0.999	0.000	0.044	14.579
	x2	1.002	0.002	0.039	10.339	1.002	0.005	0.038	14.306
	cons	-0.003	-0.001	0.056	10.647	-0.002	-0.006	0.051	14.954
CQR-Q90	x1	0.363	0.000	0.069	17.349	0.392	-0.004	0.073	16.621
	x2	1.254	0.004	0.057	16.519	1.238	0.004	0.057	16.803
	cons	1.278	-0.003	0.074	16.502	1.217	-0.015	0.081	16.015
UQR-Q10	x1	1.596	0.001	0.110	20.406	1.466	-0.022	0.104	20.321
	x2	0.581	-0.001	0.057	9.942	0.564	-0.006	0.052	9.387
	cons	-1.527	0.002	0.126	23.503	-1.388	0.022	0.115	21.728
UQR-Q50	x1	1.020	-0.004	0.067	23.466	1.041	0.016	0.067	26.614
	x2	0.870	0.002	0.044	17.169	0.870	0.010	0.044	19.589
	cons	0.143	-0.002	0.062	18.832	0.125	-0.020	0.063	20.646
UQR-Q90	x1	0.424	-0.002	0.057	8.705	0.436	0.006	0.057	6.172
	x2	1.598	-0.005	0.107	40.036	1.604	0.021	0.111	36.694
	cons	1.239	0.006	0.113	26.234	1.206	-0.025	0.118	23.795
$y = x\beta + u * e^{\gamma x}$		u~Chi2				u~uniform			
		E($\hat{\beta}_f$)	Bias	MAE	StErr Ratio	E($\hat{\beta}_f$)	Bias	MAE	StErr Ratio
CQR-Q10	x1	1.535	0.029	0.056	42.223	1.692	0.011	0.061	25.847
	x2	0.788	-0.018	0.044	40.444	0.728	-0.046	0.063	21.116
	cons	-1.071	0.049	0.069	40.039	-1.383	0.011	0.071	27.536
CQR-Q50	x1	1.101	-0.022	0.049	13.874	0.997	0.000	0.058	5.218
	x2	0.958	0.010	0.037	13.895	0.998	-0.006	0.047	5.024
	cons	-0.202	-0.055	0.068	14.043	0.003	0.008	0.066	5.155
CQR-Q90	x1	0.329	0.003	0.085	6.245	0.307	-0.011	0.058	21.818
	x2	1.261	0.006	0.070	5.772	1.271	0.014	0.048	20.132
	cons	1.346	0.015	0.095	6.053	1.385	0.037	0.070	21.289
UQR-Q10	x1	1.288	-0.054	0.084	20.882	1.724	0.103	0.124	8.051
	x2	0.654	0.011	0.049	14.901	0.671	0.073	0.091	4.870
	cons	-1.427	0.027	0.097	28.814	-1.836	-0.175	0.197	15.846
UQR-Q50	x1	1.118	-0.038	0.074	27.610	0.928	-0.072	0.091	17.477
	x2	0.870	0.027	0.050	21.249	0.846	-0.043	0.059	13.263
	cons	0.003	-0.055	0.074	21.657	0.234	0.076	0.094	16.551
UQR-Q90	x1	0.373	0.016	0.067	1.116	0.421	-0.005	0.055	13.190
	x2	1.572	0.053	0.130	25.284	1.606	-0.031	0.108	47.061
	cons	1.325	-0.049	0.133	13.514	1.241	0.032	0.112	30.753

Note: Monte Carlo Simulation Results. True coefficients represent the average quantile coefficients based on uncensored data. Bias is the average difference of the coefficients using uncensored data and Multiple imputed (MI) data. MAE ratio represents the average Mean absolute error (MAE) ratio between MI data and the uncensored data. StErr ratio represents the average coefficients standard error ratio between MI data and uncensored data. CQR: Conditional Quantile Regression; UQR: Unconditional Quantile Regression.

Table A3 Monte Carlo Simulation: N=500, Varying coefficient structure

$y = x\beta(t)$		Type 1				Type 2			
		$E(\hat{\beta}_f)$	Bias	MAE	StErr Ratio	$E(\hat{\beta}_f)$	Bias	MAE	StErr Ratio
CQR-Q10	x1	-1.133	-0.002	0.090	10.978	-0.078	-0.010	0.084	13.873
	x2	-0.036	-0.008	0.079	11.914	-0.078	-0.009	0.077	14.626
	cons	0.359	0.009	0.086	16.881	-0.094	0.041	0.087	15.181
CQR-Q50	x1	0.401	0.001	0.060	6.352	0.852	0.009	0.071	5.172
	x2	0.598	0.000	0.049	5.893	0.848	0.004	0.059	4.186
	cons	1.002	-0.001	0.054	7.040	0.847	-0.028	0.070	5.117
CQR-Q90	x1	1.933	-0.001	0.089	10.035	2.296	-0.001	0.133	4.750
	x2	1.234	0.005	0.072	9.891	2.276	0.010	0.152	6.726
	cons	1.649	-0.004	0.076	13.577	2.305	0.001	0.151	8.629
UQR-Q10	x1	-1.183	-0.003	0.108	15.229	-0.085	-0.013	0.080	16.586
	x2	-0.045	-0.003	0.057	7.473	-0.077	-0.014	0.059	14.851
	cons	0.471	0.005	0.071	6.551	-0.077	0.053	0.091	15.220
UQR-Q50	x1	0.419	0.000	0.066	12.220	0.914	0.005	0.053	3.745
	x2	0.542	0.003	0.050	10.874	0.746	0.003	0.040	3.049
	cons	0.892	-0.004	0.069	11.825	0.742	-0.015	0.054	3.150
UQR-Q90	x1	1.953	-0.003	0.135	12.268	2.215	0.000	0.163	4.969
	x2	1.291	-0.002	0.112	13.027	2.305	0.003	0.152	7.560
	cons	1.800	0.004	0.139	11.212	2.517	0.005	0.174	4.880

Note: Monte Carlo Simulation Results. True coefficients represent the average quantile coefficients based on uncensored data. Bias is the average difference of the coefficients using uncensored data and Multiple imputed (MI) data. MAE ratio represents the average Mean absolute error (MAE) ratio between MI data and the uncensored data. StErr ratio represents the average coefficients standard error ratio between MI data and uncensored data. CQR: Conditional Quantile Regression; UQR: Unconditional Quantile Regression.

Table A4. Monte Carlo Simulation: N=2000, Linear Heteroskedasticity

$y = x\beta + u * \gamma x$		u~normal				u~logistic			
		$E(\hat{\beta}_f)$	Bias	MAE	StErr Ratio	$E(\hat{\beta}_f)$	Bias	MAE	StErr Ratio
CQR-Q10	x1	2.012	-0.001	0.070	22.780	1.956	0.032	0.073	14.003
	x2	0.797	0.000	0.043	14.959	0.813	0.007	0.045	12.110
	cons	-2.379	0.002	0.090	28.314	-2.253	-0.025	0.090	18.776
CQR-Q50	x1	1.000	-0.001	0.031	7.565	1.000	0.001	0.030	9.549
	x2	1.000	0.000	0.025	7.578	1.001	0.001	0.024	9.869
	cons	-0.001	0.000	0.036	7.606	-0.001	-0.002	0.033	9.613
CQR-Q90	x1	-0.013	-0.001	0.044	9.852	0.040	-0.008	0.048	9.173
	x2	1.204	0.001	0.038	10.723	1.193	0.002	0.039	9.938
	cons	2.380	0.002	0.052	9.576	2.251	-0.008	0.054	8.823
UQR-Q10	x1	2.102	-0.004	0.077	17.398	1.928	0.018	0.071	12.349
	x2	0.611	-0.001	0.033	5.883	0.615	0.010	0.038	4.991
	cons	-2.537	0.004	0.068	12.837	-2.357	-0.014	0.072	10.949
UQR-Q50	x1	0.997	-0.002	0.040	13.649	1.014	0.010	0.040	16.228
	x2	0.918	0.000	0.028	11.177	0.914	0.003	0.027	13.060
	cons	0.144	0.001	0.040	12.948	0.134	-0.006	0.039	14.705
UQR-Q90	x1	0.051	-0.001	0.048	10.379	0.104	-0.006	0.055	11.012
	x2	1.478	0.001	0.055	20.333	1.502	0.002	0.062	21.991
	cons	2.249	0.000	0.061	14.151	2.126	-0.008	0.067	14.438
$y = x\beta + u * \gamma x$		u~Chi2				u~uniform			
		$E(\hat{\beta}_f)$	Bias	MAE	StErr Ratio	$E(\hat{\beta}_f)$	Bias	MAE	StErr Ratio
CQR-Q10	x1	1.846	-0.315	0.315	102.805	2.097	-0.254	0.254	79.038
	x2	0.831	-0.120	0.120	41.393	0.785	-0.053	0.059	34.323
	cons	-1.990	0.462	0.462	115.185	-2.576	0.261	0.262	85.905
CQR-Q50	x1	1.162	-0.020	0.033	9.024	0.997	-0.001	0.037	3.781
	x2	0.967	0.000	0.022	8.560	1.001	-0.001	0.030	3.536
	cons	-0.379	-0.024	0.036	8.650	0.002	0.002	0.041	3.955
CQR-Q90	x1	-0.059	0.011	0.056	4.530	-0.099	0.010	0.037	17.994
	x2	1.207	0.003	0.045	4.106	1.217	0.000	0.029	17.804
	cons	2.497	0.003	0.061	4.204	2.577	0.042	0.053	17.678
UQR-Q10	x1	1.849	-0.168	0.170	28.163	2.605	-0.208	0.233	50.409
	x2	0.683	-0.073	0.076	24.098	0.672	-0.018	0.047	17.640
	cons	-2.374	0.157	0.161	31.806	-3.104	0.135	0.165	41.542
UQR-Q50	x1	1.141	-0.046	0.054	17.116	0.941	-0.013	0.043	6.530
	x2	0.932	0.000	0.027	15.086	0.924	-0.007	0.034	6.317
	cons	-0.174	-0.017	0.037	14.281	0.190	0.003	0.047	7.732
UQR-Q90	x1	0.010	0.005	0.047	3.471	-0.006	0.003	0.042	16.880
	x2	1.463	-0.001	0.067	12.651	1.474	0.011	0.045	25.062
	cons	2.362	0.003	0.072	8.312	2.326	-0.008	0.048	19.042

Note: Monte Carlo Simulation Results. True coefficients represent the average quantile coefficients based on uncensored data. Bias is the average difference of the coefficients using uncensored data and Multiple imputed (MI) data. MAE ratio represents the average Mean absolute error (MAE) ratio between MI data and the uncensored data. StErr ratio represents the average coefficients standard error ratio between MI data and uncensored data. CQR: Conditional Quantile Regression; UQR: Unconditional Quantile Regression.

Table A5 Monte Carlo Simulation: N=2000, exponential Heteroskedasticity

$y = x\beta + u\ e^{\gamma x}$		u~normal				u~logistic			
		E($\hat{\beta}_f$)	Bias	MAE	StErr Ratio	E($\hat{\beta}_f$)	Bias	MAE	StErr Ratio
CQR-Q10	x1	1.639	0.000	0.033	16.242	1.603	0.009	0.037	13.033
	x2	0.744	-0.004	0.028	16.854	0.758	-0.002	0.030	14.981
	cons	-1.280	0.004	0.037	15.255	-1.210	0.013	0.042	11.389
CQR-Q50	x1	0.999	-0.001	0.024	12.710	1.001	0.000	0.022	16.966
	x2	1.000	0.000	0.020	12.194	0.999	0.003	0.019	17.030
	cons	0.000	0.001	0.028	12.898	0.000	-0.004	0.026	16.930
CQR-Q90	x1	0.363	0.002	0.034	16.461	0.397	-0.006	0.037	14.246
	x2	1.256	0.002	0.028	16.068	1.241	0.002	0.029	14.168
	cons	1.279	-0.002	0.038	15.993	1.211	-0.016	0.042	13.960
UQR-Q10	x1	1.621	0.001	0.070	31.593	1.485	-0.022	0.068	31.361
	x2	0.587	0.000	0.033	12.283	0.570	-0.006	0.029	11.513
	cons	-1.542	0.001	0.076	32.641	-1.401	0.020	0.070	30.055
UQR-Q50	x1	0.991	-0.002	0.035	25.811	1.014	0.016	0.036	28.899
	x2	0.843	0.001	0.024	19.445	0.842	0.010	0.024	21.943
	cons	0.184	-0.002	0.035	22.803	0.167	-0.019	0.037	24.307
UQR-Q90	x1	0.433	0.000	0.027	7.175	0.455	0.007	0.026	4.576
	x2	1.630	0.001	0.068	55.317	1.632	0.027	0.071	51.402
	cons	1.201	-0.001	0.073	37.912	1.167	-0.030	0.077	35.785
$y = x\beta + u * \gamma x$		u~Chi2				u~uniform			
		E($\hat{\beta}_f$)	Bias	MAE	StErr Ratio	E($\hat{\beta}_f$)	Bias	MAE	StErr Ratio
CQR-Q10	x1	1.536	0.030	0.035	54.139	1.693	0.014	0.031	39.009
	x2	0.786	-0.017	0.025	51.237	0.726	-0.044	0.046	33.306
	cons	-1.072	0.048	0.051	50.328	-1.386	0.005	0.034	42.038
CQR-Q50	x1	1.103	-0.024	0.030	15.691	1.001	0.001	0.028	3.822
	x2	0.957	0.008	0.019	16.548	1.001	-0.007	0.024	3.272
	cons	-0.203	-0.053	0.054	16.133	-0.002	0.008	0.032	3.817
CQR-Q90	x1	0.332	0.003	0.043	4.651	0.310	-0.014	0.031	33.775
	x2	1.265	0.003	0.034	2.820	1.276	0.012	0.025	31.715
	cons	1.341	0.013	0.047	4.467	1.383	0.040	0.047	33.290
UQR-Q10	x1	1.266	-0.081	0.085	32.567	1.735	0.141	0.142	10.636
	x2	0.643	-0.002	0.028	21.053	0.675	0.091	0.091	3.700
	cons	-1.406	0.052	0.072	41.204	-1.850	-0.216	0.216	18.086
UQR-Q50	x1	1.082	-0.036	0.046	29.848	0.916	-0.070	0.072	19.639
	x2	0.844	0.030	0.034	24.071	0.831	-0.041	0.043	15.592
	cons	0.046	-0.059	0.062	26.353	0.256	0.074	0.077	20.025
UQR-Q90	x1	0.388	0.023	0.038	-0.905	0.437	-0.012	0.029	13.537
	x2	1.606	0.081	0.103	34.782	1.647	-0.047	0.078	68.154
	cons	1.282	-0.083	0.106	19.940	1.194	0.049	0.082	47.202

Note: Monte Carlo Simulation Results. True coefficients represent the average quantile coefficients based on uncensored data. Bias is the average difference of the coefficients using uncensored data and Multiple imputed (MI) data. MAE ratio represents the average Mean absolute error (MAE) ratio between MI data and the uncensored data. StErr ratio represents the average coefficients standard error ratio between MI data and uncensored data. CQR: Conditional Quantile Regression; UQR: Unconditional Quantile Regression.

Table A6 Monte Carlo Simulation: N=2000, Varying coefficient structure

$y = x\beta(t)$		Type 1				Type 2			
		$E(\hat{\beta}_f)$	Bias	MAE	StErr Ratio	$E(\hat{\beta}_f)$	Bias	MAE	StErr Ratio
CQR-Q10	x1	-1.131	-0.003	0.044	10.117	-0.085	-0.010	0.044	16.671
	x2	-0.041	-0.010	0.039	11.638	-0.090	-0.011	0.040	15.811
	cons	0.358	0.012	0.045	18.744	-0.087	0.043	0.054	18.407
CQR-Q50	x1	0.400	0.000	0.030	7.158	0.849	0.009	0.036	5.888
	x2	0.599	0.002	0.025	7.239	0.845	0.006	0.029	4.707
	cons	0.999	-0.004	0.028	9.754	0.850	-0.030	0.041	6.545
CQR-Q90	x1	1.937	0.000	0.044	9.752	2.294	0.007	0.068	4.131
	x2	1.240	0.006	0.036	9.402	2.293	0.015	0.081	4.712
	cons	1.640	-0.005	0.037	13.364	2.291	-0.006	0.077	6.313
UQR-Q10	x1	-1.219	0.000	0.066	19.397	-0.089	-0.017	0.044	17.878
	x2	-0.049	-0.002	0.030	6.867	-0.079	-0.014	0.032	15.935
	cons	0.492	0.003	0.039	6.731	-0.076	0.057	0.065	16.610
UQR-Q50	x1	0.410	0.003	0.031	12.352	0.897	0.007	0.025	2.706
	x2	0.528	0.006	0.025	11.565	0.733	0.006	0.019	2.229
	cons	0.908	-0.010	0.037	12.949	0.765	-0.016	0.029	2.666
UQR-Q90	x1	1.999	-0.002	0.087	18.978	2.249	0.002	0.082	5.155
	x2	1.314	-0.002	0.062	15.866	2.348	0.003	0.082	8.132
	cons	1.747	0.003	0.093	19.129	2.458	0.002	0.104	7.188

Note: Monte Carlo Simulation Results. True coefficients represent the average quantile coefficients based on uncensored data. Bias is the average difference of the coefficients using uncensored data and Multiple imputed (MI) data. MAE ratio represents the average Mean absolute error (MAE) ratio between MI data and the uncensored data. StErr ratio represents the average coefficients standard error ratio between MI data and uncensored data. CQR: Conditional Quantile Regression; UQR: Unconditional Quantile Regression.

Figure A1 Average Monthly Earnings by Year and Gender: IPW

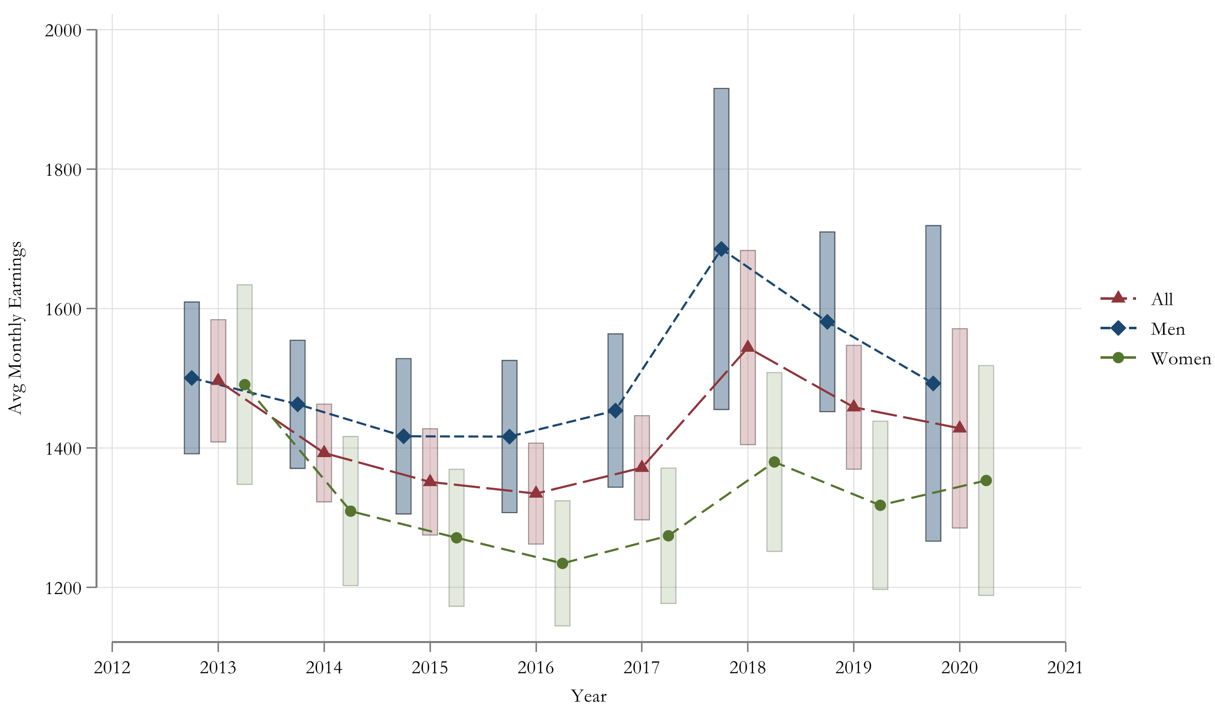


Figure A2 Selected Quantiles and Gini Coefficient across Years: IPW