Recovering income distribution in the presence of interval-censored data

Gustavo Canavire-Bacarreza[1] Fernando Rios-Avila[2]

Flavia Sacco-Capurro[3]

Abstract

We propose a novel method to analyze interval-censored data. The modeling framework follows a multiple imputation approach based on a Heteroskedastic Interval regression. The proposed model aims to obtain synthetic datasets that can be used to implement standard regression analysis. We present three 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 conditional normality. Second, we analyze income inequality using the Current Population Survey - Annual Social Economics Supplement, comparing estimates for imputed and observed data when the assumption of normality might not hold. Finally, we apply 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 all three exercises.

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

JEL Codes: C150, C340, J3

[1] The World Bank, gcanavire@worldbak.org
[2] Levy Institute at Bard College, f.rios.a@gmail.com
[3] The World Bank, fsaccocapurro@worldbank.org

1. Introduction

Labor force household surveys are a useful source to understand employment dynamics in both developing and developed countries. These surveys provide vast information of the labor market at higher frequency levels (in comparison with living condition surveys) and, in some cases, are the only source of information to describe and examine the status and structure of the labor market. In fact, in the Latin American and the Caribbean region, the OECS countries, together with other countries like Bolivia, Costa Rica, Ecuador, Jamaica, Mexico, Peru, and Uruguay, all collect their labor force surveys quarterly as oppose as 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 the wage and salaries of workers, which allow, in some cases, to estimate official poverty and inequality measures. However, many countries cannot retrieve the full income distribution due to how the collection and reporting of earned income are carried out (in brackets). This is the case of the labor force survey for all countries in the Organization of Eastern Caribbean States (OECS), for which labor income is collected in brackets to create a sense of anonymity of the data. Colombia, Germany, Australia, and New Zealand have similar data collection protocols for their micro census, along with (Walter and Weimer 2018). Moreover, the same issue arises with health and mortality data in many Demographic and Health Surveys (DHS), such as the one in Nigeria, for which, due to some practical restrictions, the survival time of the children are only recorded in months or years, which makes it an interval-censored data structure (Chen and Zhao 2021).

One argument in favor of these types of survey questions is that when respondents are asked to report amounts, these are subject to high rates of non-response in the variable of interest, in this case, income (Wang et al., 2013). In the survey literature, it is well-known that questions about income are considered "sensitive"; therefore, the non-response rate is considerably higher for these questions (Moore et al., 2000; Hagenaars and Vos, 1988). Field tests conducted in the past have shown that asking follow-up income questions in a series of unfolding brackets achieve 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

and Ekstrom 2018, Yan et al. 2018). However, even though this form of data collection solves the problem of underreporting or miss reporting, it implies a problem for recovering the full wage (income) distribution that is useful to study poverty and inequality, which are calculated in most cases through an income aggregate.

Different approaches exist to recover the full distribution and estimate indicators from interval-censored data; most rely on parametric methods. In this sense, Chen (2017) provides a generalized approach to multinomial maximum likelihood estimation for several types of grouped data and shows its consistency through complementary simulation results. Chih-Yuan et al. (2021) rely on regression analysis for quantile functions where the quantile regression coefficients are treated as functions over a continuum of quantile levels. With this method, they propose a general inference procedure for quantile regression coefficient functions with interval-censored outcome data using a survey on monthly salaries of Taiwan workers, where only parts of the salary data are exact. In contrast, the others are interval-censored according to the salary intervals prespecified in the survey questionnaire. Along this line, Zhou et al. (2017) propose an estimation method for quantile regression models with interval-censored data considering asymptotic normality's property and two bias correction methods.

Other studies, like the one proposed by Han et al. (2020), construct new measures of the income distribution and estimate poverty in the U.S. with a lag of only a few weeks using high-frequency data from the Basic Monthly Current Population Survey (CPS), to understand the impact of the Covid-19 pandemic. A similar case is proposed by Parolin and Wimer (2020), who produce monthly updates of SPM poverty rates with demographic data from the monthly Current Population Survey (CPS) and data on SPM poverty from the previous CPS ASEC. Using ten years of prior data, validation tests demonstrate that this methodology estimates poverty rates that closely track observed poverty rates released nearly ten months later. 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.

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

multiply imputed using draws from a flexible parametric model fitted to the censored distribution, yielding partially synthetic data. In order to analyze wages in the German IAB employment survey and solve the problem of censored wages, Buutner and Rassler (2008) derive new multiple imputation approaches to impute the censored wages by draws of a random variable from a truncated distribution based on Markov chain Monte Carlo techniques. Moreover, using data from the German micro census, which also reports income in brackets, Walter and Weimer (2018) propose an iterative kernel density algorithm that generates metric pseudo samples from the interval-censored income variable to estimate poverty and inequality indicators. However, most of the studies found in the literature focus on estimation of and inference about mean incomes and income regressions for a single year rather than estimates of income inequality and trends.

To address this issue, we propose an interval imputation method to retrieve the income distribution of the labor income variable using a Labor Force Survey in Grenada and the US CPS data for sensitivity analyses. The method relies on an interval regression, often known as a generalization of the censored regression estimators. We model the probability that a person's income be within the underlying income brackets with this method. Once we get the full sample of imputed data, we can then analyze the trends in labor income, the evolution of wages in a given country and perform standard inequality estimates.

The paper is organized as follows. Section 2 introduces the model and the econometric issues associated with the imputation method; Section 3 provides three exercises to assess the performance of the model; a Monte Carlo simulation using the Swiss Labor Market Survey of 1998; an analysis of income inequality using the Current Population Survey - Annual Social Economics Supplement of 2020 and finally, an analysis of labor income trends and income inequality in Grenada using the 2013-2020 series of the Labor Force Survey. Section 4 concludes.

2. Methodology

To address the problem of interval-censored data, we propose a multiple imputation approach based on a Heteroskedastic Interval regression model. An interval regression model is a

generalization of the Tobit model that allows using 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. Predicted earnings can be obtained using the estimated parameters based on random draws of the estimated conditional distributions.

2.1. Interval regression model

Assume that (log) earned income (y_i) has a data generating process such that:

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

v_i is a homoscedastic 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. Following Machado and 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. Under the assumption that v_i follows a standard normal distribution, $y_i|x_i$ is also normally distributed with mean $\mu(x_i)$ and standard deviation $\sigma(x_i)$.

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

And equation one can be estimated via maximum likelihood by maximizing the following function:

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

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

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 P or Inequality I, 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 and Trivedi (2010) ch 16 for a discussion of censored regressions), where data can be a mixture of censored-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 interval regression approach focuses on modeling the probability that an individual i reports income to be within the underlying income brackets:

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

Using the data generating process defined by equation 4, 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:

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

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

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

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's implications on the Imputation strategy are related to the appropriate draw of the imputed error.

Consider the d.g.p again. 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:

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

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:

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

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:

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

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.[4] To account for the uncertainty of the regression estimation, we obtain random draws from the following joint normal distribution:

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

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 distribution n degrees of freedom. Finally, the imputation for y_i^* will be given by:

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

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

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

In summary, the imputation algotithm is as follows:

1. Estimate the parameters associated with $\mu(x)$ and $\sigma(x)$ using a heteroskedastic interval regression approach via maximum likelihood.

[4] 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.

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 $\tilde{\tilde{y}}_i$, 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 obtains M sets of imputed samples.

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

2.3. Model estimation and inference

Once all 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 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 data with 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:

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

4. Applications

To show the performance of the proposed method, this section presents three applications illustrating the methods proposed in section 3.*4.1. A Monte Carlo Simulation*

The first application is designed to study the performance of the proposed methodology under the assumption that the d.g.p. follows a conditionally normal distribution. To capture the kind of process we may be expected to see when using real data, the simulation is constructed using an Excerpt from the Swiss Labor Market Survey 1998. This is a small dataset, 1434 observations with declared wages, readily available online (Jann, 2003). Simulated data is obtained using the following procedure.

i. Estimate a Heteroskedastic linear regression model for wages, where the conditional mean and log variance are modeled as linear functions of age, education, experience, tenure, gender, and marital status, including all their interactions. Using this, obtain point predictions for the conditional mean $\mu(x)$ and conditional log variance $ln\sigma 2(x)$.
ii. Using the predicted conditional mean and log variance, obtain draws for simulated wages from a random log-normal distribution. $\widehat{l\omega_i} \sim N\left(\hat{\mu}(x), \hat{\sigma}^2(x)\right), w_i = \exp(l\omega_i)$. This provides a dataset with simulated wages.

Once simulated wages w_i are obtained, interval censored wages are created using the following rules:

Table 1 Interval Censoring

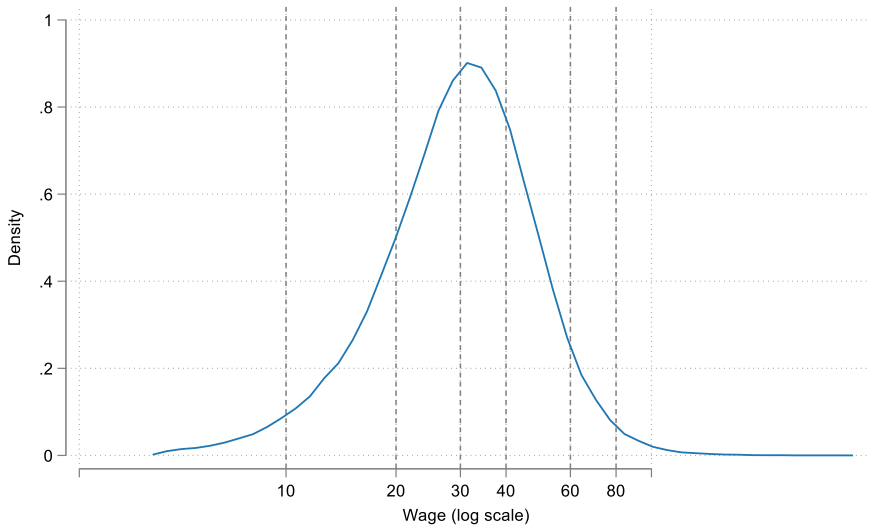
Wage Group	Lower Bound	Upper Bound	Proportion
1	0	10	3.3
2	10	20	17.1
3	20	30	28.8
4	30	40	25.0
5	40	60	20.6
6	60	80	4.0
7	80	.	1.3

Note: The last column shows the approximate distribution of observations that fall within each wage bracket.

Where Group 7 has no upper limit wage.

We repeat steps 2 and 3 to obtain 5000 independent samples. Figure 1 shows the distribution of the simulated wages (log scale) and the cuts used for the income brackets:

Figure 1 Simulated Wage distribution



We apply the proposed imputation procedure for each simulated dataset, using the log of the upper and lower bounds for each bracket, and obtain ten simulated wages. This leaves the lower bound for group 1 to be unbounded. Rather than using the same set of explanatory variables used in the wage simulation, we use a simplified specification where the $\mu(x)$ and $\ln \sigma(x)$ are modeled as a function of education, experience, tenure, gender, age, and age squared. We do this because this is a common specification used in applied labor applications. Furthermore, it simulates the scenario where the imputation model specification is misspecified because we never observe true d.g.p.[5]

To test the performance of the procedure, we estimate selected distributional coefficients using both the fully observed data and the imputed data using our proposed procedure. Population parameters are obtained by using pooled data from all 5000 simulated samples. The results are provided in table 2.

Table 2 Simulation Results: Selected Distributional Statistics

	Population	**Full Observed Data**

[5] The estimation of the Heteroskedastic Interval regression model is done using Stata's -intreg-. The imputation procedure is done using the command intreg_mi.ado, available upon request.

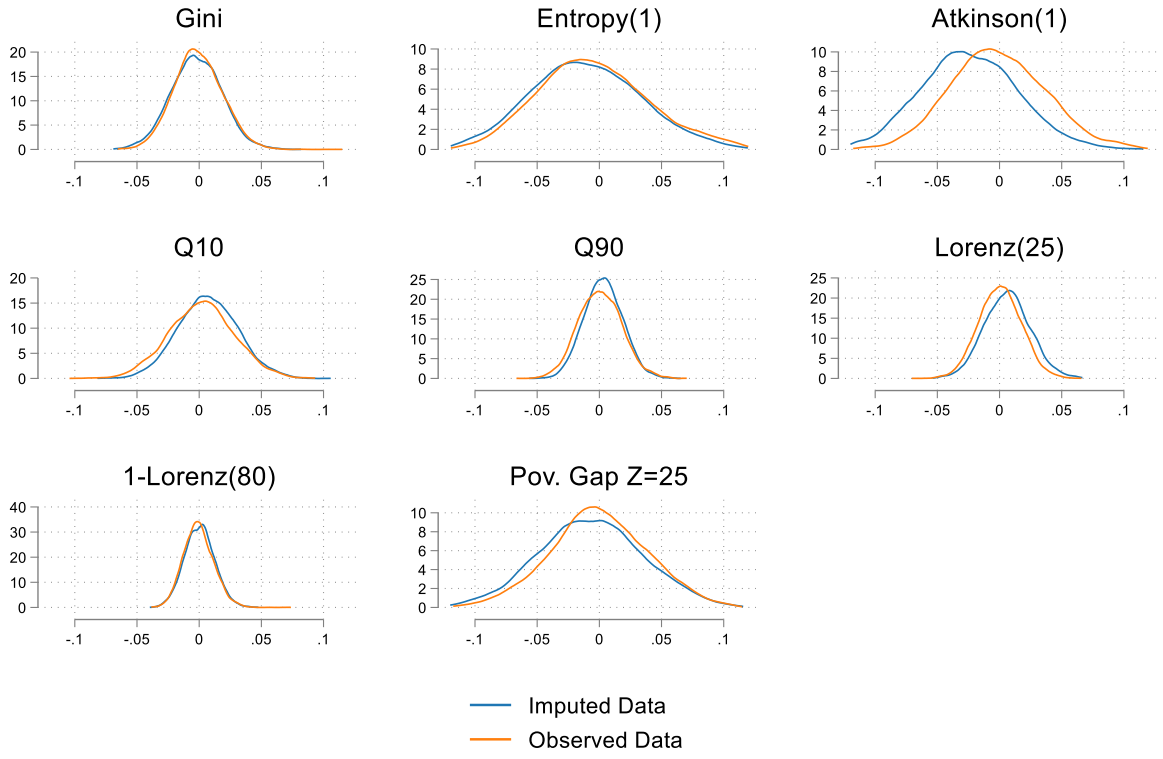
	Parameters	Mean	RMSE	Rel Bias	p2.5	p97.5
Gini	0.259	-0.0002	0.0051	-0.1%	-3.8%	3.9%
Entropy(1)	0.112	0.0000	0.0074	0.0%	-8.5%	11.2%
Atkinson(1)	0.114	-0.0001	0.0046	-0.1%	-7.4%	8.2%
10^{th} Quantile	15.094	0.0283	0.4011	0.2%	-5.0%	5.5%
90^{th} Quantile	52.017	0.0330	0.9061	0.1%	-3.3%	3.6%
Lorenz Ordinate 25p	0.120	0.0001	0.0021	0.1%	-3.3%	3.4%
1-Lor Ordinate 80	0.347	-0.0001	0.0043	0.0%	-2.3%	2.5%
Pov Gap Z=25	0.100	0.0000	0.0039	0.0%	-7.9%	7.7%
	Population			**Imputed Data**		
	Parameters	**Bias**	**RMSE**	**Rel Bias**	**p2.5**	**p97.5**
Gini	0.259	-0.0005	0.0053	-0.2%	-4.2%	3.8%
Entropy(1)	0.112	-0.0011	0.0053	-1.0%	-9.7%	8.6%
Atkinson(1)	0.114	-0.0028	0.0053	-2.5%	-10.3%	5.6%
10^{th} Quantile	15.094	0.1068	0.3784	0.7%	-4.0%	5.6%
90^{th} Quantile	52.017	0.1691	0.8248	0.3%	-2.7%	3.4%
Lorenz Ordinate 25p	0.120	0.0008	0.0023	0.6%	-2.9%	4.4%
1-Lor Ordinate 80	0.347	0.0001	0.0042	0.0%	-2.3%	2.5%
Pov Gap Z=25	0.100	-0.0007	0.0043	-0.7%	-9.2%	7.8%

In general, the distributional statistics estimated with the interval censored-imputed data seem to closely reproduce the population parameters, albeit with a larger bias than the estimates that use observed data. The expected bias is small and close to zero, with the largest relative bias observed for the Atkinson and Entropy Indices. Interestingly, the root means squared error (RMSE) for the imputed and observed data are similar, with no clear advantage between imputed and fully observed data, despite the bias.

To compare the magnitude of the bias, we also include statistics of the average, the 2.5%, and 97.5% percentiles of the relative bias for the parameters based on imputed and fully observed data. While this also shows that imputed data suffers from a small bias, particularly for the Atkinson index, the distribution of such bias is comparable across both surveys.

For a visual inspection of the bias distribution, Figure 2 provides densities plots for the distribution of relative bias for all statistics across both samples. They further confirm that the Atkinson index is estimated with the largest bias, but the bias distribution appears symmetric for all other statistics, with a small bias for the Lorenz ordinate and the poverty gap.

Figure 2 Distribution of Relative Bias



Since the analysis of this type of data usually requires an estimation of different models, we also estimate a few selected models with a limited set of explanatory variables, comparing the results using imputed and fully observed data. Because we use a miss-specified, the underlying assumption is that estimated parameters based on pooled observed data to be the truth. Table 3 summarizes these results. We compare the results for linear regression and quantile regressions at the 20th and 80th conditional quantiles.

Table 3 Regression analysis: Imputed vs. Observed Data

	Population Parameter	Observed E(Bias)	Observed RMSE	Imputed E(Bias)	Imputed RMSE
Linear Regression					
Education	1.8341	-0.0048	0.2134	-0.0066	0.2042
Experience	-0.1922	-0.0020	0.0650	-0.0025	0.0646
Tenure	0.1473	-0.0004	0.0593	-0.0019	0.0618
Female=1	-3.5227	-0.0049	0.7467	0.0037	0.7723
Age	1.9799	-0.0042	0.2698	-0.0099	0.2681

Age^2	-0.0189	0.0001	0.0034	0.0002	0.0035
Constant	-31.6772	0.1550	5.1524	0.2642	4.9147
Quantile Regression 20th					
Education	1.4619	0.0018	0.1693	-0.0328	0.1596
Experience	0.0395	-0.0001	0.0592	-0.0108	0.0569
Tenure	0.1440	-0.0016	0.0540	0.0014	0.0504
Female=1	-5.5125	-0.0007	0.6559	-0.0845	0.6222
Age	2.1012	-0.0010	0.2012	-0.0181	0.1914
Age^2	-0.0241	0.0000	0.0025	0.0004	0.0024
Constant	-36.2430	-0.0060	3.5739	0.6063	3.4892
Quantile Regression 80th					
Education	2.3336	-0.0157	0.3081	-0.0434	0.2904
Experience	-0.3333	-0.0033	0.0944	-0.0042	0.0889
Tenure	0.1307	0.0016	0.0948	-0.0018	0.0892
Female=1	-1.5361	-0.0080	1.1368	0.0670	1.1052
Age	2.0218	-0.0099	0.3805	-0.0102	0.3584
Age^2	-0.0166	0.0001	0.0048	0.0001	0.0045
Constant	-32.6795	0.4011	6.7428	0.7731	6.1467

As expected, we observe negligible bias across estimated coefficients when using observed data. The bias, however, is somewhat larger when using our proposed imputation procedure, especially when looking at the quantile regression estimates. However, the imputed data has some disadvantages in terms of RMSE and often seems better than using fully observed data. This may be explained because the Monte Carlo simulation assumes that characteristics are kept fixed across all subsamples.

4.2. Analyzing Earning Income Inequality using Current Population Survey - Annual Social Economics Supplement (CPS-ASEC)

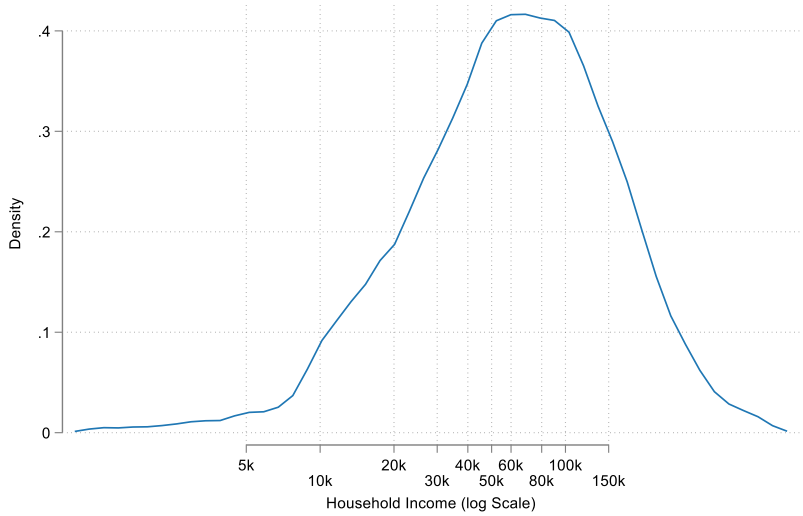
The CPS-ASEC is a monthly survey administered by the Bureau of Labor Statistics, and it is used to assess personal and labor market characteristics for people living in the U.S. In March of every year, approximately 60 thousand households are interviewed, with detailed information on income by source over the last fiscal year. In this exercise, we use the proposed methodology to analyze income inequality, comparing estimates for imputed and observed data estimates, assuming that total household income is available only in brackets, using the same thresholds that the CPS uses

to report household income in other months. In this exercise, we use sampling weights to estimate the interval regression model and use 25 imputed values per household.

In contrast with the Monte Carlo simulation exercise, we do not know the true d.g.p. which implies that the assumption of conditional (log) normality of household income may not hold, and the imputation and estimated models are likely to be miss-specified. In addition, we can only make one comparison between the imputed and observed statistics.

All imputations are done at the family level, excluding from the sample households with total income lower than 150$ per year. [6] For the interval regression approach, we use age, education, race, and job market status over the last year for the head of the household, family structure characteristics, and state-level dummies to model both the conditional mean and conditional variance. Figure 3 shows the distribution of log household income, as well as the thresholds used for the interval-censored data:

Figure 3 Distribution of Log Household Income



To assess the performance of the imputation strategy, similar to the previous section, we estimate various inequality statistics, which are provided in table 4. Overall, the statistics based on imputed

[6] This eliminates 1298 households from the sample.

values are very similar to those based on the observed data. The relative gaps are smaller than 5%, compared to the observed data statistics, across all statistics except for the Atkinson and Entropy indices, and closely followed by the Share of income held by the richest 10% of households, possibly overestimating inequality levels. Regarding the precision of the estimates, the standard errors are generally smaller when observed data is used, except for quantile statistics.

Table 4 Selected Summary Statistics

	Imputed	**Observed**	**Ratio**
Mean	43795.2	43480.9	1.007
	(210.6)	(198.1)	
10th Quantile	9393.4	9470.3	0.992
	(79.7)	(80.5)	
50th Quantile	30707.2	30982.2	0.991
	(147.9)	(154.6)	
90th Quantile	87429.0	89630.9	0.975
	(507.5)	(659.2)	
Gini Coefficient	0.4647	0.4534	1.025
	(0.0018)	(0.0016)	
Atkinson (1)	0.3373	0.3186	1.059
	(0.0021)	(0.0019)	
Entropy (1)	0.4015	0.3606	1.113
	(0.0048)	(0.0038)	
Lorenz (20)	0.0401	0.0411	0.976
	(0.0003)	(0.0003)	
1-Lorenz(90)	0.3432	0.3276	1.048
	(0.0020)	(0.0018)	

Note: Statistics correspond to total household income, weighted at the household level. Standard errors in parenthesis.

In addition to the comparison of unconditional distribution Statistics, we can also assess the performance of the imputation procedure comparing conditional distributions. Table 5 compares estimated models using both observed and imputed data, linear regression, and quantile regressions at the 10^{th} and 90^{th} quantiles.

Overall, the results are promising. Across most models, the ratios between imputed to observed models fall within 5% from each other. Across all three models, the largest divergences are observed for the Region coefficients, although the absolute magnitudes are negligible. Also, in contrast with the unconditional statistics, we observe that the standard errors are often larger when using imputed data than observed data.

Table 5 Selected Regression analysis comparison

Variable	Linear Regression			QREG 10th			QREG 90th		
	Observed	Imputed	Ratio	Observed	Imputed	Ratio	Observed	Imputed	Ratio
Age HH	0.012	0.011	0.96	0.015	0.014	0.96	0.011	0.011	1.01
	(0.0003)	(0.0003)		(0.0005)	(0.0006)		(0.0005)	(0.0006)	
Sex HH:Female=1	-0.187	-0.195	1.04	-0.260	-0.259	1.00	-0.141	-0.159	1.13
	(0.008)	(0.008)		(0.015)	(0.017)		(0.013)	(0.015)	
Race HH:white=1	0.196	0.195	0.99	0.253	0.248	0.98	0.162	0.163	1.00
	(0.011)	(0.011)		(0.021)	(0.022)		(0.015)	(0.021)	
Educ HH (Base LTHS)									
High School	0.404	0.395	0.98	0.406	0.404	0.99	0.348	0.374	1.08
	(0.018)	(0.017)		(0.028)	(0.034)		(0.030)	(0.030)	
Scoll	0.649	0.639	0.98	0.630	0.640	1.02	0.579	0.613	1.06
	(0.018)	(0.017)		(0.029)	(0.035)		(0.030)	(0.031)	
College+	1.115	1.102	0.99	1.071	1.069	1.00	1.091	1.120	1.03
	(0.017)	(0.017)		(0.027)	(0.033)		(0.030)	(0.031)	
HH Employed	0.559	0.546	0.98	0.756	0.776	1.03	0.355	0.363	1.02
	(0.011)	(0.010)		(0.021)	(0.025)		(0.015)	(0.019)	
Log (famsize)	-0.403	-0.412	1.02	-0.422	-0.425	1.01	-0.455	-0.444	0.98
	(0.008)	(0.008)		(0.013)	(0.015)		(0.014)	(0.017)	
Division (Base NorthEast)									
MidWest	-0.071	-0.073	1.03	-0.030	-0.025	0.84	-0.129	-0.138	1.08
	(0.014)	(0.014)		(0.025)	(0.027)		(0.020)	(0.027)	
South	-0.105	-0.108	1.04	-0.074	-0.072	0.97	-0.112	-0.135	1.20
	(0.012)	(0.012)		(0.022)	(0.025)		(0.019)	(0.021)	
West	-0.024	-0.025	1.02	-0.013	-0.006	0.45	-0.041	-0.049	1.19
	(0.013)	(0.013)		(0.023)	(0.026)		(0.020)	(0.023)	
Constant	8.921	8.999	1.01	7.729	7.752	1.00	10.096	10.101	1.00
	(0.030)	(0.029)		(0.054)	(0.059)		(0.046)	(0.051)	

4.3. Wage Inequality in Grenada

The final illustration focuses on an empirical application used in the Grenada Poverty Assessment of 2021 (forthcoming) to describe wage inequality trends in the country between 2013 and 2020 using the annual Labor Force Survey. This survey provides information on the labor market in the country and is the only source of information that can be used to describe the status of the labor market and the distribution of income in the country.

One major limitation of this survey, however, is the collection of earned income data. Compared to standard household surveys or labor force surveys in most developed countries, earned income recorded in the LFS in Grenada is available in brackets. Furthermore, there is a large proportion of the employed population who do not declare their income. Table 6 provides an overview of the earned income distribution across time.

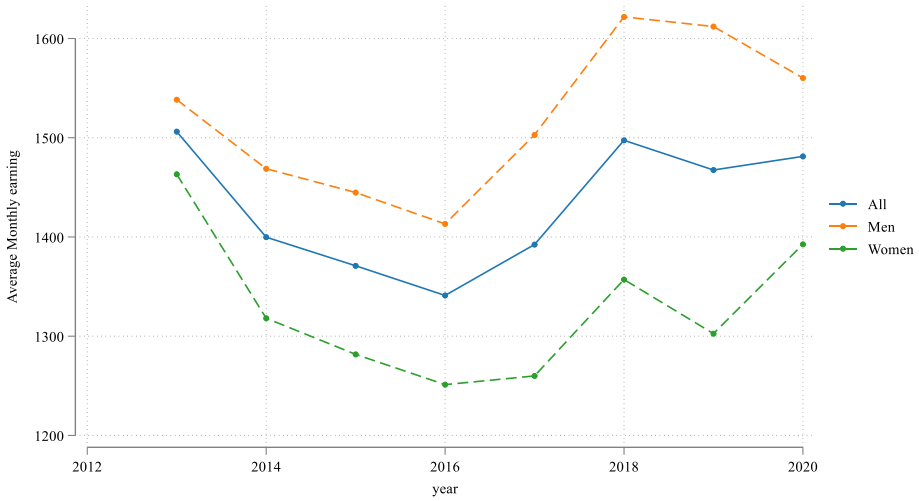
Table 6 Earned Income distribution by year

Year	2013	2014	2015	2016	2017	2018	2019	2020
>200	3.0	1.2	3.7	3.5	1.4	0.2	0.0	0.4
200-399	6.9	5.8	6.3	5.3	4.1	1.6	1.2	1.1
400-799	15.4	15.9	12.3	14.2	13.7	9.0	8.3	10.3
800-1199	19.1	20.0	18.3	18.7	21.1	20.4	23.8	24.6
1200-1999	17.7	17.4	13.9	13.1	18.4	14.7	14.9	15.9
2000-3999	15.6	11.3	11.2	11.5	10.5	9.7	12.8	11.8
4000-5999	2.6	2.4	2.4	2.2	2.2	1.6	1.2	2.1
6000+	2.0	1.2	0.6	0.6	0.7	1.0	1.0	0.5
Not stated	17.7	24.8	31.3	30.9	27.9	41.8	36.7	33.2

In this case, we face two types of problems. On the one hand, 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 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 and stated their income.

We make the simplifying assumption that not stating income is randomly distributed conditional on observed characteristics. To account for the fact that characteristics may differ across those who state and do 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 but assuming no lower and upper bounds for the imputed 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. In all cases, imputed earnings are adjusted by inflation.

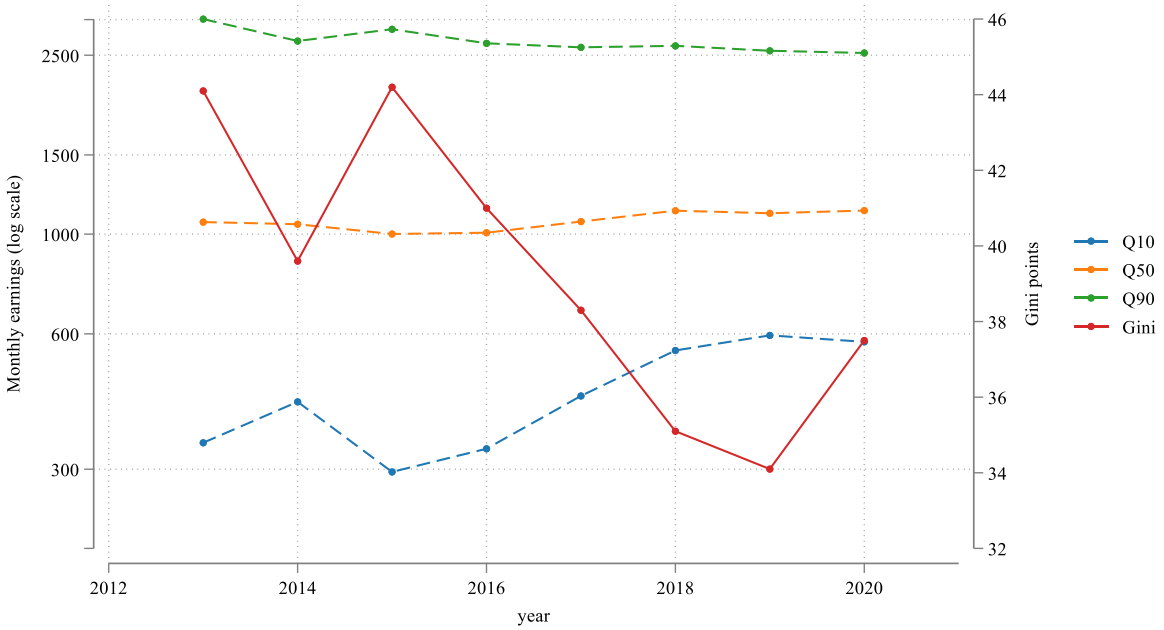
Figure 4 Average Monthly Earnings by Year and Gender



The results suggest that after a small decline in average real monthly earning 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.[7] The results also suggest that the gender earnings gap has shown a somewhat increasing trend between 2013 and 2019, although it predicted a small decline in 2020.

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

Figure 5 Selected Quantiles and Gini coefficient across Years



In terms of inequality, the estimates suggest that it has declined substantially across the years. The estimated Gini coefficient fell from 44.2 Gini points in 2015 to 34.1 in 2019, with a small 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.

4. Conclusion

A shred of vast evidence on methods used to deal with censored data in surveys is found in the literature. However, most of them are intended to estimate income or wages using the uncensored distribution of previous years, which is not always available with other data sources. In other studies, authors focus on estimation of and inference about mean incomes and income regressions for a single year rather than estimates of income inequality and trends, as we do with our analysis.

We present an imputation strategy that can be used to analyze interval-censored data from household surveys or labor market surveys. 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. We describe that when dealing with interval-censored data,

we have limited access to the observed distribution of the variable of interest, in contrast with standard multiple imputation analysis, where the variable of interest is fully unobserved. We propose a multiple imputation strategy using a heteroskedastic interval regression approach via maximum likelihood to overcome this. Once the imputed income has been obtained, statistical analysis can be done by independently implementing the desired model estimation across all imputed samples. To assess the consistency of our method, we perform simulations and analyze income trends and income inequality using three different sources of data.

The results of the simulation study can be summarized as follows:

The first application, using data from the Swiss Labor Market Survey of 1998, assesses the performance of the proposed methodology under the assumption that the d.g.p. follows a conditionally normal distribution. Simulated data is obtained using a Heteroskedastic linear regression model for wages, where the conditional mean and log variance are modeled as linear functions of observed characteristics. Using the predicted conditional mean and log variance, we obtain simulated wages from a random log-normal distribution providing a dataset of simulated wags; further, we create interval-censored wages with these simulated wages. In general, the distributional statistics estimated with the interval censored-imputed data seem to closely reproduce the population parameters, albeit with a larger bias than the estimates that use observed data.

For the second application, using the CPS-ASEC survey from the Bureau of Labor Statistics, we use sampling weights to estimate the interval regression model and use 25 imputed values per household. In this case, we do not know the true d.g.p. which implies that the assumption of conditional (log) normality of household income may not hold, and the imputation and estimated models are likely to be miss-specified. We estimate various inequality statistics to assess the performance of the imputation model, and overall, the statistics based on imputed values are very similar to those based on the observed data. Compared to the observed data statistics, the relative gaps are smaller than 5% across almost all statistics.

For the specific case of Grenadam 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, estimating the interval-censored regression for each year, with a set of household-level characteristics and job type characteristics. The results suggest that earned income inequality in this country has declined, which coincides with other economic performance indicators in the country.

The three applications we present based on different data sources and under different assumptions show that our multiple imputations approaches applied to interval-censored data can yield consistent income trends and income inequality for any country that collects income information in a censored way. The proposed model can also be applied to some datasets other than the ones we discuss in this paper and, at the same time, can be a good alternative to other imputation methods used to recover the income distribution from income intervals or top-coded data.

References

Angelov, A. G., & Ekström, M. (2018). Maximum likelihood estimation for survey data with informative interval censoring. *AStA Advances in Statistical Analysis*, *103*(2), 217-236.

Büttner, T., & Rässler, S. (2008). Multiple imputation of right-censored wages in the German IAB Employment Sample considering heteroscedasticity.

Cameron, A. Colin and Trivedi, Pravin K. (2010). Microeconometrics: Methods and Applications.Cambridgege University press.

Chen Y, Zhao Y (2021) Efficient sparse estimation on interval-censored data with approximated L0 norm: Application to child mortality. PLoS ONE 16(4): e0249359. https://doi.org/10.1371/journal.pone.0249359

Chih-Yuan, H., Chi-Chung, W. and Yi-Hau, C. (2021). Quantile function regression analysis for interval censored data, with application to salary survey data. Japanese Journal of Statistics and Data Science 72. DOI: 10.1007/s42081-021-00113-3

Demirtas, H., S. A. Freels, and R. M. Yucel. 2008. "Plausibility of Multivariate Normality Assumption When Multiply Imputing Non-Gaussian Continuous Outcomes: A Simulation Assessment." *Journal of Statistical Computation and Simulation* 78 (1): 69–84.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. 2014. Bayesian Data Analysis. 3rd ed. Boca Raton, FL: Chapman & Hall/CRC.

Hagenaars, A. and De Vos, K. (1988). The Definition and Measurement of Poverty. Journal of Human Resources, 23, 211-221. http://dx.doi.org/10.2307/145776

Han, J., Meyer, B. D., & Sullivan, J. X. (2020). *Income and Poverty in the COVID-19 Pandemic* (No. w27729). National Bureau of Economic Research.

Jann, B. (2003). The Swiss Labor Market Survey 1998 (SLMS 98). Schmollers Jahrbuch : Zeitschrift für Wirtschaftsund Sozialwissenschaften, 123(2), 329-335. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-409467

Jenkins, S., Burkhauser, R., 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.

Machado, José A.F. & Santos Silva, J.M.C., 2019. "Quantiles via moments," Journal of Econometrics, Elsevier, vol. 213(1), pages 145-173.

Moore, J. C., L. Stinson and E. Welniak. "Income Measurement Error in Surveys: A Review." *Journal of Official Statistics* 16 (2000): 331-362.

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

Rubin, D. B. 1987. Multiple Imputation for Non-response in Surveys. New York: Wiley.

Ting Yan. Liangqiang Qu. Zhaohai Li. Ao Yuan. "Conditional kernel density estimation for some incomplete data models." Electron. J. Statist. 12 (1) 1299 - 1329, 2018. https://doi.org/10.1214/18-EJS1423

Walter, P., & Weimer, K. (2018). *Estimating poverty and inequality indicators using interval censored income data from the german microcensus* (No. 2018/10). Diskussionsbeiträge.

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

Xiuqing Zhou, Yanqin Feng & Xiuli Du (2017) Quantile regression for interval censored data, Communications in Statistics - Theory and Methods, 46:8, 3848 3863, DOI: 10.1080/03610926.2015.1073317

Yi-Ting Chen (2018) A Unified Approach to Estimating and Testing Income Distributions With Grouped Data, Journal of Business & Economic Statistics, 36:3, 438-455, DOI: 10.1080/07350015.2016.1194762