Recovering Income Distribution in the Presence of Interval-Censored Data

Gustavo Canavire-Bacarreza [1]
Fernando Rios-Avila [2]
Flavia Sacco-Capurro[3]

[1]The World Bank

[2]Levy Economics Institute

[3]The World Bank

2022 Stata Conference

Motivation

- Household and labor force surveys are useful to understand employment dynamics in both developing and developed countries
- In the Latin American and the Caribbean region, many countries collect their labor force surveys quarterly as oppose to a yearly basis
- However, the higher data collection frequency comes at a cost: Wage data is often censored (in brackets).
- Thus, income distribution is difficult to analyze using standard methods.

Why reporting incomes in brackets?

- Questions to collect information on income is the higher response rate compare to questions asking to report exact amounts (Wang et al., 2013)
- Income information is considered ”sensitive”, and people are reluctant to report actual earnings, and may choose not to respond those questions at all (Moore et al., 2000; Hagenaars and Vos, 1988).
- This form of data collection solves the problem of underreporting or missreporting, it raises a problem for recovering the full wage (income) distribution

What do we do?

- This problem can be address using multiple imputation method, by simulating multiple candidates for the observations with censored data.
- What we propose is an extension on the imputation approach described in Royston (2007) (implemented in `mi impute intreg`), by explicitly allowing for heteroskedastic errors.
- The goal is to best model the conditional distribution of the censored data.
- The estimated model is then used to impute of wages.
- 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. `mi estimate`.

How does the paper fits into the literature?

- Measuring income inequality with right-censored (top-coded) data (Jenkins et al.(2011))
- Estimation of parametric income distributions using grouped data (Chen 2017)
- Pseudo - Samples from interval-censored income variable (Walter and Weimer, 2018)
- CPS imputation methods Han et al.(2020), Parolin and Wimer (2020).
- Multiple imputation software implements various methods for the treatment of missing data, in `Stata, mi impute (intreg)` implements a similar algorithm.
- The approach, however, assumes homoskedastic errors; allowing for heteroskedasticty (our approach) provides more flexibility to capture conditional distributions, and is less biased compared to `mi impute intreg`.

Methodology

- How can we handle interval bracket data?
- Use Interval Regression.

```
intreg ll uu indepvars, options
```

- This, however, only helps you to analyze one thing: conditional means.
- What if you would like to analyze something other than conditional means?
  - Quantile regression,
  - Unconditional quantile regressions,
  - distributional analysis? (Gini, variance, etc)

How does Interval Regression works?

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

$$y_i = \mu\left(x_i\right) + v_i \sigma\left(x_i\right)$$

- if v_i follows a normal distribution then

$$v_i \sim N(0,1) \rightarrow y_i | x_i \sim N\left(\mu(x), \sigma(x)\right)$$

- We could estimate this, using maximum likelihood to maximize:

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

- This model can then be used to impute missing data.

Interval Regression

- If your data is available in brackets, Interval regression can be used to analyze it. We simply change the objective function:

$$P\left(ll_i \leq y_i < uu_i | x_i\right)$$

- Which changes the Log Likelihood to the following v_i

$$L_i\left(\mu(x), \sigma(x)\right) = \Phi\left(\frac{uu_i - \mu\left(x_i\right)}{\sigma\left(x_i\right)}\right) - \Phi\left(\frac{ll_i - \mu\left(x_i\right)}{\sigma\left(x_i\right)}\right) \text{ if interval - censored}$$

$$L_i\left(\mu(x), \sigma(x)\right) = \Phi\left(\frac{uu_i - \mu\left(x_i\right)}{\sigma\left(x_i\right)}\right) \text{ if left-censored}$$

$$L_i\left(\mu(x), \sigma(x)\right) = 1 - \Phi\left(\frac{ll_i - \mu\left(x_i\right)}{\sigma\left(x_i\right)}\right) \text{ if right - censored}$$

$$L_i\left(\mu(x), \sigma(x)\right) = \frac{1}{\sigma\left(x_i\right)}\phi\left(\frac{ll_i - \mu\left(x_i\right)}{\sigma\left(x_i\right)}\right) \text{ if fully observed}$$

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

Model Imputation

- Once the model is estimated, the imputation process is similar to the one implemented in `mi impute intreg`
- First: we obtain a random draw for v_i assuming that:

$$v_i^* \in \left[\frac{ll_i - \mu(x_i)}{\sigma(x_i)}, \frac{uu_i - \mu(x_i)}{\sigma(x_i)} \right]$$

  Which is simple a draw from a truncated normal distribution.
- Second, we obtain an imputation for the unobserved y_i, simply by using the d.g.p. implied by the model estimation:

$$\widetilde{y}_i = \hat{\mu}(x_i) + \widetilde{v}_i \hat{\sigma}(x_i)$$

Model Imputation

- However, because the population parameters $\hat{\mu}(x_i)$ and $\hat{\sigma}(x_i)$ are measured with error, we obtain draws based on the MLE estimates for the coefficients and the corresponding Variance Covariance matrix Ω.
- This is also what `mi impute intreg` does, but with 2 differences
  - We allow for modeling $\sigma(x_i)$ to be a function of characteristics
  - We also allow for added stochastic variation, assuming that:

$$\tilde{\Omega} = \Omega * \frac{n}{\chi^2_n}$$

- The rest of the imputation follows the standard approach.
  - Obtain draws for $\tilde{\mu}(x_i)$ and $\tilde{\sigma}(x_i)$
  - Obtain draws for $\tilde{v}_i$ given $\tilde{\mu}(x_i)$ and $\hat{\sigma}(x_i)$
  - Obtain draw for $\tilde{y}_i = \tilde{\mu}(x_i) + \tilde{\sigma}(x_i) * \tilde{v}_i$

Inference

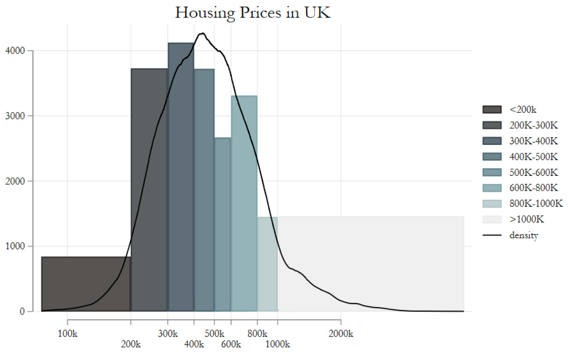
- For the analysis and statistical inference, we use `mi suit` in `Stata`, by simply importing the data in Wide Format, and use `mi estimate` for the analysis.
- What this command does in the background is estimate the model of interest using all imputations, gathers all estimated coefficients and their Variance Covariance matrix. And summarizes them as follows:

$$\widehat{\beta}_M = \frac{1}{M}\sum_{m=1}^{M}\widehat{\beta}_m$$

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

How to implement the method

```
* Setup: House Sales in King County, USA
* www.kaggle.com/code/burhanykiyakoglu/predicting-house-prices/
use pricehouse, clear
* Create censored Data
gen price_1k=price/1000
recode price_1k (0/200 = 1) (200/300=2) (300/400=3) ///
       (400/500=4) (500/600=5) (600/800=6) (800/1000=7) ///
       (1000/999999=8), gen(price_g)
```



How to implement the method

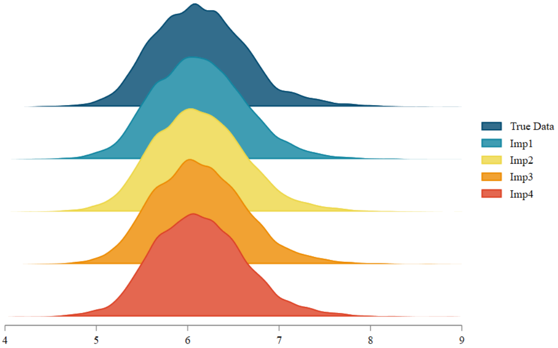
```
* Model Estimation:
* step 1: Use intreg and model a "normal" variable
intreg log_ll log_uu  /// bracket thresholds
        bedrooms bathrooms log_liv log_lot floors /// E(Y|X)
        waterfront view condition grade age_hs  renov, ///
       het(bedrooms bathrooms log_liv log_lot floors /// V(Y|X)
       waterfront view condition grade age_hs  renov)
#Notice we model the conditional mean and variance

******* Step 2: Use intreg_mi, Syntax:
intreg_mi prefix /// Prefix to be used for the new variables
            , replace /// Request Replacing variables
              reps(#) /// Request # of imputations (default 10)
              seed(str)  // And to set the seed for replication
intreg_mi lw, reps(10)

******* Step 3: Importing as MI
gen lnwage_h=.   // create the "missing" variable as anchor
tempfile s1
save `s1'         // Save a temp file
mi import wide, impute(lnwage_h = lw*)  // import into MI
* Done! Proceed as usual
```

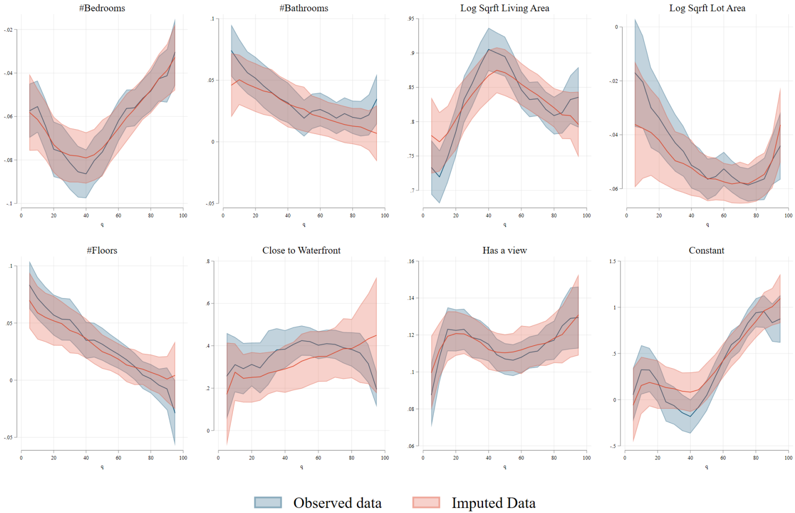
Results

```
mi passive:gen price_1k_hat=exp(log_price)
* Compare densities (ssc install joy_plot)
joy_plot logprice log_price1 log_price2 log_price3 log_price4, ///
   dadj(2) notext range(4 9) ///
   legend(order(1 "True Data" 2 "Imp1" 3 "Imp2" 4 "Imp3" 5 "Imp4"))
```

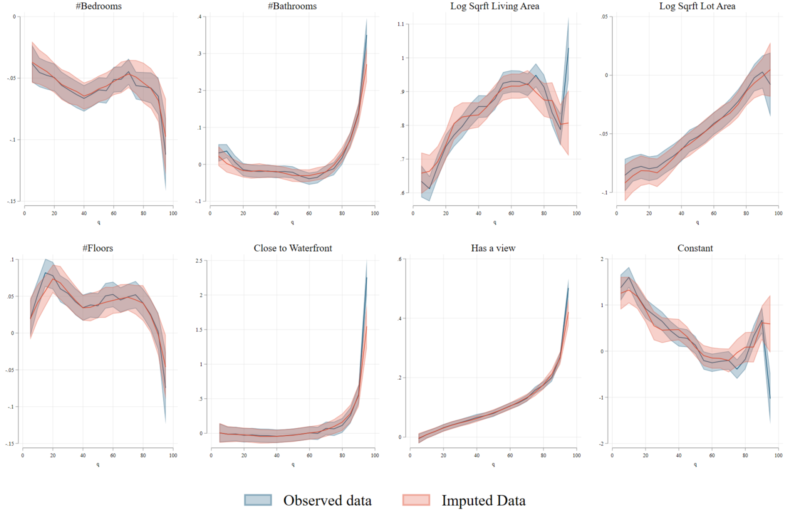


An example using CQR:Koenker and Bassett (1978)

Stylized model $Price =$
$f(bedrooms, bathrooms, log\_liv, log_lot, floors, waterfront, view)$



An example using UQR: Firpo, Fortin and Lemieux (2009)



Monte Carlo Simulations

Data Generating Process

$$y = \beta_0(\theta) + \beta_1(\theta)x_1 + \beta_2(\theta)x_2 \; \forall \; \theta \in (0,1)$$

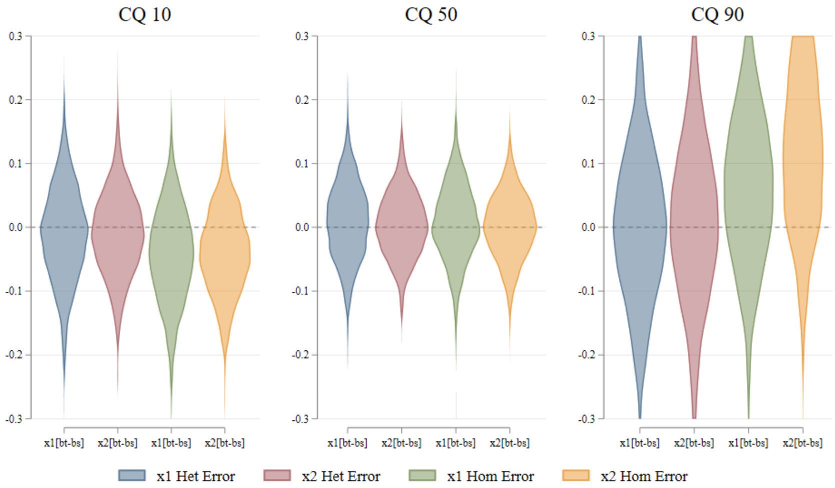
$$\beta_0(\theta) = \beta_1(\theta) = \beta_2(\theta) = 0.5 * \left(1 + \Phi^{-1}(\theta) - log(1-\theta)\right)$$

$$x_1 \sim Bernulli(0.5)$$

$$x_2 \sim \chi^2(5)/5$$

$$N = 1000$$

`intreg_mi` vs `mi impute intreg`



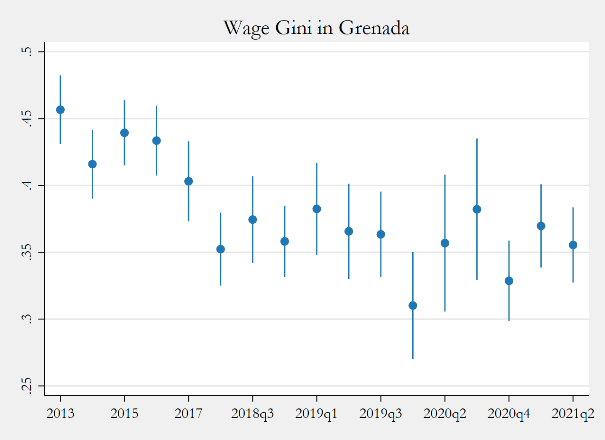
An example with real data: Grenada

Wages in Grenada are reported in brackets only.

Table 4 Labor 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

An example with real data: Grenada



Conclusions

- 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 interval-censored data
- The main limitation of our strategy is the assumption of conditional normality, which is required for the estimation of the interval regression model using standard software. This can be relaxed.
- For the specific case of Grenada the results suggest that earned income inequality in this country has declined, which coincides with other economic performance indicators in the country.