Tutorial 11: Censored and Truncated Regression

Hammad Shaikh

April 1, 2021

Truncation and Censoring

Resembles the missing data problem

- Censoring: outcome value is only partially known

$\hookrightarrow Y_i = \begin{cases} Salary_i, & Salary_i < 200{,}000 \\ 200{,}000, & Salary_i \geq 200{,}000 \end{cases}$ ⇒ Data can still be a random sample of population

$\hookrightarrow X_i = Educ_i$ can have full information

- Truncation: observations for which outcome value is outside a range is not recorded

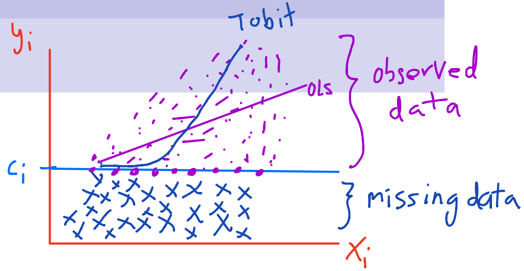
Resembles the problem of sample selection

$Y_i = Salary_i$ if $Salary_i < 100{,}000$ ⇒ Sampling based on outcome value falling in some range

$\hookrightarrow X_i = Educ_i$ only observed when $Salary_i < 100{,}000$

Censored regression model

- Let y_i be left censored by c_i $\Rightarrow$



- Observe $w_i = max(y_i, c_i)$ $= \begin{cases} y_i & \text{if } y_i > c_i \\ c_i & \text{if } y_i \le c_i \end{cases}$

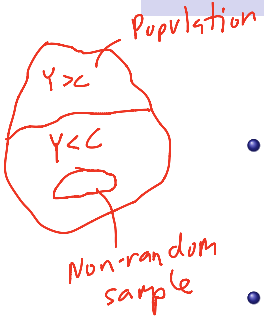
- $y_i = \beta x_i + u_i, u_i \sim N(0, \sigma^2)$ $\Rightarrow y_i | X_i \sim N(\beta X_i, \sigma^2)$

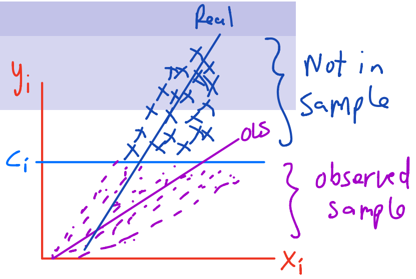
  $\hookrightarrow f(y_i | X_i, c_i) = \begin{cases} N(\beta X_i, \sigma^2), & y_i > c_i \\ Pr(w_i = c_i), & y_i \le c_i \end{cases} \Rightarrow Pr(w_i = c_i) = \Phi\left(\frac{c_i - x_i\beta}{\sigma}\right)$

- Estimate using (β, σ) use MLE

  $\hookrightarrow \log(\mathcal{L}(\beta, \sigma)) = \sum_{i=1}^{N} \log(f(y_i | X_i, c_i)) \Rightarrow$ MLE for $\hat{\beta}_{MLE}, \hat{\sigma}_{MLE}$

Truncated regression model





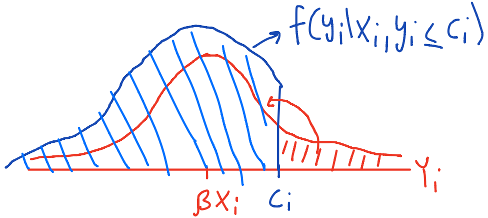
- Data (x_i, y_i) only observed when $y_i \leq c_i \Rightarrow$

  ↳ X_i only observed when $y_i \leq c_i$

- $y_i = \beta x_i + u_i$, $u_i \sim N(0, \sigma^2)$, only for $y_i \leq c_i$

  ↳ $y_i | X_i \sim N(\beta X_i, \sigma^2)$ $\Longrightarrow$

  ↳ not truncated

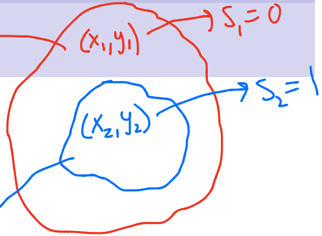


- Likelihood will use truncated distribution $f(y_i|x_i, y_i \leq c_i)$

$$f(y_i|x_i, y_i \le c_i) = \frac{f(y_i|x_i)}{Pr(y_i \le c_i)} = \frac{\frac{1}{\sigma}\phi\left(\frac{y_i - X_i\beta}{\sigma}\right)}{\Phi\left(\frac{X_i\beta}{\sigma}\right)} \Rightarrow \log(\mathcal{L}(\beta,\sigma)) = \sum_{i=1}^{N} \log(f(y_i|X_i, c_i))$$

↳ MLE $\hat{\beta}^{MLE}$, $\hat{\sigma}^{MLE}$

Sample Selection



- Consider a population $\{(x_i, y_i)\}_{i=1}^{N}$
- Let s_i be a indicator for the selected sample

$$s_i = \begin{cases} 1, & \text{if } (x_i, y_i) \text{ is in the sample} \\ 0, & \text{if } (x_i, y_i) \text{ is not in sample} \end{cases}$$

- OLS on selected sample unbiased if selection is random

$$y_i = \beta X_i + \varepsilon_i \text{ if } s_i = 1 \Leftrightarrow s_i y_i = \beta(s_i X_i) + (s_i \varepsilon_i) \Rightarrow E[s_i \varepsilon_i | s_i X_i] = s_i E[\varepsilon_i | s_i X_i]$$

$$= s_i E[\varepsilon_i | X_i] = 0$$

$$s_i \perp \varepsilon_i$$

- OLS can be biased when sample non-random

$$Cov(s_i, \varepsilon_i) \neq 0 \Rightarrow s_i E[\varepsilon_i | s_i X_i] \neq 0 \Rightarrow \hat{\beta}^{OLS} \text{ is biased}$$