Selection Bias

Ethan Levien

November 4, 2025

1.1 Heckman model

Consider a standard linear regression model

$$Y = \sum_{j=1}^{K} \beta_j X_j + \epsilon. \tag{1}$$

Given observed values $\mathbf{Y}$ and design matrix X, we obtain the usual estimate $\hat{\beta} = \hat{\Sigma}^{-1} X^T \mathbf{Y}$. When fitting a regression model, the predictor values can be sampled in any way and it will not change our estimator of $\hat{\beta}$, it being defined by expectations conditioned on X. What concerns is situations where the response variable is filtered for censured in some way.

Heckman proposed a famous model for this. This idea is that there is an axuilary variable W defined by

$$W = \sum_{j=1}^{K} \alpha_j X_j + \eta \tag{2}$$

We assume the covariance matrix of the noise terms is

$$\begin{bmatrix} \sigma_\epsilon^2 & \rho \\ \rho & \sigma_\eta^2 \end{bmatrix} \tag{3}$$

The observed Y are then selected based on the auxililary process. We consider data points to be randomly selected based on W (in Heckman's formulation there is a sharp cutoff). We therefore introduce a new variable S which indicate whether data point i is selected and take

$$S \sim \text{Bernoulli}(h(W)) \tag{4}$$

for $h : \mathbb{R} \to \mathbb{R}_{>0}$. Combing everything,

$$Y|X \sim \text{Normal}\left(\sum_{j=1}^{K} \beta_j X_j, \sigma_\epsilon^2\right) \tag{5}$$

$$W|X = \text{Normal}\left(\sum_{j=1}^{K} \alpha_j X_j, \sigma_\eta^2\right) \tag{6}$$

$$S|W \sim \text{Bernoulli}(h(W)) \tag{7}$$

Also note that

$$\text{cov}(Y, W|X) = \rho \tag{8}$$

in fact the covariance matrix of $Y|X$ and $W|X$ are the same as ϵ and η. We then define the selected sample as What if we have observations of Y_s. Let's assume we have an estimate $\hat{h}(W)$ of the conditional probability of being selected conditioned on the auxiliary variable W. The idea is to define a regression model for Y among the selected samples. That is, for $Y|X, W\{S = 1\}$. We note that

$$Y = \sum_{j=1}^{K} \beta_j X_j + \epsilon = \sum_{j=1}^{K} \beta_j X_j \tag{9}$$

$$W = \sum_{j=1}^{K} \alpha_j \tag{10}$$