Selection Bias

Ethan Levien

November 10, 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.

Let S denote whether in the data is selected for. Heckman's model $S = 1_{\{W>0\}}$ where W is the auxility variable

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

and the noise terms are correlated

$$\begin{bmatrix} \mathrm{var}(\epsilon) & \mathrm{cov}(\epsilon,\eta) \\ \mathrm{cov}(\epsilon,\eta) & \mathrm{var}(\eta) \end{bmatrix} = \begin{bmatrix} \sigma_\epsilon^2 & \sigma_{\epsilon,\eta} \\ \sigma_{\epsilon,\eta} & \sigma_\eta^2 \end{bmatrix} \tag{3}$$

This is also the covariance matrix of $(Y|X, W|X)$. This we may write

$$Y = \sum_{j=1}^{K} \beta_j X_j + \frac{\sigma_{\epsilon,\eta}}{\sigma_\eta^2}\eta + \epsilon', \quad \epsilon' \sim \mathrm{Normal}(0, \sigma_\epsilon^2 - \sigma_{\epsilon,\eta}^2/\sigma_\eta^2) \tag{4}$$

This gives our regression model with η as a predictor, but we want to condition on the event $W > 0$, which is obtained by averaging the regression model above:

$$E[Y|\{X_j\}_{j=1}^K, W > 0] = E[Y|\{X_j\}_{j=1}^K, \sum_{j=1}^{K} \alpha_j X_j > \eta] \tag{5}$$

$$= E[E[Y|\{X_j\}_{j=1}^K, \eta]|\{X_j\}_{j=1}^K, \sum_{j=1}^{K} \alpha_j X_j > \eta] \tag{6}$$

$$= \sum_{j=1}^{K} \beta_j X_j + \frac{\sigma_{\epsilon,\eta}}{\sigma_\eta^2} E[\eta| \sum_{j=1}^{K} \alpha_j X_j > \eta] \tag{7}$$

We write $\tilde{\beta} = \sigma_{\epsilon,\eta}/\sigma_\eta^2$ and $\tilde{X} = E[\eta| \sum_{j=1}^{K} \alpha_j X_j > \eta]$ so we now have the regression model with $K+1$ predictors

$$Y = \sum_{j=1}^{K} \beta_j X_j + \epsilon = \sum_{j=1}^{K} \beta_j X_j + \tilde{\beta}\tilde{X} + \epsilon' \tag{8}$$

1.2 Alternating iteration

We can consider the approach of finding the latent variable