Heckman Correction

Shivank Goel, Email : sg2359@cornell.edu

INTRODUCTION

Heckman correction is used when there is self-selection bias in the sample dataset [1]. Suppose we want to estimate i.e. $y = X\beta + \epsilon$. However, since selection is biased, let s_i be the variable denoting if the i^{th} entry is sampled, i.e., $s = \mathbb{1}(\bar{X}\delta + u)$, where, $\mathbb{1}(\cdot)$ indicates the indicator function. Often, $\bar{X}$ is a superset of the matrix X.

ASSUMPTIONS

We assume that, (ϵ, u) is independent of $\bar{X}$ and $\mathbb{E}(\epsilon, u) = 0$. Further, we assume, $u \sim N(0, 1)$ and the two error terms are linerly related, i.e., $\mathbb{E}(\epsilon|u) = \lambda u$. In other words, heckman assumes that, $(\epsilon, u) \sim N(0, 0, \sigma_\epsilon^2, \sigma_u^2 = 1, \rho_{\epsilon u})$.

DETAILS

$\mathbb{E}(y|\bar{X}, u) = X\beta + \mathbb{E}(\epsilon|u) = X\beta + \lambda u$, where λu is the bias caused due to correlation between the error terms. However, since we don't observe u we calculate the expectation conditional on s, i.e., whether the entry was self selected into the datset. Thus we get, $\mathbb{E}(y|\bar{X}, s) = \mathbb{E}(\mathbb{E}(y|\bar{X}, u)|\bar{X}, s) = \mathbb{E}(X\beta + \lambda u|\bar{X}, s) = X\beta + \lambda\mathbb{E}(u|\bar{X}, s)$. To get the unbiased coeeficient of X, i.e., β, we need to include the term $\mathbb{E}(u|\bar{X}, s)$ in the regression analysis. We calculate $\mathbb{E}(u|\bar{X}, s)$ for the two cases $(s = 0, s = 1)$ separately. $\mathbb{E}(u|\bar{X}, s = 1) = \mathbb{E}(u|\bar{X}, u > -\bar{X}\delta) = \phi(\bar{X}\delta)/\Phi(\bar{X}\delta)$. The ratio, $\phi(\bar{X}\delta)/\Phi(\bar{X}\delta)$ is called inverse Mill's ratio. We can do a similar analysis for, $s = 0$ case.

EXCLUSION RESTRICTION

In general it is not necessary for $\bar{X}$ to be a super set of X. However, if all the covariates in X and $\bar{X}$ are same then the mills ratio $\phi(\bar{X}\delta)/\Phi(\bar{X}\delta)$ can be just a linear function of $\bar{X}$ with a significantly high probability, because the inverse Mill's ratio function is almost linear function of it's inputs for a large part of it's domain (as shown in the figure 1). This can cause the problem of severe multicollinearity and large standard errors. Hence, in practice $\bar{X}$ contains few covariates that are not highly correlated with the existing covariates in the matrix X. That is why, the heckman selection model in *Stata* imposes the condition on the selection equation, that it should contain at least one variable that is not included in the outcome equation.

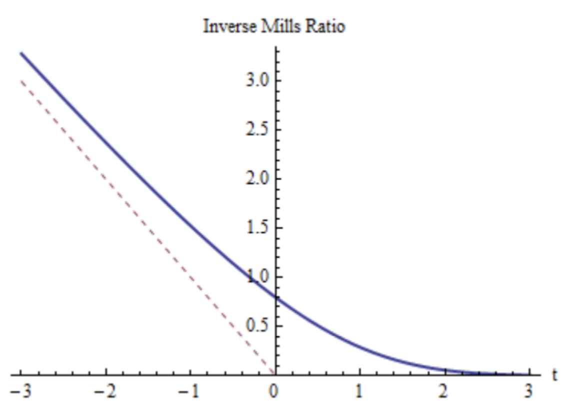

Figure 1. Inverse Mill's ratio function

[1] This article has been adapted from http://www.yichijin.com/files/heckman.pdf