1 Selection bias

1.1 Heckman’s two-step model

Consider the model in Heckman (1979)

\[ Y_i = X_i' \beta + \varepsilon_i, \]
\[ D_i = I \{ Z_i' \gamma + \eta_i > 0 \}. \]

For a random sample from the population, we observe \( D_i, Z_i \), and \( X_i \), but \( Y_i \) only if \( D_i = 1 \).

The parametric form of the model assumes that

\[ \begin{pmatrix} \varepsilon_i \\ \eta_i \end{pmatrix} | X_i, Z_i \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_\varepsilon & \rho \sigma_\varepsilon \\ \rho \sigma_\varepsilon & 1 \end{pmatrix} \right) \]

Note that the variance for \( \eta_i \) is normalized to one since we only observe the sign of \( Z_i' \gamma + \eta_i \). Therefore \( \gamma \) can only be identified up to a scale.

A classical example of this two-equation model is the wage regression model, where \( Y_i \) is the observed wage rate and \( D_i \) is the labor supply decision, which is decided by the latent calculation that one’s expected wage rate is larger than her reservation wage rate such that she chooses to work. This is the ‘selection bias’ problem, where individuals with observable wage are those who self-select into working since the expected payoff of working offers a higher utility than not working.

Under the assumption of bi-variate normality assumption, one can estimate the model
efficiently using the MLE. However, the log-likelihood function is complicated and the computational problem does not behave ‘nicely’. Instead, Heckman proposed a two-step method. Thanks to the bi-variate normality assumption, one can show that the conditional mean of \( \varepsilon_i \) given \( \eta_i \) is a linear function of \( \eta_i \). In fact,

\[
E[\varepsilon_i | \eta_i] = \delta \eta_i,
\]

where \( \delta = \rho \sigma \varepsilon \). Thus we have

\[
E[Y_i | X_i, \eta_i] = X_i' \beta + \delta \eta_i.
\]

Since \( E \eta_i \neq 0 \) in the wage equation, the OLS estimate \( Y_i \) on \( X_i \) is biased unless \( \rho = 0 \), where the selection process is independent of the expected outcome. This is the so called ‘selection bias’ or ‘endogenous variable bias’.

Note that

\[
E[\eta_i | X_i, Z_i, D_i = 1] = E[\eta_i | X_i, Z_i, Z_i' \gamma + \eta_i > 0] = E[\eta_i | \eta_i > -Z_i' \gamma] = \lambda(Z_i' \gamma),
\]

which is the truncated (from below) mean of the normal variate \( \eta_i \). One can show that the truncated mean of a standard normal variate from below at a constant \( a \) takes the form

\[
\lambda(a) = \phi(a) / \Phi(a),
\]

which is the inverse Mill’s ratio, wherein \( \phi \) and \( \Phi \) are the pdf and cdf of standard normal.

Therefore, we have

\[
E[Y_i | X_i, \eta_i] = X_i' \beta + \delta \lambda(Z_i' \gamma).
\]

Heckman’s two-step estimator is the following. First estimate \( \gamma \) by probit (recall the normality assumption), which gives consistent estimate of \( \gamma \). Then for each observation with \( D_i = 1 \), calculate the estimated inverse Mill’s ratio \( \hat{\lambda}_i = \lambda(Z_i' \hat{\gamma}) \), and regress \( Y_i \) on \( X_i \) and \( \hat{\lambda}_i \).
Compared to the MLE, this two-step estimate is not necessarily efficient. However, to test if selection problem is present \((\delta = 0)\), the OLS standard error suffices here.

### 1.2 Exclusion Restriction

So far, we do not use exclusion restriction, and \(Z_i\) is allowed to be identical to \(X_i\). However, this practice may lead to nearly perfect collinearity. In this case, the identification is called ‘identification based on functional form assumptions’ in the sense that \(\lambda_i\) is a nonlinear function of \(X_i\) (when \(Z_i = X_i\)) and the conditional expectation of \(Y_i\) given \(X_i\) ends up being nonlinear in \(X_i\), where the nonlinear part is interpreted as the selection bias. Therefore, the identification depends crucially on the functional form assumptions.

Instead, when we use exclusion restrictions (variables in \(Z_i\) that are not in \(X_i\)), the identification is not that fragile. In this case, there is variation in \(\lambda_i\) conditional on \(X_i\), so the selection bias coefficient is separately identified, even when the normality assumption does not hold. However, sometimes it is difficult to justify an exclusion restriction that a certain variable affects the participation decision but not the outcome.

### 2 Causal Inference using Instrumental Variable


Summary: This paper estimates the effect of veteran status in the Vietnam war on mortality, using the lottery number that assigned priority for the draft as an instrument.

Consider the problem of inferring the effect of veteran status on health outcome. For person \(i\), let \(Y_i\) be the observed health outcome, \(D_i\) be the observed treatment (veteran status) and \(Z_i\) be the observed draft status. For simplicity, \(Z_i\) is a binary variable such that those with a low lottery number \((Z_i = 1)\) would have served in the military \((D_i = 1)\) and those with a high lottery number \((Z_i = 0)\) would not have served \((D_i = 0)\).
2.1 Structural model

Goldberger (1972) defined structural equation models as “stochastic models in which each
equation represents a causal link, rather than a mere empirical association.”

Consider the standard dummy endogenous variables model

\[ Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i, \]
\[ D_i^* = \alpha_0 + \alpha_1 Z_i + v_i, \]

and

\[ D_i = 1, \text{ if } D_i^* > 0, \]
\[ = 0, \text{ if } D_i^* \leq 0. \]

In this model, \( \beta_1 \) represents the causal effect of treatment of \( D \) on \( Y \). The underlying assumptions include:

1. \( E[Z_i \varepsilon_i] = 0 \) and \( E[Z_i v_i] = 0 \). This suggests that any effect of \( Z \) on \( Y \) must be through an effect of \( Z \) on \( D \).

2. \( \text{cov}(D_i, Z_i) \neq 0 \), implying that \( \alpha_1 \neq 0 \).

Under this two conditions, \( Z \) is considered a valid IV. One can show that the IV estimator can be defined as the ratio of sample covariance

\[ \hat{\beta}_1^{IV} = \frac{\hat{\text{cov}}(Y_i, Z_i)}{\hat{\text{cov}}(D_i, Z_i)} = \frac{\sum_{i=1}^{N} Y_i Z_i / \sum_{i=1}^{N} Z_i - \sum_{i=1}^{N} Y_i (1 - Z_i) / \sum_{i=1}^{N} (1 - Z_i)}{\sum_{i=1}^{N} D_i Z_i / \sum_{i=1}^{N} Z_i - \sum_{i=1}^{N} D_i (1 - Z_i) / \sum_{i=1}^{N} (1 - Z_i)}. \]

This IV estimand depends crucially on the functional form (linearity) and distributional assumptions.
2.2 Causal inference using potential outcome

Let the vector $Z = [Z_i], i = 1, \ldots, N$. The vector $D$ and $Y$ are similarly defined. Then $D_i(Z)$ is the indicator for whether person $i$ would serve given the random vector of draft assignment $Z$. [note that this is a potential outcome, instead of realized one]. In an ideal world, $D_i(Z) = Z_i$ for all $i$ if everyone complied with the draft lottery. However, in real world, we observe $D_i(Z)$ differs from $Z_i$ for various reasons.

Similarly, $Y_i(Z, D)$ is another potential outcome given the draft and treatment vectors. The two potential outcomes $D_i(Z)$ and $Y_i(Z, D)$ are fixed but unknown parameters to estimate. Generally, differences in these potential outcomes due to assigned and received treatment will be revealed by analyzing data obtained by randomly assigning $Z$ to the sample.

• **Assumption 1**: Stable Unit Treatment Value Assumption (SUTVA)
  
  - If $Z_i = Z'_i$, then $D_i(Z) = D_i(Z')$
  - If $Z_i = Z'_i$ and $D_i = D'_i$, then $Y_i(Z, D) = Y_i(Z', D')$.

• **Assumption 2**: Random assignment (of $Z$).

• **Definition 1**: causal effects of $Z$ on $D$ for person $i$ is $D_i(1) - D_i(0)$ and on $Y$ is $Y_i(1, D_i(1)) - Y_i(0, D_i(0))$.

Given assumptions 1 and 2, the average causal effect of $Z$ on $Y$ is

$$\frac{\sum Y_i Z_i}{\sum Z_i} - \frac{\sum Y_i (1 - Z_i)}{\sum (1 - Z_i)}$$

and the average causal effect of $Z$ on $D$ is

$$\frac{\sum D_i Z_i}{\sum Z_i} - \frac{\sum D_i (1 - Z_i)}{\sum (1 - Z_i)}$$.

Obviously, these two estimands are simple average differences between those with $Z_i = 1$ and with $Z_i = 0$. Next, one can show the ratio of these two estimands is the IV estimand.
Note that due to the presence of imperfect compliance, we cannot evaluate the causal effect of $D$ on $Y$ by taking simple average difference between those with $D_i = 1$ and with $D_i = 0$ since the assignment is not ignorable. We need some extra assumptions to identify the causal effect.

- **Assumption 3:** Exclusion restriction $Y(Z, D) = Y(Z', D)$ for all $Z, Z'$ and for all $D$.

  This assumption implies that $Y_i(1, d) = Y_i(0, d)$ for $d = 0, 1$. In other words, any effect of $Z$ on $Y$ must be through an effect of $Z$ on $D$. Now we can define the potential outcome

  \[ Y(D) = Y(Z, D) \text{ for all } Z \text{ and } D. \]

  Combined with assumption 1, we can write $Y_i(D_i)$ instead of $Y_i(Z, D)$.

- **Definition 2:** the causal effect of $D$ on $Y$ for person $i$ is $Y_i(1) - Y_i(0)$.

Next we focus on average causal effects in groups of people who can be induced to change treatments.

- **Assumption 4:** Nonzero causal effects of $Z$ on $D$: the causal effects of $Z$ on $D$,

  \[ E[D_i(1) - D_i(0)] 
eq 0. \]

- **Assumption 5:** Monotonicity:

  \[ D_i(1) \geq D_i(0) \text{ for all } i \]

  This implies that no one does the opposite of his assignment.

- **Definition 3:** Instrumental variable for the causal effect of $D$ on $Y$: a variable satisfies assumption 1 – 5 above.
Now one can show that the causal effect of $Z$ on $Y$ for person $i$

\[
Y_i (1, D_i (1)) - Y_i (0, D_i (0))
\]

\[
= Y_i (D_i (1)) - Y_i (D_i (0)) \quad \text{(by exclusion restriction)}
\]

\[
= [Y_i (1) D_i (1) + Y_i (0) (1 - D_i (1))] \quad \text{(since } D_i (1) = \Pr (D_i = 1|Z_i = 1))
\]

\[
- [Y_i (1) D_i (0) + Y_i (0) (1 - D_i (0))]
\]

\[
= [Y_i (1) - Y_i (0)] [D_i (1) - D_i (0)].
\]

Therefore, the causal effect of $Z$ on $Y$ is the product of: (i) causal effect of $D$ on $Y$ and (ii) causal effect of $Z$ on $D$.

Note that $D_i (1) - D_i (0)$ only takes three values: 1, 0 and -1. When $D_i (1) - D_i (0) = 0$, the causal effect of $Z$ on $Y$ is clearly zero. The case $D_i (1) - D_i (0) = -1$ is ruled out by the monotonicity assumption. Therefore, we obtain

\[
E [Y_i (1, D_i (1)) - Y_i (0, D_i (0))]
\]

\[
= E \{[Y_i (1) - Y_i (0)] [D_i (1) - D_i (0)]\}
\]

\[
= E [Y_i (1) - Y_i (0) | D_i (1) - D_i (0) = 1] \Pr [D_i (1) - D_i (0) = 1].
\]

Finally, by Assumption 4, we obtain the causal interpretation of IV estimand:

\[
\frac{E [Y_i (1, D_i (1)) - Y_i (0, D_i (0))]}{E [D_i (1) - D_i (0)]}
\]

\[
= \frac{E [Y_i (1, D_i (1)) - Y_i (0, D_i (0))]}{\Pr [D_i (1) - D_i (0) = 1]}
\]

\[
= E [Y_i (1) - Y_i (0) | D_i (1) - D_i (0) = 1].
\]

This treatment effect is called Local Average Treatment Effect. Therefore, the IV estimand captures the effect of the treatment of those who would be affected by the treatment, rather than the generally population.