Causal Inference2

Weiheng Zhang

2022-04-16

Regression: another method to eliminate the selection bias

- The CEF Decomposition Property

$$y_i = E[y_i|X_i] + \varepsilon_i$$

- The CEF Prediction Property

Let $m(X_i)$ be any function of X_i, then the CEF solves

$$E[y_i|X_i] = \underset{m(X_i)}{\operatorname{argmax}} E[(y_u - m(X_i))^2]$$

so it is the MMSE predictor of y_i given X_i.

- The ANOVA Theorem(Conditional variance theorem)

$$Var(y_i) = Var(E[y_i|X_i]) + E[Var(y_i|X_i)]$$

- The linear CEF theorem (Regression Justification I)

Suppose the CEF is linear, then the population regression function is it.

- The Best Linear Predictor Theorem (Regression Justification II)

The function $X'\beta$ is the best linear predictor of y_i given X_i in a MMSE sense. (based on The CEF Prediction Property)

RCT Theory

$$\begin{aligned} Y_{1i} &: \text{Potential outcome for unit i with treatment} \\ Y_{0i} &: \text{Potential outcome for unit i without treatment} \end{aligned}$$

The Causal effect is

$$Y_{1i} - Y_{0i}$$

- A *constant-effects* assumption allows us to write:

$$Y_{1i} = Y_{0i} + k \quad or \quad Y_{1i} - Y_{0i} = k$$

Thus, we have

$$
\begin{aligned}
E[y_i|D_i=1]-E[y_i|D_i=0] &= E[y_{1i}|D_i=1]-E[y_{0i}|D_i=0] \\
&= k + \underbrace{E[y_{0i}|D=1]-E[y_{0i}|D=0]}_{Bias}
\end{aligned}
$$

When D_i is randomly assigned, $E[y_{0i}|D=1]-E[y_{0i}|D=0]=0$, so

$$E[y_i|D_i=1]-E[y_i|D_i=0]=k$$

Random assignment eliminates selection bias.

Application (potential income ~ schooling)

$$
\begin{aligned}
Y_{1i} &: \text{gradutate i's earnings haveing gone private} \\
Y_{0i} &: \text{gradutate i's counterfactual} \\
P_i &: \text{the dummy variable for private school}
\end{aligned}
$$

- initial assumption

  $$Y_{0i}=\alpha+\eta_i$$

  assume $Y_{1i}-Y_{0i}=\beta$. Though $E[\eta_i|P_i]\neq 0$

- second try

  Assume controls satisfy a conditional independence assumptiion:

  $$E[\eta_i|P_i,X_i]=E[\eta_i|X_i]=\gamma X_i$$

$$Y_i=\alpha+\gamma X_i+\beta P_i+u_i$$

where

X_i is the control variable to make β causal and eliminate selection bias; β is causal, γ is meaningless and the assumption $E[u_iX_i]=0$.

Ommited Variables Bias

The omitted variables bias (OVB) formula descries the relationship between regression coefficients in models with different controls

- Go long: wages on schooling, s_i, controlling for ability (A_i)

  $$Y_i=\alpha+\rho s_i+A_i'\gamma+\varepsilon_i$$

  $$\frac{Cov(y_i,s_i)}{Var(s_i)}=\rho+\gamma'\delta_{As}$$

where δ_{As} is the vector of coefficients from regressions of the elements of A_i on s_i.

Short equals long when omitted and included are uncorrelated.

Bad Control

Bad controls are variables that are also affected by treatment.

Table 3.2.1

Controls:	None	Age Dummies	Col(2) + additional family background controls	Col(3) + AFQT score	Col(4) + Occupation dummy
-	.132	.131	.114	.087	.066

1. adding age the coefficient did not change a lot: age is a good predict for earning but for worked people, their age is uncorrelated with their schooling.
2. adding family background, the coefficient decreasing: because parents' year of schooling is positively correlated with children's schooling, which explained part of effect of schooling on age.
3. adding AFQT score: AFQT is similar to IQ test, so the effect was explained by part of test score.
4. adding occupation, the coefficient decrease further: this is positive correlated with earning.

Why col (4) is more appropriate, i.e. col(5) is over control. Earning correlated with position and if we control them, bad control creates selection bias(Table 6.1 MM).