Causal Effects in Policy - Lecture 2

Alice Wu

UW Madison Econ 695

This lecture

- Prove the Law of Iterated Expectations (LIE)
- Conditional Expectation Function (CEF) - Definition & 3 Properties
- Best Linear Predictor as an approximation to CEF
- Regressions with Discrete Covariates (Dummies)

Law of Iterated Expectations

- (X, Y) are two r.v.'s with joint pdf $f(X, Y)$
  - Marginal densities: $f(X) = \int_Y f(X, Y)dY$, and conditional density: $f(Y|X) = \frac{f(X,Y)}{f(X)}$.
- Expectation: $E[Y] = \int_Y f(Y)dY$, conditional $E[Y|X] = \int_Y Yf(Y|X)dY$, which is a function of X.
- Law of iterated expectations (LIE):

$$E[Y] = E[E[Y|X]] \tag{1}$$

Proof (continuous r.v.'s):

$$\begin{aligned} E[E[Y|X]] = \int_X E[Y|X]f(X)dx &= \int_X \left(\int_Y Yf(Y|X)dY \right) f(X)dX \\ &= \int_Y Y \left(\int_X f(Y|X) f(X)dX \right) dY \\ &= \int_Y Yf(Y)\, dY = E[Y] \end{aligned}$$

Conditional Expectation Function

- CEF $E[Y|X]$ is a function of X that represents the expectation of Y conditional on random variables X.
- Let $\epsilon = Y - E[Y|X]$. *Two excellent properties of CEF:*
  1. $E[\epsilon] = 0$, $E[\epsilon|X] = 0$ and $E[\epsilon\, h(X)] = 0$ for *any* function of X.
     Proof:

$$E[\epsilon] = E[Y] - E[E[Y|X]] = 0 \text{ by LIE} \tag{2}$$

$$E[\epsilon|X] = E[Y|X] - E[E[Y|X]|X] = E[Y|X] - E[Y|X] = 0$$

     Remark: $E[\epsilon|X] = 0$ means ϵ is mean independent of X. With $E[\epsilon|X] = 0$, by LIE,

$$E[\epsilon h(X)] = E[E[\epsilon h(X)|X]] = E[E[\epsilon|X]\, h(X)] = 0 \tag{3}$$

Conditional Expectation Function

2. the function $E[Y|X]$ minimizes $E[(Y - m(X))^2]$
Proof: given any function m of r.v. X,

$$\begin{aligned} Y - m(X) &= Y - E[Y|X] + E[Y|X] - m(X) \\ \Rightarrow (Y - m(X))^2 &= (Y - E[Y|X])^2 + (E[Y|X] - m(X))^2 \\ &\quad + 2\,(Y - E[Y|X])\,\underbrace{(E[Y|X] - m(X))}_{\text{funx of X: } h} \\ \Rightarrow E[(Y - m(X))^2] &= E[\epsilon^2] + \underbrace{E[(E[Y|X] - m(X))^2]}_{\geq 0} + \underbrace{2E[\epsilon h(X)]}_{=0} \end{aligned}$$

The minimizing choice is $m(X) = E[Y|X]$!
So: we've established that among the functions of X, the CEF $E[Y|X]$ gives the "best guess" for Y in the sense of minimizing $E[(Y - m(X))^2]$.

Linear Approximation to CEF

- *Problem: we don't know the conditional density $f(Y|X)$.*
- Solution: we'll use the "linear regression function": combination $x_i\beta$.
- Recall: given a sample of size N the OLS regression coefficients β solve:

$$\min_{\beta} \sum_{i=1}^{N} (y_i - x_i'\beta)^2$$

- Consider WLLN for the r.v. $(y_i - x_i'\beta)^2$: we have $\frac{1}{N}\sum_{i=1}^{N}(y_i - x_i'\beta)^2 \to^p E[(y_i - x_i'\beta)^2]$
- The "infeasible" (or population) OLS estimator solves:

$$\min_{\beta} E[(y_i - x_i'\beta)^2] \tag{4}$$

$$\Rightarrow \beta^* = E[x_i x_i]^{-1} E[x_i y_i]$$

This gives us the best linear predictor of Y given X: $E^*[y_i|x_i] = x_i'\beta^*$ (proof below).

Linear Approximation to CEF

How does $x_i'\beta^*$ relate to $E[y_i|x_i]$?

1. If the CEF is linear $E[y_i|x_i] = x_i'\beta^e$, then $\beta^* = \beta^e$.
   Why? Recall that if we define the CEF error $\varepsilon_i = y_i - E[y_i|x_i]$,

   $$E[x_i\varepsilon_i] = 0 \Rightarrow E[x_i(y_i - x_i'\beta^e)] = 0 \Rightarrow \beta^e = \beta^*$$

   This means that *if the true CEF is linear*, then the infeasible OLS represents the CEF.

   - This happens when $x's$ are dummies since $E[y_i|x_i]$ is $E[y_i|i$ in group k]

Linear Approximation to CEF

2. $E^*[y_i|x_i] = x_i'\beta^*$ is the "best" linear approx. to $E[y_i|x_i]$ (best as in *minimum-MSE*)

$$\beta^* = argmin_\beta E[(E[y_i|x_i] - x_i'\beta)^2] \tag{5}$$

Proof:

$$\begin{aligned} y_i - x_i'\beta &= y_i - E[y_i|x_i] + E[y_i|x_i] - x_i'\beta \\ \Rightarrow E[(y_i - x_i'\beta)^2] &= E[\varepsilon_i^2] + E[(E[y_i|x_i] - x_i'\beta))^2] \\ &\quad +2\underbrace{E[\varepsilon_i(E[y_i|x_i] - x_i'\beta)]}_{=0} \\ E[(E[y_i|x_i] - x_i'\beta))^2] &= E[(y_i - x_i'\beta)^2] - E[\epsilon_i^2] \end{aligned}$$

So $x_i'\beta^*$ that minimizes the mean squared error also minimizes $E[(E[y_i|x_i] - x_i'\beta))^2]$.

We can think of the "population regression" as:

$$y_i = x_i'\beta^* + u_i$$

which satisfies $E[x_i u_i] = 0$ (why?). Unless the CEF is linear, we won't have $E[u_i|x_i] = 0$.

Feasible OLS

- We have established that $E^*[y_i|x_i] = x_i'\beta^*$ is the best linear approximation to the CEF $E[y_i|x_i]$.

$$\begin{aligned}\beta^* &= argmin_\beta\, E[(y_i - x_i'\beta)^2] \\ &= E[x_i x_i]^{-1} E[x_i y_i]\end{aligned}$$

- How do find an estimate for β^*? The feasible OLS minimizes the sum of squared residuals (SSR) in the sample:

$$\hat{\beta} = argmin_\beta \sum_{i=1}^{N} (y_i - x_i'\beta)^2 \tag{6}$$

The FOC (in vector form) are:

$$\sum_{i=1}^{N} -2x_i(y_i - x_i'\beta) = 0$$

Feasible OLS

which implies that

$$\begin{aligned}\sum_{i=1}^{N} x_i x_i' \beta &= \sum_{i=1}^{N} x_i y_i \\ \Rightarrow \hat{\beta} &= [\sum_{i=1}^{N} x_i x_i']^{-1} \sum_{i=1}^{N} x_i y_i \\ &= [\frac{1}{N}\sum_{i=1}^{N} x_i x_i']^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} x_i y_i\right)\end{aligned}$$

Now plug in the population regression: $y_i = x_i'\beta^* + u_i$, where $E[u_i x_i] = 0$,

$$\begin{aligned}\hat{\beta} &= [\frac{1}{N}\sum_{i=1}^{N} x_i x_i']^{-1} \frac{1}{N}\sum_{i=1}^{N} x_i(x_i'\beta^* + u_i) \\ &= \beta^* + [\frac{1}{N}\sum_{i=1}^{N} x_i x_i']^{-1} [\frac{1}{N}\sum_{i=1}^{N} x_i u_i] \qquad (7)\end{aligned}$$

Feasible OLS

- Consistency: $\hat{\beta} \rightarrow^p \beta^*$ by WLLN and Slutsky Theorem.
- Asymptotic distribution:
  - by the Central Limit ThM and $E[x_i u_i] = 0$:

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N} x_i u_i - E[x_i u_i]\right) \rightarrow^d N(0, E[x_i x_i' u_i^2])$$

    note the variance of K-by-1 vector $x_i u_i$ (u_i is a scalar): $var(x_i u_i) = E[x_i u_i u_i x_i'] - E[x_i u_i]E[x_i u_i]' = E[x_i x_i' u_i^2] - 0_{K\times K} = E[x_i x_i' u_i^2]$.
  - By Slutsky ThM, we have:

$$\sqrt{N}(\hat{\beta} - \beta^*) \rightarrow^d N(0, E[x_i x_i']^{-1} E[x_i x_i' u_i^2] E[x_i x_i']^{-1}) \tag{8}$$

    when do we have $E[x_i x_i' u_i^2] = E[x_i x_i'] \times var(u_i)$? Apply LIE:

$$\begin{aligned} E[x_i x_i' u_i^2] &= E_x[E[x_i x_i' u_i^2 | x]] \\ &= E_x[x_i x_i' E[u_i^2 | x]] \end{aligned}$$

    Homoskedasticity: $E[u_i^2|x] = E[u_i^2]$. Otherwise the variance of residual vary by x. We need to compute robust heteroskedasticity-robust standard errors.

When is CEF linear? $E[y|x] = E^*[y|x]$

- Recall in the linear normal model, we assume:

$$y_i = \beta_0 + \beta_1 x_i + u_i,\ u_i \sim N(0, \sigma^2) \tag{9}$$

in which the normally distributed U is independent of X: $E[u_i|x_i] = 0$. This is a special case where the CEF is linear:

$$E[y_i|1, x_i] = \beta_0 + \beta_1 x_i$$

note the constant 1 is often merged into x_i:

$$E[y_i|x_i] = x_i'\beta = \begin{bmatrix} 1 & x_i \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

- Another import case: regressions with discrete regressors! x_i are indicators/dummies...

Regressions with Dummy Variables

- Consider a (population) regression with a binary independent variable $D_i \in \{0, 1\}$:

$$\begin{aligned} y_i &= \alpha + \beta_1 D_i + \epsilon_i \\ E[\epsilon_i] &= E[\epsilon_i D_i] = 0 \end{aligned} \tag{10}$$

Recall this represents the best linear projection of y on D, $E^*[y_i|D_i] = \alpha + \beta_1 D_i$.

$$\begin{aligned} (\alpha, \beta_1) &= argmin_{(a,b)} E[(y_i - a - bD_i)^2] \\ \alpha &= E[y_i] - \beta_1 E[D_i] \\ 0 &= E[(y_i - \alpha - \beta_1 D_i)D_i] = E[\epsilon_i D_i] \\ \rightarrow \beta_1 &= \frac{cov(y_i, D_i)}{var(D_i)} \text{ shown in lecture 1 (univariate reg)} \end{aligned}$$

Regressions with Dummy Variables

- Since D_i is a dummy variable/indicator, we have $E[D_i^2] = E[D_i] = Pr(D_i = 1)$. Denote by $p = E[D_i]$.

$$\begin{aligned} E[y_i D_i] =& E_{D_i}[E[y_i D_i | D_i]] \\ =& p \times E[y_i D_i | D_i = 1] + (1-p) \times \underbrace{E[y_i D_i | D_i = 0]}_{0} \\ =& p \times E[y_i | D_i = 1] \end{aligned}$$

therefore,

$$\begin{aligned} cov(y_i, D_i) &= E[y_i D_i] - E[y_i]E[D_i] = E[y_i | D_i = 1] - E[y_i] * p \\ var(D_i) &= E[D_i^2] - E[D_i]^2 = p(1-p) \end{aligned}$$

- denote by ,prove:

$$\begin{aligned} \beta_1 &= E[y_i | D_i = 1] - E[y_i | D_i = 0] \\ \alpha &= E[y_i | D_i = 0] \end{aligned} \tag{11}$$

Regressions with Dummy Variables

- In summary, with a single binary variable, the best linear projection of y_i on $(1, D_i)$ is:

$$\begin{aligned} E^*[y_i|D_i] &= E[y_i|D_i = 0] + D_i \times (E[y_i|D_i = 1] - E[y_i|D_i = 0]) \\ &= E[y_i|D_i] \text{ CEF!!} \end{aligned}$$

  we have shown that in this case the conditional expectation of any outcome y_i on a discrete variable is Linear! The constant is the mean of the outcome among $D_i = 0$, and the slope is the difference in means for the two groups.

Regressions with Dummy Variables

- In a randomized control trial, individuals are In the RCT we have 2 groups: the treatment group, with $D_i = 1$, and the control group, with $D_i = 0$. So we could fit a regression model to the "pooled" data

$$y_i = \alpha + \beta D_i + \epsilon_i$$

The treatment effect of interest is $E[y_i|D_i = 1] - E[y_i|D_i = 0]$. (We will discuss this in the potential outcome framework.)

- The (feasible) OLS estimator: let $\bar{Y}_0$ denote the sample mean of the outcome for the control group, and $\bar{Y}_1$ for the treatment group,

$$\begin{aligned} \hat{\alpha} &= \overline{Y}_0 \ \to^P E[y_i|D_i = 0] \\ \hat{\beta} &= \overline{Y}_1 - \overline{Y}_0 \ \to^p E[y_i|D_i = 1] - E[y_i|D_i = 0] \end{aligned} \tag{12}$$

So we get the estimated mean of the control group in the intercept, and the estimated treatment effect in the coefficient of D_i.

Regressions with Multiple Dummies

- Now we generalize the result to a regression with multiple indicators/dummies.
- Suppose individuals belong to G mutually exclusive groups. For example, education level: less than HS, HS, College and above.
- Define $D_{gi} = 1[i \in g]$ for $g = 1, 2, ..., G$. Because of the mutual exclusivity, $D_{gi} \times D_{ki} = 0$ for any $g \neq k$.
- We posit a population regression:

$$y_i = x_i'\beta + u_i \tag{13}$$

$$x_i := \begin{bmatrix} D_{1i} \\ ... \\ D_{Gi} \end{bmatrix}$$

Given a random sample, let $N_g = \sum_i D_{gi}$ denote the num. individuals in group g. The fraction of individuals in g: $\bar{p}_g = \frac{1}{N}\sum_i D_{gi} = \frac{N_g}{N}$. Recall the OLS estimator:

$$\hat{\beta} = [\frac{1}{N}\sum_i x_i x_i']^{-1}\frac{1}{N}\sum_i x_i y_i$$

Regressions with Multiple Dummies

- Now let's break it down:

$$
\frac{1}{N}\sum_i x_i x_i' = \begin{pmatrix} \frac{1}{N}\sum D_{1i}^2 & \frac{1}{N}\sum D_{1i}D_{2i} & \dots & \frac{1}{N}\sum D_{1i}D_{Gi} \\ \frac{1}{N}\sum D_{2i}D_{1i} & \frac{1}{N}\sum D_{2i}^2 & \dots & \\ \dots & & & \\ & & & \frac{1}{N}\sum D_{Gi}^2 \end{pmatrix}
$$

$$
= \begin{pmatrix} \overline{p}_1 & 0 & \dots & 0 \\ 0 & \overline{p}_2 & \dots & 0 \\ \dots & & & \\ 0 & & & \overline{p}_G \end{pmatrix}
$$

And with
$\frac{1}{N}\sum_{i=1}^{N} D_{gi}y_i = \frac{1}{N}\sum_{i\in g} y_i = \frac{N_g}{N} \times \left(\frac{1}{N_g}\sum_{i\in g} y_i\right) = \overline{p}_g \times \overline{y}_g,$

$$
\frac{1}{N}\sum_i x_i y_i = \begin{pmatrix} \frac{1}{N}\sum D_{1i}y_i \\ \frac{1}{N}\sum D_{2i}y_i \\ \dots \\ \frac{1}{N}\sum D_{Gi}y_i \end{pmatrix} = \begin{pmatrix} \overline{p}_1\overline{y}_1 \\ \overline{p}_2\overline{y}_2 \\ .. \\ \overline{p}_G\overline{y}_G \end{pmatrix}
$$

Regressions with Multiple Dummies

So

$$\hat{\beta} = \begin{pmatrix} \overline{p}_1 & 0 & \dots & 0 \\ 0 & \overline{p}_2 & \dots & 0 \\ \dots & & & \\ 0 & & & \overline{p}_G \end{pmatrix}^{-1} \begin{pmatrix} \overline{p}_1 \overline{y}_1 \\ \overline{p}_2 \overline{y}_2 \\ .. \\ \overline{p}_G \overline{y}_G \end{pmatrix} = \begin{pmatrix} \overline{y}_1 \\ \overline{y}_2 \\ .. \\ \overline{y}_G \end{pmatrix} \tag{14}$$

Also:

$$\overline{x} = \frac{1}{N} \sum_i x_i = \begin{pmatrix} \frac{1}{N} \sum D_{1i} \\ \dots \\ \frac{1}{N} \sum D_{Gi} \end{pmatrix} = \begin{pmatrix} \overline{p}_1 \\ .. \\ \overline{p}_G \end{pmatrix}$$

So obviously

$$\overline{y} = \overline{x}' \hat{\beta} = \sum_g \overline{p}_g \overline{y}_g \tag{15}$$

So the OLS regression is just a way to get the group-specific means.

Regressions with Multiple Dummies

The population version:

$$\begin{aligned} E^*[y_i|x_i] &= E[x_i]'\beta \\ &= \sum_g E[D_{gi}] \times E[y_i|D_{gi} = 1] \\ &= E[y_i|x_i] \end{aligned}$$

the linear projection is also the CEF!

Regressions with Multiple Dummies

- What if there is a constant in x_i: $x_i = [1, D_{1i}, D_{2i}, ..., D_{Gi}]$? $\sum_{g=1}^{G} D_{gi} = 1$ (the first element of x_i)
  - Multicollinearity: The matrix $\frac{1}{N}\sum_i x_i x_i'$ is singular (not invertible)! the sum of columns 2-$(G+1)$ equals the first column (1's).

$$\frac{1}{N}\sum_i x_i x_i' = \begin{pmatrix} 1 & \bar{p}_1 & 0 & ... & 0 \\ 1 & 0 & \bar{p}_2 & ... & 0 \\ ... & ... & ... & ... & \\ 1 & 0 & 0 & & \bar{p}_G \end{pmatrix}$$

  that is, the covariates $x_i = [1, D_{1i}, D_{2i}, ..., D_{Gi}]$ with dummies for every possible group and a constant are linearly dependent!

Regressions with Multiple Dummies

- In Stata (see reg_dummy.do), you may run "reg y ibn.group" which automatically generate $D_{1i}, D_{2i}, \ldots, D_{Gi}$ for you. By default it will keep constant, but drop a group as base to ensure $\frac{1}{N}\sum_i x_i x_i'$ is nonsingular:

```
. reg tempjuly ibn.region, robust  // West as base
note: 4.region omitted because of collinearity.

Linear regression                               Number of obs     =        954
                                                F(3, 950)         =     392.53
                                                Prob > F          =     0.0000
                                                R-squared         =     0.4247
                                                Root MSE          =     4.1746

------------------------------------------------------------------------------
             |               Robust
    tempjuly | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      region |
         NE  |   1.241406   .4451532     2.79   0.005     .3678092    2.115004
     N Cntrl |    1.35866   .4451531     3.05   0.002     .4850625    2.232257
      South  |   8.881006   .4468209    19.88   0.000     8.004136    9.757876
       West  |          0  (omitted)
             |
       _cons |   72.10859    .405254   177.93   0.000      71.3133    72.90389
------------------------------------------------------------------------------
```

Regressions with Multiple Dummies

- And if you suppress constant, Stata will keep all groups and the coef on D_{gi}: $\hat{\beta}_g = \bar{y}_g$,

```
. reg tempjuly ibn.region, robust nocons

Linear regression                               Number of obs     =        954
                                                F(4, 950)         >   99999.00
                                                Prob > F          =     0.0000
                                                R-squared         =     0.9969
                                                Root MSE          =     4.1746
```

tempjuly	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
region						
NE	73.35	.1842025	398.20	0.000	72.98851	73.71149
N Cntrl	73.46725	.1842024	398.84	0.000	73.10576	73.82874
South	80.9896	.1881971	430.34	0.000	80.62027	81.35893
West	72.10859	.405254	177.93	0.000	71.3133	72.90389

- In Python (see reg_dummy.py), I use statsmodel-OLS. By default it will keep constant, but drop a group as base to ensure $\frac{1}{N}\sum_i x_i x_i'$ is nonsingular: to be consistent with Stata, I use "West" as the base/reference group.

```
==============================================================================
Dep. Variable:               tempjuly   R-squared:                       0.425
Model:                            OLS   Adj. R-squared:                  0.423
Method:                 Least Squares   F-statistic:                     392.5
Date:                Thu, 04 Sep 2025   Prob (F-statistic):          8.61e-166
Time:                        10:21:30   Log-Likelihood:                -2715.0
No. Observations:                 954   AIC:                             5438.
Df Residuals:                     950   BIC:                             5457.
Df Model:                           3
Covariance Type:                  HC1
=======================================================================================================
                                              coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------
Intercept                                  72.1086      0.405    177.934      0.000      71.314      72.903
C(region, Treatment(reference='West'))[T.NE]      1.2414      0.445      2.789      0.005       0.369       2.114
C(region, Treatment(reference='West'))[T.N Cntrl]  1.3587      0.445      3.052      0.002       0.486       2.231
C(region, Treatment(reference='West'))[T.South]   8.8810      0.447     19.876      0.000       8.005       9.757
==============================================================================
```

Regressions with Multiple Dummies

- Suppress the constant by adding "0+" before other covariates. Python - smf.ols also keeps all groups and the coef on D_{gi}: $\hat{\beta}_g = \bar{y}_g$,

```
==============================================================================
Dep. Variable:               tempjuly   R-squared:                       0.425
Model:                            OLS   Adj. R-squared:                  0.423
Method:                 Least Squares   F-statistic:                       nan
Date:                Thu, 04 Sep 2025   Prob (F-statistic):                nan
Time:                        10:21:30   Log-Likelihood:                -2715.0
No. Observations:                 954   AIC:                             5438.
Df Residuals:                     950   BIC:                             5457.
Df Model:                           3
Covariance Type:                  HC1
======================================================================================
                         coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------
C(region)[NE]         73.3500      0.184    398.203      0.000      72.989      73.711
C(region)[N Cntrl]    73.4673      0.184    398.840      0.000      73.106      73.828
C(region)[South]      80.9896      0.188    430.344      0.000      80.621      81.358
C(region)[West]       72.1086      0.405    177.934      0.000      71.314      72.903
```