Microeconometrics Module

Lecture 6: Regression

Swapnil Singh

Lietuvos Bankas and KTU

Course Link

Introduction

- Running randomized control experiments require time, and more importantly, money
- We generally have time, but not so much money
- There are other tools, in the absence of randomization, which can help us for causal identification
- For now we focus on regression

Regression: Mathematical Details

Population Model

- Random sample of (y, x) from the population
- **Objective:** Understand how y changes with x?
- Linear model: $y = \beta_0 + \beta_1 x + u$
- Note that this model specification is an assumption
- What we mean by writing this type of specification?
    - x is affecting y linearly, **but**
    - A host of other factors, captured by u are also affecting

Population Model: Assumption 1

Assumption

In population, $\mathbb{E}(u) = 0$*.*

- This is an innocuous assumption

$$\begin{aligned} y &= \beta_0 + \beta_1 x + u \\ &\equiv \beta_0 + \alpha_0 + \beta_1 x + u - \alpha_0 \end{aligned}$$

- Note that changing intercept has no effect on β_1

Population Model: Assumption 2

Assumption

$\mathbb{E}(u|x) = \mathbb{E}(u)$ *for all values of* x

- Crucial assumption
- Not verifiable. Why?

Example

Let's say y is wage, and x is years of schooling. What we don't observe is the ability of a person, which is subsumed in u. Essentially, with Assumption 2 we are saying:

$$\mathbb{E}(u|x = 1) = \mathbb{E}(u|x = 16)$$

Population Model: Assumption 2

Assumption

$\mathbb{E}(u|x) = \mathbb{E}(u)$ *for all values of* x

- Crucial assumption
- Not verifiable. Why?

Example

Let's say y is wage, and x is years of schooling. What we don't observe is the ability of a person, which is subsumed in u. Essentially, with Assumption 2 we are saying:

$$\mathbb{E}(u|x = 1) = \mathbb{E}(u|x = 16)$$

- If both assumptions hold,

$$\mathbb{E}(u|x) = \mathbb{E}(u) = 0$$

- Conditional expectation function [Very important concept. Coming back to it later.]

$$\mathbb{E}(y|x) = \beta_0 + \beta_1 x$$

Ordinary Least Squares (OLS)

- **Question:** We have data on x and y. How can we estimate β_0 and β_1?
- We have the following model

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + u_i, \quad i = \{1, \cdots, n\} \\ \mathbb{E}(u) &= 0 \\ \mathbb{E}(u|x) &= 0 \end{aligned}$$

- We **observe** y's and x's, but u is never going to be observed

Lemma

$\mathbb{E}(u \mid x) = 0$ *implies* $\mathbb{E}(ux) = 0$

Proof.

$$\begin{aligned} \mathbb{E}(ux) &= \mathbb{E}[\mathbb{E}(ux \mid x)] \\ &= \mathbb{E}[x\mathbb{E}(u \mid x)] \\ &= 0 \end{aligned}$$

Ordinary Least Squares (OLS)

- We have two unknowns – β_0, β_1 – and two equations

$$\mathbb{E}(u) = 0$$
$$\mathbb{E}(ux) = 0$$

- We can further write

$$\mathbb{E}(y - \beta_0 - \beta_1 x) = 0$$
$$\mathbb{E}(x[y - \beta_0 - \beta_1 x]) = 0$$

- Let's say we have n observations, indexed by i, (y_i, x_i)
- Sample counterpart of two conditions is

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right) = 0$$
$$\frac{1}{n}\sum_{i=1}^{n}\left(x_i\left[y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right]\right) = 0$$

- Notice the switch from β_0, β_1 to $\widehat{\beta}_0, \widehat{\beta}_1$

Ordinary Least Squares

- Solution of two equations will give

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

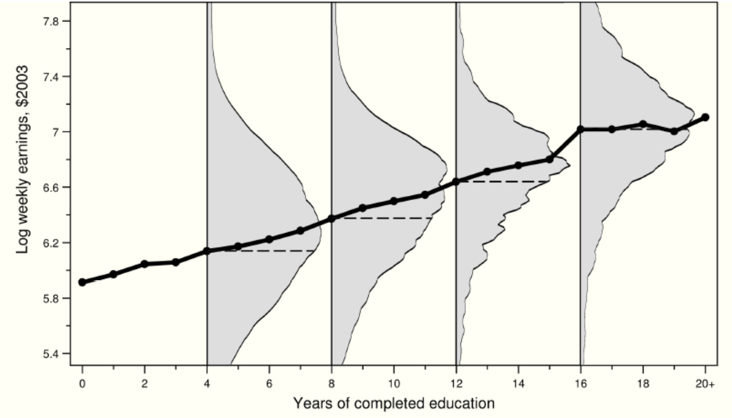
$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

- You cannot identify $\widetilde{\beta}_1$ if $\sum_{i=1}^{n}(x_i - \overline{x})^2 = 0$
  - When will this be the case?

Conditional Expectation Function (CEF)

- CEF is given as $\mathbb{E}[Y_i|X_i]$ where X_i is a $K \times 1$ vector of covariates
- Interpretation: population average of Y_i keeping X_i fixed
- Population average: mean in an infinitely large sample
- Note that expectation is a population concept.
  - What is the difference between population and sample?

Conditional Expectation Function (CEF)



Conditional Expectation Function (CEF)

- CEF at $X_i = x$ is given as

$$\mathbb{E}[Y_i|X_i = x] = \int t f_y(t|X_i = x)dt$$

- Using the law of iterated expectations, the unconditional average of Y_i can be derived as the unconditional average of CEF

$$\mathbb{E}[Y_i] = \mathbb{E}\left\{\mathbb{E}[Y_i|X_i]\right\}$$

  where the outer expectation is on the distribution of X_i

Proof.
Assume (X_i, Y_i) are continuously distributed with $f_{xy}(u,t)$ where $f_y(t|X_i = u)$ is the conditional distribution of Y_i given $X_i = u$ and $g_y(t)$ and $g_x(u)$ are marginal densities $\square$

Conditional Expectation Function (CEF)

$$
\begin{aligned}
\mathbb{E}\left\{\mathbb{E}[Y_i|X_i]\right\} &= \int \mathbb{E}[Y_i|X_i = u]g_x(u)\,\mathrm{d}\,u \\
&= \int \left[\int tf_y(t|X_i = u)\,\mathrm{d}\,t\right] g_x(u)\,\mathrm{d}\,u \\
&= \int \int tf_y(t|X_i = u)g_x(u)\,\mathrm{d}\,u \\
&= \int t \left[\int f_y(t|X_i = u)g_x(u)\,\mathrm{d}\,u\right] \mathrm{d}\,t \\
&= \int t\,[f_{xy}(u,t)\,\mathrm{d}\,u]\,\mathrm{d}\,t \\
&= \int tg_y(t)\,\mathrm{d}\,t \\
&= \mathbb{E}[Y_i]
\end{aligned}
$$

Three Theorems of CEF

1. The CEF Decomposition Property.

$$Y_i = \mathbb{E}[Y_i|X_i] + \varepsilon_i$$

where (i) ε_i is mean dependent of X_i i.e. $\mathbb{E}[\varepsilon_i|X_i] = 0, and(ii)\varepsilon_i$ is uncorrelated with any function of X_i

Proof.

For the first point:

$$\begin{aligned}\mathbb{E}[\varepsilon_i|X_i] &= \mathbb{E}[Y_i - \mathbb{E}[Y_i|X_i]|X_i] \\ &= \mathbb{E}[Y_i|X_i] - \mathbb{E}[Y_i|X_i] = 0\end{aligned}$$

For the second point let $h(X_i)$ be any function of X_i. Then

$$\begin{aligned}\mathbb{E}[h(X_i)\varepsilon_i] &= \mathbb{E}\left\{\mathbb{E}[h(X_i)\varepsilon_i|X_i]\right\} \\ &= \mathbb{E}\left\{\mathbb{E}[h(X_i)]\mathbb{E}(\varepsilon_i|X_i)\right\} \\ &= 0\end{aligned}$$

Three Theorems of CEF

- Intuitively, Theorem 1 says that any random variable Y_i can be decomposed into two parts
  1. one which is explained by X_i i.e. $\mathbb{E}[Y_i|X_i]$
  2. and the left over piece which is orthogonal to X_i by construction

Three Theorems of CEF

2. The CEF Prediction Property. *Let $m(X_i)$ be any function of X_i. The CEF solves*

$$\mathbb{E}[Y_i|X_i] = \arg min_{m(X_i)} \mathbb{E}[(Y_i - m(X_i))^2]$$

Hence CEF is the minimum mean square estimator of Y_i given X_i

Proof.

$$\begin{aligned}[Y_i - m(X_i)]^2 &= ([Y_i - \mathbb{E}(Y_i|X_i)] + [\mathbb{E}(Y_i|X_i) - m(X_i)])^2 \\ &= (Y_i - \mathbb{E}[Y_i|X_i])^2 + 2(\mathbb{E}[Y_i|X_i] - m(X_i))(Y_i - \mathbb{E}[Y_i|X_i]) + \\ &\qquad (\mathbb{E}[Y_i|X_i] - m(X_i))^2\end{aligned}$$

The second term can be written as $h(X_i)\varepsilon_i$ where $h(X_i) = 2(\mathbb{E}[Y_i|X_i] - m(X_i))$ which will have expectation zero by Theorem 1. The last term is zero when $m(X_i) = \mathbb{E}[Y_i|X_i]$ □

Three Theorems of CEF

3. ANOVA Theorem.

$$\text{var}[Y_i] = \text{var}[\mathbb{E}(Y_i|X_i)] + \mathbb{E}[\text{var}(Y_i|X_i)]$$

Proof.

By theorem 1:

$$Y_i = \mathbb{E}[Y_i|X_i] + \varepsilon_i$$
$$\text{var}(Y_i) = \text{var}(\mathbb{E}[y_i|X_i]) + \text{var}(\varepsilon_i)$$

Now $\text{var}(\varepsilon_i)$ can be written as

$$\begin{aligned}\text{var}(\varepsilon_i) &= \mathbb{E}[\varepsilon_i^2] \\ &= \mathbb{E}[\mathbb{E}(\varepsilon_i^2|X_i)] \\ &= \mathbb{E}[\mathbb{E}([Y_i - \mathbb{E}(Y_i|X_i)]^2|X_i)]\end{aligned}$$

Regression and CEF

- Regression function: the best fitting line generated by minimizing expected square errors
- So what is the relationship between the regression function (which has a restricted functional form) and the CEF (which does not)

Regression Anatomy

- Let β be a $K \times 1$ vector of regression coefficients obtained by solving

$$\beta = \arg\min_{b} \quad \mathbb{E}[(Y_i - X_i' b)^2] \tag{1}$$

- The first order condition for this problem is

$$\mathbb{E}[X_i(Y_i - X_i' b)] = 0$$

  which gives $\beta = \mathbb{E}[X_i X_i']^{-1} \mathbb{E}[X_i Y_i]$
- In a simple bivariate case i.e. $K = 1$, slope coefficient is $\beta_1 = \frac{\text{cov}(Y, X_i)}{\text{var}(X_i)}$ and constant is given as $\alpha = \mathbb{E}[Y_i] - \beta_1 \mathbb{E}[X_i]$
- For $K > 1$, the k^{th} coefficient is given as

$$\beta_k = \frac{\text{var}(Y_i, \widetilde{x}_{ki})}{\text{var}(\widetilde{x}_{ki})} \tag{2}$$

  where $\widetilde{x}_{ki}$ is the residual from a regression of x_{ki} on all other covariates

Regression Anatomy

- Regression anatomy formula (2) is more intuitive than matrix notation (1)
  - Each coefficient in a multivariate regression is the bivariate slope coefficient for the corresponding regressor after partialling out all other covariates
- Now we move onto the
  point which Mostly Harmless Econometrics captures very nicely

  > *you should be interested in the regression parameters if you are interested in the CEF*

Three theorems (without proofs)

1. The Linear CEF Theorem [Regression Justification 1].
Suppose the CEF is linear. Then the population regression function is it.

- This means if the CEF is linear then it will be captured by the regression function.
- Question is: when CEF is linear?
  1. When vector (Y_i, X_i') has a multivariate normal distribution
  2. When regression is saturated:
     - a saturated regression model has a separated parameter for every possible combination of values that the set of regressors take on

Three Theorems (without proofs)

2. The Best Linear Predictor Theorem [Regression Justification 2].

The function $X_i'\beta$ is the best linear predictor of Y_i given X_i in a MMSE sense.

- Among **all** functions, CEF $E[Y_i|X_i]$ is the best predictor of Y_i given X_i
- Among **linear** functions, the regression function $X_i'\beta$ is the best predictor of Y_i given X_i

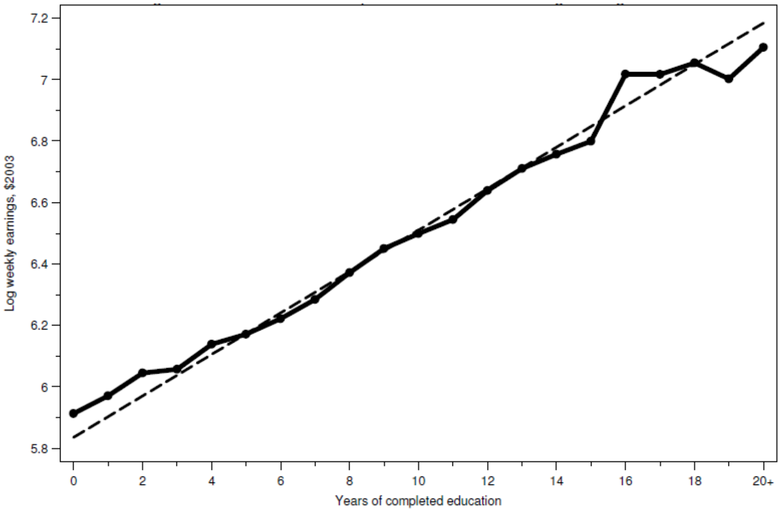
Three Theorems (without proofs)

3. The Regression CEF Theorem [Regression Justification 3].
The regression function provides minimum mean square error (MMSE) approximation to $\mathbb{E}[Y_i|X_i]$*, that is,*

$$\beta = \arg\min_b \mathbb{E}[(E[Y_i|X_i] - X_i'b)^2]$$

- This theorem says that even if actual CEF is non-linear, regression provides the best linear approximation to it

CEF and Regression



Sample is limited to white men, age 40-49. Data is from Census IPUMS 1980, 5% sample.

Regression and Causality

- From previous discussion: regression gives the best MMSE linear approximation to the CEF
- However, we still don't know when regression has a causal interpretation
- Let's go to the example of earnings and education
- Assume that the schooling is a binary decision:
  - Going to the college, $C_i = 1$
  - Not going to the college, $C_i = 0$
- Potential outcome

$$= \begin{cases} Y_{1i} & \text{if } \; C_i = 1 \\ Y_{0i} & \text{if } \; C_i = 0 \end{cases}$$

Regression and Causality

- Blind comparison of earnings of college goers (not-goers) will lead to

$$\begin{aligned}
\mathbb{E}[Y_i|C_i = 1] - \mathbb{E}[Y_i|C_i = 0] &= \mathbb{E}[Y_{1i}|C_i = 1] - \mathbb{E}[Y_{0i}|C_i = 0] \\
&= \underbrace{\mathbb{E}[Y_{1i}|C_i = 1] - \mathbb{E}[Y_{0i}|C_i = 1]}_{\text{ATET}} + \\
&\quad \underbrace{\mathbb{E}[Y_{0i}|C_i = 1] - \mathbb{E}[Y_{0i}|C_i = 0]}_{\text{selection bias}}
\end{aligned}$$

- Here selection bias is positive, why?

Conditional Independence Assumption (CIA)

- How can we eliminate the selection bias?
- Invoke the conditional independence assumption (CIA)
- CIA asserts that conditional on observed characteristics X_i, selection bias disappears

$$\{Y_{0i}, Y_{1i}\} \perp\!\!\!\perp C_i | X_i$$

- Hence,

$$\begin{aligned}\mathbb{E}[Y_i|X_i, C_i = 1] - \mathbb{E}[Y_i|X_i, C_i = 0] &= \mathbb{E}[Y_{1i}|X_i, C_i = 1] - \mathbb{E}[Y_{0i}|X_i, C_i = 0] \\ &= \mathbb{E}[Y_{1i} - Y_{0i}|X_i]\end{aligned}$$

Causal Interpretation under continuous variable

- Denote $Y_{si} = f_i(s)$ as the **potential** earnings that person i would receive for s years of education
- The CIA will become

$$Y_{si} \perp\!\!\!\perp s_i | X_i$$

- Hence

$$\begin{aligned}
&\Rightarrow \mathbb{E}[Y_i|X_i, s_i = s] - \mathbb{E}[Y_i|X_i, s_i = s-1] \\
&\Rightarrow \mathbb{E}[Y_{si}|X_i, s_i = s] - \mathbb{E}[Y_{(s-1)i}|X_i, s_i = s-1] \\
&\Rightarrow \mathbb{E}[f_i(s) - f_i(s-1)|X_i]
\end{aligned}$$

Regression and CIA

- Regression provides a way to turn CIA into causal estimate
- For now, assume, $f_i(s)$ is both linear in s and same for everyone except for an additive error term
- In this case, linear constant effects causal model is given as

$$f_i(s) = \alpha + \rho s + \eta_i$$

- Two points to note:
  1. functional relationship is same for everyone, except for η_i
  2. this equation is a causal model in the sense that it is relating s to potential outcomes
- If we replace s with observed value we will get

$$Y_i = \alpha + \rho s_i + \eta_i$$

- Due to selection bias, s_i and η_i may be correlated

Regression and CIA

- Suppose CIA holds given a vector of observed co-variates X_i
- Decompose the random part of potential earnings η_i into a linear function of observable characteristics X_i and an error term ν_i

  $$\eta_i = X_i'\gamma + \nu_i$$

  such that $\mathbb{E}[\eta_i|X_i] = X_i'\gamma$
- Then

$$\begin{aligned}
\mathbb{E}[f_i(s)|X_i, s_i] &= \mathbb{E}[f_i(s)|X_i] \\
&= \alpha + \rho s + \mathbb{E}[\eta_i|X_i] \\
&= \alpha + \rho s + X_i'\gamma
\end{aligned}$$

Regression and CIA

- Hence, residual in the linear causal model,

$$Y_i = \alpha + \rho s + X_i'\gamma + \nu_i$$

  is uncorrelated with s_i and X_i and ρ is the causal effect of interest
- Take a note of an important assumption: X_i is the reason why s_i and η_i are uncorrelated
- What if some observable characteristics are missing?
  - If missing characteristic correlated with inlcuded $X_i \Rightarrow$ omitted variable bias