Linear Regression

Po-Chun Huang

National Chengchi University

April 9, 2021

Matching, Regression and Omitted Variable Bias

- Randomized controlled trial is the gold standard for causal inference, but it is rare.
- More often than not, we have only observational data, where treatment is not randomly assigned.
- In the example of return to education, comparison of those who do and don't go to college are likely to be a poor measure of the causal effect of college attendance. Why?
  - Students who go to college tend to be smarter in the first place
  - Even if smarter students didn't go to college, they might earn higher wages anyway
  - If so, we will observe students who go to college earn higher earnings than those who don't even if the return to education is zero.

Matching, Regression and Omitted Variable Bias

- The above example suggests we can estimate the return to college for students who are of the same intelligence.
- This comparison identifies the return to college attendance if "conditional independence" or "selection on observable" assumption holds:

  students who are equally smart are comparable to each other
- There are two ways to implement this estimator
  1. Matching: Take every member of the treatment group and match them to a member of the control group based on X
  2. Regression: Add X as control variables

Conditional Independence Assumption (CIA)

- The CIA asserts that conditional on observed characteristics, X_i, treatment is independent of potential outcomes

$$\{Y_{0i}, Y_{1i}\} \amalg D_i | X_i$$

- Therefore, selection bias disappears

$$E(Y_0|D=1,X) - E(Y_0|D=0,X)$$

$$= E(Y_1|D=1,X) - E(Y_1|D=0,X) = 0$$

Conditional Independence Assumption (CIA)

- CIA ensures $E(Y_0|D=1,X) - E(Y_0|D=0,X) = 0$

$$E(Y|D=1,X) - E(Y|D=0,X)$$

$$=E(Y_1|D=1,X) - E(Y_0|D=0,X)$$

$$=E(Y_1|D=1,X) - E(Y_0|D=1,X) + [E(Y_0|D=1,X) - E(Y_0|D=0,X)]$$

$$=E(Y_1 - Y_0|D=1,X) + [E(Y_0|D=1,X) - E(Y_0|D=0,X)]$$

$$=E(Y_1 - Y_0|D=1,X)$$

$$=E(Y_1 - Y_0|X)$$

The Return of Attending a Private College

- Students who went to public universities paid less than $9,000.
- Those who went to private colleges pay $29,000 per year in tuition and fees. Is it worthy?
- Comparison of earnings between these two groups of students reveal large gaps in favor of elite-college alumni.
- But, students who went to elite colleges also have better high school grades and SAT scores, and are more motivated.

Dale and Krueger (2002)

- Dale and Krueger (2002), "Estimating the Payoff to Attending a More Selective College," Quarterly Journal of Economics
- Since college attendance decisions are not randomly assigned, we must control for all factors that determine both attendance decsions and later earnings.
- There are too many factors to to control for.
- Instead of identifying everything that might matter for college choice and earnings, they work with a key summary measure: the characteristics of colleges to which students applied and admitted.

Dale and Krueger (2002)

Regression and Causality

Identification

- Start with the assumption that one variable y is a linear function of x plus an error term

$$y = \beta_0 + \beta_1 x + u$$

  - x: independent variable (e.g. years of education)
  - y: dependent variable (e.g. salary of age 40)
  - u: all the stuff that affect y besides x

- β_1 is the effect of a one-year increase in years of education on person i's salary of age 40 holding other things (in u) fixed

Regression and Causality

Identification

- Our goal is to figure out what we can do about β_1 (and β_0).
- Write

$$
\begin{aligned}
& Cov(x, y) \\
= & Cov(x, \beta_0 + \beta_1 x + u) \\
= & Cov(x, \beta_0) + \beta_1 Cov(x, x) + Cov(x, u) \\
= & Cov(x, \beta_0) + \beta_1 Var(x) + Cov(x, u).
\end{aligned}
$$

- Then,

$$\frac{Cov(x, y)}{Var(x)} = \beta_1 + \frac{Cov(x, u)}{Var(x)}$$

- Therefore, $\frac{Cov(x,y)}{Var(x)}$ identifies β_1 iff $Cov(x, u) = 0$

Regression and Causality

Estimation

- OK. We know $\frac{Cov(x,y)}{Var(x)}$ identifies β_1 iff $Cov(x, u) = 0$
- But, we don't have access to the whole population, so we have no idea what β_1 are.
- Estimate $\frac{Cov(x,y)}{Var(x)}$ using its sample counterpart:

$$\frac{\frac{1}{n}\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_1^n (x_i - \bar{x})^2}$$

- With random sampling,

$$\frac{\frac{1}{n}\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_1^n (x_i - \bar{x})^2} \to \frac{Cov(x, y)}{Var(x)}$$

Regression and Causality

Estimation

- $\frac{\frac{1}{n}\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_1^n (x_i - \bar{x})^2}$ is ordinary least square (OLS) estimator for β_1
- Write
$$y_i = \hat{\beta}_{0,OLS} + \hat{\beta}_{1,OLS} x_i + \hat{u}_i$$
where $\hat{u}_i$ is the residual, the difference between observed value and fitted value of y_i
- OLS estimators, $\hat{\beta}_{0,OLS}$ and $\hat{\beta}_{1,OLS}$, minimize the sum of squared residuals
$$\min_{\hat{\beta}_{0,OLS}, \hat{\beta}_{1,OLS}} \sum_{i=1}^{n} \hat{u}_i^2$$
- $\hat{\beta}_{1,OLS} = \frac{\frac{1}{n}\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_1^n (x_i - \bar{x})^2}$; $\hat{\beta}_{0,OLS} = \bar{y} - \hat{\beta}_{1,OLS} \cdot \bar{x}$

Regression and Causality

Standard Error

- Homoskedasticity:

$$Var(u|x) = \sigma^2$$

- Under homoskesdaticty,

$$Var(\hat{\beta}_{1,OLS}|x) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- σ^2 is unknown. Estimate it using

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{1}^{n}\hat{u}_i^2$$

- Sampling distribution of $\hat{\beta}_{1,OLS}$:

$$t = \frac{\hat{\beta}_{1,OLS} - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} \rightarrow z$$

Regression and Causality

Hypothesis Testing and Confidence Interval

- Step 1:

$$H_0 : \beta_1 = \beta_{1,0}$$

$$H_a : \beta_1 \neq \beta_{1,0}$$

- Step 2:

$$t = \frac{\hat{\beta}_{1,OLS} - \beta_1, 0}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} \to z$$

- Step 3: Reject the null if $|t| > 1.96$
- More often than not, the null value is zero, so $\beta_{1,0} = 0$. But you should think about question—the null can be something other than zero.

- If $Var(u|x)$ varies over individuals, we call the error term exhibits heteroskesdaticity.
- e.g. heteroskesdaticity in a wage equation:

  $$wage = \beta_0 + \beta_1 edu + u, \; Var(u|x = 16) > Var(u|x = 12) > Var(u|x = 8)$$

- People with more education have a wide variety of interest and job opportunities, which leads to more wage variability.

Regression and Causality

Standard Error

- Whether $Var(u|x)$ is constant has nothing to do with the OLS estimator of β is biased or inconsistent
- So what's the problem if you have heteroskedasticity?
- The t (or F) statistics are no longer distributed as t (or F).

Regression and Causality

Standard Error

- Under heteroskedasticity,

$$Var(\hat{\beta}_{1,OLS}|x) = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})^2\sigma_i^2]}{[\sum_{i=1}^{n}(x_i - \bar{x})^2]^2}$$

- Heteroskedasticity-robust standard error (RSE) for $\hat{\beta}_{1,OLS}$:

$$\sqrt{\frac{\sum_{i=1}^{n}[(x_i - \bar{x})^2\hat{u}_i^2]}{[\sum_{i=1}^{n}(x_i - \bar{x})^2]^2}}$$

- Then,

$$t = \frac{\hat{\beta}_{1,OLS} - \beta_1}{RSE(\hat{\beta}_{1,OLS})} \to z$$

Regression and Causality

Identification

- $Cov(x, u) = 0$ does not hold in general.
- In the example of return to education, u contains ability, which increases earnings. Therefore,

$$\frac{Cov(x, u)}{Var(x)} > 0$$

- People with more schooling have more ability and they would have earned more even without the additional schooling.

Regression and Causality

Identification

- $\frac{Cov(x,y)}{Var(x)}$ identifies the additional earnings from
  - an increased schooling
  - the added ability that goes with the additional schooling
- Remember that $\hat{\beta}_{1,OLS}$ is always consistent to $\frac{Cov(x,y)}{Var(x)}$ as long as you have a random sample, but $\frac{Cov(x,y)}{Var(x)}$ is often not the causal effect of interest.
- Identification precedes estimation!

Regression and Causality
Omitted Variable Bias

- Let's say the right regression model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u,$$

  where x_1 is years of education and x_2 is ability

- AND, let's say that $Cov(x_1, u) = 0$ and $Cov(x_2, u) = 0$, which means that we could get consistent estimate of β_1 and β_2 using a regression

- What if we left x_2 out and estimate

$$y = \beta_0 + \beta_1 x_1 + u?$$

Regression and Causality

Omitted Variable Bias

- Recall that

$$\hat{\beta}_{1,OLS} \rightarrow \frac{Cov(x_1, y)}{Var(x_1)}$$

- We can write

$$\frac{Cov(x_1, y)}{Var(x_1)}$$

$$= \frac{Cov(x_1, \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u)}{Var(x_1)}$$

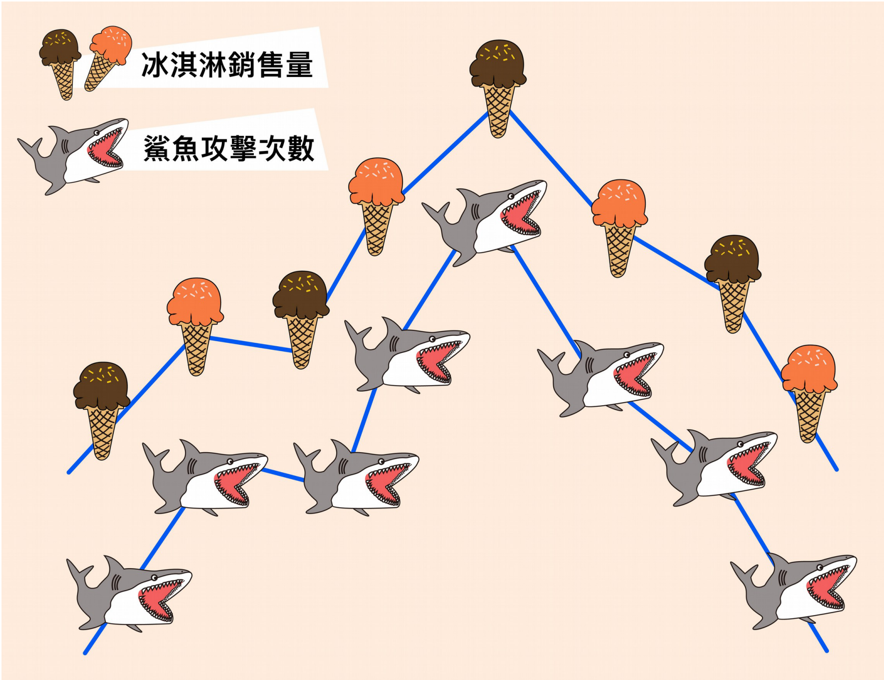
$$= \beta_1 + \beta_2 \cdot \frac{Cov(x_1, x_2)}{Var(x_1)}$$

$$= \beta_1 + \beta_2 \cdot \pi_{21}$$

- Omitted Variable Bias $= \beta_2 \cdot \frac{Cov(x_1,x_2)}{Var(x_1)}$
  - β_2: the effect of ability (x_2) on earnings (y)
  - π_{21}: the slope coefficient of the regression of ability (x_2) on years of education (x_1)
- Omitted Variable Bias = Relationship between x_2 and x_1 $\cdot$ The effect of x_2 on earnings y

Dumb Statistical Mistake 1

Source



Dumb Statistical Mistake 2

- Consider $y = \beta_0 + \beta_1 x + u$
  - y: some measure of whether you have cancer
  - x: number of times/week you brush your teeth
  - u is all the stuff that affects y besides x (e.g. smoking)
- $\hat{\beta}_1, OLS < 0$
- Does it mean brushing your teeth prevents cancer?
- $Cov(x, u) \neq 0$. Whether you brush your teeth or not is correlated with other stuff that probably influences cancer

Multiple Linear Regression

Interpretation

- Suppose $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$
  - k independent variables
  - $k + 1$ parameters
  - still one dependent variable, y and one error term, u
- β_k: the effect of x_k on $E(y)$ holding the other $k - 1$ x's constant

Multiple Linear Regression ($k = 2$)

Interpretation

- $E(y|X) = \beta_0 + \beta_1 D + \beta_2 IQ$
- $E(y|D = 1, IQ = 160) = \beta_0 + \beta_1 + \beta_2 \cdot 160$,
  $E(y|D = 0, IQ = 160) = \beta_0 + \beta_2 \cdot 160$
- $E(y|D = 1, IQ = 160) = \beta_0 + \beta_1 + \beta_2 \cdot 160 - E(y|D = 0, IQ = 160) = \beta_0 + \beta_2 \cdot 160 = \beta_1$
- β_1 in the multiple regression measures the average earnings of people who went to college, relative to people who didn't go, but were of the same intelligence

Multiple Linear Regression

Estimation

- How do you estimate $\beta_1, \beta_2, ..., \beta_k$?
- Again, minimize sum of squared residuals:

$$\min \sum_{i=1}^{n} \hat{u}_i^2$$

- OLS estimator for β_k

$$\hat{\beta}_{k,OLS} = \frac{\hat{Cov}(\tilde{x}_k, y)}{\hat{Var}(\tilde{x}_k)},$$

where $\tilde{x}_k$ is the residual from a regression of x_{ki} on $k-1$ other covariates in the model

Multiple Linear Regression

Standard Error

- A valid estimator for $Var(\hat{\beta}_{k,OLS})$

$$Var(\hat{\beta}_{k,OLS}) = \frac{\sum_{i=1}^{n} \hat{r}_{ij}^2 \hat{u}_i^2}{SSR_j^2}$$

where $\hat{r}_{ij}$ denotes the *ith* residual from regressing x_j on all other independent variables, and SSR_j is the sum of squared residuals from this regression

- Robust standard error for $\hat{\beta}_{k,OLS}$:

$$RSE(\hat{\beta}_{k,OLS}) = \sqrt{\frac{\sum_{i=1}^{n} \hat{r}_{ij}^2 \hat{u}_i^2}{SSR_j^2}}$$

- The effect of adding more covariates on $RSE(\hat{\beta}_{k,OLS})$ is ambiguous

Multiple Linear Regression

R^2...

- R^2: the percentage of variance of y that can be explained by the model

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})}{\sum_{i=1}^{n} (y_i - \bar{y})} = 1 - \sum_{i=1}^{n} \hat{u}_i^2 = Corr^2(y, \hat{y})$$

- R^2 always increases when more independent variables are included
- More variables are included, less varaition left in the error term
- A high R^2 doe not mean the regression has a causal interpretation!
- A low R^2 does not mean the regression is useless!

Regression and Matching

- Regression estimands can be viewed as matching estimators.
- They differ only in the weights used to sum the covariate-specific effects, X into a single effect.
- Matching uses the distribution of covariates among the treated to weight covariate-specific estimates into an estimate of the effect of treatment on the treated, while regression produces a variance-weighted average of these effects.

Dale and Krueger (2002)

Dale and Krueger (2002)

Dale and Krueger (2002)

- Dale and Krueger (2002)'s within-group estimates suggest that much of the shortfall in earnings for public school attendants is unrelated to students' college attendance decisions.
- Rather, the cross-group differential is explained by a combination of ambition and ability, as reflected in application decisions and the set of schools to which students were admitted.