ECON 340
Economic Research Methods

Div Bhagia

Lecture 16: Prediction vs. Causal Inference

Ordinary Least Squares (OLS)

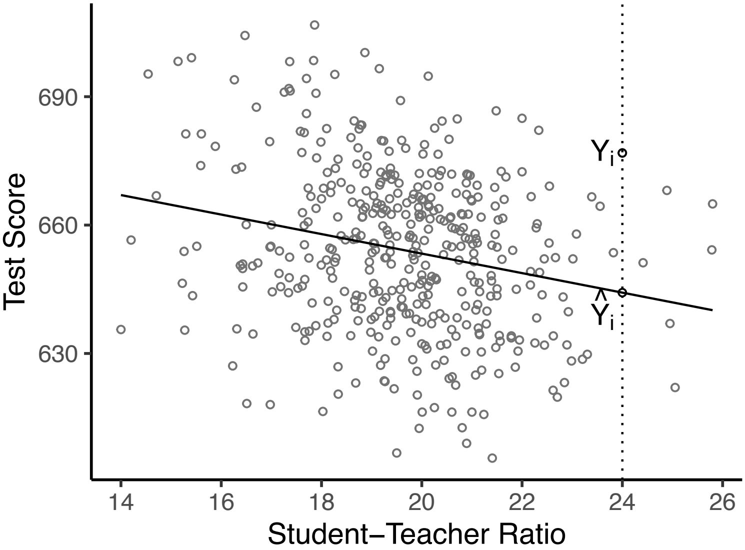
What is the main goal of Ordinary Least Squares (OLS)?

(a) Choose the line that passes through as many data points as possible

(b) Choose the values for slope and intercept that minimize the sum of squared residuals

(c) Choose the line that minimizes the absolute distance between the predicted values and data

Ordinary Least Squares (OLS)



Best fit line minimizes the sum of squared errors:

$$\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Fitted line:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X$$

Goodness of Fit: The R^2

Total Sum of Squares: $TSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$

Explained Sum of Squares: $ESS = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$

Residual Sum of Squares: $RSS = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}\hat{u}_i^2$

$$TSS = ESS + RSS$$

A measure of goodness of fit:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

Goodness of Fit: The R^2

Are the following statements true or false?

(a) R^2 ranges from 0 to 1.

(b) A higher R^2 indicates that the regression line is a better fit.

(c) A higher R^2 indicates that X explains a large percent of variation in Y.

(d) If the slope $\hat{\beta}_1 = 0$, then $R^2 = 1$.

How to interpret the coefficients?

Fitted line:

$$\widehat{testscr} = 698.93 - 2.28 \cdot str$$

How to interpret the coefficients?

Fitted line:

$$\widehat{testscr} = 698.93 - 2.28 \cdot str$$

- Intercept: Predicted test score is 698.93 for a school with $str = 0$. (Doesn't always make sense!)
- Slope: One more student per teacher lowers the predicted test score by 2.28. How?

  Alternatively: Schools in our sample that had one more student per teacher on average had an average test score that was 2.28 points lower.

Two Different Questions

- I am trying to figure out what are the test scores for a particular school, but I can only observe it's class size. If my linear model captures the data well, I could use it to *predict* the test score for this school.

- But now, what if the Department of Education wants to know whether reducing class size across schools will *lead* to an improvement in test scores. Can my model answer this question?

Two Different Questions

- First question concerns *prediction*: using the observed value of some variable to predict the value of another variable
- The second concerns *causal inference*: using data to estimate the *effect* of changes in one variable on another variable
- To attach a causal interpretation to β_1, we need additional assumptions

Simple Linear Regression Model

Assumption 1 (Linearity): The relationship between X and Y is given by:

$$Y = \beta_0 + \beta_1 X + u$$

Here, u is the mean zero error term, $E(u) = 0$.

There is a linear (in parameters) relationship between X and Y with some error that is on average zero.

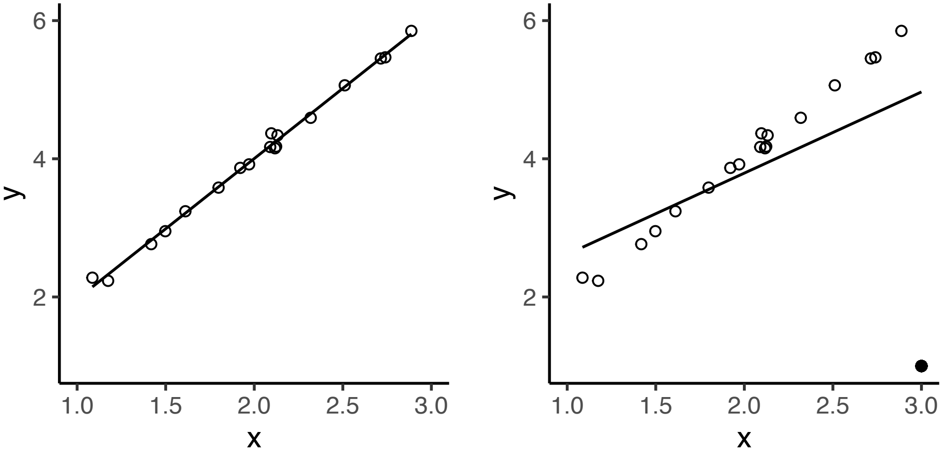
Can think of u as the impact of omitted factors on Y.

Assumptions for Causal Inference

Assumption 2 (Random Sample): The observed data (Y_i, X_i) for $i = 1, 2, ..., n$ represent a random sample of size n from the above population model.

Assumption 3 (No large outliers): Fourth moments (or Kurtosis) of X and Y are finite.

Why we don't want outliers



Assumptions for Causal Inference

Assumption 4 (Mean Independence/Exogeneity): The expected value of the error term is the same conditional on any value of the explanatory variable.

$$E(u|X) = E(u) = 0$$

This assumption is crucial for attaching a causal interpretation to our regression coefficients.

Reminder: Independence and Uncorrelatedness

- Two random variables are *independent* if $f(y|x) = f(y)$ for all x and y or equivalently $E(Y|X) = E(Y)$.
- Two random variables are *uncorrelated* if the correlation between them is 0.
- Independence $\rightarrow$ uncorrelatedness, if two variables are independent then they are uncorrelated as well

The Exogeneity Assumption

$$Y = \beta_0 + \beta_1 X + u \qquad \mathit{Exogeneity} : E(u|X) = E(u) = 0$$

- Omitted factors do not dependent on values of X
- In other words, the error term is uncorrelated with the independent variable X
- Why do we need this assumption to attach a causal interpretation to β_1?

When the exogeneity assumption fails

$$Y = \beta_0 + \beta_1 X + u$$

- Y: test scores, X: class-size, u : teacher quality
- If schools with higher student-teacher ratios have worse teachers ($\uparrow X, \downarrow u$)
- Then, if we see test scores decline with class size ($\uparrow X, \downarrow Y$), hard to say if it's due to teacher quality or class size.

The Exogeneity Assumption

$$Y = \beta_0 + \beta_1 X + u$$

Let's take the expectation of Y conditional on X:

$$E(Y|X) = \beta_0 + \beta_1 X + E(u|X)$$

If the exogeneity assumption holds, $E(u|X) = 0$, then

$$E(Y|X) = \beta_0 + \beta_1 X$$

If X increases by 1 unit, on average, Y increases by β_1.

$$\beta_1 = E(Y|X = x + 1) - E(Y|X = x)$$

When the exogeneity assumption fails

$$E(Y|X = x) = \beta_0 + \beta_1 x + E(u|X = x) \qquad (1)$$
$$E(Y|X = x + 1) = \beta_0 + \beta_1(x + 1) + E(u|X = x + 1) \qquad (2)$$

Subtracting equation (1) from (2):

$$E(Y|X = x+1) - E(Y|X = x) = \beta_1 + [E(u|X = x+1) - E(u|X = x)]$$

So, in this case:

$$\beta_1 = \underbrace{[E(Y|X = x + 1) - E(Y|X = x)]}_{\text{Impact of } X \text{ on } Y} - \underbrace{[E(u|X = x + 1) - E(u|X = x)]}_{\text{Confounding effect of } u}$$

Linear Regression Model

Assumptions 1-4 imply that:

1. OLS estimators are unbiased, that is

$$E(\hat{\beta}_0) = \beta_0, \quad E(\hat{\beta}_1) = \beta_1$$

2. In large samples, OLS estimators are normally distributed due to the Central Limit Theorem (CLT)

Sampling Distribution for OLS Estimators

Under Assumptions 1-4, in large samples ($n > 100$),

$$\hat{\beta}_0 \sim N(\beta_0, \sigma^2_{\hat{\beta}_0}), \qquad \hat{\beta}_1 \sim N(\beta_1, \sigma^2_{\hat{\beta}_1})$$

where

$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n} \frac{Var[(X_i - \mu_X)u_i]}{Var(X_i)}$$

Sampling Distribution for OLS Estimators

Under Assumptions 1-4, in large samples ($n > 100$),

$$\hat{\beta}_0 \sim N(\beta_0, \sigma^2_{\hat{\beta}_0}), \qquad \hat{\beta}_1 \sim N(\beta_1, \sigma^2_{\hat{\beta}_1})$$

where

$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n} \frac{Var[(X_i - \mu_X)u_i]}{Var(X_i)}$$

Can you think why the variance of $\hat{\beta}_1$ decreases as the variance of X increases?

Variance of $\hat{\beta}_1$ and X