

[Lesson1.pdf](#)

[Lesson10.pdf](#)

[Lesson11.pdf](#)

[Lesson12.pdf](#)

[Lesson13.pdf](#)

[Lesson13_examples.pdf](#)

[Lesson14.pdf](#)

[Lesson15.pdf](#)

[Lesson16.pdf](#)

[Lesson17.pdf](#)

[Lesson18.pdf](#)

[Lesson19.pdf](#)

[Lesson2.pdf](#)

[Lesson20.pdf](#)

[Lesson21.pdf](#)

[Lesson3.pdf](#)

[Lesson4.pdf](#)

[Lesson5.pdf](#)

[Lesson6.pdf](#)

[Lesson7.pdf](#)

[Lesson8.pdf](#)

[Lesson9.pdf](#)

[Midterm 1 Solutions.pdf](#)

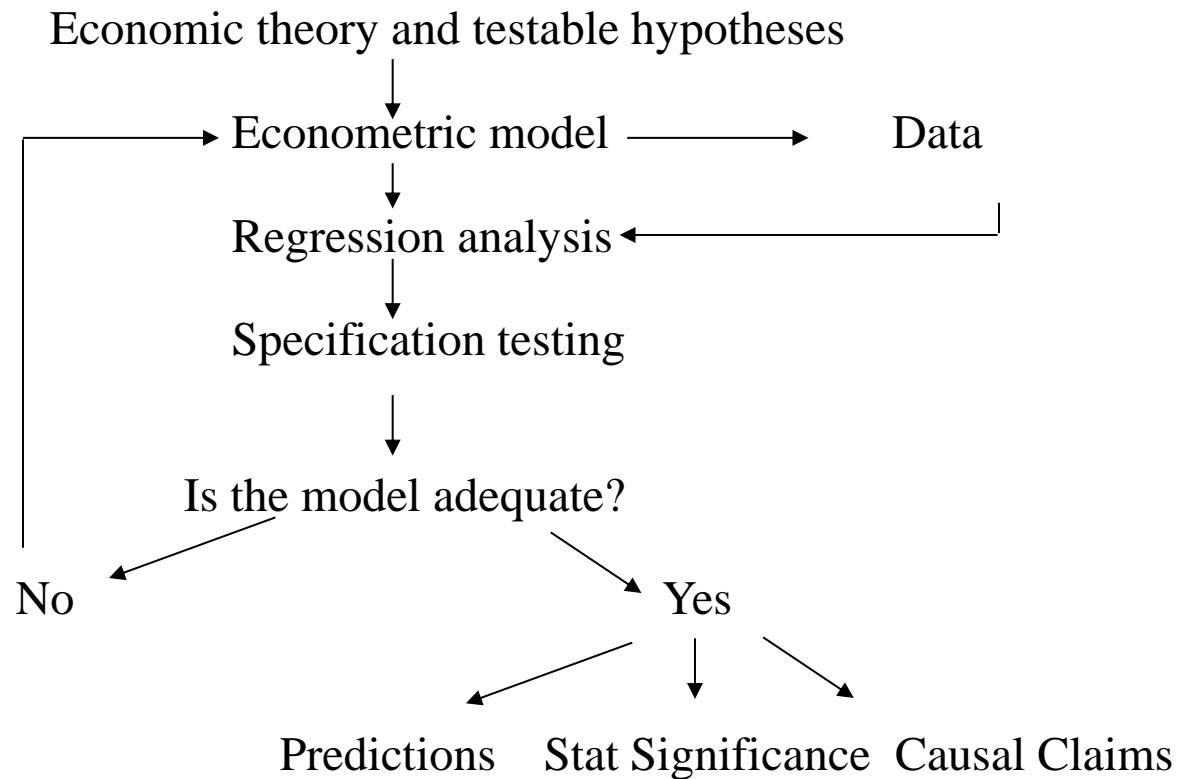
Lesson 1

Introduction to Econometrics

Econometrics Defined

“Econometrics is the application of mathematics, statistical methods, and computer science to economic data.”

Stages of Applied Econometric Analysis



Econometrics Defined

1. Recurring Examples
2. Process:
 - a. Economic Theory
 - b. Econometric Model
 - c. Acquire Data
 - d. Regression Analysis
3. Objectives:
 - a. Test Hypotheses
 - b. Make Predictions
 - c. Make Causal Claims
 - d. Compare Models

Recurring Examples

1. Demand Function – Examine how the quantity purchased depends on the price of a good in period t. What kind of hypotheses might we have about a_1 ?

$$q_t = a_0 + a_1 p_t + u_t$$

2. Cobb-Douglas Production – Examines how firm output depends on the amount of labor and capital used. How can we make this linear?

$$Y_i = AK_i^\alpha L_i^\beta$$

3. Okun's Law – Examines the relationship between the unemployment rate and a country's Gross National Product. What sign do we expect b to have?

$$\Delta U_t = a + b\Delta GNP_t + u_t$$

Recurring Examples

4. Keynesian Consumption – Examines how consumption depends on disposable income. What is disposable income?

$$C_i = a + \delta Y_i^d$$

5. Wages and Education – Examines how future income depends on years of education attained. What is Exp? Why use $\ln(Y)$?

$$\ln Y_{it} = a_0 + rS_i + \beta Exp_i$$

6. Hedonic Pricing – Value various community and environmental characteristics using house price. What environment characteristics might we want to value? Why is their value/price not obvious?

$$P_i = a + \sum \alpha_c X_{ic} + \beta Env_i$$

Theory & Econometric Model

Theory:

Our hypothesis is that as the price of a car increases, the quantity demanded will decrease. In economics we call this our “prior”.

Econometric Model:

$$q_t = a_0 + a_1 p_t + u_t$$

Testable hypothesis:

$$a_1 < 0$$

Note: We could have a more specific hypothesis. For example, for each 1,000 increase in price we predict that the number of cars demanded will decrease by 100,000. [i.e. $a_1 = -100$]

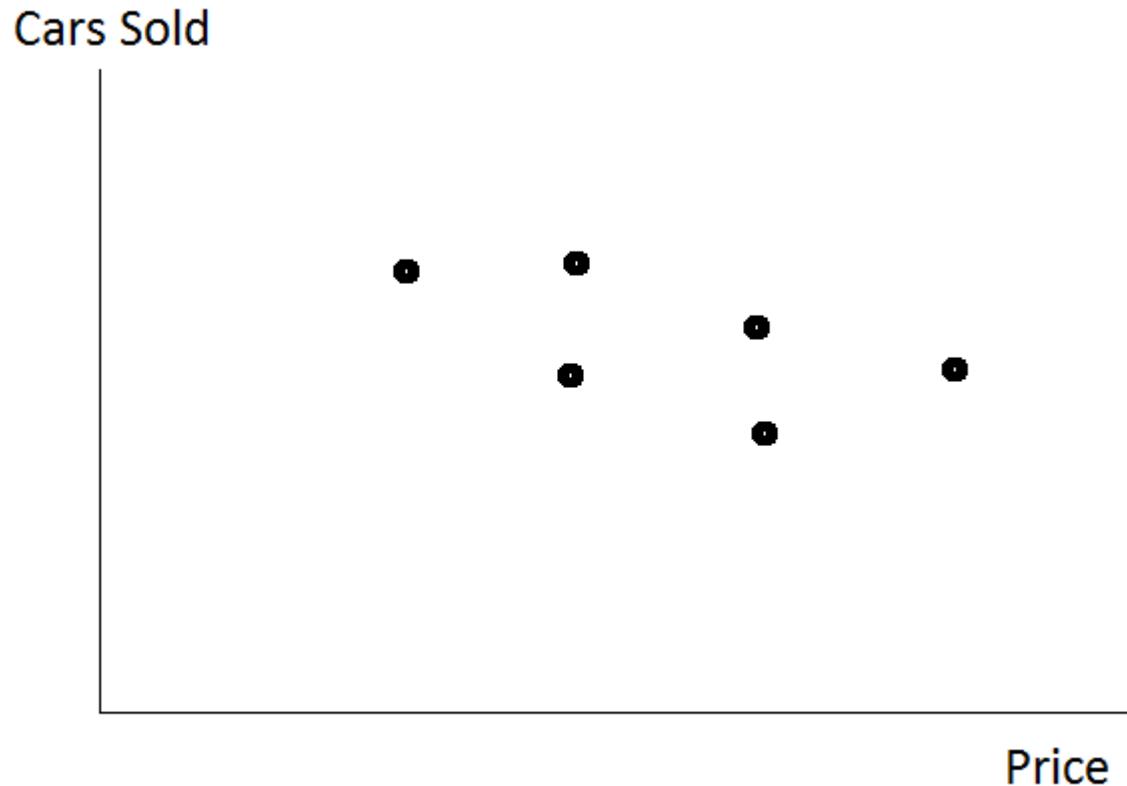
Data

What variables do we need?

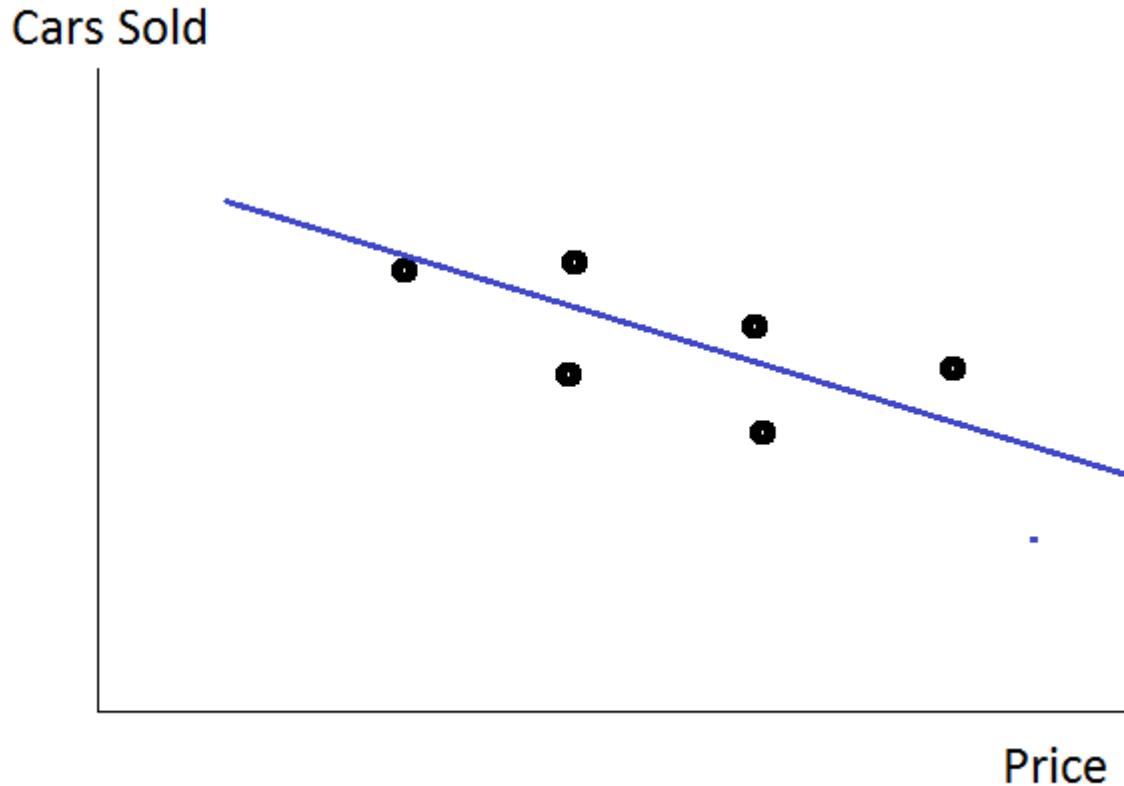
- price of the car: p
- number of cars sold: q
- time period: t

What other variables might we want to know?

Regression Analysis



Regression Analysis



Regression Analysis

Regression Results:

$$\hat{q}_t = 1,400,000 - 50p_t$$

Interpret the regression:

$$\frac{\partial \hat{q}_t}{\partial p_t} = -50$$

For each 1 dollar increase in price, the number of cars purchased decreases by 50. Equivalently, for each 1,000 increase in price, the number of cars purchased decreases by 50,000.

Note that these results might be more useful if they were in terms of percents (i.e. an x percent increase in price results in y percent reduction in demand for cars).

Specification Testing

We might think that there should be other variables in the model:

$$q_t = a_0 + a_1 p_t + a_2 GDP_t + u_t$$

Is the model adequate?

- Did the estimate of a_1 change significantly when we added GDP?
- Was a_1 statistically significant in the simple model? Is it significant in the model with GDP included?

What other variables might we want to add as a specification test?

Prediction

$$\hat{q}_t = 1,400,000 - 50p_t$$

Prediction:

- ❑ Predict the number of cars that will be sold for prices that are not in the data or that have never occurred.
- ❑ For example, if the price is 15,000 dollars, then we predict that 650,000 cars will be sold.
- ❑ How many cars do we predict will be sold if the price is 0? Does this make sense?

Make Causal Claims

- We found a negative correlation between price and quantity sold.
We would like to claim that the change in the price caused the number of cars sold to change.
- What if we had found that lower prices were associated with lower sales? Could we explain this?
- Ceteris Paribus – "all other things being equal or held constant."
- First half of the course: focus on estimation and testing
- Second half of the course: focus on causality

To do:

- Buy the textbook and read Chapters 1 & 2
- Make sure you are registered on Canvas
- Make sure you have easy access to STATA
- If you did not take the math methods course, make sure you review basic probability, statistics, calculus, and matrix algebra

Lesson 10

Omitted Variables
Measurement Error
Reverse Causality

Outline

Previous Lessons:

1. Heteroskedasticity: Detection & Correction
2. Autocorrelation: Detection & Correction

This Lesson:

1. Omitted Variables
2. Measurement Error
3. Reverse Causality

Next Lesson:

1. Functional Form
2. Simultaneous equations

Outline

1. The last few classes we have focused on issues that cause problems for inference (i.e. getting the wrong standard errors):
 - A. Heteroskedasticity
 - B. Autocorrelation
2. Now we are going to discuss problems that cause the estimates to be biased. This occurs when the explanatory variables (X) are correlated with the error term (u or ε). This can happen in 3 possible ways:
 - A. Omitted variables
 - B. Reverse causality
 - C. Measurement error

These three issues are call “endogeneity” problems.

Outline

Intuitively, why is there bias if X is correlated with u ?

$$Y_i = a + \beta X_i + u_i$$

Recall the univariate proof of unbiasedness:

$$\begin{aligned}\hat{\beta} &= \frac{Cov(X, Y)}{Var(X)} = \frac{Cov(X, \alpha + \beta X + u)}{Var(X)} \\ &= \frac{\beta Var(X) + Cov(X, u)}{Var(X)} = \beta + \boxed{\frac{Cov(X, u)}{Var(X)}}\end{aligned}$$

Recall the multivariate proof of unbiasedness:

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{u}) \\ &= \beta + \boxed{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}}\end{aligned}$$

Omitted Variable Bias

Omitted variable bias: occurs when an important causal factor is omitted from a regression, resulting in the effect of the other factor being over or underestimated.

An omitted variable creates bias if:

1. it is correlated with the outcome variable
2. it is correlated with an explanatory variable

Why are important variables omitted:

1. they are not available in the data
2. the researcher does not think of them

Omitted Variable Bias: Univariate

True equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

The equation we estimate:

$$Y = \beta_0^* + \beta_1^* X_1 + u$$

And suppose that X_2 and X_1 are correlated, so we can write:

$$X_2 = \delta_0 + \delta_1 X_1 + \eta$$

Omitted Variable Bias: Univariate

Let us start with the true population equation and examine what happens when we omit X_2 and just run a regression of Y on X_1 . The omitted variable is in the error term:

$$Y = \beta_0 + \beta_1 X_1 + [\beta_2 X_2 + \epsilon]$$

$$Y = \beta_0 + \beta_1 X_1 + [\beta_2(\delta_0 + \delta_1 X_1 + \eta) + \epsilon]$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 \delta_0 + \beta_2 \delta_1 X_1 + \beta_2 \eta + \epsilon$$

$$Y = \underbrace{(\beta_0 + \beta_2 \delta_0)}_{\beta_0^*} + \underbrace{(\beta_1 + \beta_2 \delta_1)}_{\beta_1^*} X_1 + \underbrace{(\beta_2 \eta + \epsilon)}_{u^*}$$

So, instead of estimating B_1 , we estimate:

$$\boxed{\beta_1^* = \beta_1 + \beta_2 \delta_1}$$

Omitted Variable Bias: Univariate

So, instead of estimating B_1 , we estimate: $\beta_1^* = \beta_1 + \beta_2 \delta_1$

$$Bias = \beta_2 \delta_1$$

So, the amount of bias depends on B_2 , which is how the omitted variable affects the outcome variables, and δ_1 , which is the correlation of the omitted variable with X_1 .

$$\delta_1 > 0$$

$$\delta_1 < 0$$

	$\text{Corr}(x_1, x_2) > 0$	$\text{Corr}(x_1, x_2) < 0$
$\beta_2 > 0$	Positive bias	Negative bias
$\beta_2 < 0$	Negative bias	Positive bias

Omitted Variable Bias: Univariate

- Suppose we estimate the effect of years of education on annual income.

$$\widehat{income} = 10,000 + 1,500 * educ$$

A person's IQ is omitted from the equation. IQ is positively correlated with their years of education and their income. Thus 1,500 is positively biased by omitting IQ. When we estimate the new equation we get:

$$\widehat{income} = 5,000 + 800 * educ + 1,000 * IQ$$

- Suppose we estimate the following regression:

$$wage_i = \beta_0 + \beta_1 age_i + u_i$$

How is the coefficient of interest likely to be affected by omitting the following variables?

A. Experience?

B. Education?

Omitted Variable Bias: Multivariate

General Case – Suppose the following is the true relationship:

$$\mathbf{Y}_i = \mathbf{X}'_i \boldsymbol{\beta} + e_i$$

But we include the first k_1 explanatory variables but not all of the k variables.

We can divide the vector of explanatory variables into the included \mathbf{X}^* and the omitted \mathbf{X}^O . That is, we partition the matrix:

$$X_i = \begin{bmatrix} 1 \\ X_{i,2} \\ X_{i,3} \\ \vdots \\ X_{i,k1} \\ \hline \cdots \\ X_{i,k1+1} \\ \vdots \\ X_{t,k} \end{bmatrix} = \begin{bmatrix} X_i^* \\ \hline \cdots \\ X_i^O \end{bmatrix}$$

Omitted Variable Bias: Multivariate

We also split the corresponding coefficients in those on the included and omitted variables:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{k1} \\ \hline \beta_{k1+1} \\ \vdots \\ \beta_k \end{bmatrix} = \begin{bmatrix} \beta^* \\ \hline \beta^O \end{bmatrix}$$

So, the right hand side of the true regression can be written as follows:

$$\mathbf{X}'_i \boldsymbol{\beta} = [X_i^{*'} \quad X_i^{O'}] \begin{bmatrix} \beta^* \\ \beta^O \end{bmatrix} = X_i^{*'} \beta^* + X_i^{O'} \beta^O$$

Thus we, for one individual, we can write:

$$\mathbf{Y}_i = X_i^{*'} \beta^* + X_i^{O'} \beta^O + \epsilon_i$$

Omitted Variable Bias: Multivariate

The matrix of all the X's can be written as (where each row is a person):

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_N' \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^{*' & \mathbf{x}_1^O'} \\ \mathbf{x}_2^{*' & \mathbf{x}_2^O'} \\ \vdots & \vdots \\ \mathbf{x}_N^{*' & \mathbf{x}_N^O'} \end{bmatrix} = [\mathbf{X}^* \quad \mathbf{X}^O]$$

So, the right hand side of the true regression can be written as follows:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} = [\mathbf{X}^* \quad \mathbf{X}^O] \begin{bmatrix} \boldsymbol{\beta}^* \\ \boldsymbol{\beta}^O \end{bmatrix} + \mathbf{u} = \mathbf{X}^*\boldsymbol{\beta}^* + \mathbf{X}^O\boldsymbol{\beta}^O + \mathbf{u}$$

Note: this is all just notation that splits the true regression into the included and omitted components.

However, we run the regression with omitted variables.

Omitted Variable Bias: Multivariate

The true regression equation follows. We are interested in estimating β^* :

$$\mathbf{Y} = \mathbf{X}^* \boldsymbol{\beta}^* + \mathbf{X}^O \boldsymbol{\beta}^O + \mathbf{u}$$

But we run the regression with omitted variables:

$$\mathbf{Y} = \mathbf{X}^* \hat{\boldsymbol{b}}_R^* + \mathbf{u}$$

So, the coefficient estimates we get is:

$$\begin{aligned}\hat{\boldsymbol{b}}_R^* &= (\mathbf{X}^{*\prime} \mathbf{X}^*)^{-1} \mathbf{X}^{*\prime} \mathbf{Y} \\ &= (\mathbf{X}^{*\prime} \mathbf{X}^*)^{-1} \mathbf{X}^{*\prime} (\mathbf{X}^* \boldsymbol{\beta}^* + \mathbf{X}^O \boldsymbol{\beta}^O + \mathbf{u}) \\ &= \boldsymbol{\beta}^* + (\mathbf{X}^{*\prime} \mathbf{X}^*)^{-1} \mathbf{X}^{*\prime} \mathbf{X}^O \boldsymbol{\beta}^O + (\mathbf{X}^{*\prime} \mathbf{X}^*)^{-1} \mathbf{X}^{*\prime} \mathbf{u}\end{aligned}$$

The expected value of this expression is as follows:

$$\begin{aligned}E(\hat{\boldsymbol{b}}_R^* | \mathbf{X}) &= \boldsymbol{\beta}^* + (\mathbf{X}^{*\prime} \mathbf{X}^*)^{-1} \mathbf{X}^{*\prime} \mathbf{X}^O \boldsymbol{\beta}^O + (\mathbf{X}^{*\prime} \mathbf{X}^*)^{-1} \mathbf{X}^{*\prime} E(\mathbf{u} | \mathbf{X}) \\ &= \boldsymbol{\beta}^* + (\mathbf{X}^{*\prime} \mathbf{X}^*)^{-1} \mathbf{X}^{*\prime} \mathbf{X}^O \boldsymbol{\beta}^O\end{aligned}$$

Which is generally a biased estimate of $\boldsymbol{\beta}^*$.

Omitted Variable Bias: Multivariate

So, OLS that omits X^O produces a potentially biased B^* :

$$E(\hat{b}_R^* | \mathbf{X}) = \boldsymbol{\beta}^* + (\mathbf{X}^{*'} \mathbf{X}^*)^{-1} \mathbf{X}^{*'} \mathbf{X}^O \boldsymbol{\beta}^O$$

This will be biased if the second term is not equal to 0:

1. The omitted X^* 's are correlated with the included X^O 's. That is, $\mathbf{X}^{*'} \mathbf{X}^O \neq 0$.
2. The omitted X^* 's are correlated with the outcome Y : That is, $\mathbf{X}^O \boldsymbol{\beta}^O \neq 0$.

Note: that these are the exact same conditions for there to be an omitted variable bias in the univariate case.

Note: the expression $(\mathbf{X}^{*'} \mathbf{X}^*)^{-1} \mathbf{X}^{*'} \mathbf{X}^O$ is just the OLS regression of the omitted variables on the included variables (with X^O instead of Y). This is an auxiliary regression of the omitted X^O 's on the included X^* 's.

Omitted Variable Bias: Multivariate

The relationships become clearer when considering the multivariate notation with one omitted variable k.

$$\mathbf{X}^* = \begin{bmatrix} 1 & X_{12} & X_{13} & \cdots & X_{1k-1} \\ 1 & X_{22} & X_{23} & \cdots & X_{2k-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{N2} & X_{N3} & \cdots & X_{Nk-1} \end{bmatrix} \quad \mathbf{X}^k = \begin{bmatrix} X_{1k} \\ X_{2k} \\ \vdots \\ X_{Nk} \end{bmatrix}$$

We estimate:

$$E(\hat{\boldsymbol{b}}_R^* | \mathbf{X}) = \boldsymbol{\beta}^* + (\mathbf{X}^{*\prime} \mathbf{X}^*)^{-1} \mathbf{X}^{*\prime} \mathbf{X}^k \beta_k$$

We can define the following vector of coefficients since the second term is just an auxiliary regression of X_k on the other X's:

$$\hat{\boldsymbol{\alpha}} = (\mathbf{X}^{*\prime} \mathbf{X}^*)^{-1} \mathbf{X}^{*\prime} \mathbf{X}^k$$

Thus the bias is:

$$E(\hat{\boldsymbol{b}}_R^* | \mathbf{X}) = \boldsymbol{\beta}^* + \hat{\boldsymbol{\alpha}} \beta_k$$

Omitted Variable Bias: Multivariate

- So, the general case is completely analogous to the univariate case.
- The bias in the estimate is a function of two things:
 1. The relationship between the omitted variable and the included variable of interest, which is represented by α .
 2. The true relationship between the omitted variable and the outcome variable, which is given by β_k .
- Every regression has omitted variable bias:
 1. No matter how many variables you add, it may not be enough
 2. There may be omitted variables that are unobserved or can not be measured
 3. This is the fundamental problem for establishing unbiased causal relationships in econometrics

Measurement Error

Measurement error – occurs when the observed values of a variable are larger or smaller than the true values.

This causes a problem even when the measurement error has mean 0 and is not correlated with X or Y.

$$x_i^* = x_i + \epsilon_i$$

Measurement error in an explanatory variable will cause “attenuation bias” – the estimated coefficient will be too small in magnitude (too close to 0).

Measurement Error

First, let us review some mathematical properties. Assume that X is measured with random error:

$$\epsilon \sim N(0, \sigma_\epsilon)$$

$$Cov(X, \epsilon) = 0$$

$$\begin{aligned} Var(X + \epsilon) &= Var(X) + Var(\epsilon) + 2Cov(X, \epsilon) \\ &= Var(X) + Var(\epsilon) \end{aligned}$$

$$Cov(X + \epsilon, Y) = Cov(X, Y) + Cov(\epsilon, Y) = Cov(X, Y)$$

Measurement Error

Error in X:

$$X^* = X + \epsilon$$

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

$$Y = \beta_0^* + \beta_1^* (X + \epsilon) + u$$

$$\begin{aligned}\hat{\beta}_1^* &= \frac{Cov(X + \epsilon, Y)}{Var(X + \epsilon)} = \frac{Cov(X, Y) + Cov(\epsilon, Y)}{Var(X) + Var(\epsilon)} \\ &= \frac{Cov(X, Y)}{Var(X) + Var(\epsilon)} < \frac{Cov(X, Y)}{Var(X)} = \beta_1\end{aligned}$$

So, the measurement error in X causes the estimated coefficient to be too small in magnitude $|B_1^*| < |B_1|$. This is **attenuation bias**.

Measurement Error

Another way to show that measurement error causes the explanatory variables to be correlated with the error term:

$$X^* = X + \epsilon$$

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

$$Y = \beta_0 + \beta_1(X^* - \epsilon) + u$$

$$Y = \beta_0 + \beta_1 X^* + u - \beta_1 \epsilon$$

$$Y = \beta_0 + \beta_1 X^* + (u - \beta_1 \epsilon)$$

So, both X^* and the error term are dependent on ϵ , so the X^* is correlated with the error term.

Measurement Error

Example 1: Surveys

- recall bias – people misremember what they have done in the past

National Longitudinal Survey of Youth (NLSY) – survey young people about their schooling, training, job search, income, hours worked, commute times, etc.

$$Income = \beta_0 + \beta_1(Schooling + \epsilon) + u$$

Fix: Collect data from each person's high school and employer to verify the responses. Administrative data is much less likely to have measurement error.

Measurement Error

Example 2: Administrative data

- data entry mistakes

Fix:

Data cleaning – looking for data that is likely to be a mistake and omitting it from the analysis.

- worked 410 hours in a week
- college student was born in 1919

Double measurement. This involves getting two measurements of the same variable and dropping/investigating observations where they do not match.

Measurement Error

Error in Y:

$$Y^* = Y + \epsilon$$

$$\begin{aligned} Y^* &= \beta_0^* + \beta_1^* X + u \\ Y + \epsilon &= \beta_0^* + \beta_1^* X + u \end{aligned}$$

$$\begin{aligned} \hat{\beta}_1^* &= \frac{Cov(X, Y + \epsilon)}{Var(X)} = \frac{Cov(X, Y) + Cov(X, \epsilon)}{Var(X)} \\ &= \frac{Cov(X, Y)}{Var(X)} = \beta_1 \end{aligned}$$

So, the measurement error in Y does not cause bias in the estimated coefficient.

Note that this is only true if the error in measuring Y is not correlated with X. Also, the variance of the estimator will be larger.

Reverse Causality

Reverse causality (simultaneity)- if the explanatory variable is a function of the outcome variable, then it will be correlated with the error term (and therefore biased).

Example: Suppose we estimate the effect of police on the crime rate.

$$CrimeRate_c = \beta_0 + \beta_1 PolicePerCap_c + u$$

The problem is that the number of police in a city may be a function of the crime rate.

For example, a city may hire more police if it has a high crime rate.

Reverse Causality

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

$$X = \gamma_0 + \gamma_1 Y + \epsilon$$

- The outcome variable Y is a function of X.
- However, the explanatory variable X is a function of Y.
- This creates a circular issue where X is causing Y and Y is causing X.
- This causes X to be correlated with the error term and thus biased.

Logic:

- Suppose “u” is large due to some unexplained factor.
- That is, Y is larger than expected.
- Thus X is larger because it is a function of Y.
- So a large “u” is correlated with a large X, so they are correlated.

Reverse Causality

Let's show that X is a function of u . That is, that X is correlated with the error term and thus the coefficient on X will be biased.

$$X = \gamma_0 + \gamma_1 Y + \epsilon$$

$$X = \gamma_0 + \gamma_1(\beta_0 + \beta_1 X + u) + \epsilon$$

$$X = \gamma_0 + \gamma_1\beta_0 + \gamma_1\beta_1 X + \gamma_1 u + \epsilon$$

$$(1 - \gamma_1\beta_1)X = (\gamma_0 + \gamma_1\beta_0) + \gamma_1 u + \epsilon$$

$$X = \frac{\gamma_0 + \gamma_1\beta_0}{1 - \gamma_1\beta_1} + \frac{\gamma_1 u}{1 - \gamma_1\beta_1} + \frac{\epsilon}{1 - \gamma_1\beta_1}$$

Endogeneity

$$CeoSalary = \beta_0 + \beta_1 CompanyProfits + u$$

What omitted variables might we have?

Is there likely to be reverse causality?

Is there likely to be measurement error?

Endogeneity

$$Consumption = \beta_0 + \beta_1 Income + u$$

We have data from 100 poor families in rural India. We want to estimate B_1 – how much food consumption would increase if we increased income by 1 dollar.

What omitted variables might we have?

Is there likely to be reverse causality?

Is there likely to be measurement error?

Review

Endogeneity:

- Omitted variables cause bias in the estimates.
- Measurement error causes bias in the estimates.
- Reverse causality causes bias in the estimates.

How to fix omitted variable bias:

- Instrumental variables
- Regression discontinuity
- Experiments

We will dedicate the second half of the class to these techniques.

Lesson 11

Practical Considerations Functional Form

Outline

Previous Lesson:

1. Omitted Variables
2. Measurement Error
3. Reverse Causality

This Lesson Lesson:

1. Some practical considerations
2. Functional Form

Next Lessons:

1. Maximum Likelihood
2. Probit, Logit, Tobit

Outlier Data

When should we be concerned about outliers?

As a general rule, we should be concerned about outliers any time that the exclusion of a single or small fraction of observations results in a very different regression line.

This is most likely to occur when two conditions are satisfied:

1. our data has a small number of observations
2. the outlier is very different from the typical observation

This does not mean that the regression line is wrong, or that the outlier should be excluded, but rather that careful inspection and consideration is needed.

Outlier data occur for two reasons:

1. There is a mistake in the data (e.g. an extra zero).
2. An observation is legitimately different (e.g. Bill Gates).

Outlier Data

Why should we be concerned about outliers?

Ordinary least squares is the only tool we have used in this course to estimate a regression line. OLS, by nature, minimizes the sum of the squared errors (RSS):

$$\text{Min } \sum_{i=1}^N (Y_i - \hat{Y}_i)^2$$

Because this expression is squared, it gives greater weight to outliers:

$$(cY_i - c\hat{Y}_i)^2 = c^2(Y_i - \hat{Y}_i)^2$$

Suppose we have an outlier observation: $Y_7 = 20$ and $\hat{Y}_7 = 10$. What happens to the SSR if we change the regression line such that $\hat{Y}_7 = 11$?

Now consider a smaller observation with so $Y_4 = 6$ and $\hat{Y}_4 = 3$. What happens to the SSR if we change the regression line such that $\hat{Y}_4 = 4$.

Including all observations:

$$\widehat{rdintens} = 2.625 + .000053 \text{ sales} + .0446 \text{ profmarg}$$

(0.586) (0.000044) (.0462)

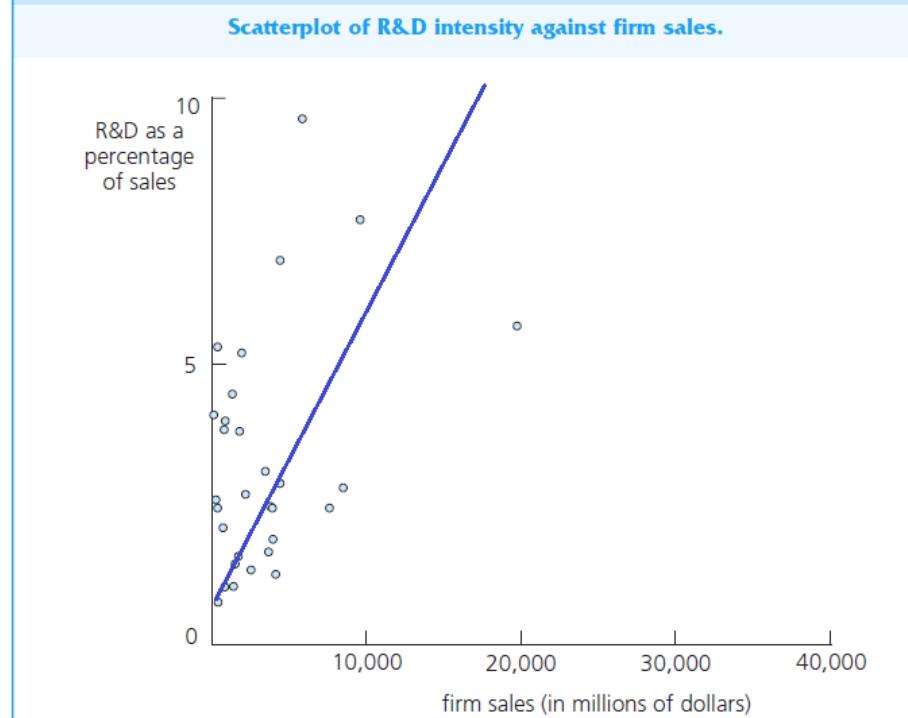
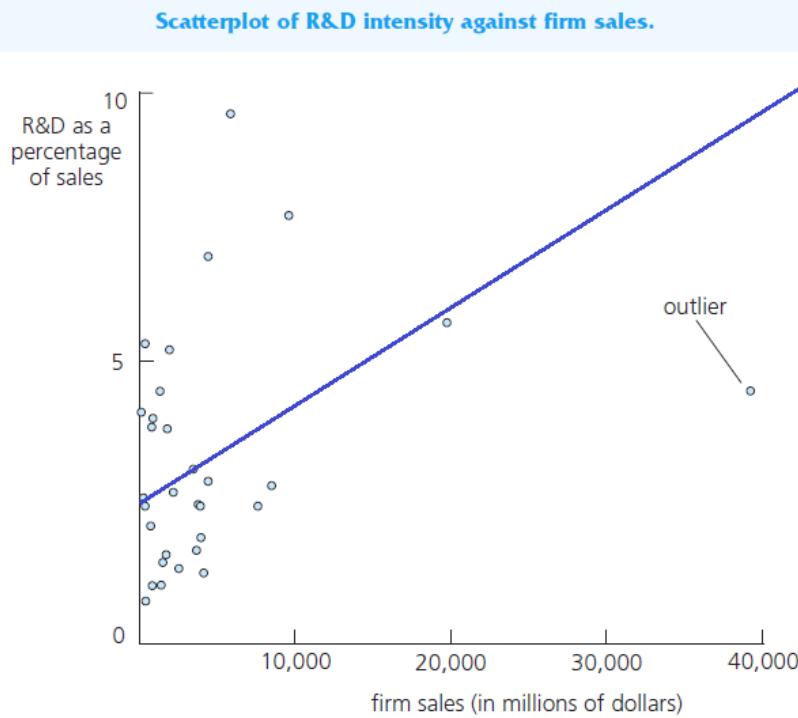
$n = 32, R^2 = .0761$.

Dropping the outlier:

$$\widehat{rdintens} = 2.297 + .000186 \text{ sales} + .0478 \text{ profmarg}$$

(0.592) (0.000084) (.0445)

$n = 31, R^2 = .1728$



Outlier Data

How do we address outliers in practice?

1. Trimming:
 - dropping outlier observations
 - often the highest and lowest x% of observations
2. Winsorising:
 - changing outlier observations
 - set highest x% to the (100 – x)% value and lowest x% to the x% value
3. Alternative functional form:
 - take the natural log of the outcome
 - recall our outlier example: $\ln(20) - \ln(10) = .69$ and $\ln(6) - \ln(3) = .69$
4. Use a method other than Ordinary Least Squares:
 - a popular method now is quantile regression
 - can estimate the median response (or any other percentile) rather than the mean response
 - also useful if you are interested in certain parts of the distribution

Outlier Data

Consider a log-log transformation:

With all 32 companies:

$$\widehat{\log(rd)} = -4.378 + 1.084 \log(sales) + .0217 profmarg,$$
$$(.468) \quad (.062) \quad (.0128)$$
$$n = 32, R^2 = .9180$$

With the outlier dropped:

$$\widehat{\log(rd)} = -4.404 + 1.088 \log(sales) + .0218 profmarg,$$
$$(.511) \quad (.067) \quad (.0130)$$
$$n = 31, R^2 = .9037$$

Out of sample prediction

Prediction:

- Prediction is one of the primary objectives of regression analysis: given values of the Xs, predict the Y value.
- This works well, when we make a prediction for X values that are similar to the X values in our data.
- However, the prediction may be poor (large \hat{u}) or illogical when the X values are very different from those in our data. This occurs for two reasons (typically both play a role):
 1. functional form may be wrong (e.g. line may be curved but we assume it is linear).
 2. omitted variables (e.g. observations with very different X values may be very different in other ways).

Out of sample prediction

Functional form: We use data on 100 employees to estimate the effect of experience on wages. Nearly all of the workers in our sample have 5 or fewer years of experience.

$$\ln(\widehat{wage}_i) = 6.73 + 0.17Experience_i$$

How much higher wages would we predict for someone who has 40 years of experience versus someone with 30 years of experience?

Omitted variables: We use data on 100 high school drop-outs to estimate the returns to a an additional year of schooling. No one in the sample attended college.

$$\ln(\widehat{wage}_i) = 5.87 + 0.085Education_i$$

How much higher wages would we predict for someone who has 16 years of education versus someone with 12 years of education?

Functional Forms

Reciprocal:
$$Y = \beta_1 + \beta_2 \frac{1}{X} + u$$

$$\frac{\partial Y}{\partial X} = -\frac{\beta_2}{X^2}$$

As X increases, its effect decreases. Note that $X \rightarrow \infty$, $Y \rightarrow \beta_1$.

Example: Demand curves – as the price increases, there may be some minimum amount that people still buy (e.g. food).

Functional Forms

log – log: $\ln(Y) = \beta_1 + \beta_2 \ln(X) + u$

$$\frac{d\ln(Y)}{dY} = \frac{1}{Y} \Rightarrow d\ln(Y) = \frac{dY}{Y}$$

$$\frac{d\ln(X)}{dX} = \frac{1}{X} \Rightarrow d\ln(X) = \frac{dX}{X}$$

So:

$$\beta_2 = \frac{d\ln(Y)}{d\ln(X)} = \frac{dY/Y}{dX/X} = \frac{\% \Delta Y}{\% \Delta X}$$

A 1 percent change in X generates a β_2 percent change in Y.

Example: Cobb-Douglas – A percent change in labor is likely to generate a percent change in output.

Functional Forms

log – linear $\ln(Y) = \beta_1 + \beta_2 X + u$

$$\beta_2 = \frac{d\ln(Y)}{dX} = \frac{dY/Y}{dX} = \frac{\% \Delta Y}{\Delta X}$$

A 1 unit change in X generates a β_2 percent change in Y.

Example:

$$GDP_t = GDP_0(1 + g)^t$$

$$\ln(GDP_t) = \ln(GDP_0) + (1 + g)t$$

$$\ln(GDP_t) = \beta_1 + \beta_2 X + \epsilon$$

So, each year X increases GDP by a percent.

Functional Forms

linear – log $Y = \beta_1 + \beta_2 \ln(X) + u$

$$\beta_2 = \frac{dY}{d\ln(X)} = \frac{dY}{dX/X} = \frac{\Delta Y}{\% \Delta X}$$

A 1 unit percent change in X generates a β_2 unit change in Y.

Example:

$$wage_i = \beta_1 + \beta_2 \ln(exper_i) + \epsilon_i$$

This would imply that a change in experience from 5 to 6 years would have the same effect on wages as an increase in experience from 10 to 12 years.

Functional Forms

Interaction terms:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_2 X_3 + u$$

$$\frac{\partial Y}{\partial X_2} = \beta_2 + \beta_4 X_3 \quad \beta_4 = \frac{\partial Y}{\partial X_2 \partial X_3}$$

The effect of a one unit change in X_2 is larger when X_3 is larger. Note that β_4 is the additional effect of a one unit change in X_2 when X_3 is one unit larger.

When one of the variables is a dummy:

$$wage = \beta_1 + \beta_2 black + \beta_3 exper + \beta_4 black * exper + u$$

When both variables are continuous:

$$wage = \beta_1 + \beta_2 educ + \beta_3 exper + \beta_4 educ * exper + u$$

Functional Forms

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_2 X_3 + u$$

Note that excluding a main term results in β_4 suffering from omitted variable bias. Suppose X_3 is excluded:

1. It is correlated with $X_2 X_3$
2. It is correlated with Y

For example:

$$wage = \beta_1 + \beta_2 black + \beta_3 exper + \beta_4 black * exper + u$$

$$wage = \beta_1^* + \beta_3^* exper + \beta_4^* black * exper + u$$

The coefficient β_4 could be negative even if it should be positive if the value of β_2 is negative.

Comparing Models

Comparing models: We have seen a number of different options for models. There are various ways of testing which models are the best.

Comparing nested models with same dependent variable:

$$Y = \beta_1 + \beta_2 X_2 + \dots + \beta_{k1} X_{k1} + \beta_{k1+1} X_{k1+1} + \dots + \beta_k X_k + u$$

$$Y = \beta_1 + \beta_2 X_2 + \dots + \beta_{k1} X_{k1} + u$$

As we have seen, this simply requires an F-test comparing the restricted and unrestricted models using either the RSS or R-squared.

But many times the models we want to compare are not nested:

- Explanatory variables of one model not a subset of the other
- Different form of outcome variable Y

Comparing Models

Comparing non-nested models with same outcome variable:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

$$Y = \beta_1 + \beta_2 \ln(X_2) + \beta_3 \ln(X_3) + u$$

Test 1: Run the following model as the unrestricted model:

$$Y = \delta_1 + \delta_2 X_2 + \delta_3 X_3 + \delta_4 \ln(X_2) + \delta_5 \ln(X_3) + u$$

And then run one of the equations above as the restricted model.

Test 2:

1. Run one of the models and get predicted Y.
2. Run the other model adding the predicted Y as an extra control.
3. Conduct t-test on the predicted Y.

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \delta \hat{Y} + u$$

Comparing Models

$$Y = \beta_1 + \beta_2 X + u \quad \ln(Y) = \beta_1 + \beta_2 \ln(X) + u$$

Comparing these two model is problematic because they have different outcome variables. We can transform the outcome variable to make the RSS's comparable.

Box-Cox transformation:

1. Compute the geometric mean for each set of Ys:

$$\tilde{Y} = (Y_1 * Y_2 * \dots * Y_N)^{1/N}$$

2. Transform the Y's by dividing by the geometric means:

$$Y_i^* = Y_i / \tilde{Y}$$

3. Run the original regressions using the new Y^* 's.
4. We can now compare the RSS's: The following statistic has a chi-squared dist with 1 deg freedom. Check if the larger RSS is worse:

$$(\frac{1}{2}n) \ln\left(\frac{RSS_2}{RSS_1}\right)$$

Comparing Models

OLS depends on the fact that the errors are normally distributed. That is, the shape of errors is normal (not just that the variance is homoskedastic and that errors are uncorrelated). One way to check for a misspecified model is to check if the errors are not normally distributed.

Jarque-Berra Test

1. Compute the moments of the predicted error from the model (mean, skewness, kurtosis):

$$\mu_2 = \frac{\sum \hat{u}^2}{N} \quad \mu_3 = \frac{\sum \hat{u}^3}{N} \quad \mu_4 = \frac{\sum \hat{u}^4}{N}$$

2. Compute the Jarque-Berra statistic:

$$JB = N \left[\frac{\mu_3^2}{6} + \frac{(\mu_4 - 3)^2}{24} \right]$$

3. This statistic is distributed as chi-squared with 2 degrees of freedom. If we reject, then it means that we reject that the errors are normally distributed and may suspect that the model is misspecified.

Comparing Models

Ramsey Reset Test: Popular test for a misspecified model. Can detect if there is a non-linear combinations of the X's that helps explain Y.

1. Run the regression and get the predicted Y.
2. Run the regression again adding powers of the predicted Y as additional explanatory variables.

$$Y = \delta_1 + \delta_2 X_2 + \delta_3 X_3 + \delta_4 \hat{Y}^2 + \delta_5 \hat{Y}^3 + u$$

3. Use this as the unrestricted model in an F-test.

This test is very useful when there are many X's (otherwise we could just add powers of the X's). The shortcoming is that rejecting the model does not tell us which X is the problem.

Lesson 12

Maximum Likelihood
LPM, Logit, Probit

Outline

Previous Lesson:

1. Practical considerations
2. Functional Form

This Lesson Lesson:

1. Maximum Likelihood
2. LPM, Logit, Probit

Next Lessons:

1. Ordered Probit
2. Tobit

Maximum Likelihood

Maximum likelihood – an alternative method of estimating parameters (i.e. not OLS). This method selects the parameters (the β s or Θ s) that maximize the likelihood of getting the observed data X .

Consider the probability density function (pdf) – the probability of observing x_i given Θ : $f_i(x_i|\theta)$

Joint density function (the probability of observing each of the X s given the Θ s):

$$\begin{aligned} f(x_1, \dots, x_N | \theta) &= \prod_{i=1}^n f(x_i | \theta) \\ &= f(x_1 | \theta) * f(x_2 | \theta) * \dots * f(x_n | \theta) \end{aligned}$$

The likelihood function (the probability of the Θ s given the X s):

$$\begin{aligned} l(\theta | x_1, \dots, x_N) &= \prod_{i=1}^n f(x_i | \theta) \\ &= f(x_1 | \theta) * f(x_2 | \theta) * \dots * f(x_n | \theta) \end{aligned}$$

Maximum Likelihood

Simple example:

Suppose we have an unfair coin. And suppose the true probability of getting heads is either:

$$P_0 = \Pr(H) = \begin{cases} \frac{3}{4} \\ \frac{1}{4} \end{cases}$$

These are the possible parameters. We want to determine which parameter is correct.

We toss the coin three times and get the following data:

H, H, T.

[intuition: $\frac{3}{4}$ is correct]

The density function in this example is simply the probability of tossing heads or tails:

$$f(H|P_0) = P_0$$

$$f(T|P_0) = 1 - P_0$$

Maximum Likelihood

The likelihood function given the data is:

$$l(P_0) = f(H|P_0) * f(H|P_0) * f(T|P_0) = P_0 * P_0 * (1 - P_0)$$

Now we can compute the likelihood function for each of the parameter options ($\frac{1}{4}$ and $\frac{3}{4}$):

$$l\left(\frac{1}{4}\right) = \frac{1}{4} * \frac{1}{4} * \frac{3}{4} = \frac{3}{64}$$

$$l\left(\frac{3}{4}\right) = \frac{3}{4} * \frac{3}{4} * \frac{1}{4} = \frac{9}{64}$$

So, the likelihood function is larger when the parameter is $\frac{3}{4}$. Thus, based on the data, we think we know which P_0 is correct.

The same concept is used to choose the best β s to explain a large amount X and Y data.

Maximum Likelihood

Taking the log of the likelihood function can make it much easier to evaluate:

$$\begin{aligned} L(\theta|x_1, \dots, x_N) &= \ln[l(\theta)] \\ &= \ln\left[\prod_{i=1}^n f(x_i|\theta)\right] \\ &= \sum_{i=1}^n \ln f(x_i|\theta) \\ &= \ln f(x_1|\theta) + \ln f(x_2|\theta) + \dots + \ln f(x_N|\theta) \end{aligned}$$

Note that this transformation never changes the choice of the best parameters:

$$c_1 > c_2 \Rightarrow \ln(c_1) > \ln(c_2)$$

So:

$$l(\theta_1) > l(\theta_2) \Rightarrow \ln(l(\theta_1)) > \ln(l(\theta_2)) \Rightarrow L(\theta_1) > L(\theta_2)$$

Maximum Likelihood

$$L(\theta|x_1, \dots, x_N) = \ln f(x_1|\theta) + \ln f(x_2|\theta) + \dots + \ln f(x_N|\theta)$$

The likelihood function can be maximized by taking the first order conditions:

$$\frac{\partial L(\hat{\theta})}{\partial \theta_j} = 0 \text{ for every } j$$

Which will produce an estimate of the parameters Θ .

This is just like OLS.

In fact, under the assumption of normality, it is identical to OLS.

Maximum Likelihood

The pdf for a normally distributed variables is:

$$X \sim N(\mu_X, \sigma_X^2) \quad f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Consider maximum likelihood in the case of a linear regression.

$$Y_i = \beta_1 + \beta_2 X_{2i} + u_i$$

Recall that the error term is normally distributed:

$$u_i | X \sim N(0, \sigma^2)$$

This also implies that:

$$Y_i | X \sim N(\beta_1 + \beta_2 X_{2i}, \sigma^2)$$

Our goal is to choose the parameters β that maximize the likelihood of the data under the assumption of this normal distribution.

Maximum Likelihood

The pdf of Y given X for an individual i can be expressed as follows:

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y_i - \beta_1 - \beta_2 X_{2i})^2}{2\sigma^2}}$$

Thus, the joint pdf across all individuals is:

$$\prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y_i - \beta_1 - \beta_2 X_{2i})^2}{2\sigma^2}}$$

Equivalently:

$$\prod_{i=1}^N (2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{(Y_i - \beta_1 - \beta_2 X_{2i})^2}{2\sigma^2}}$$

So, the log-likelihood function is:

$$\sum_{i=1}^N -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} (Y_i - \beta_1 - \beta_2 X_{2i})^2$$

Maximum Likelihood

Log-likelihood function is:

$$\sum_{i=1}^N -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} (Y_i - \beta_1 - \beta_2 X_{2i})^2$$

Which simplifies to:

$$-\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^N (Y_i - \beta_1 - \beta_2 X_{2i})^2$$

Note that the term on the right is just the sum of the squared errors.

Taking the FOCs is just like OLS.

$$\frac{\partial L(\hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}^2)}{\partial \hat{\beta}_1} = 0$$

$$\frac{\partial L(\hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}^2)}{\partial \hat{\beta}_2} = 0$$

$$\frac{\partial L(\hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}^2)}{\partial \hat{\sigma}^2} = 0$$

Maximum Likelihood

However, in many cases, the functional form is not so simple. In these cases, it is not possible to simply solve first order conditions or to multiply a few matrices. So, an optimization procedure is run to choose the best parameters.

The basic idea of these maximum likelihood optimization procedures:

1. Start with some initial parameter values.
2. Determine whether increasing or decreasing each β_1, \dots, β_k increases the likelihood function. Adjust the parameters accordingly.
3. Repeat this process until you have converged to values of the β s where no further improvement is possible (the likelihood function is maximized).

Note: There are economists who spend all of their time figuring out the best optimization procedures and ways to make sure the optimal choice is unique.

LPM, Logit, Probit

Summary:

- As an alternative to OLS, we can write a likelihood function using the pdf. Often it is helpful to consider the log-likelihood.
- The pdf typically requires some functional form assumption about the error term.
- An optimization procedure is used to maximize the likelihood function.

Now we are going to consider a context in which maximum likelihood procedures are frequently used: when the outcome variable is binary.

- Linear probability model – OLS
- Logit model – MLE
- Probit model – MLE

Linear Probability Model

The Model

$$Y_i = \beta_1 + \beta_2 X_{2i} + u_i$$

$$D_i = \beta_1 + \beta_2 X_{2i} + u_i$$

For example:

$$D_i = \begin{cases} 0 & \text{i not working} \\ 1 & \text{i working} \end{cases}$$

The predicted value of D is the probability that the person is working (which should range from 0 to 1).

$$\hat{D} = 0.25 + 0.05educ$$

In this example, an additional year of schooling increases the probability that a person is working by 5 percentage points.

Linear Probability Model

Note that the predicted probabilities can be written as follows:

$$P_i = Pr(D_i = 1) = \beta_1 + \beta_2 X_{2i}$$

$$1 - P_i = Pr(D_i = 0) = 1 - \beta_1 - \beta_2 X_{2i}$$

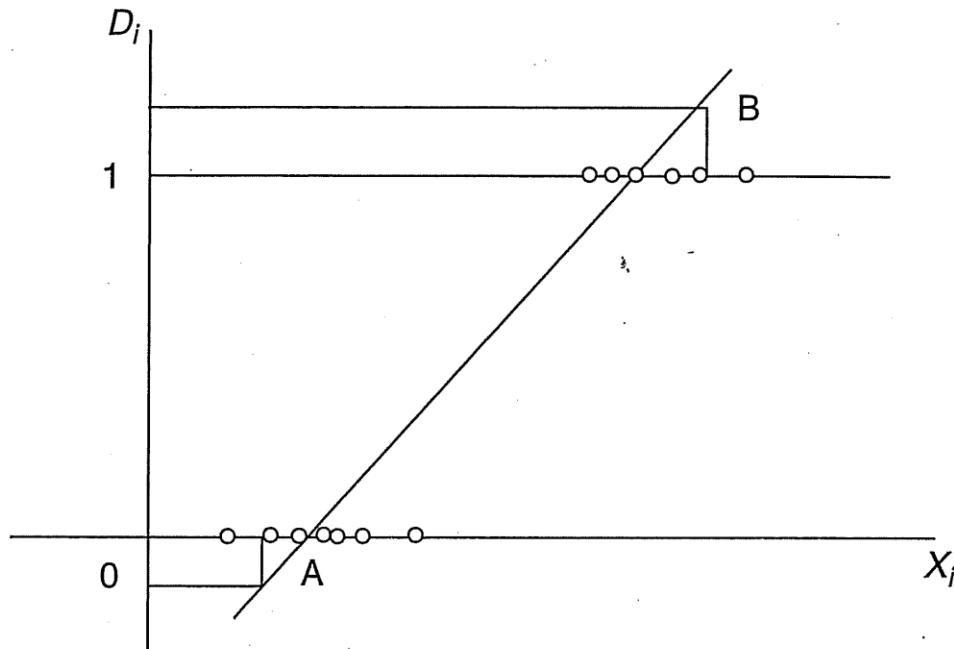


Figure 12.1 The linear probability model

Linear Probability Model

Shortcomings of the linear probability model:

Problem 1: It can produce predicted probabilities that are negative or greater than 1.

$$\hat{D} = 0.25 + 0.05\text{educ}$$

Consider someone with 17 years of education (12+4+1). The predicted probability of attending college is 1.1 (or 110%) which doesn't make any sense.

This is most likely to occur when predicting values for X that are relative outliers.

Linear Probability Model

Problem 2: Heteroskedastic standard errors.

Consider an individual with value X_{2i} . Then the error can only take on two values:

$$\text{if } D_i = 1 \Rightarrow u_i = Y_i - \hat{Y}_i = 1 - \beta_1 - \beta_2 X_{2i} = 1 - P_i$$

$$\text{if } D_i = 0 \Rightarrow u_i = Y_i - \hat{Y}_i = -\beta_1 - \beta_2 X_{2i} = -P_i$$

If, conditional on the X 's, the error term can only take on two values, then it is definitely not being drawn from the normal distribution (an assumption of OLS). Consider the variance of u_i explicitly:

$$\begin{aligned} Var(u_i) &= E(u_i)^2 \\ &= P_i(u_i|D_i = 1)^2 + (1 - P_i)(u_i|D_i = 0)^2 \\ &= P_i(1 - P_i)^2 + (1 - P_i)(-P_i)^2 \\ &= P_i(1 - P_i)(1 - P_i) + (1 - P_i)(P_i)^2 \\ &= (1 - P_i)[P_i(1 - P_i) + (P_i)^2] \\ &= (1 - P_i)P_i \end{aligned}$$

Thus the variance varies across individuals (based on their X 's).

Logit Model

We want an alternate with predicted values that lie between 0 and 1:

Logit Model:

The ratio of the probability of success to the probability of failure (i.e. probability of working to not working).

$$odds_i = \frac{P_i}{1-P_i}$$

Now take the natural log of the odds ratio:

$$L_i = \ln\left(\frac{P_i}{1-P_i}\right)$$

So, the relationship that will be estimated is:

$$L_i = \beta_1 + \beta_2 X_2 + \dots + u_i$$

This solves the 0-1 boundary problem:

- As P_i approaches 0, then the logit approaches $\ln(0)$, which is equal to $-\infty$.
- As P_i approaches 1, then the logit approach $\ln(\infty)$, which is equal to ∞ .

That is, very large (or small) values of X will still be mapped to probabilities between 0 and 1.

Logit Model

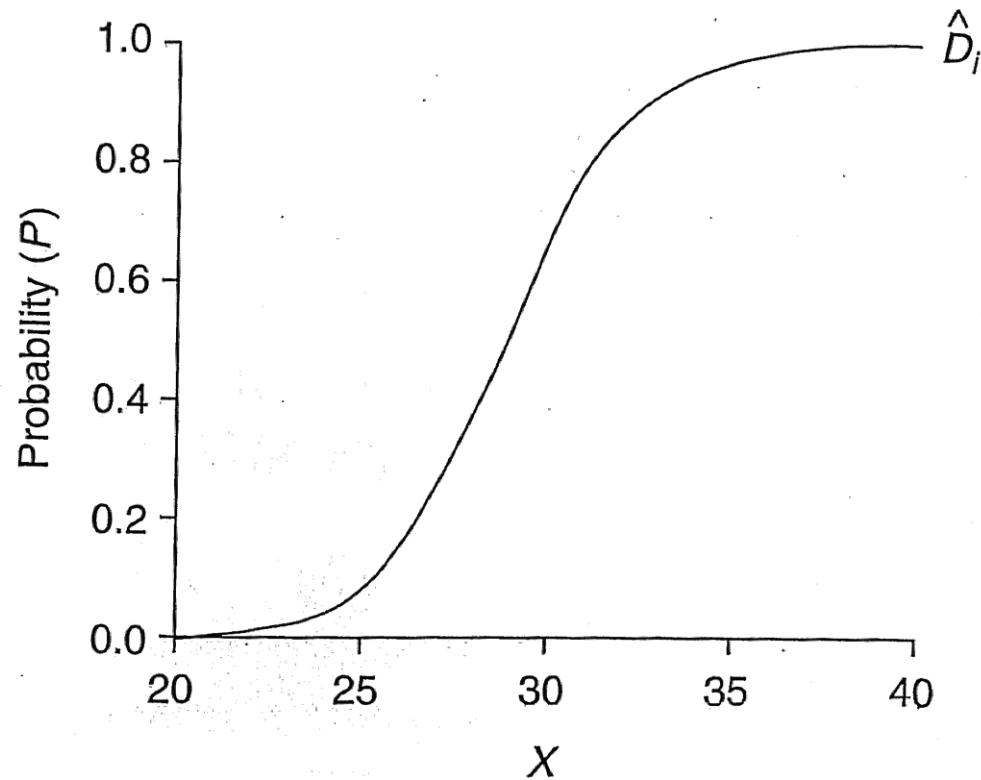


Figure 12.2 The logit function

Logit Model

$$\ln\left(\frac{P_i}{1 - P_i}\right) = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u_i$$

Note: The coefficients β have no easy, logical interpretation.

What we really want is the effect of X_2 on the probability of $D=1$ (i.e. P_i).
Here are two methods to convert the Logit estimate to what we want:

1. Compute the effect of a 1 unit change in X_2 for the average person:
 - A. Insert the average X values into the estimated equation and get \hat{D} .
 - B. Repeat this using X_2+1 instead of the average X_2 .
 - C. Difference is the marginal effect of X_2 for the average individual.

2. Take the partial derivative:
$$\frac{\partial \hat{D}_i}{\partial X_{ji}} = \hat{\beta}_j \hat{D}_i (1 - \hat{D}_i)$$

Thus, we simply insert the estimate β and \hat{D} to get the marginal effect. We will see how this is derived shortly.

Logit Model

$$\ln\left(\frac{P_i}{1 - P_i}\right) = \beta_1 + \beta_2 X_{2i} + u_i$$

$$\ln(P_i) - \ln(1 - P_i) = \beta_1 + \beta_2 X_{2i} + u_i$$

Let's derive the marginal effect of X_2 :

$$\frac{\partial}{\partial X_2} [\ln(P_i) - \ln(1 - P_i)] = \frac{\partial}{\partial X_2} [\beta_1 + \beta_2 X_{2i} + u_i]$$

$$\frac{1}{P_i} \frac{\partial P_i}{\partial X_2} + \frac{1}{1 - P_i} \frac{\partial P_i}{\partial X_2} = \beta_2$$

$$\frac{\partial P_i}{\partial X_2} \left[\frac{1}{P_i} + \frac{1}{1 - P_i} \right] = \beta_2$$

$$\frac{\partial P_i}{\partial X_2} = \beta_2 P_i (1 - P_i)$$

So, this is the why the marginal effect of interest can be computed using the equation on the previous slide.

Logit Model

Let's solve for P_i :

$$\ln\left(\frac{P_i}{1 - P_i}\right) = \beta_1 + \beta_2 X_{2i} + u_i$$

$$\frac{P_i}{1 - P_i} = e^{(\beta_1 + \beta_2 X_{2i} + u_i)}$$

$$P_i = e^{(\beta_1 + \beta_2 X_{2i} + u_i)}(1 - P_i)$$

$$P_i[1 + e^{(\beta_1 + \beta_2 X_{2i} + u_i)}] = e^{(\beta_1 + \beta_2 X_{2i} + u_i)}$$

$$P_i = \frac{e^{(\beta_1 + \beta_2 X_{2i} + u_i)}}{1 + e^{(\beta_1 + \beta_2 X_{2i} + u_i)}}$$

$$P_i = \frac{1}{1 + e^{-(\beta_1 + \beta_2 X_{2i} + u_i)}}$$

Logit Model

The Likelihood Function:

$$\begin{aligned} L &= \prod_{i=1}^n P_i^{Y_i} (1 - P_i)^{1 - Y_i} \\ &= \prod_{i=1}^n \left[\frac{1}{1 + e^{-(\beta_1 + \beta_2 X_{2i})}} \right]^{Y_i} \left[1 - \frac{1}{1 + e^{-(\beta_1 + \beta_2 X_{2i})}} \right]^{1 - Y_i} \end{aligned}$$

This can be estimated with maximum likelihood.

```
. logit married age educ
```

```
Iteration 0:  log likelihood = -317.9885
Iteration 1:  log likelihood = -311.0301
Iteration 2:  log likelihood = -310.85059
Iteration 3:  log likelihood = -310.85041
Iteration 4:  log likelihood = -310.85041
```

```
Logistic regression                                         Number of obs     =      935
                                                               LR chi2(2)       =     14.28
                                                               Prob > chi2      =    0.0008
Log likelihood = -310.85041                                Pseudo R2        =    0.0224
```

married	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
age	.1181068	.0362492	3.26	0.001	.0470597 .1891539
educ	-.0889811	.0483671	-1.84	0.066	-.1837787 .0058166
_cons	-.5197703	1.317726	-0.39	0.693	-3.102466 2.062925

```
. mfx
```

```
Marginal effects after logit
y = Pr(married) (predict)
= .8992384
```

variable	dy/dx	Std. Err.	z	P> z	[95% C.I.]	X
age	.0107015	.00315	3.40	0.001	.004528 .016875	33.0802
educ	-.0080625	.00433	-1.86	0.063	-.016547 .000422	13.4684

Probit Model

Whether or not the dummy outcome variable is equal to 0 or 1 can be motivated as a latent variable problem:

Suppose there is a latent variable that determines whether or not someone will work:

$$D = \begin{cases} 1 & \text{if } Y^* > 0 \\ 0 & \text{if } Y^* \leq 0 \end{cases}$$

For example, you could think of Y^* as capturing the benefit of working. A person will work if the benefit exceeds 0 and will not work if it is less than 0.

$$\begin{aligned} Pr(D_i = 1) &= Pr(Y^* > 0) \\ &= Pr(\beta_1 + \beta_2 X_{2i} + u_i > 0) \\ &= Pr(u_i > -\beta_1 - \beta_2 X_{2i}) \\ &= Pr(u_i < \beta_1 + \beta_2 X_{2i}) \\ &= \Phi(\beta_1 + \beta_2 X_{2i}) \\ &= \Phi(X'\beta) \end{aligned}$$

Where the last step comes from u_i being distributed as normal.

Probit Model

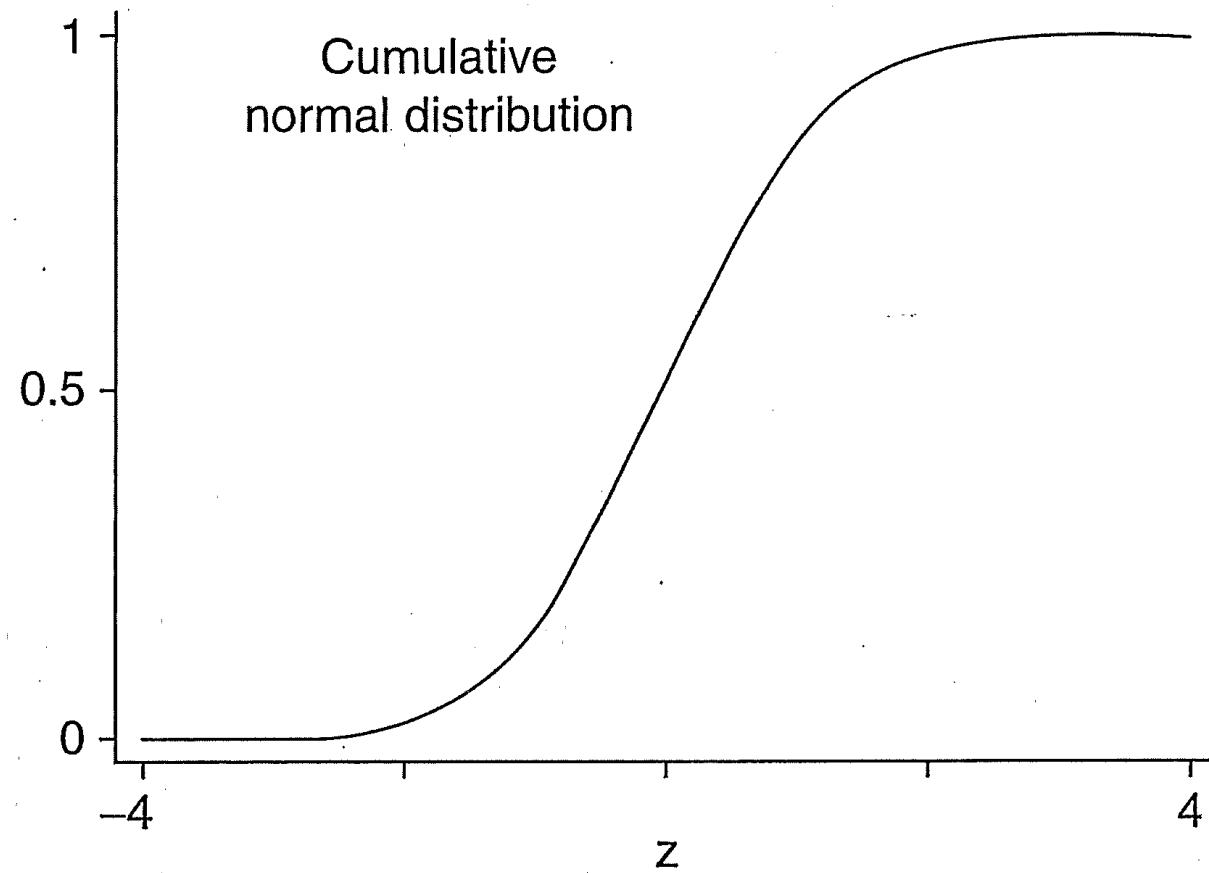


Figure 12.3 Cumulative normal distribution

Probit Model

The Likelihood Function:

$$\begin{aligned} L &= \prod_{i=1}^n P_i^{Y_i} (1 - P_i)^{1 - Y_i} \\ &= \prod_{i=1}^n \Phi^{Y_i} (1 - \Phi)^{1 - Y_i} \end{aligned}$$

$$\ln L = \sum_{i=1}^n Y_i \ln(\Phi) + (1 - Y_i) \ln(1 - \Phi)$$

This can be estimated with maximum likelihood.

```
. probit married age educ
```

```
Iteration 0: log likelihood = -317.9885  
Iteration 1: log likelihood = -310.81677  
Iteration 2: log likelihood = -310.76214  
Iteration 3: log likelihood = -310.76213
```

```
Probit regression  
Number of obs = 935  
LR chi2(2) = 14.45  
Prob > chi2 = 0.0007  
Log likelihood = -310.76213  
Pseudo R2 = 0.0227
```

married	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
age	.061467	.0185625	3.31	0.001	.0250852 .0978489
educ	-.0465612	.0251888	-1.85	0.065	-.0959303 .0028079
_cons	-.1342543	.6832534	-0.20	0.844	-1.473406 1.204898

```
. mfx
```

```
Marginal effects after probit
```

```
y = Pr(married) (predict)  
= .89831016
```

variable	dy/dx	Std. Err.	z	P> z	[95% C.I.]	X
age	.01092	.00324	3.37	0.001	.004572 .017268	33.0802
educ	-.0082719	.00445	-1.86	0.063	-.016996 .000452	13.4684

Probit Model

The Probit and Logit estimates will produce similar results in most cases.

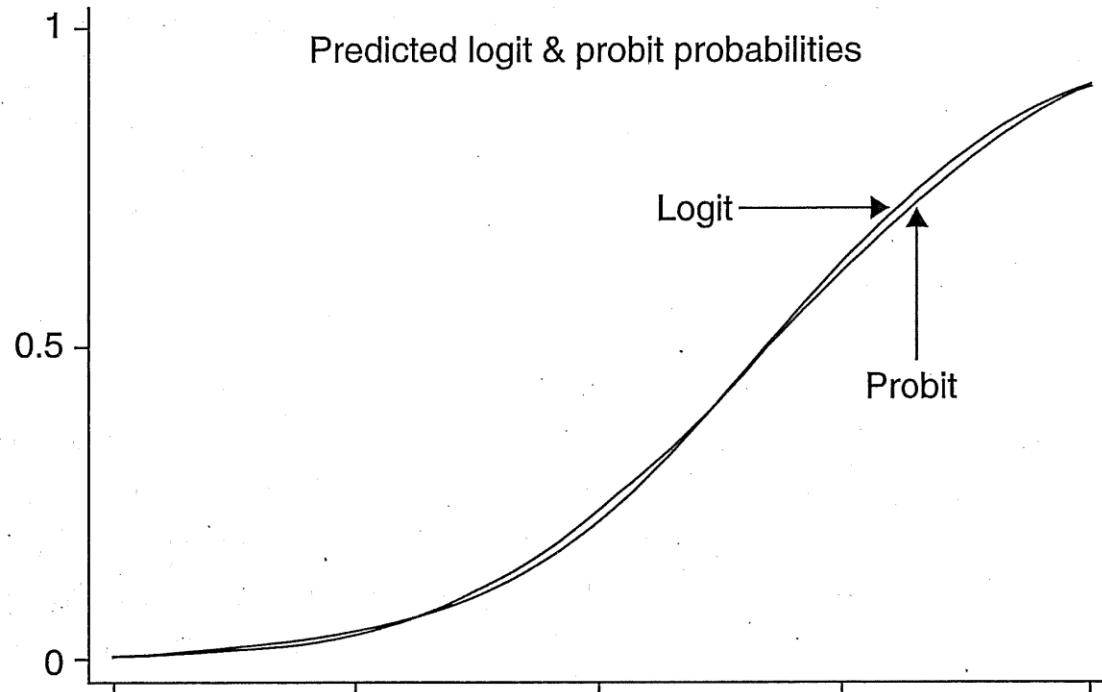


Figure 12.4 Differences between logit and probit probabilities

Probit Model

Which model should you use when you have a binary outcome variable?

1. Linear Probability Model
2. Logit Model
3. Probit Model

Pro and Con of the LPM:

Pro: Easy to interpret the coefficients.

Con: Can produce predictions outside of the logical range

Note: Can use robust errors to account for heteroskedasticity

Pro and Con of Probit and Logit:

Pro: Predicted values always bounded by 0 and 1

Con: Difficult to interpret the coefficients

Lesson 13

Ordered Probit Tobit

Outline

Previous Lesson:

1. Maximum Likelihood
2. Probit, Logit, Probit

This Lesson Lesson:

1. Ordered Probit Model
2. Tobit Model

Next Lesson:

1. The Ideal Experiment

When Y is not normal

We have considered the case where Y is free to take on any value. We have shown that in this case, OLS is the best choice.

We then looked at the case of a binary outcome (0,1) and showed that OLS has some shortcomings. So we proposed Logit and Probit models as alternatives to the linear probability model.

But, as you can imagine, there are a lot of other cases that exist in practice:

1. Ordered: Y values can be one of several values (e.g. 0,1,2).
2. Truncated: Y is not observed if it falls above/below some value.
3. Censored: Y takes on a fixed value if it falls above/below some value.
4. Categorical: Y takes on discrete outcomes that are not ordered.

When Y is not normal

- There are many methods that have been developed for each of these cases.
- In specific fields, some of them are very common.
- Focusing on these methods is not really the direction that applied economics has gone.
- So, we are going to discuss a few common cases, the implications, and the approaches that have been proposed.

Ordered Probit

Recall the generic form of a likelihood function:

$$l = \prod_{i=1}^n P_i^{Y_i} (1 - P_i)^{1-Y_i}$$

And the log-likelihood function:

$$L = \sum_{i=1}^n Y_i \ln(P_i) + (1 - Y_i) \ln(1 - P_i)$$

- When $Y=1$ a lot, then L will be larger when P_i is large.
- When $Y=0$ a lot, then L will be larger when P_i is small.

Thus a lot of 1 outcomes will support a high value of P_i .

Ordered Probit

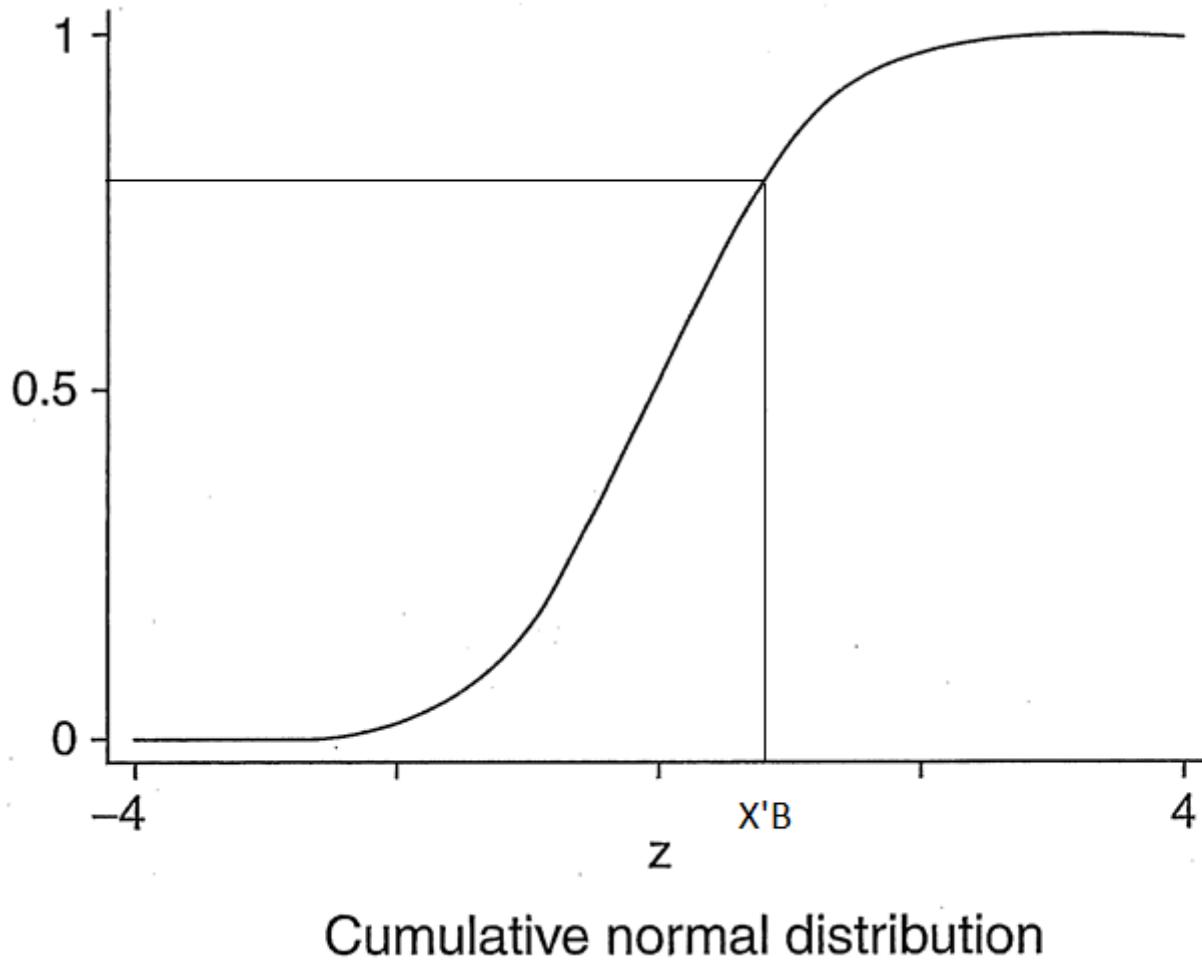
The Probit model likelihood function:

$$L = \prod_{i=1}^n \Phi(X'\beta)^{Y_i} (1 - \Phi(X'\beta))^{1-Y_i}$$

- When $Y=1$ a lot then L will be larger when the cdf $\Phi(X\beta)$ is larger, which will occur if β is positive.
- When $Y=0$ a lot then L will be larger when $1 - \Phi(X\beta)$ is larger, which will occur when β is negative.

Thus larger values of β indicate that X causes the outcome $Y=1$.

Ordered Probit



Ordered Probit

Consider a case where an individual is choosing between several options (rather than just Yes or No) and these options can be ordered.

For example:

1. A student decides between the following educational options: high school only, community college, four-year college.
2. A patient experiences the following outcomes: fully cured, remains sick, death.

In order to evaluate these in a regression model, we need to assign numerical values to each outcome (e.g. 0, 1, 2, 3). However:

- A. OLS treats the actual numbers as meaningful even though they are not.
- B. OLS might predict outside of the logical range (e.g. <0 or >3).
- C. The errors will not be normally distributed and will be heteroskedastic.

Ordered Probit

An Ordered Probit model is just like a binary Probit but with several outcomes.

For example, if someone chooses HS, then $Y=0$, if community college, then $Y=1$, if four-year college, then $Y=2, \dots$

$$Y = \begin{cases} 0 & \text{if } Y^* \leq \mu_0 \\ 1 & \text{if } \mu_0 < Y^* \leq \mu_1 \\ 2 & \text{if } \mu_1 < Y^* \leq \mu_2 \\ \vdots & \\ N & \text{if } \mu_N < Y^* \end{cases}$$

The latent variable Y^* captures the underlying decision variable. You might think of it as the desire for education in this example.

Note that it must be the case the outcomes can be ordered in a meaningful way.

Ordered Probit

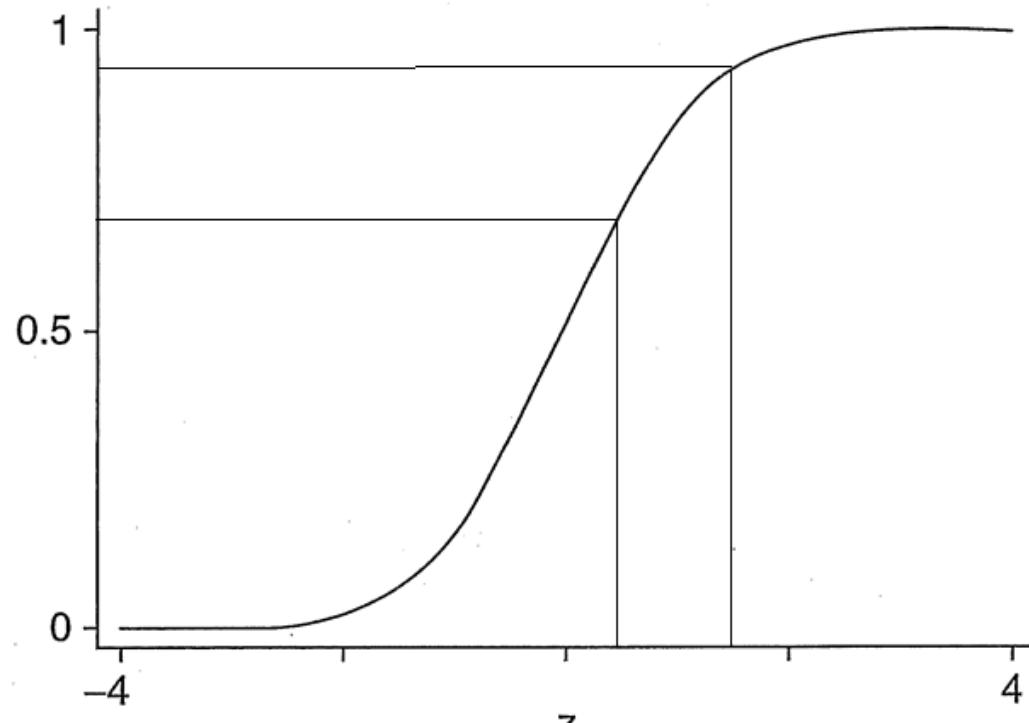
In the case of the Probit model:

$$\begin{aligned} Pr(D_i = 1) &= Pr(Y^* > 0) \\ &= \Phi(X'\beta) \end{aligned}$$

In the case of the ordered Probit:

$$\begin{aligned} Pr(Y_i = 2) &= Pr(\mu_1 < Y^* < \mu_2) \\ &= Pr(\mu_1 < X'\beta + \epsilon_i < \mu_2) \\ &= Pr(\mu_1 - X'\beta < \epsilon_i < \mu_2 - X'\beta) \\ &= \Phi(\mu_2 - X'\beta) - \Phi(\mu_1 - X'\beta) \\ &= \Phi_2 - \Phi_1 \end{aligned} \tag{1}$$

Ordered Probit



Cumulative normal distribution

Ordered Probit

The ordered Probit likelihood function:

$$L = \prod_{i=1}^n (\Phi_0)^{Z_0} (\Phi_1 - \Phi_0)^{Z_1} (\Phi_2 - \Phi_1)^{Z_2} \dots$$

And the loglikelihood function:

$$L = \sum_{i=1}^n Z_0 \ln(\Phi_0) + Z_1 \ln(\Phi_1 - \Phi_0) + Z_2 \ln(\Phi_2 - \Phi_1) + \dots$$

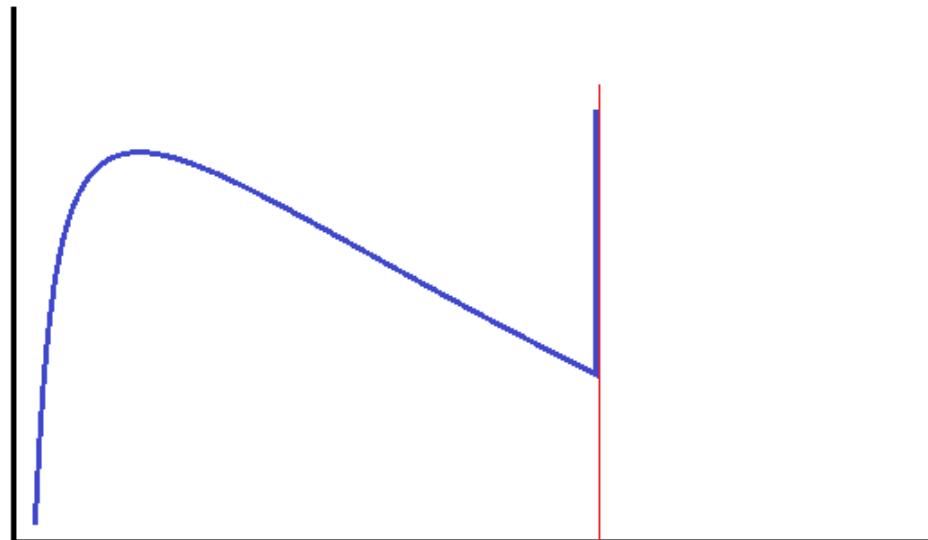
Where Z_j is a dummy that equals 1 if $Y_i=j$.

- When Y takes on the value 2 a lot, then L will be larger when the β s take on values such that $\Phi_2 - \Phi_1$ is large.

Censored Data

Censored data occurs when Y is replaced by some constant if it is very large or small. For example:

1. A data set that caps reported income at \$200,000, so everyone that earns more than this is nonetheless listed as earning \$200k:
2. An exam that does not allow scores below 200 or above 800, so those who could have scored higher (or lower) are lumped at these two scores:



Censored Data

Why is censored data a problem? It causes biased estimates when using OLS.

Consider an estimate of the effect of years of experience on income:

$$Y_i = \beta_0 + \beta_1 Exp_i + u_i$$

Income is censored at \$200k:

$$Y_i = \begin{cases} Y_i^* & \text{if } Y_i^* \leq 200,000 \\ 200,000 & \text{if } Y_i^* \geq 200,000 \end{cases}$$

Suppose we just ran OLS on the censored version of the data. Note that β will be biased if causes Experience to be correlated with the error term:

- Exp is correlated with Y , then it is correlated with being censored.
- Being censored is correlated with the error.

Tobit

Tobit model – The outcome variable takes on the value of the latent variable when it exceeds some threshold and is otherwise 0:

$$Y_i = \begin{cases} Y_i^* & \text{if } Y_i^* > 0 \\ 0 & \text{if } Y_i^* \leq 0 \end{cases}$$

There are many variations on this:

Truncation lower: $Y_i = \begin{cases} Y_i^* & \text{if } Y_i^* > Y_L \\ Y_L & \text{if } Y_i^* \leq Y_L \end{cases}$

Truncation upper: $Y_i = \begin{cases} Y_i^* & \text{if } Y_i^* < Y_U \\ Y_U & \text{if } Y_i^* \geq Y_U \end{cases}$

Truncation both: $Y_i = \begin{cases} Y_i^* & \text{if } Y_L < Y_i^* < Y_U \\ Y_L & \text{if } Y_i^* \leq Y_L \\ Y_U & \text{if } Y_i^* \geq Y_U \end{cases}$

Tobit

The Tobit model for censored data can be solved using maximum likelihood. We will not derive the likelihood model.

The Tobit model can be estimated in STATA as long as you specify the type of censorship.

Truncated Data

Truncated data occurs when Y is not observed if it is very large or small. For example:

Data is collected on companies but only for companies with revenue of at least \$500,000.

This is actually a very difficult problem to solve.

James Heckman has proposed several sample selection correction methods to address it.

If this is something of interest, you should check out Heckman Correction.

Order Probit

Tobit

Examples

Example Tobit – censored data

Let's do an exercise where see what a Tobit does and test its performance.

Start with data on individual wages, education, experience, tenure.

1. Run the regression of wages on education
[this produces the “correct” coefficients”]
2. Artificially censor wages at some level
[to replicate what could occur in data]
3. Rerun the regression using OLS
4. Rerun the regression using a Tobit
5. Compare the performance of #3 and #4 against #1.

Command

tobit depvar [indepvars] [if] [in] [weight] , ll[(#)] ul[(#)]

Example Tobit – censored data

```
. reg wage educ exper
```

Source	SS	df	MS	Number of obs	=	900
Model	33337864.5	2	16668932.3	F(2, 897)	=	130.88
Residual	114238893	897	127356.625	Prob > F	=	0.0000
Total	147576757	899	164156.571	R-squared	=	0.2259

wage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ	84.67465	5.825556	14.54	0.000	73.24135	96.10796
exper	34.32022	2.861293	11.99	0.000	28.70461	39.93583
_cons	-597.5725	97.88015	-6.11	0.000	-789.6733	-405.4717

Example Tobit – censored data

```
. summ wage, detail
```

monthly earnings

	Percentiles	Smallest		
1%	321.5	115		
5%	433	200		
10%	505	233	Obs	900
25%	675	260	Sum of Wgt.	900
50%	912.5		Mean	964.2644
		Largest	Std. Dev.	405.1624
75%	1161.5	2668		
90%	1443	2771	Variance	164156.6
95%	1704.5	3078	Skewness	1.203925
99%	2309	3078	Kurtosis	5.746906

```
gen wage_censored= wage  
replace wage_censored=1500 if wage_censored>1500
```

Example Tobit – censored data

```
. reg wage_censored educ exper
```

Source	SS	df	MS	Number of obs	=	900
Model	21093989	2	10546994.5	F(2, 897)	=	126.80
Residual	74612484	897	83180.0267	Prob > F	=	0.0000
Total	95706473	899	106458.813	R-squared	=	0.2204

wage_censo~d	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ	67.44233	4.707996	14.33	0.000	58.20236	76.68231
exper	27.23335	2.31239	11.78	0.000	22.69503	31.77168
_cons	-310.5101	79.10308	-3.93	0.000	-465.7588	-155.2614

Example Tobit – censored data

```
. tobit wage_censored educ exper , ul(1500)
```

Tobit regression

Number of obs = 900

LR chi2(2) = 233.61

Prob > chi2 = 0.0000

Log likelihood = -5945.2811

Pseudo R2 = 0.0193

wage_censo~d	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
educ	75.58838	5.213729	14.50	0.000	65.35587 85.8209
exper	30.48323	2.54168	11.99	0.000	25.4949 35.47155
_cons	-445.5704	87.51488	-5.09	0.000	-617.3279 -273.8129
/sigma	311.9708 7.90098				296.4643 327.4773

Obs. summary: 0 left-censored observations

817 uncensored observations

83 right-censored observations at wage_censo~d>=1500

Example Tobit – censored data

How many observations were censored?

How did the Tobit perform compared to OLS?

Can you interpret the coefficients the way you usually would?

Why doesn't the Tobit perfectly replicate the uncensored OLS?

What are some reasons why data might be censored?

Ordered Probit

```
oprobit depvar [indepvars] [if] [in] [weight] [, options]
```

Description (from Stata): The actual values taken on by the dependent variable are irrelevant, except that larger values are assumed to correspond to "higher" outcomes.

When this is run in Stata, the program chooses both the Bs and the cutoff values.

You can still compute marginal effects, but you must tell it one outcome that you are interested in (e.g. outcome 5).

Consider an example where people assess their happiness with some product on a scale of 1 to 5.

Ordered Probit

Suppose people answer a question about how happy they are where 1=“very unhappy” and 5=“very happy”:

. tab happy

happy	Freq.	Percent	Cum.
1	101	11.22	11.22
2	400	44.44	55.67
3	260	28.89	84.56
4	90	10.00	94.56
5	49	5.44	100.00
Total	900	100.00	

Ordered Probit

Ordered probit regression

Number of obs = 900

LR chi2(2) = 30.23

Prob > chi2 = 0.0000

Log likelihood = -1202.8676

Pseudo R2 = 0.0124

happy	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
wage	.3547298	.0927554	3.82	0.000	.1729325	.5365272
educ	.0432524	.0170221	2.54	0.011	.0098897	.0766151
/cut1	-.3130921	.2221235			-.7484461	.1222618
/cut2	1.07117	.2233136			.6334833	1.508856
/cut3	1.958456	.226716			1.514101	2.402811
/cut4	2.557311	.2334965			2.099667	3.014956

Ordered Probit

```
. mfx compute, predict(outcome(4))
```

Marginal effects after oprobit

```
y = Pr(happy==4) (predict, outcome(4))
= .09940064
```

variable	dy/dx	Std. Err.	z	P> z	[95% C.I.]	x
wage	.0456211	.01263	3.61	0.000	.020861	.070381	.964264	
educ	.0055626	.00225	2.47	0.013	.00115	.009975		13.48

This is computed using the average characteristics of those in this range of Y.

You may get very different marginal effects looking at different outcomes.

Lesson 14

Introduction to Causality: Ideal Experiment

Outline

Previous Lesson:

1. Ordered Probit
2. Tobit

This Lesson:

1. Selection Problem
2. Ideal Experiment

Next Lesson:

1. Difference-in-Differences

[Note: I have moved a thorough discussion of individual fixed effects to a later lesson]

Topics Covered

1. Univariate OLS
2. Multivariate OLS
3. Dummy Variables and Interaction Terms
4. Hypothesis Testing
5. Multicollinearity
6. Heteroskedasticity
7. Autocorrelation
8. Practical Considerations & Functional Form
9. **Endogeneity: 3 Sources
10. Maximum Likelihood
11. Linear Probability Model
12. Logit, Probit, Order Probit, Tobit

Asteriou and Hall: Chapters 1 – 9, 11-12, 20

Topics Covered

1. The Ideal Experiment
2. Fixed Effects
3. Single Difference
4. Difference-in-Difference
5. Matching
6. Experiments
7. Instrumental Variables
8. Regression Discontinuity

Mostly Harmless Econometrics: Chapters 2, 4, 5, 6

The Causality Problem

Sources of Endogeneity (any time the Xs are correlated with u):

- A. Omitted variables
- B. Reverse causality
- C. Measurement error

The most common (and intuitive) is omitted variable bias.

We have seen this mathematically:

$$\beta_1^* = \beta_1 + \beta_2 \delta_1$$

$$E(\hat{\boldsymbol{b}}_R^* | \mathbf{X}) = \boldsymbol{\beta}^* + (\mathbf{X}^{*\prime} \mathbf{X}^*)^{-1} \mathbf{X}^{*\prime} \mathbf{X}^O \boldsymbol{\beta}^O$$

The underlying sources of omitted variable bias is frequently selection bias (that people sort into various treatments of interest – e.g. college).

The Selection Problem

We are interested in estimating whether or not going to the hospital improves one's health. Suppose that 1 is excellent health and 5 is poor health.

$$Health_i = \beta_0 + \beta_1 HospitalVisit_i + u_i$$

We expect β to be negative in this regression. We observe the following data:

Group	Sample Size	Mean health status	Std. Error
Hospital	7774	2.79	0.014
No Hospital	90049	2.07	0.003

What is the problem?

What omitted variables result from the selection process?

Ideal Experiment

$$Y_i = \beta_0 + \beta_1 D_i + u_i$$

It is useful to think about “hospital visit” as a dummy treatment:

$$D = \begin{cases} 0 & \text{no hospital} \\ 1 & \text{hospital} \end{cases}$$

We can think of each person i as having two potential outcomes. One that occurs when they go the hospital and one that occurs when they do not:

$$\text{Potential Outcome} = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases}$$

What we would like to observe is:

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

That is, we want to observe the outcome (health) for the same person under two scenarios: when the person is treated and untreated (the counterfactual).

Ideal Experiment

Unfortunately, we can only observe a person in one state of the world. That is, we only observe Y_{0i} or Y_{1i} but not both.

What we observe when comparing the means:

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0]$$
$$= \underbrace{E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1]}_{\text{Treatment on Treated}} + \underbrace{E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]}_{\text{Selection Bias}}$$

We observe Y_i for those who went to a hospital and those who did not.

This contains two effects: the effect of going the hospital for those who went (which would like to know) and the difference between the two groups ($D=1$ and $D=0$) in the absence of treatment (selection bias).

Ideal Experiment

$$E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]$$

The selection bias problem stems from the fact that those who go to the hospital ($D=1$) do not have the same health as those who do not go to the hospital ($D=0$) [in the absence of medical care Y_0].

The ideal randomized experiment fixes this problem.

Suppose treatment D is assigned randomly.

Individuals in each group $D=0$ and $D=1$ should be, on average, identical.

Most importantly, they can be expected to have the same health Y_0 in the absence of a hospital visit:

$$E[Y_{0i}|D_i = 1] = E[Y_{0i}|D_i = 0]$$

Ideal Experiment

So, in a randomized experiment:

$$\begin{aligned} E[Y_i|D_i = 1] - E[Y_i|D_i = 0] &= E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] \\ &= E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] \end{aligned}$$

Where the second equality holds because, in expectation, the two groups have the same health. The expected selection bias is 0.

Further, because treatment is random, the expected treatment effect should be the same for the two groups.

So, instead of getting just getting the treatment on the treated effect (i.e. the benefit for people who choose to go to the hospital), we get the average effect for all individuals:

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = E[Y_{1i} - Y_{0i}]$$

Ideal Experiment

A classic example examines the effect of government employment training programs:

$$Wage_i = \beta_0 + \beta_1 Training_i + u_i$$

Frequently these studies find a negative β_1 .

What selection bias problems are likely?

What omitted variables would be useful to include?

Program evaluations with random assignment do not (generally) produce negative estimates.

Ideal Experiment

Another (too) frequently studied question is the effect of class size on student performance. A famous study was the Tennessee STAR program:

$$Score_i = \beta_0 + \beta_1 SmallClass_i + u_i$$

Students who entered STAR in kindergarten					
Variable	Small	Regular	Regular/Aide	Joint	P-value
1. Free lunch	.47	.48	.50		.09
2. White/Asian	.68	.67	.66		.26
3. Age in 1985	5.44	5.43	5.42		.32
4. Attrition rate	.49	.52	.53		.02
5. Class size in kindergarten	15.10	22.40	22.80		.00
6. Percentile score in kindergarten	54.70	48.90	50.00		.00

Students were randomly assigned to three groups. Why do the researchers present comparisons of the observable characteristics of students?

What is the treatment effect in this study?

Ideal Experiment

The regression equations for the experiment with no controls:

$$Score_i = \beta_0 + \beta_1 SmallClass_i + u_i$$

As long as the randomization worked, β_1 will be an unbiased estimate of the effect of being enrolled in a smaller class.

If we are concerned that there is some imbalance in the estimates, then we should rerun the regression with additional controls:

$$Score_i = \beta_0 + \beta_1 SmallClass_i + \beta_2 White_i + \beta_3 Girl_i + \beta_4 FreeLunch_i + u_i$$

We should expect to get a similar estimate of β_1 when we add controls. If the estimate changes a lot, then we may be concerned that the randomization process did not work well.

Ideal Experiment

Table 2.2.2: Experimental estimates of the effect of class-size assignment on test scores

Explanatory variable	(1)	(2)	(3)	(4)
Small class	4.82 (2.19)	5.37 (1.26)	5.36 (1.21)	5.37 (1.19)
Regular/aide class	.12 (2.23)	.29 (1.13)	.53 (1.09)	.31 (1.07)
White/Asian (1 = yes)	—	—	8.35 (1.35)	8.44 (1.36)
Girl (1 = yes)	—	—	4.48 (.63)	4.39 (.63)
Free lunch (1 = yes)	—	—	-13.15 (.77)	-13.07 (.77)
White teacher	—	—	—	-.57 (2.10)
Teacher experience	—	—	—	.26 (.10)
Master's degree	—	—	—	-0.51 (1.06)
School fixed effects	No	Yes	Yes	Yes
R ²	.01	.25	.31	.31

Ideal Experiment

The generic experimental regression equation:

$$Y_i = \beta_0 + \beta_1 D_i + u_i$$

Then the two groups have the following expected means:

$$E[Y_i | D_i = 0] = \beta_0$$

$$E[Y_i | D_i = 1] = \beta_0 + \beta_1$$

The difference in the means is the coefficient of interest in our regression:

$$E[Y_i | D_i = 1] - E[Y_i | D_i = 0] = \beta_1$$

Lesson 15

Single Difference Difference-in-Differences

Outline

Previous Lesson:

1. Selection Problem
2. Ideal Experiment

This Lesson:

1. Single Difference
2. Difference-in-Differences

Next Lesson:

1. Matching

Single Difference

So far in this class we have been using cross-sectional data for our regressions:

- across people: wages vs educ
- across companies: CEO salary vs sales

The biggest concern with this approach is omitted variables:

- differences in aptitude, motivation
- differences in local economy, CEO ability

Often times it will be difficult or impossible to locate all of these omitted variables. As a result, we probably get a biased estimate.

Now we are going to switch to panel data:

- same person over time: change in wage vs change in educ
- same company over time: change in salary vs change in sales

Single Difference

Ideal Experiment: Comparing the means in the after period if assignment was random:

$$\begin{aligned} E[Y_i|D_i = 1] - E[Y_i|D_i = 0] &= E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] \\ &= \underbrace{E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1]}_{\text{Assignment bias}} + \underbrace{E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]}_{\text{Treatment effect}} \\ &= E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] \end{aligned}$$

That is, if assignment into treatment was random, then simply comparing the means of the two groups in the after period will result in an unbiased estimate (free from selection bias).

Unfortunately, the ideal experiment is often not an option: cost, ethics, politics, etc.

Single Difference

Suppose we are interested in estimating the effect of a state policy that bans smoking in bars and restaurants. We want to estimate if there were fewer hospital admissions for breathing related issues.

In 2003, DC passed such a ban. We observe residents' hospital admissions before (2000) and after the ban (2005).

	Before	After
DC	0.121	0.093

$$\text{Difference} = \text{After} - \text{Before} = 0.093 - 0.121 = -0.028$$

So the first difference estimate is simply the difference between the after and before averages.

Single Difference

We can get this difference measure by estimating the following regression equation:

$$Hospital = \beta_0 + \beta_1 Yr2005 + u$$

“Hospital” is a binary variables, so we can interpret the coefficients as percentage points.

β_0 – the rate of hospitalization for breathing issues in 2000
(i.e. when $Yr2005 = 0$)

β_1 – the change in the rate of hospitalization rate in 2005
(i.e. when $Yr2005=1$). We are interested in β_1 .

$$\text{Before} = \beta_0 + \beta_1 * 0 = \beta_0$$

$$\text{After} = \beta_0 + \beta_1 * 1 = \beta_0 + \beta_1$$

$$\text{After} - \text{Before} = (\beta_0 + \beta_1) - (\beta_0) = \beta_1$$

Single Difference

$$\widehat{Hospital} = 0.121 - 0.028 * Yr2005$$

(0.017) (0.021)

Why are we doing this using a regression rather than just taking the difference in means?

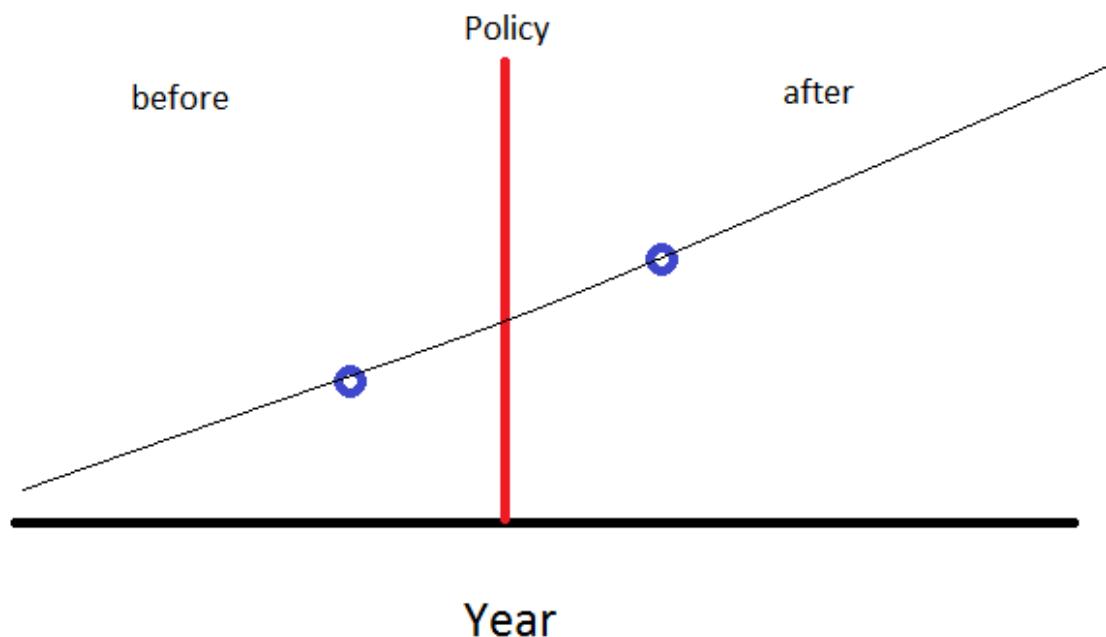
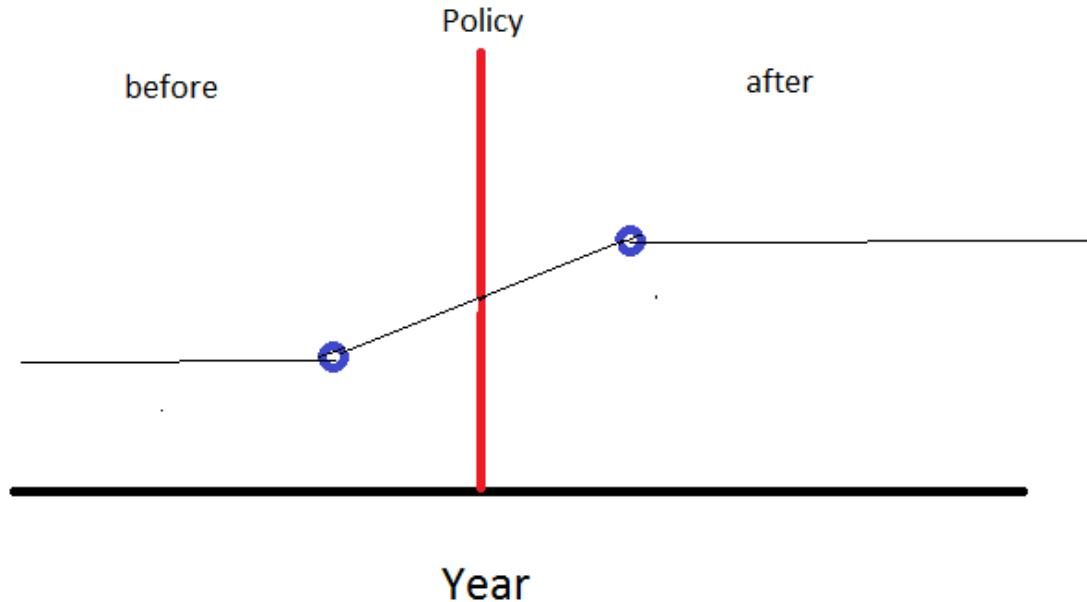
Because we might want to control for additional factors that could change over time.

For example, perhaps more people worked in factories in 2000, or perhaps people were more likely to have insurance in 2000:

$$\widehat{Hospital} = 0.077 - 0.022 * Yr2005 + 0.048 * Insured + 0.191 * Factory$$

(0.008) (0.010) (0.012) (0.045)

How do these controls change the outcome?



Single Difference

Advantages:

- not across people/places/companies
- within same person/place/company over time
- can add control for compositional changes

Disadvantages:

- other important factors might change over time
 - policies
 - economic conditions
 - environmental conditions
- pre-existing trends

It may be hard to argue that the policy change (smoking ban, minimum wage law, training program) was the only thing that affected people.

Difference-in-Differences

Difference-in-Differences – a method of estimation that involves comparing the change in a treatment group over time against the change in a control group over time.

Treatment group – the group that is affected by the policy

Control group – the group that is not affected by the policy

Purpose of difference-in-differences:

- By taking the difference over time within a group, we control for omitted variables that do not change over time.
- The control group helps us to control for factors that do change over time.
- Assumption – the control group would follow the same time trend as the treated group would have in the absence of the policy. This is often called the “counter-factual”

Difference-in-Differences

Recall our examination of the effect of the smoking ban on breathing related hospital visits in Washington, D.C. Suppose we also have data on hospitalizations for Maryland.

	Before	After
DC	0.121	0.093
MD	0.125	0.117

Difference for DC: $\text{After} - \text{Before} = 0.093 - 0.121 = -0.028$

Difference for MD: $\text{After} - \text{Before} = 0.117 - 0.125 = -0.008$

Difference-in-Differences: $-0.028 - (-0.008) = -0.028 + 0.008 = -0.020$

Difference-in-Differences

Difference-in-Differences – the change in the treated group minus the change in the control group.

$$(\text{Treated_After} - \text{Treated_Before}) - (\text{Control_After} - \text{Control_Before})$$

- The control group “controls” for the omitted factors that would otherwise cause us to get a biased estimate of the effect of the policy.
- This will not work if the control group is not a good match for the treated group (not a good counterfactual).

Difference-in-Differences

$$Outcome = \beta_0 + \beta_1 Treated + \beta_2 After + \beta_3 Treated * After + u$$

B_0 – mean for the control group in before period ($T=0, A=0$)

B_0+B_1 – mean for the treatment group in before period ($T=1, A=0$)

B_0+B_2 – mean for the control group in after period ($T=0, A=1$)

$B_0+B_1+B_2+B_3$ – mean for the treatment group in after period ($T=1, A=1$)

B_3 is the difference-in-differences effect of the policy:

$$(Treated_After - Treated_Before) - (Control_After - Control_Before)$$

$$= [(B_0+B_1+B_2+B_3) - (B_0+B_1)] - [(B_0+B_2) - (B_0)]$$

$$= [B_2+B_3] - [B_2]$$

$$= B_3$$

Difference-in-Differences

There is an alternative way of writing (and thinking about difference-in-differences):

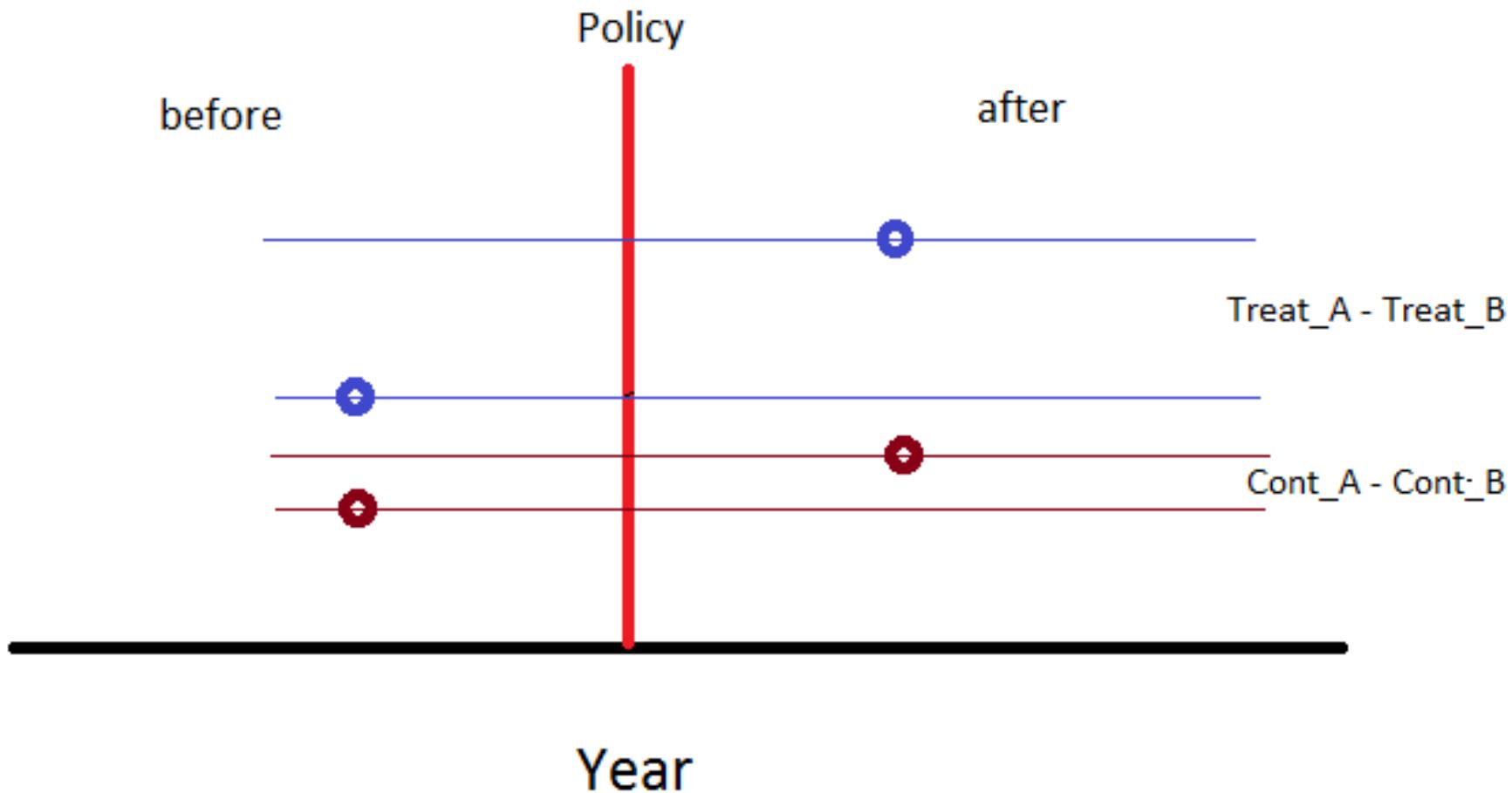
$$(\text{Treated_Aft} - \text{Treated_Bef}) - (\text{Control_Aft} - \text{Control_Bef})$$

$$(\text{Treated_Aft} - \text{Control_Aft}) - (\text{Treatment_Bef} - \text{Control_Bef})$$

So, the difference-in-differences can also be thought of as the difference between the treated and control group in the after period minus the difference between the treated and control group in the before period.

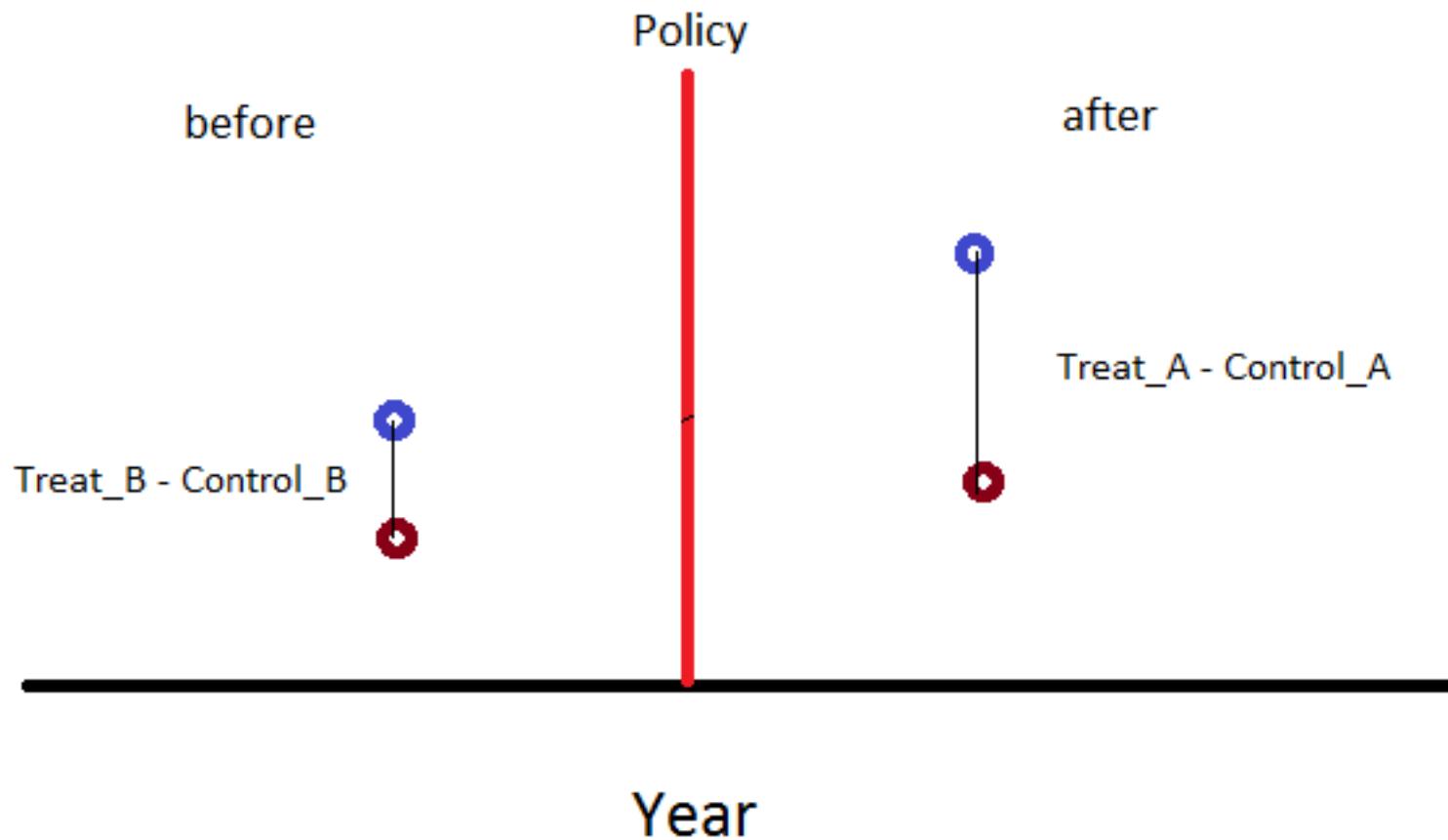
That is, it measures if the gap between the two groups gets wider over time.

Difference-in-Differences



$$(\text{Treated_Aft} - \text{Treated_Bef}) - (\text{Control_Aft} - \text{Control_Bef})$$

Difference-in-Differences



$$(\text{Treated}_{\text{Aft}} - \text{Control}_{\text{Aft}}) - (\text{Treatment}_{\text{Bef}} - \text{Control}_{\text{Bef}})$$

Difference-in-Differences

$$Hospital = \beta_0 + \beta_1 DC + \beta_2 Yr2005 + \beta_3 DC * Yr2005 + u$$

Note that this is the same as the regression with Treated (DC) and After (Yr2005).

β_3 is the difference-in-differences estimate of the effect of the smoking ban.

β_1 is the difference between DC and MD in the pre-period.

β_2 is the difference between MD in the pre and post-period.

	Before	After
DC	0.121	0.093
MD	0.125	0.117

What are the values of β_0 , β_1 , β_2 , and β_3 ?

Difference-in-Differences

How do we choose a control group?

Company

- same or similar industry
- similar size

Wages

- similar education levels
- similar geographic location
- similar industries

Health outcomes:

- similar age
- similar gender
- similar health outcomes in baseline

The process of choosing the best control group is called **matching**.

We will learn about the most common algorithms for matching in the next class.

Difference-in-Differences: Example

Let's revisit difference-in-difference using the notation of the ideal experiment from MHE:

On April 1, 1992, New Jersey raised the state minimum from \$4.25 to \$5.05. Card and Krueger collected data on employment at fast food restaurants in New Jersey in February 1992 and again in November 1992. These restaurants (Burger King, Wendy's, and so on) are big minimum-wage employers. Card and Krueger collected data from the same type of restaurants in eastern Pennsylvania, just across the Delaware river. The minimum wage in Pennsylvania stayed at \$4.25 throughout this period. They used their data set to compute differences-in-differences (DD) estimates of the effects of the New Jersey minimum wage increase. That is, they compared the change in employment in New Jersey to the change in employment in Pennsylvania around the time New Jersey raised its minimum.

Difference-in-Differences: Example

Y_{1ist} = fast food employment at restaurant i and period t

if there is a high state minimum wage

Y_{0ist} = fast food employment at restaurant i and period t

if there is a low state minimum wage

For a restaurant in state s and time period t , employment can be expressed as the state baseline and period adjustment:

$$E(Y_{0ist}|s, t) = \gamma_s + \lambda_t$$

We can add a dummy variable for having a high minimum wage:

$$Y_{ist} = \gamma_s + \lambda_t + \beta D_{st} + \varepsilon_{ist}$$

Difference-in-Differences: Example

The After – Before for Pennsylvania:

$$E(Y_{ist}|s = PA, t = Nov) - E(Y_{ist}|s = PA, t = Feb) = \lambda_{Nov} - \lambda_{Feb}$$

The After – Before for New Jersey:

$$E(Y_{ist}|s = NJ, t = Nov) - E(Y_{ist}|s = NJ, t = Feb) = \lambda_{Nov} - \lambda_{Feb} + \beta.$$

The Difference-in-Difference Estimate:

$$[E(Y_{ist}|s = NJ, t = Nov) - E(Y_{ist}|s = NJ, t = Feb)]$$

$$- [E(Y_{ist}|s = PA, t = Nov) - E(Y_{ist}|s = PA, t = Feb)] = \beta$$

Difference-in-Differences: Example

Table 5.2.1: Average employment per store before and after the New Jersey minimum wage increase

Variable	PA (i)	NJ (ii)	Difference, NJ-PA (iii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	-2.89 (1.44)
2. FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	-0.14 (1.07)
3. Change in mean FTE employment	-2.16 (1.25)	0.59 (0.54)	2.76 (1.36)

Difference-in-Differences: Example

We can write the difference-in-difference as follows:

$$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \beta(NJ_s \cdot d_t) + \varepsilon_{ist}$$

Where NJ is Treat and d is After.

$$\alpha = E(Y_{ist} | s = PA, t = Feb) = \gamma_{PA} + \lambda_{Feb}$$

$$\gamma = E(Y_{ist} | s = NJ, t = Feb) - E(Y_{ist} | s = PA, t = Feb) = \gamma_{NJ} - \gamma_{PA}$$

$$\lambda = E(Y_{ist} | s = PA, t = Nov) - E(Y_{ist} | s = PA, t = Feb) = \lambda_{Nov} - \lambda_{Feb}$$

$$\beta = \{E(Y_{ist} | s = NJ, t = Nov) - E(Y_{ist} | s = NJ, t = Feb)\}$$

$$-\{E(Y_{ist} | s = PA, t = Nov) - E(Y_{ist} | s = PA, t = Feb)\}.$$

Difference-in-Differences: Example

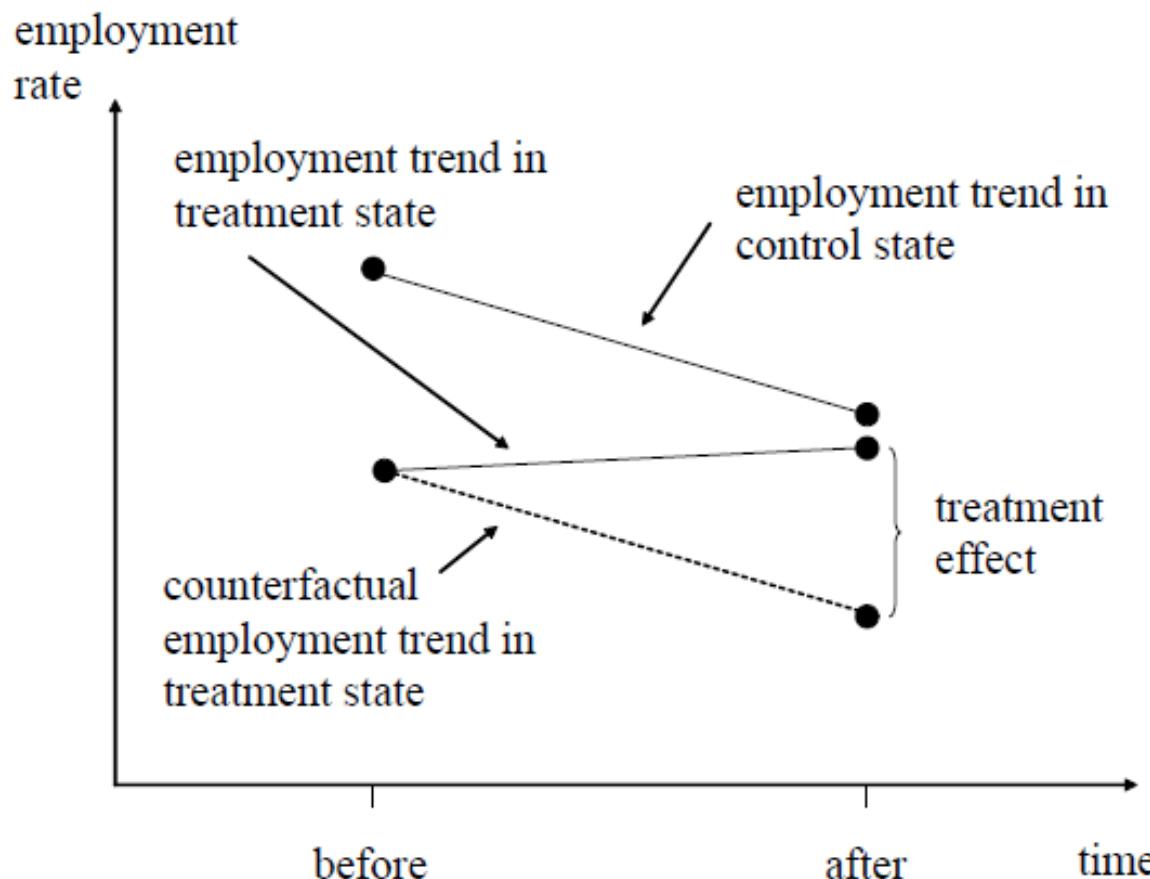
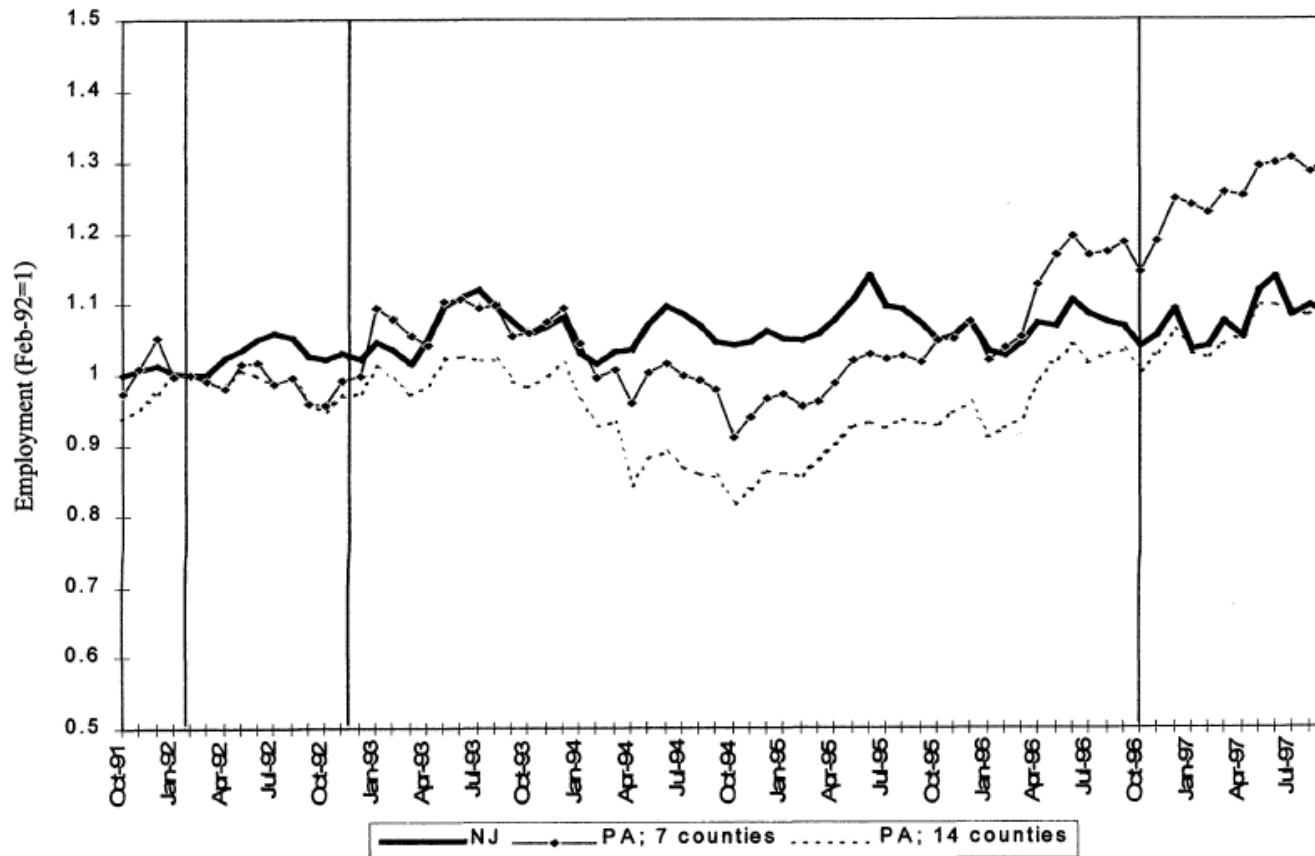


Figure 5.2.1: Causal effects in the differences-in-differences model

Difference-in-Differences: Example

Does Pennsylvania look like a good control group for New Jersey?

The lines on the left are the before and after dates of data collection. The line on the right corresponds to an increase in the federal minimum wage (when PA is treated and NJ is not since it is already high).



Fixed Effects as Repeated D-in-D

Difference-in-Difference is based on two groups (Treat and Control) and two time periods (Before and After).

$$Y_{i,t} = \beta_0 + \beta_1 Treat_i + \beta_2 After_t + \beta_3 Treat_i * After_t + \epsilon_{i,t}$$

Panel data may have many individuals and many time periods. Using individual fixed effects and time period fixed effects is analogous to running many difference-in-difference estimates.

$$Y_i = \gamma_1 + \dots + \gamma_N + \delta_1 + \dots + \delta_T + \beta X_{i,t} + \epsilon_{i,t}$$

Instead of there being two groups, there are N fixed effects (or however many groups).

Instead of there being two periods, there are T time period effects.

The explanatory variable can vary from year-to-year. It could be a dummy variable for “treatment” or a continuous variable that varies from year-to-year.

Fixed Effects as Repeated D-in-D

Consider the analysis of the effects of minimum wage.

Suppose we expanded the analysis to include all states, where many experienced changes in minimum wage laws at different times.

Then we may want to include a fixed effect for each state and a year effect for each year.

$$Y_{ist} = \gamma_1 + \dots + \gamma_{50} + \delta_{1980} + \dots + \delta_{2012} + \beta MinWage_{i,s,t} + \epsilon_{i,s,t}$$

This basically runs a series of difference-in-difference style estimates in different states in different years.

This may not be a good strategy if states are not good control groups for each other on average.

Lesson 16

Matching

Outline

Previous Lesson:

1. Single Difference
2. Difference-in-Differences

This Lesson:

1. Matching
2. Synthetic Controls

Next Lesson:

1. Instrumental Variables

References:

Mostly Harmless Econometrics

Caliendo and Kopeinig (2005): “Some Practical Guidance...”

Abadie, Diamong, Hainmueller (2010): “Synthetic Control...”

Difference-in-Differences

Difference-in-Differences:

Treatment group – the group that is affected by the policy

Control group – the group that is not affected by the policy

Diff-in-Diff specification:

$$Y_{i,t} = \beta_0 + \beta_1 Treat_i + \beta_2 After_t + \beta_3 Treat_i * After_t + \epsilon_{i,t}$$

Identifying assumption:

The control group capture the counterfactual. That is, the control group experiences the change between the before and after periods that the treatment group would have experienced.

Why might a control group not satisfy this assumption?

Difference-in-Differences

Is the identification assumption likely to hold in an (ideal) randomized experiment?

The results of a randomized experiment can be estimated as a diff-in-diff:

$$Y_{i,t} = \beta_0 + \beta_1 Treat_i + \beta_2 After_t + \beta_3 Treat_i * After_t + \epsilon_{i,t}$$

They can also be estimated using just the after data (why?):

$$Y_i = \beta_0^* + \beta_1^* Treat_i + \epsilon_i$$

Note that:

$$\beta_1 = 0$$

$$\beta_0^* = \beta_0 + \beta_2$$

$$\beta_1^* = \beta_3$$

Matching

Goal: The goal is to find the best possible counterfactual. This will allow the difference-in-difference approach in a non-experimental setting to be as similar as possible to the randomized experiment setting.

Example:

Given a state minimum wage policy, we may want to find the best possible control state. What characteristics would these states share?

Example:

Given individuals selecting into a worker training program, we may want to find the most similar people who did not participate. What characteristics would these individuals share?

We may want to match: individuals, businesses, schools, counties, states, countries....

Matching

The parameter of interest:

$$\beta = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1]$$

Comparing the means of treated and untreated leaves a selection bias term:

$$\begin{aligned}\hat{\beta} &= E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] \\ &= \beta + E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]\end{aligned}$$

Independence Assumption – the selection bias term disappears if being treated ($D=1$) is uncorrelated with the outcome.

$$E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0] = 0$$

Conditional Independence Assumption (CIA) – the selection bias term disappears if being treated ($D=1$) is uncorrelated with the outcome after controlling for observables (i.e. conditional on the Xs):

$$E[Y_{0i}|X_i, D_i = 1] - E[Y_{0i}|X_i, D_i = 0] = 0$$

Matching

$$E[Y_{0i}|X_i, D_i = 1] - E[Y_{0i}|X_i, D_i = 0] = 0$$

The CIA is a pretty strong assumption. It says that treatment is as good as randomly assigned after conditioning on the observable characteristics. We will make this assumption.

Note that the observable characteristics may interact with each other in complicated ways. Thus we would like to match treated individuals with untreated individuals who have the same Xs (so we don't need to account for these interactions).

But if there are k different binary characteristics, then we would have 2^k different combinations of characteristics, which could be a huge number.

To reduce this dimensionality problem, we compute the propensity for individuals to be treated using the Xs:

$$P(D = 1|X) = P(X)$$

Matching

Matching: Compute a propensity score (propensity to be treated). Then match each treated individual with an untreated one that has a similar propensity.

Step 1: Estimate a propensity score for each individual: $P(X)$

$$T_i = \beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki} + \epsilon_i$$

The propensity is the probability that an individual is treated based on his or her characteristics. Because T is binary (0,1), it should be estimated using a probit or a logit model.

Step 2: Match each treated individuals to untreated individuals.

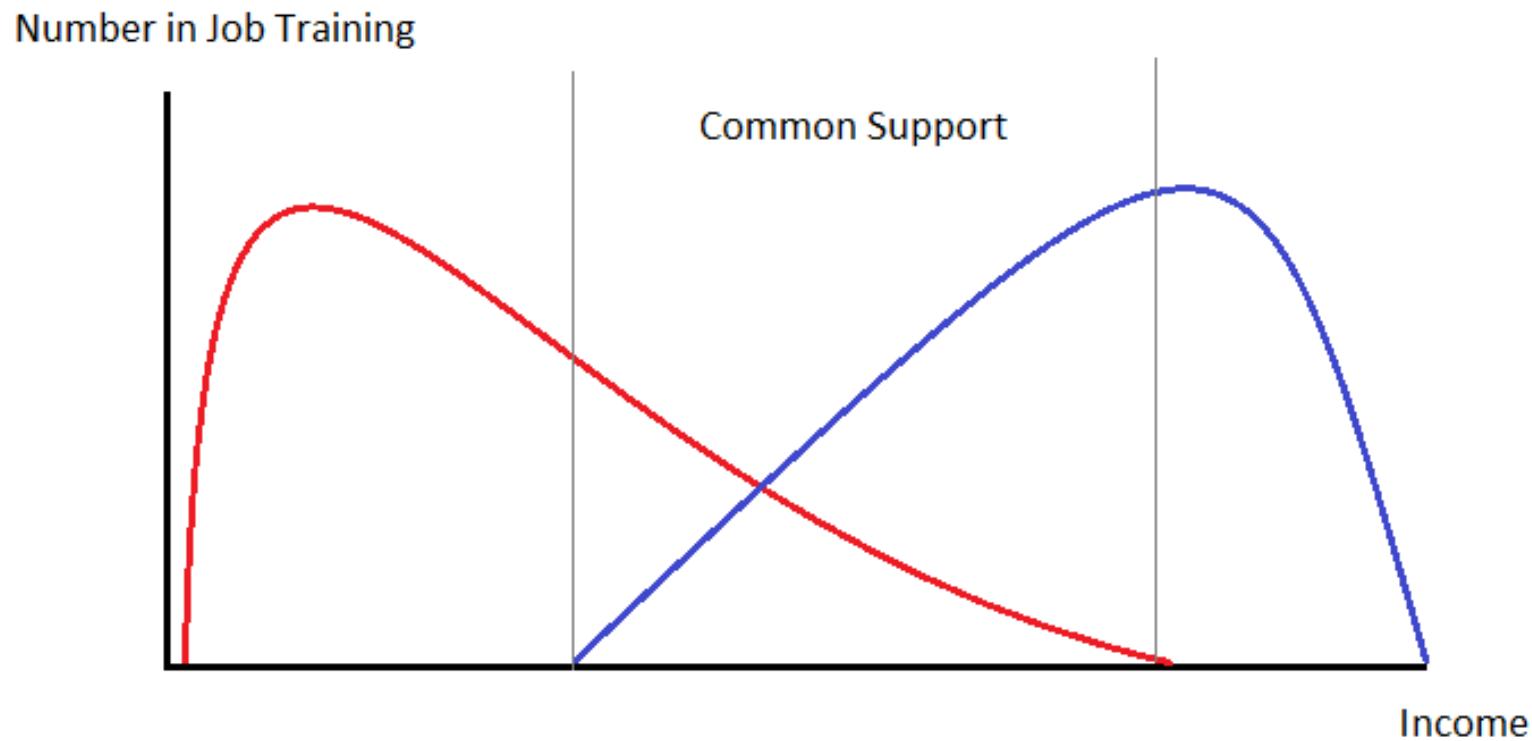
$$\text{Min } |P(X_i) - P(X_j)|$$

Each treated individual i is matched to one or more control individuals j that have the closest propensity score.

Step 3: Keep only the matched individuals for the regression.

Matching

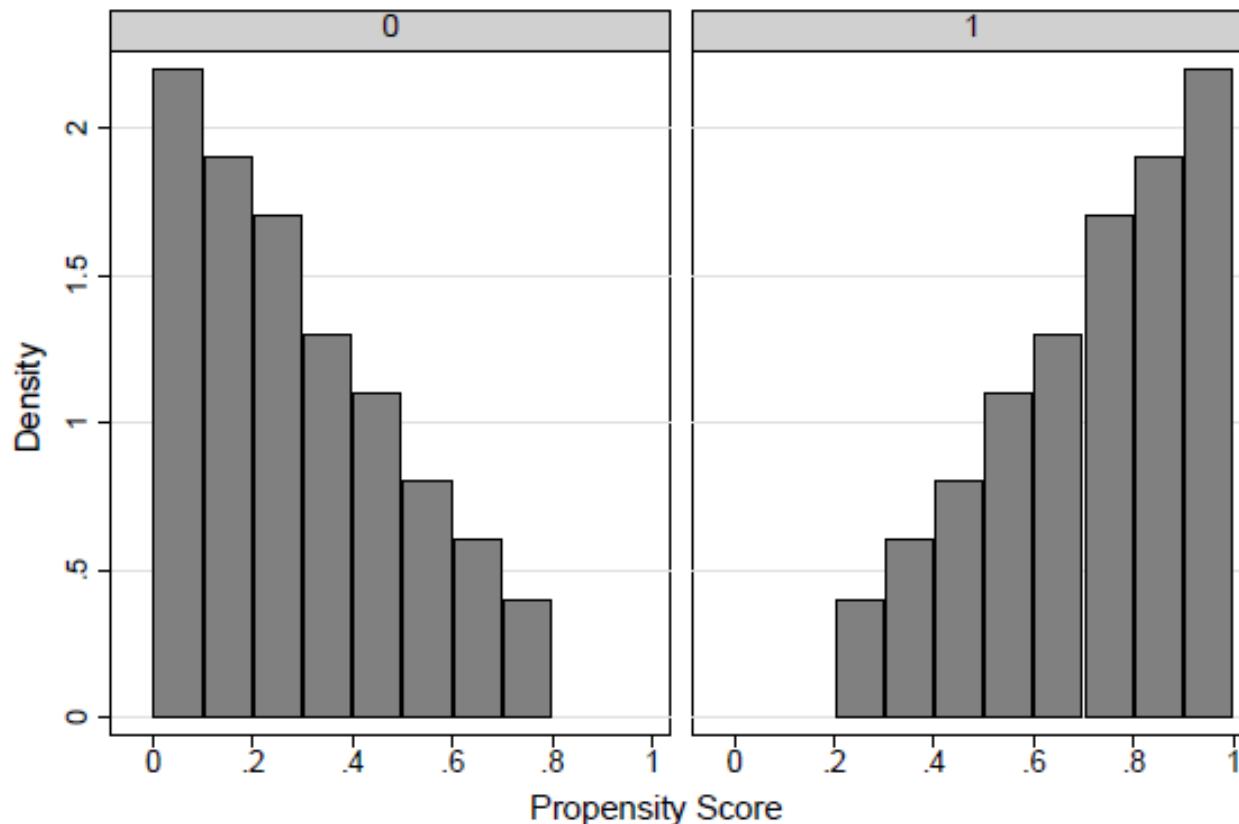
What does matching achieve? It keeps only those individuals who share a common support in terms of their observables.



Matching

It does this by matching individuals with the same propensity score.
Which individuals below will be dropped?

Example 1



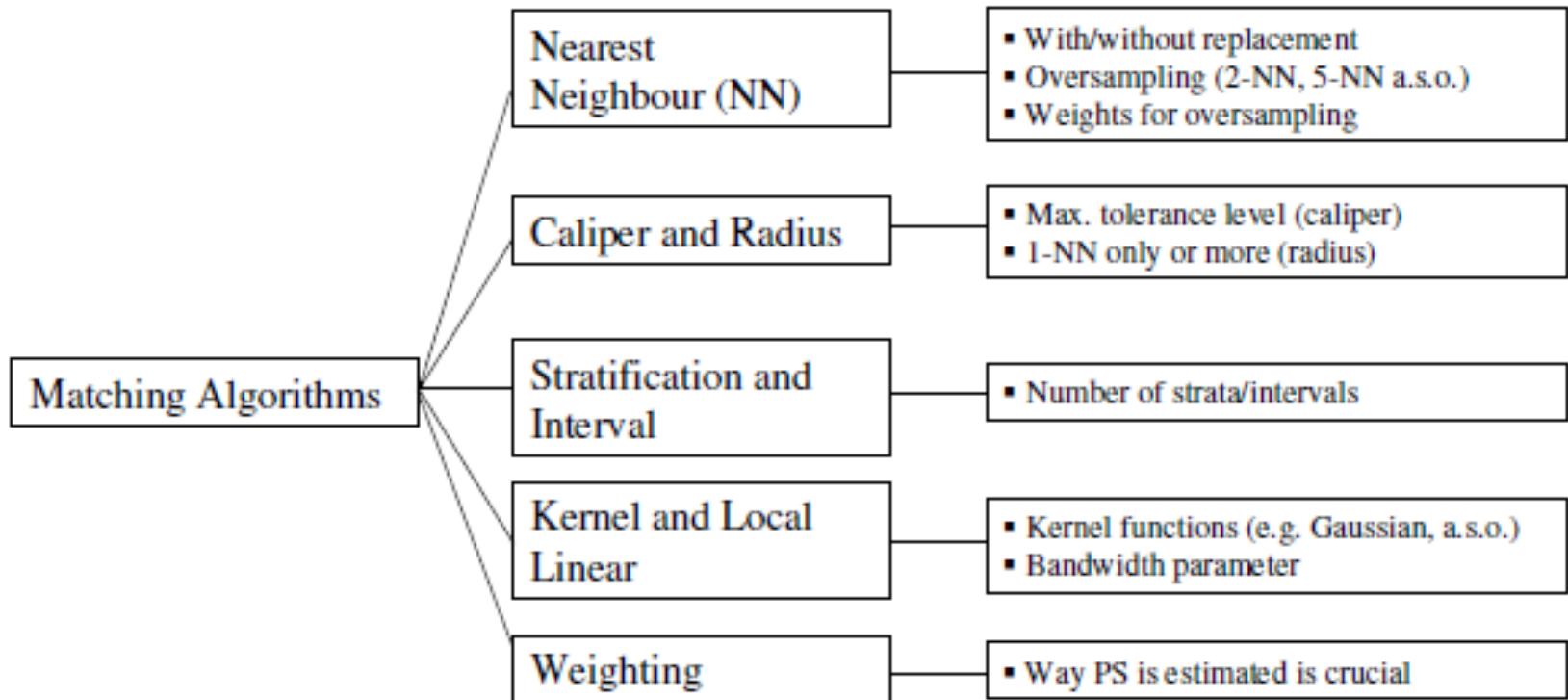
Matching

What variables should be used to compute the propensity score?

1. Fixed characteristics:
 - individuals: race, age
 - companies: industry, location
2. Characteristics measured before treatment.
 - characteristics may change as a result of treatment
 - for example, we wouldn't want to match individuals based on their income after the job training program
3. Use knowledge of institutional details
 - age requirements (e.g. medicare, soc security)
 - citizenship requirements
 - employment status

Matching

After getting the propensity scores, how should the matching be conducted?
There are many options:



Matching

Nearest neighbor (NN) – match each treated individual to only one control based on the propensity score.

By matching to only the nearest neighbor, we minimize bias (individuals are most similar). Useful when there are few good matches.

Caliper matching – match each individual to all individuals within some bandwidth of propensity score:

$$\text{all } j \text{ such that } |P(X_i) - P(X_j)| < .05$$

By including more data, we minimize the variance. Useful when there are many good matches (no need to throw away good matches).

Without replacement – each control used only one time. This reduces variance (more unique individuals).

With replacement – each control can be used repeatedly. This reduces bias.

Matching

Table 1: Trade-Offs in Terms of Bias and Efficiency

Decision	Bias	Variance
Nearest neighbour matching: multiple neighbours / single neighbour with caliper / without caliper	(+)/(-) (-)/(+)	(+)/(-)
Use of control individuals: with replacement / without replacement	(-)/(+) (+)/(-)	
Choosing method: NN-matching / Radius-matching KM or LLM / NN-methods	(-)/(+) (+)/(-)	(+)/(-)
Bandwidth choice with KM: small / large	(-)/(+) (+)/(-)	

What do we do with treated individuals that do not have a good match?

Matching

Why not just use all of the data and use control variables?

- The treatment and control individuals are fundamentally mismatched.
- Controlling for observables may not be sufficient to make the two groups comparable.
- Functional forms would need to be correctly specified.

How to code this in STATA?

- Doing it manually provides a lot of flexibility:
 - i. run Probit regression
 - ii. compute propensity
 - iii. match treat to closest control (some work)
 - iv. can restrict options (e.g. within state, policy rule)
- There are also shortcuts that compute matches: psmatch2

Matching

Evaluating the validity of match:

1. Examine if there is common support of treat and control:
 - summarize the propensity scores
 - graph the propensity scores
2. Balance test – compare means (t-test) of match variables
3. Balance test – compare means (t-test) of other variables

Ultimately: matching can not correct a selection problem.

In order to make a causal claim we need an experiment, policy experiment, other natural experiment (e.g. a smoking ban or minimum wage law).

Matching simply ensures that we are controlling for time variation more accurately by limiting the control group to the most similar individuals.

Matching

Example: Randomization vs Matching in a Job Training Program

National Supported Work (NSW) Program randomly assigned individuals to training. The data associated with this are only from the randomly chosen treatment and control groups.

Current Population Survey (CPS) includes data for a wide range of individuals that can be matched to treated individuals as controls.

We know that the NSW data should produce the correct estimates of the treatment effect.

We can examine how well using CPS individuals as controls performs with and without using a matching method. We can also compare across matching methods.

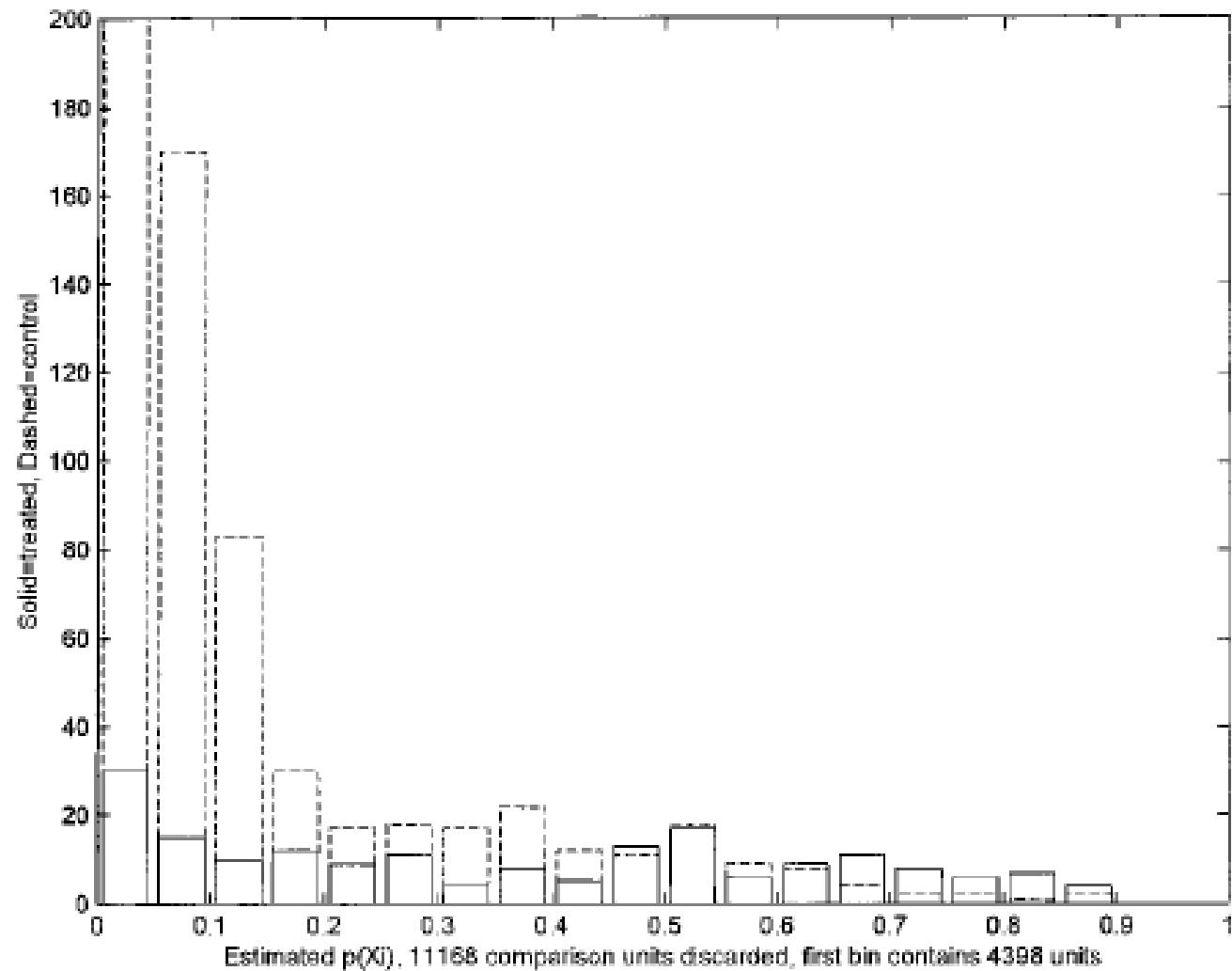
Matching

TABLE 1.—SAMPLE MEANS AND STANDARD ERRORS OF COVARIATES
FOR MALE NSW PARTICIPANTS

National Supported Work Sample (Treatment and Control)		
Variable	Dehejia-Wahba Sample	
	Treatment	Control
Age	25.81 (0.52)	25.05 (0.45)
Years of schooling	10.35 (0.15)	10.09 (0.1)
Proportion of school dropouts	0.71 (0.03)	0.83 (0.02)
Proportion of blacks	0.84 (0.03)	0.83 (0.02)
Proportion of Hispanic	0.06 (0.017)	0.10 (0.019)
Proportion married	0.19 (0.03)	0.15 (0.02)
Number of children	0.41 (0.07)	0.37 (0.06)
No-show variable	0 (0)	n/a
Month of assignment (Jan. 1978 = 0)	18.49 (0.36)	17.86 (0.35)
Real earnings 12 months before training	1,689 (235)	1,425 (182)
Real earnings 24 months before training	2,096 (359)	2,107 (353)
Hours worked 1 year before training	294 (36)	243 (27)
Hours worked 2 years before training	306 (46)	267 (37)
Sample size	185	260

Matching

FIGURE 1.—HISTOGRAM OF ESTIMATED PROPENSITY SCORE,
NSW AND CPS



Matching

TABLE 2.—SAMPLE CHARACTERISTICS AND ESTIMATED IMPACTS FROM THE NSW AND CPS SAMPLES

Control Sample	No. of Observations	Mean Propensity Score ^A	Age	School	Black	Hispanic	No Degree	RE74	U74	Treatment Effect (Diff. in Means)	Regression Treatment Effect
NSW	185	0.37	25.82	10.35	0.84	0.06	0.71	2095	0.29	1794 ^B (633)	1672 ^C (638)
Full CPS	15992	0.01 (0.02) ^D	33.23 (0.53)	12.03 (0.15)	0.07 (0.03)	0.07 (0.02)	0.30 (0.03)	14017 (367)	0.88 (0.03)	-8498 (583) ^E	1066 (554)
Without replacement:											
Random	185	0.32 (0.03)	25.26 (0.79)	10.30 (0.23)	0.84 (0.04)	0.06 (0.03)	0.65 (0.05)	2305 (495)	0.37 (0.05)	1559 (733)	1651 (709)
Low to high	185	0.32 (0.03)	25.23 (0.79)	10.28 (0.23)	0.84 (0.04)	0.06 (0.03)	0.66 (0.05)	2286 (495)	0.37 (0.05)	1605 (730)	1681 (704)
High to low	185	0.32 (0.03)	25.26 (0.79)	10.30 (0.23)	0.84 (0.04)	0.06 (0.03)	0.65 (0.05)	2305 (495)	0.37 (0.05)	1559 (733)	1651 (709)
With replacement:											
Nearest neighbor	119	0.37 (0.03)	25.36 (1.04)	10.31 (0.31)	0.84 (0.06)	0.06 (0.04)	0.69 (0.07)	2407 (727)	0.35 (0.07)	1360 (913)	1375 (907)
Caliper, $\delta = 0.00001$	325	0.37 (0.03)	25.26 (1.03)	10.31 (0.30)	0.84 (0.06)	0.07 (0.04)	0.69 (0.07)	2424 (845)	0.36 (0.06)	1119 (875)	1142 (874)
Caliper, $\delta = 0.00005$	1043	0.37 (0.02)	25.29 (1.03)	10.28 (0.32)	0.84 (0.05)	0.07 (0.04)	0.69 (0.06)	2305 (877)	0.35 (0.06)	1158 (852)	1139 (851)
Caliper, $\delta = 0.0001$	1731	0.37 (0.02)	25.19 (1.03)	10.36 (0.31)	0.84 (0.05)	0.07 (0.04)	0.69 (0.06)	2213 (890)	0.34 (0.06)	1122 (850)	1119 (843)

Synthetic Controls

Suppose California is our treated state:

In 1988 California adopted a \$0.25 cigarette tax and used the proceeds to fund anti-smoking campaigns (Prop 99).

We want to match California with another state. What is the problem?

Any ideas about how we might be able to construct a better match for California?

Synthetic Controls

Synthetic Control Procedure:

$$X_1 = (Z_1, Y_1^{K_1}, \dots, Y_1^{K_M})$$

X_i is the vector of pre-treatment characteristics and outcomes for possible control i .

This procedure weights the potential controls in order to minimize the distance from the treated group (i.e. it creates a control group by weighting several different controls in order to match the treated group).

$$\|X_1 - X_0 W\|$$

Synthetic control matching chooses W to minimize the distance from the treatment group X_1 .

When to use : few potential controls. No control is a good match.

Synthetic Controls

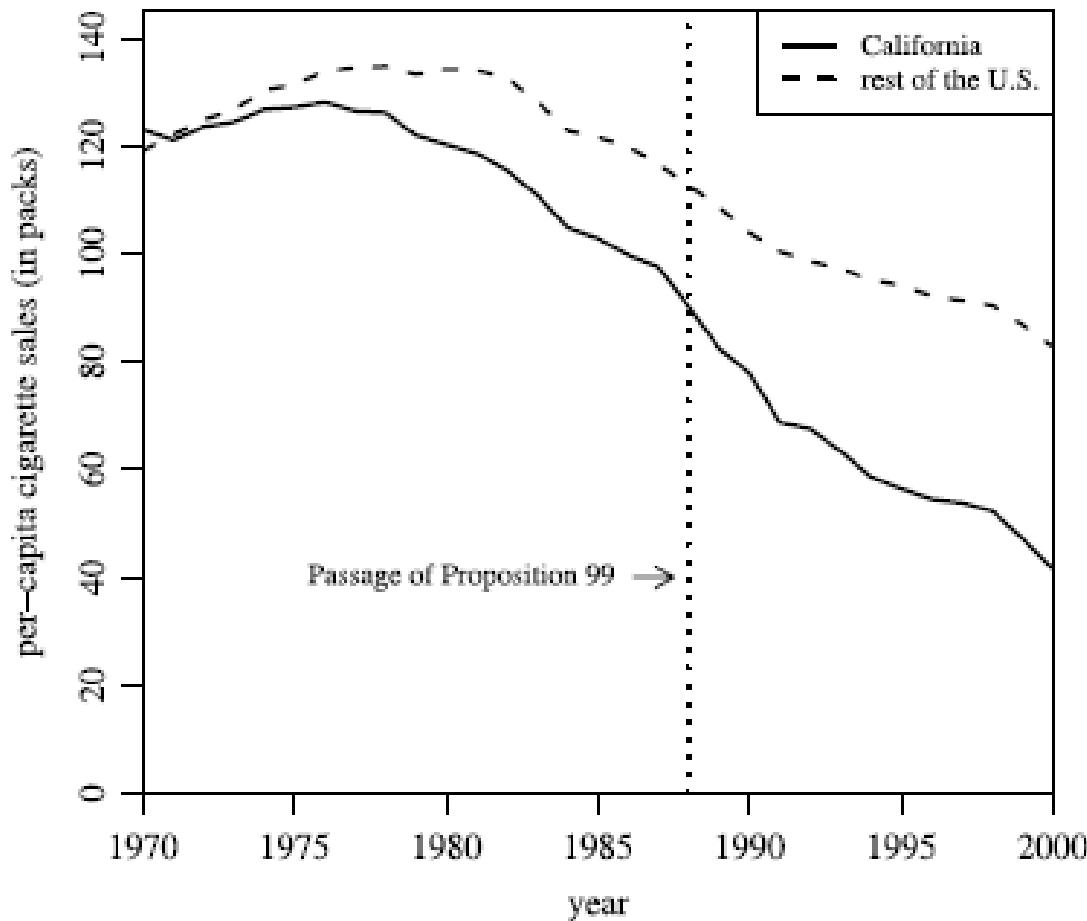


Figure 1. Trends in per-capita cigarette sales: California vs. the rest of the United States.

Synthetic Controls

Table 2. State weights in the synthetic California

State	Weight	State	Weight
Alabama	0	Montana	0.199
Alaska	–	Nebraska	0
Arizona	–	Nevada	0.234
Arkansas	0	New Hampshire	0
Colorado	0.164	New Jersey	–
Connecticut	0.069	New Mexico	0
Delaware	0	New York	–
District of Columbia	–	North Carolina	0
Florida	–	North Dakota	0
Georgia	0	Ohio	0
Hawaii	–	Oklahoma	0
Idaho	0	Oregon	–
Illinois	0	Pennsylvania	0
Indiana	0	Rhode Island	0
Iowa	0	South Carolina	0
Kansas	0	South Dakota	0
Kentucky	0	Tennessee	0
Louisiana	0	Texas	0
Maine	0	Utah	0.334
Maryland	–	Vermont	0
Massachusetts	–	Virginia	0
Michigan	–	Washington	–
Minnesota	0	West Virginia	0
Mississippi	0	Wisconsin	0
Missouri	0	Wyoming	0

Synthetic Controls

Table 1. Cigarette sales predictor means

Variables	California		Average of 38 control states
	Real	Synthetic	
Ln(GDP per capita)	10.08	9.86	9.86
Percent aged 15–24	17.40	17.40	17.29
Retail price	89.42	89.41	87.27
Beer consumption per capita	24.28	24.20	23.75
Cigarette sales per capita 1988	90.10	91.62	114.20
Cigarette sales per capita 1980	120.20	120.43	136.58
Cigarette sales per capita 1975	127.10	126.99	132.81

NOTE: All variables except lagged cigarette sales are averaged for the 1980–1988 period (beer consumption is averaged 1984–1988). GDP per capita is measured in 1997 dollars, retail prices are measured in cents, beer consumption is measured in gallons, and cigarette sales are measured in packs.

Synthetic Controls

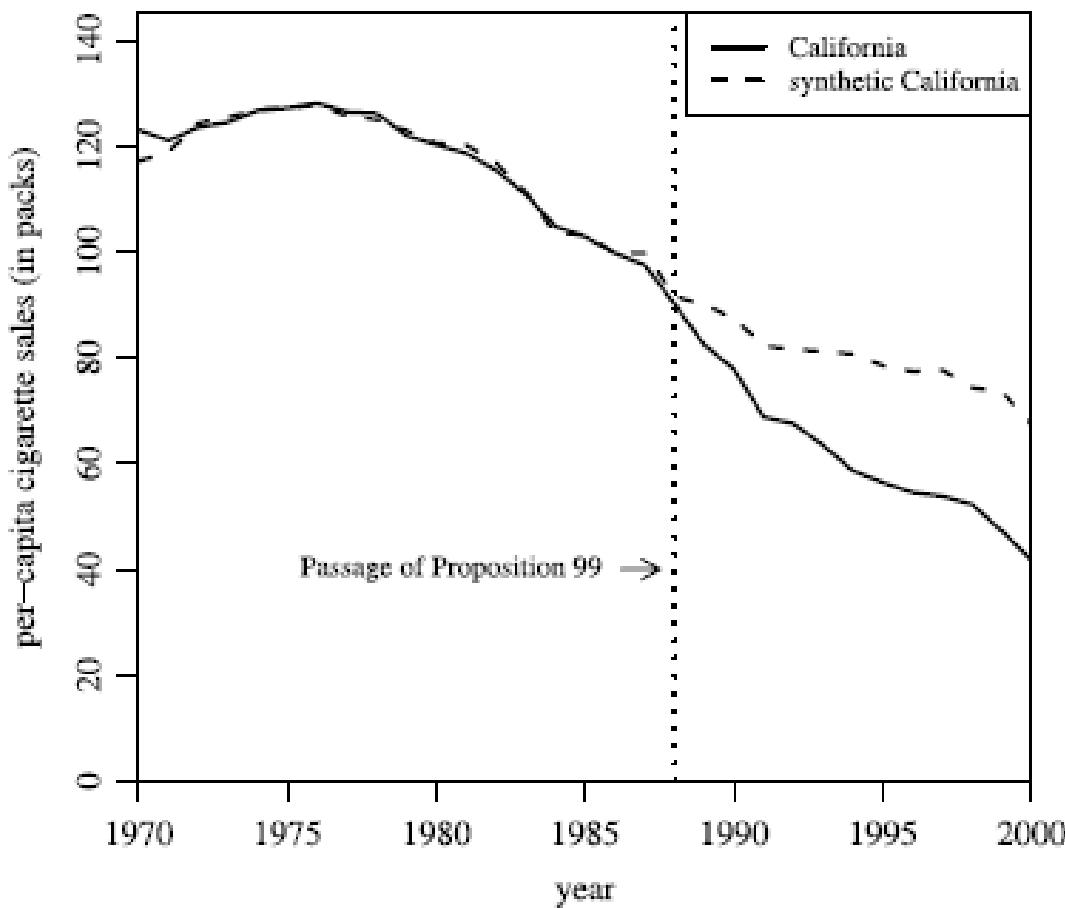


Figure 2. Trends in per-capita cigarette sales: California vs. synthetic California.

Lesson 17

Fixed Effects
Between and Within Variation

Outline

Previous Lessons:

1. Difference-in-Differences
2. Matching

This Lesson:

1. Group Fixed Effects
2. Time Fixed Effects

Next Lesson:

1. Instrumental Variables

Outline

Consider cross-sectional data for car company sales (in millions) and prices (in thousands).

Maker	Year	Quantity	Price
Ford	2004	21.2	32.4
Chevy	2004	8.4	17.4
Toyota	2004	18.2	15.6
Honda	2004	15..0	17.6
Mercedes	2004	37.1	40.7

What do we conclude about the relationship about price and quantity?

What are your primary concerns?

Outline

Consider panel data for car company sales (in millions) and prices (in thousands).

Maker	Year	Quantity	Price
Ford	2004	21.2	32.4
Ford	2005	22.4	31.4
Chevy	2004	8.4	17.4
Chevy	2005	5.6	22.8
Toyota	2004	18.2	15.6
Toyota	2005	18.8	15.3
Honda	2004	15..0	17.6
Honda	2005	16.4	17.2
Mercedes	2004	37.1	40.7
Mercedes	2005	32.2	45.1

What do we conclude about the relationship between price and quantity sold.

Outline

This difference is due to *across* versus *within* variation:

Across-variation – differences in the outcome and explanatory variables across different groups.

Within-variation – differences in the outcome and explanatory variables within the groups

In our example:

Across-variation – comparisons across different auto-makers. We think there are important omitted difference (like car quality).

Within-variation – comparisons within auto-makers over time. These differences can not be driven by fixed characteristics .

Outline

Time-invariant characteristics – characteristics that do not change over time.

Time invariant characteristics of the automakers can generate bias in cross-sectional estimates (e.g. if Mercedes makes better cars than Chevy).

Time invariant characteristics of the automakers can not generate bias if we only make comparisons within groups (e.g. if we compare Mercedes in one year to Mercedes in another year).

Note: this is the idea behind single-difference estimates.

In a single difference estimate we compare after – before. This eliminates the effects of time-invariant characteristics.

Outline

How do we use only within-group variation econometrically?

1. Take first differences and run regression on differences.

Maker	Year	Quantity	Price	ΔQ	ΔP
Ford	2004	21.2	32.4		
Ford	2005	22.4	31.4	1.2	-1
Chevy	2004	8.4	17.4		
Chevy	2005	5.6	22.8	-2.8	5.4
Toyota	2004	18.2	15.6		
Toyota	2005	18.8	15.3	0.6	-0.3
Honda	2004	15.0	17.6		
Honda	2005	16.4	17.2	1.4	-0.4
Mercedes	2004	37.1	40.7		
Mercedes	2005	32.2	45.1	-4.9	4.4

Outline

An alternative is to examine how much each observation deviates from the average value:

Maker	Year	Quantity	Price	Avg Q	Avg P	Q-AvgQ	P-AvgP
Ford	2004	21.2	32.4	21.8	31.9	-0.6	0.5
Ford	2005	22.4	31.4	21.8	31.9	0.6	-0.5
Chevy	2004	8.4	17.4	7	20.1	1.4	-2.7
Chevy	2005	5.6	22.8	7	20.1	-1.4	2.7
Toyota	2004	18.2	15.6	18.5	15.45	-0.3	0.2
Toyota	2005	18.8	15.3	18.5	15.45	0.3	-0.1
Honda	2004	15.0	17.6	15.7	17.4	-0.7	0.2
Honda	2005	16.4	17.2	15.7	17.4	0.7	-0.2
Mercedes	2004	37.1	40.7	34.65	42.9	2.5	-2.2
Mercedes	2005	32.2	45.1	34.65	42.9	-2.5	2.2

Outline

A fixed effect model involves including a dummy variable (fixed effect) for each group. Thus, in our example, we would include a fixed effect for each car company.

As a result, the only comparisons that we would use would be within-company.

The simple regression uses within and between variation:

$$Q_{ct} = \alpha + \beta P_{ct} + \epsilon_{ct}$$

The fixed effect regression only uses within-variation:

$$Q_{ct} = \alpha_c + \beta P_{ct} + \epsilon_{ct}$$

Note that this does not fix time-variant issues. For example, if the price goes up because of some change in car quality, then this is not corrected with a fixed effect.

Stata

We want to estimate the effect of the probability of being arrested on the crime rate. We have data for many counties (so a county is an individual unit) :

```
. reg crmrte prbarr
```

Source	SS	df	MS	Number of obs	=	630
Model	.026553065	1	.026553065	F(1, 628)	=	92.65
Residual	.179989001	628	.000286607	Prob > F	=	0.0000
Total	.206542066	629	.000328366	R-squared	=	0.1286

crmrte	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
prbarr	-.0379504	.0039428	-9.63	0.000	-.045693	-.0302078
_cons	.0432523	.0013869	31.19	0.000	.0405287	.0459759

Stata

```
. xi: reg crmrte i.county prbarr
i.county          _Icounty_1-197      (naturally coded; _Icounty_1 omitted)
```

Source	SS	df	MS	Number of obs	=	630
Model	.179848763	90	.00199832	F(90, 539)	=	40.35
Residual	.026693303	539	.000049524	Prob > F	=	0.0000
Total	.206542066	629	.000328366	R-squared	=	0.8708
				Adj R-squared	=	0.8492
				Root MSE	=	.00704

crmrte	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
_Icounty_3	-.0211038	.0037822	-5.58	0.000	-.0285335	-.0136741
_Icounty_5	-.0227439	.0038042	-5.98	0.000	-.0302167	-.015271
_Icounty_7	-.0125058	.00377	-3.32	0.001	-.0199114	-.0051002
_Icounty_195	.0097064	.0037717	2.57	0.010	.0022973	.0171155
_Icounty_197	-.0209701	.003779	-5.55	0.000	-.0283935	-.0135467
prbarr	-.0020232	.0026692	-0.76	0.449	-.0072666	.0032202
_cons	.0363976	.0027972	13.01	0.000	.0309028	.0418924

Fixed Effects: Time

Time as an omitted variable:

$$Wage_{it} = \beta_0^* + \beta_1^* Educ_{it} + u_{it}$$

Time is positively correlated with Education: (δ_1)

- people have higher average levels of education than in the past.

Time is positively correlated with Wages: (B_2)

- technology has improved (computers, machines)
- and many other reasons

Omitted variable bias: $\beta_1^* = \beta_1 + \beta_2 \delta_1$

So, the coefficient on education is too large.

Fixed Effects: Time

We need to either:

- include everything that changes over time in our regression
(computers, technology, capital, access to resources, inflation...)

or

- include time as a control.

Since we can't control for everything, including time makes a lot of sense:

$$Wage_{it} = \beta_0 + \beta_1 Year + \beta_2 Educ_{it} + u_{it}$$

What shortcomings does controlling for time in this way have?

Fixed Effects: Time

If we want to control for time in a more flexible way, we can:

- include a higher order polynomial for time

$$Wage_{it} = \beta_0 + \beta_1 Year + \beta_2 Year^2 + \beta_3 Educ_{it} + u_{it}$$

This will allow the effect of the year on wages to increase or decrease over time.

This would make sense if, for example, wages were increasing rapidly during the 1990s and then more slowly during the 2000s.

But this regression still forces the time trend to have a nice smooth shape, which may or not reflect reality.

Fixed Effects: Time

Suppose we are interested in estimating the effect of the size of a house, h , on the sale price of that house:

$$Price_{ht} = \beta_0 + \beta_1 SqrFt_{ht} + u_{ht}$$

Suppose we have data for houses sold between 1990 and 2012.

We know that houses have gotten larger during this period:

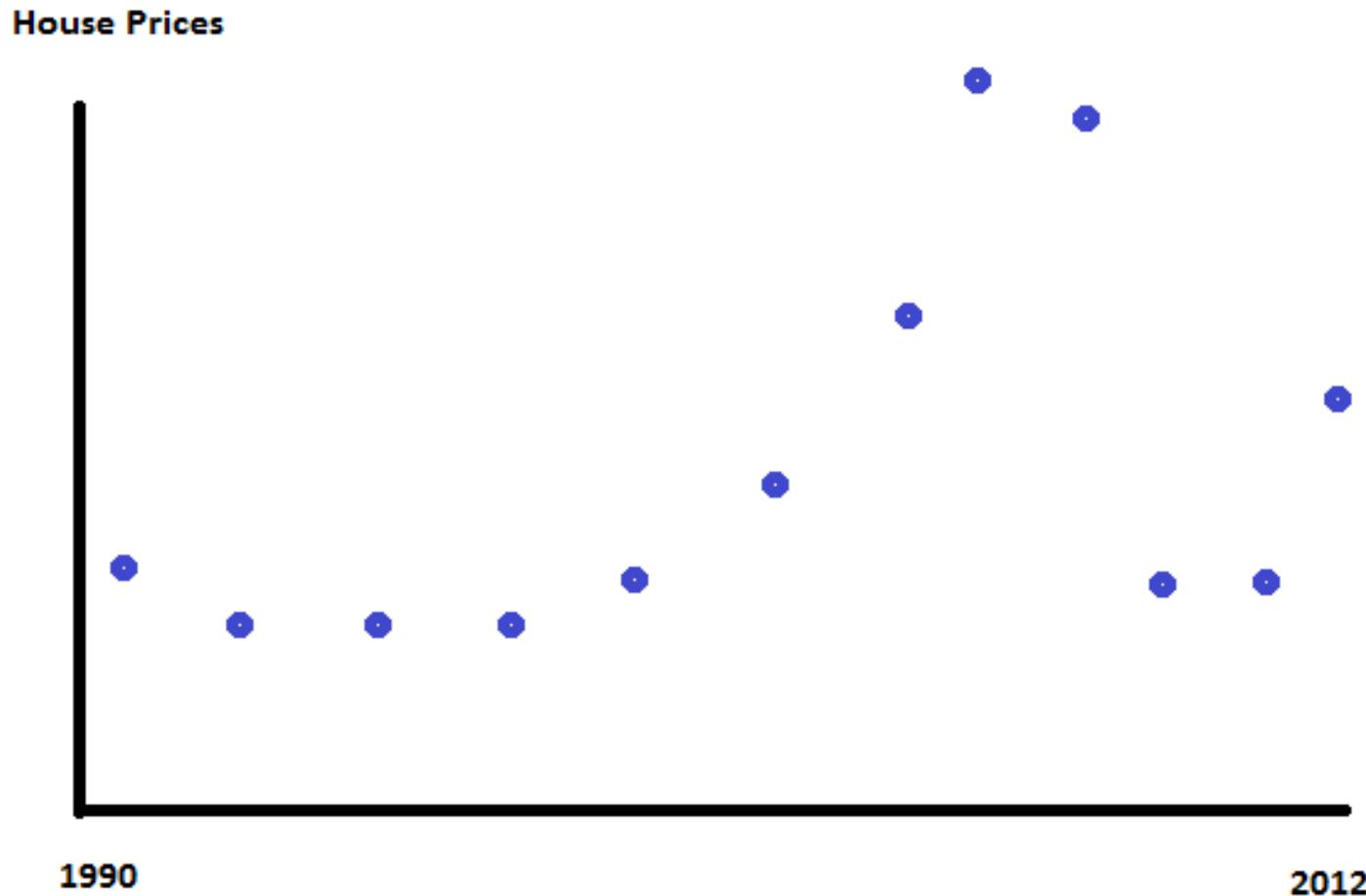
- time is correlated with house size

We also know that the average sale price has changed over time:

- time is correlated with price

Fixed Effects: Time

How are we going to control for the time trend when it looks like this?



Fixed Effects: Time

We can include a year fixed effect:

$$Price_{ht} = \alpha_t + \beta_1 SqrFt_{ht} + u_{ht}$$

The year fixed effect adjusts the predicted price by a specific amount for each year:

- large year fixed effect in 2006 (the peak of the bubble)
- small year fixed effect in 2009 (after the bubble burst)
- flexible adjustment from year-to-year

So, the resulting estimate B_1 is based on “within-year” comparisons:

- a 2,000 square foot house vs a 1,500 square foot house in same year (an apples to apples comparison).

Fixed Effects: Combined

You can use two kinds of fixed effects in the same regression:

A state, s, fixed effect (compare “within-state”):

$$Price_{hst} = \alpha_s + \beta_1 SqrFt_{hst} + u_{hst}$$

A year, t, fixed effect (compare “within-year”):

$$Price_{hst} = \alpha_t + \beta_1 SqrFt_{hst} + u_{hst}$$

Both (“within-state” and “within-year”):

$$Price_{hst} = \alpha_s + \alpha_t + \beta_1 SqrFt_{hst} + u_{hst}$$

Lesson 18

Instrumental Variables

Outline

Previous Lessons:

1. Matching
2. Synthetic Controls
3. Fixed Effects

This Lesson:

1. Instrumental Variables
2. Two-Stage Least Squares

Next Lesson

1. Regression Discontinuity

Outline

Intuitively, why is there bias if X is correlated with u?

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

Recall the univariate proof of unbiasedness:

$$\begin{aligned}\hat{\beta}_1 &= \frac{Cov(X, Y)}{Var(X)} = \frac{Cov(X, \beta_0 + \beta_1 X + u)}{Var(X)} \\ &= \frac{\beta_1 Var(X) + Cov(X, u)}{Var(X)} = \beta_1 + \boxed{\frac{Cov(X, u)}{Var(X)}}\end{aligned}$$

Bias occurs when the explanatory variable (X) is correlated with the error term (u). That is, when $Cov(X, u) \neq 0$. This can happen for 3 reasons:

- A. Omitted variables
- B. Reverse causality
- C. Measurement error

Omitted Variable Bias: Univariate

Omitted variable bias:

$$Y = \beta_0 + \beta_1 X_1 + [\beta_2 X_2 + \epsilon]$$

$$Y = \beta_0 + \beta_1 X_1 + [\beta_2(\delta_0 + \delta_1 X_1 + \eta) + \epsilon]$$

$$Y = \underbrace{(\beta_0 + \beta_2 \delta_0)}_{\text{true } \beta_0} + \underbrace{(\beta_1 + \beta_2 \delta_1)}_{\text{true } \beta_1} X_1 + \underbrace{(\beta_2 \eta + \epsilon)}_{\text{error}}$$

Measurement error bias:

$$\begin{aligned}\hat{\beta}_1^* &= \frac{Cov(X + \epsilon, Y)}{Var(X + \epsilon)} = \frac{Cov(X, Y) + Cov(\epsilon, Y)}{Var(X) + Var(\epsilon)} \\ &= \frac{Cov(X, Y)}{Var(X) + Var(\epsilon)} < \frac{Cov(X, Y)}{Var(X)} = \beta_1\end{aligned}$$

Instrumental Variables

Suppose we want to estimate the effect of skipped classes on a student's final test score:

$$score_i = \beta_0 + \beta_1 \text{skipped}_i + u_i$$

What sign do we expect β_1 to have?

Omitted variable: It is probably safe to assume that motivation is:

- a. positively correlated with score ($B_2 > 0$)
- b. negatively correlated with number skipped ($\delta_1 < 0$)

$$\beta_1^* = \beta_1 + \beta_2 \delta_1 < \beta_1$$

The estimated effect of skipping class is downward biased when we omit motivation. Our estimate of β_1 is too negative.

It captures both the negative effect of skipping class (the true B_1) and the fact that less motivated students skip more classes ($B_2 \delta_1$).

Instrumental Variables

$$\widehat{score}_i = 84.33 - 2.82skipped_i$$

Our estimate may be biased because “skipped” is correlated with u.

The solution is to find an instrumental variable. An instrumental variable Z satisfies two properties:

1. Inclusion restriction: instrument is correlated with endogenous X

$$Cov(Z, X) \neq 0$$

2. Exclusion restriction: instrument is uncorrelated with the error u

$$Cov(Z, u) = 0$$

Proposed instrument: distance from dorm to the classroom. If “distance” satisfies 1 & 2, then we can use it to estimate B_1 without bias.

Instrumental Variables

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

Instrumental variables:

$$\text{Cov}(Z, Y) = \text{Cov}(Z, (\beta_0 + \beta_1 X + u))$$

$$\text{Cov}(Z, Y) = \text{Cov}(Z, \beta_0) + \text{Cov}(Z, \beta_1 X) + \text{Cov}(Z, u)$$

$$\text{Cov}(Z, Y) = \beta_1 \text{Cov}(Z, X) + \text{Cov}(Z, u)$$

$$\beta_1 = \frac{\text{Cov}(Z, Y)}{\text{Cov}(Z, X)} - \frac{\text{Cov}(Z, u)}{\text{Cov}(Z, X)}$$

$$\beta_1 = \frac{\text{Cov}(Z, Y)}{\text{Cov}(Z, X)}$$

So, we can get an unbiased measure of β_1 if $\text{Cov}(Z, u) = 0$.

Instrumental Variables

Summary: least squares is biased because X is correlated with the error term, but instrumental variables is not because Z is not correlated with the error term.

To actually get the instrumental variables estimate:

$$\beta_1^{IV} = \frac{Cov(Z, Y)}{Cov(Z, X)} = \frac{\sum(Z_i - \bar{z})(Y_i - \bar{y})}{\sum(Z_i - \bar{z})(X_i - \bar{x})}$$

It is very similar to $Cov(X, Y)/Var(X)$, except that Z is substituted for X.

Numerator – measures how much the outcome variable changes with Z

Denominator – measures how much the explanatory variable changes with Z

Instrumental Variables

$$wage_i = \beta_0 + \beta_1 education_i + u_i$$

Do we think that β_1 is biased upward or downward?

Proposed instrument: “father education”

1. Correlated with the explanatory variable (education)
2. Uncorrelated with the error (u – motivation, connections, etc)

Do you think that the inclusion restriction holds?

Can we test it?

Do you think that the exclusion restriction holds?

Can we test it?

Two-Stage Least Squares

Two-Stage Least Squares – a way of implementing instrumental variables. It provides some good intuition about IV works. Suppose X_1 is endogenous (i.e. correlated with u) but we have an instrument Z :

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

First Stage: Regress the endogenous variable X_1 on the instrument. Then predict X using only the instrument Z :

$$X_1 = \alpha_0 + \alpha_1 Z + n$$

$$\hat{X}_1 = \hat{\alpha}_0 + \hat{\alpha}_1 Z$$

Second Stage: Use the predicted value of X_1 in the regression we are interested in:

$$Y = \beta_0 + \beta_1 \hat{X}_1 + \epsilon$$

The resulting estimate of β_1 is the instrumental variable estimate.

Two-Stage Least Squares

$$Y = \beta_0 + \beta_1 \hat{X}_1 + \epsilon$$

Two-Stage Least Squares intuition – We only exploit variation in X_1 that is generated by the instrumental variable Z .

Show two step procedure gives us the IV estimate: $\text{Cov}(Z, Y)/\text{Cov}(Z, X)$.

$$\hat{X}_1 = \hat{\alpha}_0 + \hat{\alpha}_1 Z$$

$$Y = \beta_0 + \beta_1 \hat{X}_1 + \epsilon$$

Start with the second stage and insert the first:

$$\begin{aligned}\beta_1^{2SLS} &= \frac{\text{Cov}(\hat{\alpha}_0 + \hat{\alpha}_1 Z, Y)}{\text{Var}(\hat{\alpha}_0 + \hat{\alpha}_1 Z)} = \frac{\hat{\alpha}_1 \text{Cov}(Z, Y)}{\hat{\alpha}_1^2 \text{Var}(Z)} \\ &= \frac{\text{Cov}(Z, Y)}{\hat{\alpha}_1 \text{Var}(Z)} = \frac{\text{Cov}(Z, Y)}{\frac{\text{Cov}(X, Z)}{\text{Var}(Z)} \text{Var}(Z)} = \frac{\text{Cov}(Z, Y)}{\text{Cov}(Z, X)}\end{aligned}$$

Instrumental Variables: Multivariate

Let's take a look at the multivariate setting:

$$Y = \beta_1 + \beta_2 X_2 + \dots + \beta_k X_k + u$$

Suppose that all of the explanatory variables are exogenous except X_k :

$$E(1 \cdot u) = 0$$

$$E(X_2 u) = 0$$

$$E(X_3 u) = 0$$

⋮

$$E(X_{k-1} u) = 0$$

However: $E(X_k u) \neq 0$

That is, X_k is endogenous. Note that all of the β s will be biased (not just β_k).

Instrumental Variables: Multivariate

We have an instrument Z_1 that is exogenous:

$$E(Z_1 u) = 0$$

And it is correlated with the endogenous variable.

Matrix notation:

$$\begin{aligned}\mathbf{X} &= [1 \quad \mathbf{X}_2 \quad \dots \quad \mathbf{X}_{k-1} \quad \mathbf{X}_k] \\ \mathbf{Z} &= [1 \quad \mathbf{X}_2 \quad \dots \quad \mathbf{X}_{k-1} \quad \mathbf{Z}_1]\end{aligned}$$

Each matrix element is an $N \times 1$ vector:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

And the following condition is satisfied:

$$E(\mathbf{Z}'\mathbf{u}) = 0$$

Instrumental Variables: Multivariate

$$E(\mathbf{Z}'\mathbf{u}) = \mathbf{0}$$

Now we can derive the instrumental variables estimator:

$$E(\mathbf{Z}'\mathbf{u}) = \mathbf{0}$$

$$E(\mathbf{Z}'(\mathbf{Y} - \mathbf{X}\beta)) = \mathbf{0}$$

$$E(\mathbf{Z}'\mathbf{X})\beta = E(\mathbf{Z}'\mathbf{Y})$$

$$\beta = [E(\mathbf{Z}'\mathbf{X})]^{-1}E(\mathbf{Z}'\mathbf{Y})$$

So, our IV estimate is:

$$\hat{\beta}^{IV} = (\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{Z}'\mathbf{Y})$$

Instrumental Variables: Multivariate

$$\hat{\beta}^{IV} = (\mathbf{Z}' \mathbf{X})^{-1} \mathbf{Z}' \mathbf{Y}$$

Note that this is identical to the OLS estimate, but with matrix \mathbf{Z} instead of matrix \mathbf{X} .

Note that the only difference between matrix \mathbf{Z} and matrix \mathbf{X} is that the endogenous variable X_k has been replaced with the instrument Z_1 .

It is easy to show that this is unbiased by plugging in for \mathbf{Y} and taking the expected value.

Two-Stage Least Squares: Multivariate

Two-stage least squares:

1. Regress X on Z and get predicted X.
2. Regress Y on predicted X.

$$\mathbf{X} = \mathbf{Z}\boldsymbol{\gamma} + \mathbf{n}$$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

So, in the first stage we get: $\hat{\boldsymbol{\gamma}} = (\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{Y})$

And thus:

$$\begin{aligned}\hat{\mathbf{X}} &= \mathbf{Z}\hat{\boldsymbol{\gamma}} \\ &= \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\end{aligned}$$

Now let's derive the 2SLS estimator:

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{2SLS} &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}(\hat{\mathbf{X}}'\mathbf{Y}) \\ &= [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y} \\ &= [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y} \\ &= [\mathbf{X}'\mathbf{P}_Z\mathbf{X}]^{-1}\mathbf{X}'\mathbf{P}_Z\mathbf{Y}\end{aligned}$$

Instrumental Variables: Multivariate

1. Inclusion restriction: instrument is correlated with endogenous X
2. Exclusion restriction: instrument is uncorrelated with the error u

Consider a randomized experiment. Is assignment to the treatment group a good instrument?

Consider a policy being implemented. Is the policy a good instrument?

Instrumental Variables: Multivariate

If you have more than one endogenous variable, then you need more than one instrument.

That is, one instrument fixes one endogenous variable.

But let's have a reality check here:

1. Good instruments are hard to find.
2. It is impossible to prove the exclusion restriction.
3. The most common instruments are: policies, experiments.
4. The best instruments are one's that are easy to understand.
Specifically, when it is easy to see that they affect X.

IV in Practice

Example:

$$\widehat{score}_i = 84.33 - 2.82skipped_i$$
$$(12.83) \quad (0.75)$$

Test the hypothesis that number of skipped classes has no effect on test score against the alternative that it does at the 95% level (assume we have 2,000 students in our data)?

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$
$$t = \frac{-2.82}{0.75} = -3.76$$
$$t_{crit} = 1.96$$

So, we reject that skipped classes have no effect on test score. But, we think that -2.82 is too big a negative number due to the omitted variable.

IV on Practice

Test the instrument (correlated with explanatory variable)

test condition #1: $\widehat{skipped}_i = 3.11 + 2.18distance_i$
(1.31) (0.77)

Is distance a statistically significant determinant of classes skipped at the 95% level?

$$\begin{aligned} H_0 : \beta_1 &= 0 & t &= \frac{2.18}{0.77} = 2.83 & t_{crit} &= 1.96 \\ H_A : \beta_1 &\neq 0 \end{aligned}$$

So, the instrument is correlated with the explanatory variable.

IV in Practice

Instrumental variable estimate

$$\widehat{score}_i = 83.72 - 1.43skipped_i$$

(12.84) (1.12)

This is the estimate based on $\text{Cov}(Z, Y)/\text{Cov}(Z, X)$, which in this case is $\text{Cov}(\text{dist}, \text{score})/\text{Cov}(\text{dist}, \text{skipped})$.

How does it compare to the OLS estimate?

- Smaller in magnitude
- Not statistically significant (test this)

So, when we use an instrumental variable, we find that the effect of skipping class is smaller in magnitude than when we just use ordinary least squares. This is probably because of omitted variable bias in the OLS estimate.

IV in Practice

Example:

$$\ln(\text{earn}_i) = \beta_0 + \beta_1 \text{veteran}_i + u_i$$

We are concerned that whether or not someone joins the military is correlated with other characteristics that may affect earnings (e.g. passing the AFQT, work ethic, parental income).

Proposed instrument: Vietnam draft lottery number (lower = drafted)

Must satisfy two conditions:

1. (inclusion) correlated with education
2. (exclusion) not correlated with error term

Do you think it satisfies these two conditions?

IV in Practice

OLS estimate

$$\ln(\widehat{earn})_i = .143 + .075veteran_i$$

(.101) (0.011)

Test the instrument

$$veteran_i = .091 - .143lotterynum_i$$

(.022) (0.037)

Instrumental variable (IV) estimate

$$\ln(\widehat{earn})_i = .245 + .044veteran_i$$

(.446) (0.021)

How do the OLS and IV estimates compare?

IV in Practice

Example:

$$\ln(wage_i) = \beta_0 + \beta_1 educ_i + u_i$$

We are concerned that ability (omitted) is correlated with education level and wages.

Proposed instrument: father's education level

Must satisfy two conditions:

1. (inclusion) correlated with education
2. (exclusion) not correlated with error term (i.e. omitted factors)

Do you think it satisfies these two conditions?

IV in Practice

OLS estimate

$$\ln(\widehat{wage})_i = -.185 + .109 educ_i$$

(.184) (0.014)

Test the instrument

$$\widehat{educ}_i = 10.24 + .269 fatheduc_i$$

(.280) (.029)

Instrumental variable (IV) estimate

$$\ln(\widehat{wage})_i = .441 + .059 educ_i$$

(.446) (.035)

How do the OLS and IV estimates compare?

Lesson 19

Regression Discontinuity

Outline

Previous Lesson:

1. Instrumental Variables
2. Two-Stage Least Squares

This Lesson

1. Regression Discontinuity

Next Lesson

1. Regression Kink
2. Summary of course

Topics

- Regression discontinuity
 - Intuition
 - Regression
- Sharp RD vs Fuzzy RD
- Regression Choices:
 - Polynomial order
 - Bandwidth of data
- Testing for Manipulation:
 - Continuous observables
 - Density test
 - Donut hole test
- Examples

Regression Discontinuity: Intuition

What makes a randomized experiment good:

- treatment is not correlated with omitted variables
(people do not select into treatment)
- have similar people in treatment and control group

What limits the usefulness of experiments:

- many important policies/questions can not be randomized

Regression discontinuity provides an opportunity to examine some of these policies in a way that is almost as good as a randomized experiment.

Regression Discontinuity: Intuition

Example: Suppose that a hospital provides “special care” to babies who are “low weight”. Low weight is defined as the baby weighing less than 1,500 grams. Special care involves admitting the baby to the intensive care unit, keeping the baby in the hospital for a longer period of time after birth, and more frequent appointments over the next year.

We wish to estimate the effect of getting special care on a baby’s health one year later (at 12 months old).

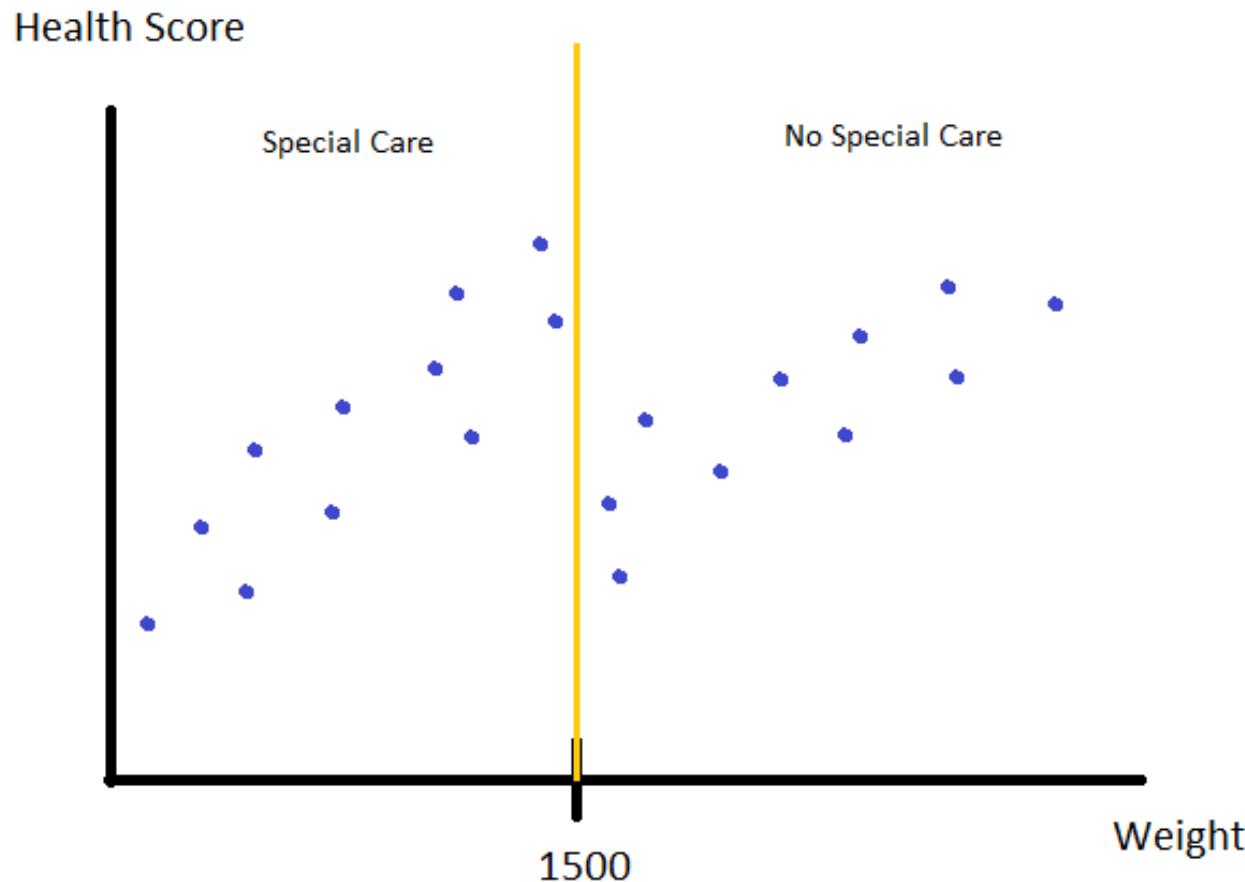
$$HealthScore_i = \beta_0 + \beta_1 * SpecialCare_i + u$$

Why are we concerned that B_1 might be biased?

What sign do you think B_1 would have if we actually estimated this regression for all newborns?

Regression Discontinuity: Intuition

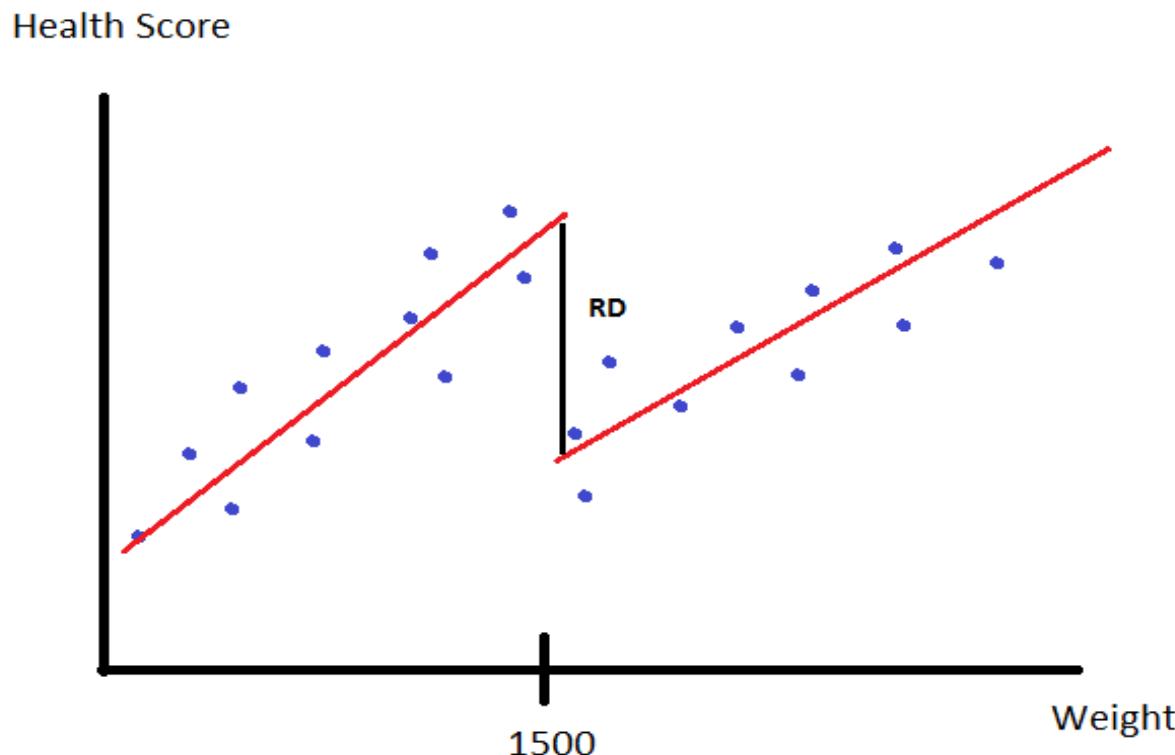
Is there any evidence that getting “Special Care” improved the health outcome of infants?



Regression Discontinuity: Intuition

The RD estimate is the gap between the predicted value at the cutoff as determined separately on the left and right hand side.

Essentially, we want to estimate a regression line on each side of the cutoff and predict outcome at the cutoff. The difference between these two predicted points is the regression discontinuity estimate.



Regression Discontinuity: Intuition

regression discontinuity design – a quasi-experimental design that measures the causal effect by assigning a cutoff above or below which an intervention is assigned.

Assumptions for regression discontinuity:

1. There is a cutoff in an assignment (or running) variable
 - the assignment variable in our example is weight
 - the cutoff needs to be discontinuous (not a steep slope)
2. Subjects can not manipulate which side of the cutoff they are on
 - specifically, they can not precisely control which side they are on
 - thus they are as good as randomly assigned

Regression Discontinuity: Intuition

How is RD similar to and different from a randomized experiment?

Similarities:

1. subjects do not choose whether to be treated
2. close to the cutoff, subjects are as good as randomly assigned
3. can check for a “first-stage” effect: treated group gets treated

Differences:

1. treated and untreated groups are not balanced on average
(we know that those to the left and right do differ by a little)
2. only get an estimate of treatment for subjects near the cutoff:
Local Average Treatment Effect (LATE)

Regression Discontinuity: Regression

What does an RD regression equation look like:

$$Y_i = \beta_0 + \beta_1 Dist_{run < cut} + \beta_2 Dist_{run \geq cut} + \beta_3 Treat_i + u_i$$

- Y_i is the outcome;
- β_1 is the slope on the left
- β_2 is the slope on the right
- β_0 is the point on the untreated side (the right in this example)
- β_3 is the RD estimate (the gap)

Another common way to write this is:

$$Y_i = \beta_0 + \beta_1(1 - T_i)(X_i - c) + \beta_2 T_i(X_i - c) + \beta_3 T_i + u_i$$

Where $X - c$ is the distance from the cutoff.

Regression Discontinuity: Regression

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1(1 - T)(X - c) + \hat{\beta}_2T(X - c) + \hat{\beta}_3T$$

Right: $X \rightarrow c$, Treat=0

$$\hat{Y}_{Untreat} = \hat{\beta}_0 + \hat{\beta}_1 * 1 * (c - c) + \hat{\beta}_2 * 0 * (c - c) + \hat{\beta}_3 * 0 = \hat{\beta}_0$$

Left: $X \rightarrow c$, Treat=1

$$\hat{Y}_{Treat} = \hat{\beta}_0 + \hat{\beta}_1 * 0 * (c - c) + \hat{\beta}_2 * 1 * (c - c) + \hat{\beta}_3 * 1 = \hat{\beta}_0 + \hat{\beta}_3$$

RD estimate: (treated side – untreated side)

$$\hat{Y}_{Treat} - \hat{Y}_{Untreat} = (\hat{\beta}_0 + \hat{\beta}_3) - \hat{\beta}_0 = \hat{\beta}_3$$

Regression Discontinuity: Regression

Example: Newborns assigned to special care if weight < 1,500 grams.

$$Health_i = 4.32 + 0.013(1 - T_i)(X - 1500) + 0.004T_i(X - 1500) + 1.64T_i$$

(0.26) (0.003) (0.003) (0.59)

Does this satisfy the two assumptions:

1. there is a cutoff in the assignment variable?
2. subjects cannot manipulate which side of the cutoff they are on?

What is the regression discontinuity estimate of the effect of Special Care?
Is it statistically significant?

How do we interpret each of the coefficients in the regression?

Do you think we would get the same estimate if the weight cutoff was higher or lower?

Regression Discontinuity: Regression

Treatment is assigned according to a threshold:

$$D = \begin{cases} 0 & X \geq c \\ 1 & X < c \end{cases}$$

What we would like to observe is:

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

What we are able to estimate is:

$$B - A = \lim_{\varepsilon \downarrow 0} E[Y_i | X_i = c + \varepsilon] - \lim_{\varepsilon \uparrow 0} E[Y_i | X_i = c + \varepsilon]$$

Which is equal to the local average treatment effect at the cutoff c :

$$E [Y_i(1) - Y_i(0) | X = c]$$

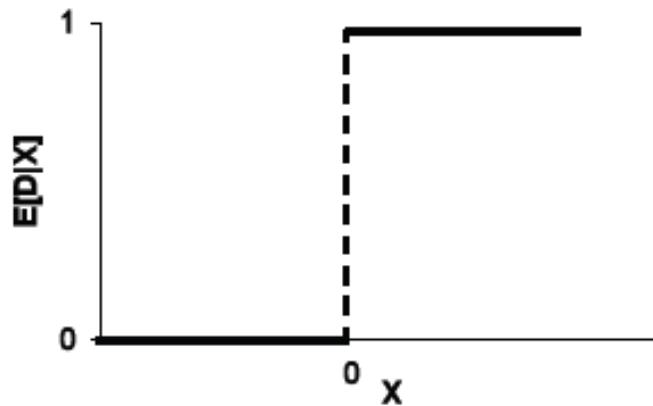
Regression Discontinuity: Fuzzy

Fuzzy RD – treatment is only partly determined by crossing the cutoff point.

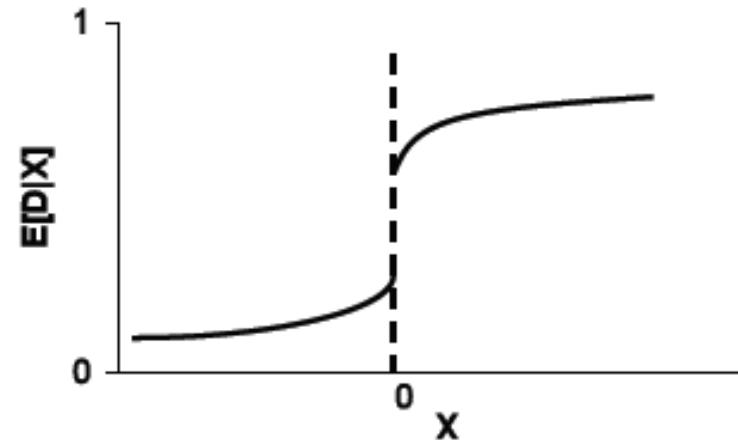
In “sharp” RD, treatment jumps from 0 to 1. In “fuzzy” RD, treatment jumps by less than 1. For example, in the graph below, from 0.25 to 0.75.

Some observations on the left are treated (always-takers) and some on the right are not (non-compliers). Examples?

Sharp RD



Fuzzy RD



Regression Discontinuity: Fuzzy

In a sharp RD, we know that treatment goes from 0 to 1, so we can interpret the change in Y_i as the local average treatment effect.

In fuzzy RD, there is a less than perfect change in treatment when we cross the threshold. However, there is a discontinuity:

$$\lim_{\varepsilon \downarrow 0} \Pr(D = 1 | X = c + \varepsilon) \neq \lim_{\varepsilon \uparrow 0} \Pr(D = 1 | X = c + \varepsilon)$$

The change in Y_i is only driven by the difference in the fraction who are treated on either side.

In order to get the desired treatment effect, we need to scale the change in Y by the change in the fraction treated:

$$\tau_F = \frac{\lim_{\varepsilon \downarrow 0} \mathbb{E}[Y | X = c + \varepsilon] - \lim_{\varepsilon \uparrow 0} \mathbb{E}[Y | X = c + \varepsilon]}{\lim_{\varepsilon \downarrow 0} \mathbb{E}[D | X = c + \varepsilon] - \lim_{\varepsilon \uparrow 0} \mathbb{E}[D | X = c + \varepsilon]}$$

Regression Discontinuity: Fuzzy

Example of fuzzy RD:

$$Earnings_i = \beta_0 + \beta_1 Flagship_i + \epsilon_i$$

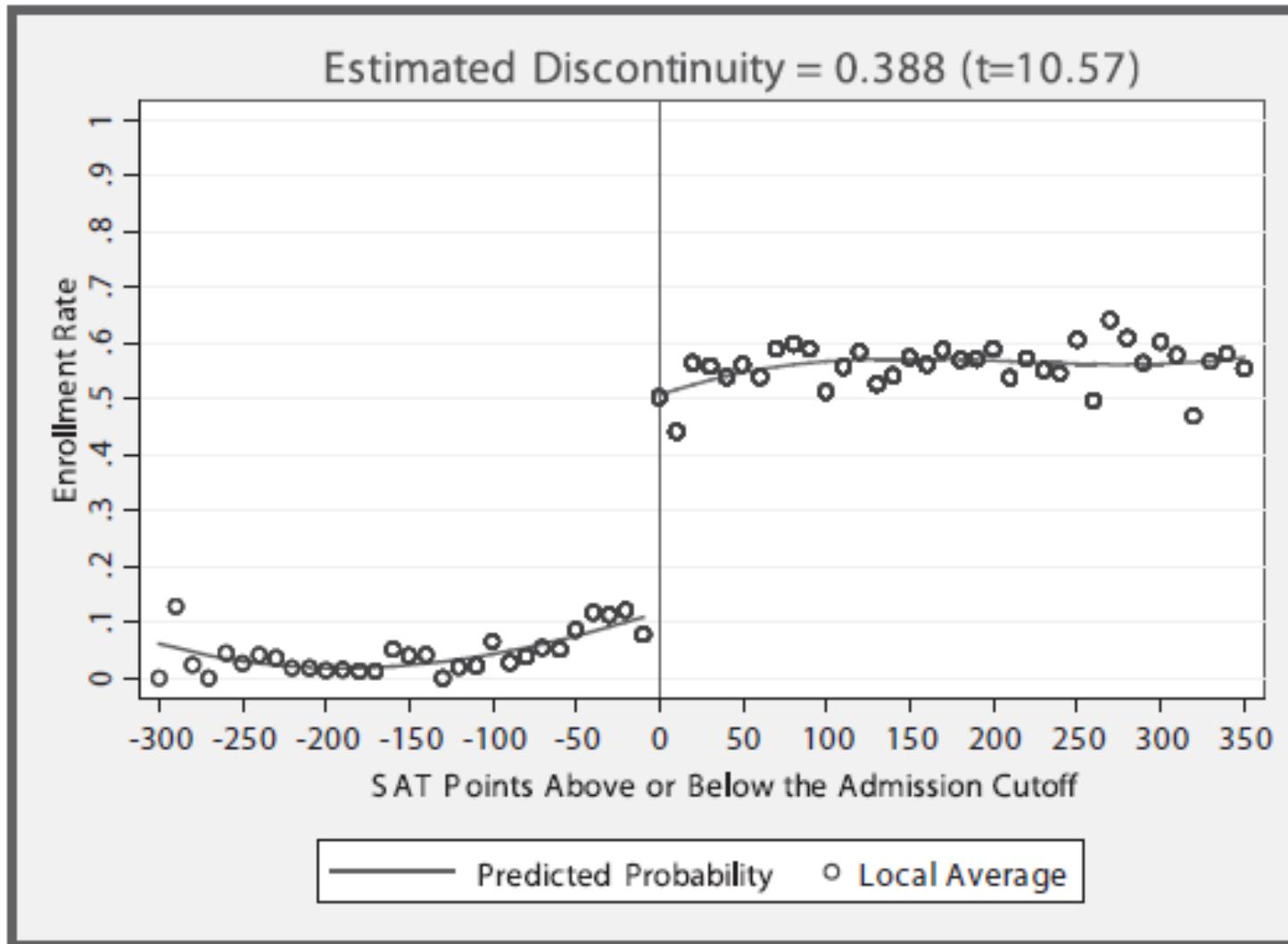
Admission to a major (unnamed) state university is determined by a student's SAT score.

However, not all students below the cutoff are rejected from the school. Why?

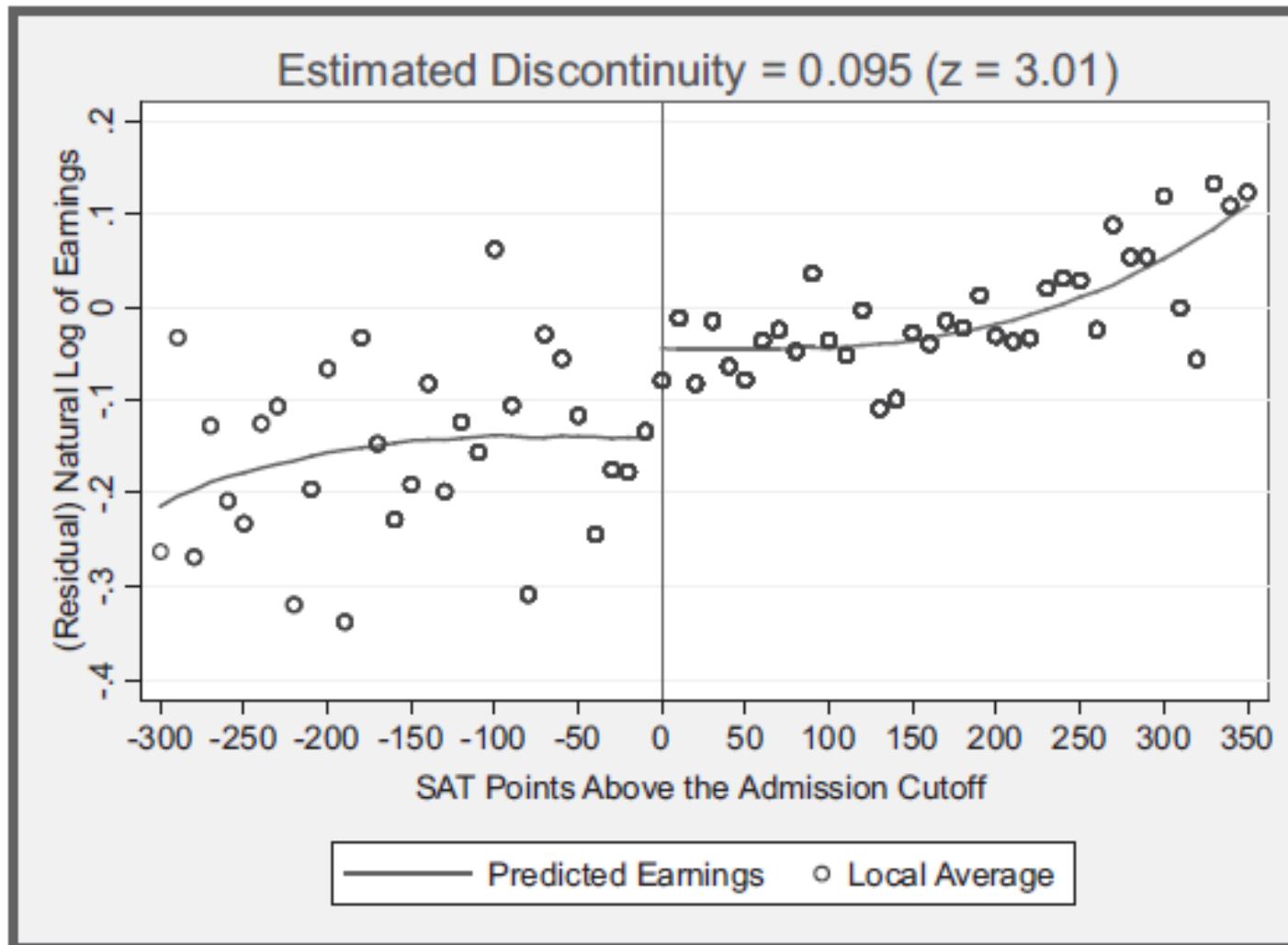
Likewise, not all students above the cutoff are admitted or attend. Why?

Regression Discontinuity: Fuzzy

FIGURE 1.—FRACTION ENROLLED AT THE FLAGSHIP STATE UNIVERSITY



Regression Discontinuity: Fuzzy



Regression Discontinuity: Fuzzy

So, in the fuzzy RD case, we estimate two equations:

one that estimates the change in fraction treated (with $\alpha_3 = .1$):

$$\ln(\text{earnings}_i) = \alpha_0 + \alpha_1 \text{Dist}_{\text{run} < \text{cut}} + \alpha_2 \text{Dist}_{\text{run} \geq \text{cut}} + \alpha_3 T_i + \eta_i$$

and another that indicates the change in the outcome (with $\delta_3 = .4$):

$$\text{attend}_i = \delta_0 + \delta_1 \text{Dist}_{\text{run} < \text{cut}} + \delta_2 \text{Dist}_{\text{run} \geq \text{cut}} + \delta_3 T_i + u_i$$

We can then take the ratio of these to properly scale the outcome:

$$\hat{\beta}_1 = \frac{\hat{\alpha}_3}{\hat{\delta}_3} = \frac{.1}{0.4} = 0.25$$

Thus, the estimated effect of attending the flagship university is 25% higher earnings.

Regression Discontinuity: Choices

Choice of Polynomial:

So far we have assumed a linear slope on either side of the cutoff. This is unlikely to hold in practice. Thus we may want to include higher order polynomials in the running variable:

$$Y_i = \beta_0 + \beta_1(1 - T_i)(X_i - c) + \beta_2(1 - T_i)(X_i - c)^2 + \beta_3 T_i(X_i - c) + \beta_4 T_i(X_i - c)^2 + \beta_5 T_i + u_i$$

Using 3rd, 4th, and 5th order polynomials is quite common.

Another consideration is whether to use a separate polynomial on each side (as we have been, or a single polynomial:

$$Y_i = \beta_0 + \beta_1(X_i - c) + \beta_2(X_i - c)^2 + \beta_3(X_i - c)^3 + \beta_4 T_i + u_i$$

What advantages does using one polynomial have?

What advantages does using separate polynomials have?

Regression Discontinuity: Choices

Choice of bandwidth:

We have been assuming that we use all of the data on either side of the cutoff.

However, it might be the case that we want to restrict attention to those close to the cutoff, since they may be more similar to those at the cutoff.

Specifically, choosing bandwidth h implies that we keep an observation if $|X-c| < h$.

The tradeoff is:

Precision: including more data on either side allows a more precise estimate of the regression.

Bias: including more data introduces bias since the observations further away are less like those at the cutoff.

Regression Discontinuity: Choices

Many rules of thumb and procedures have been developed for choosing the optimal bandwidth and polynomial.

Let's discuss on such class of procedures called "cross-validation" procedure for the case of optimal bandwidth:

1. Consider one side of the cutoff and a bandwidth to test (say h)
2. Consider the first observation on that side. Run a regression using bandwidth h while excluding that point. Predict that point using the regression. Obtain the error of the prediction and square it.
3. Repeat this process for a large number of points and sum all of the squared errors.
4. Now repeat the whole process again using a different bandwidth.
5. The bandwidth that does the best job of predicting these points (minimizes the error) is the preferred bandwidth.

Regression Discontinuity: Choices

Choice to include baseline covariates:

One way to control for changes in the composition of individuals is to include their characteristics (covariates) in the regression equation:

$$Y_i = \beta_0 + \beta_1(1 - T_i)(X_i - c) + \beta_2 T_i(X_i - c) + \mathbf{W}\boldsymbol{\delta} + \beta_3 T_i + u_i$$

Where \mathbf{W} may include things like race, gender, age for individuals, or industry, company size... for a firm.

The irrelevance of covariates:

Our assumption is that people are as good as randomly assigned to either side right around the cutoff. Thus the covariates should change smoothly at the cutoff. Thus including covariates \mathbf{W} should have no effect.

Regression Discontinuity: Choices

Best practice:

- Show your results for a range of polynomial orders.
- Show your results for one and separate polynomials.
- Show your results for a range of bandwidths.
- Show your results with and without covariates added.

Let's see how the analysis of being admitted to the flagship university handled this.

Regression Discontinuity

TABLE 1.—EARNINGS DISCONTINUITIES AND CORRESPONDING INTENT-TO-TREAT AND ENROLLMENT ESTIMATES FOR WHITE MEN

Regression Specification	Function of Adjusted SAT	Flexible Polynomial?	Additional Controls	Discontinuity		Treatment Effect	
				Estimated Earnings Discontinuity	Intent-to- Treat Effect	Enrollment Effect	
(1) Plotted in Figure 2	Cubic	No	No	0.095*** (0.032) [0.003]	0.135*** (0.046) [0.004]	0.223*** (0.079) [0.005]	
(2)	Cubic	No	Yes	0.092*** (0.033) [0.005]	0.131*** (0.048) [0.006]	0.216*** (0.081) [0.008]	
(3)	Quadratic	Yes	Yes	0.111** (0.045) [0.014]	0.170** (0.073) [0.019]	0.281** (0.121) [0.021]	
(4) (includes only applicants within 200 points of cutoff)	Quadratic	No	Yes	0.081** (0.038) [0.034]	0.116** (0.056) [0.038]	0.192** (0.094) [0.041]	
(5) (includes only applicants within 100 points of cutoff)	Linear	No	Yes	0.074** (0.038) [0.050]	0.110* (0.058) [0.060]	0.181* (0.099) [0.067]	

Regression Discontinuity

A fundamental assumption is that there is no manipulation where individuals determine which side of the cutoff they are on. We may be concerned that this is violated.

There are several ways to examine this:

1. Density: Examine the density of observations around the cutoff point: if there is bunching on one side of the cutoff, then it may indicate manipulation.
2. Continuous covariates: Estimate the RD regression but use an observable characteristic as the outcome variable. A discontinuity in characteristics may indicate manipulation.
3. Donut hole test: Omit the observations closest to the cutoff (i.e. a donut hole). If the results change dramatically, then it may suggest manipulation.

Regression Discontinuity

Political scientists are very interested in whether the incumbent party (Republican or Democrat) has an advantage in elections.

This is untestable in the cross-section, since the incumbent party may simply be more popular in an area.

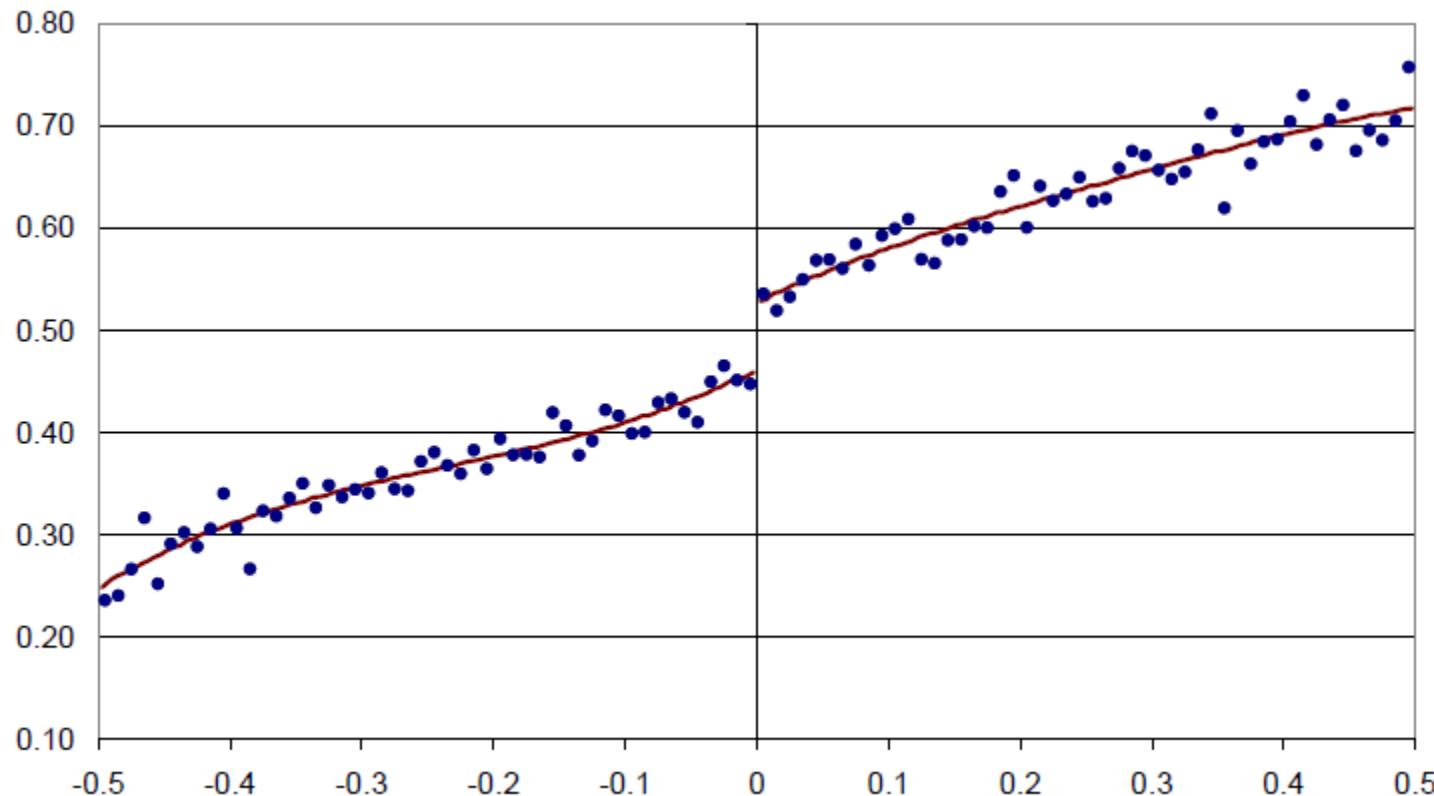
That is observing that incumbents win does not indicate that there is an incumbent advantage.

Exploit close votes:

- Consider very close votes.
- One party wins with 50.1% of votes and the other loses with 49.9%.
- Now one party is “treated” as the incumbent and we can examine the results in the next election.

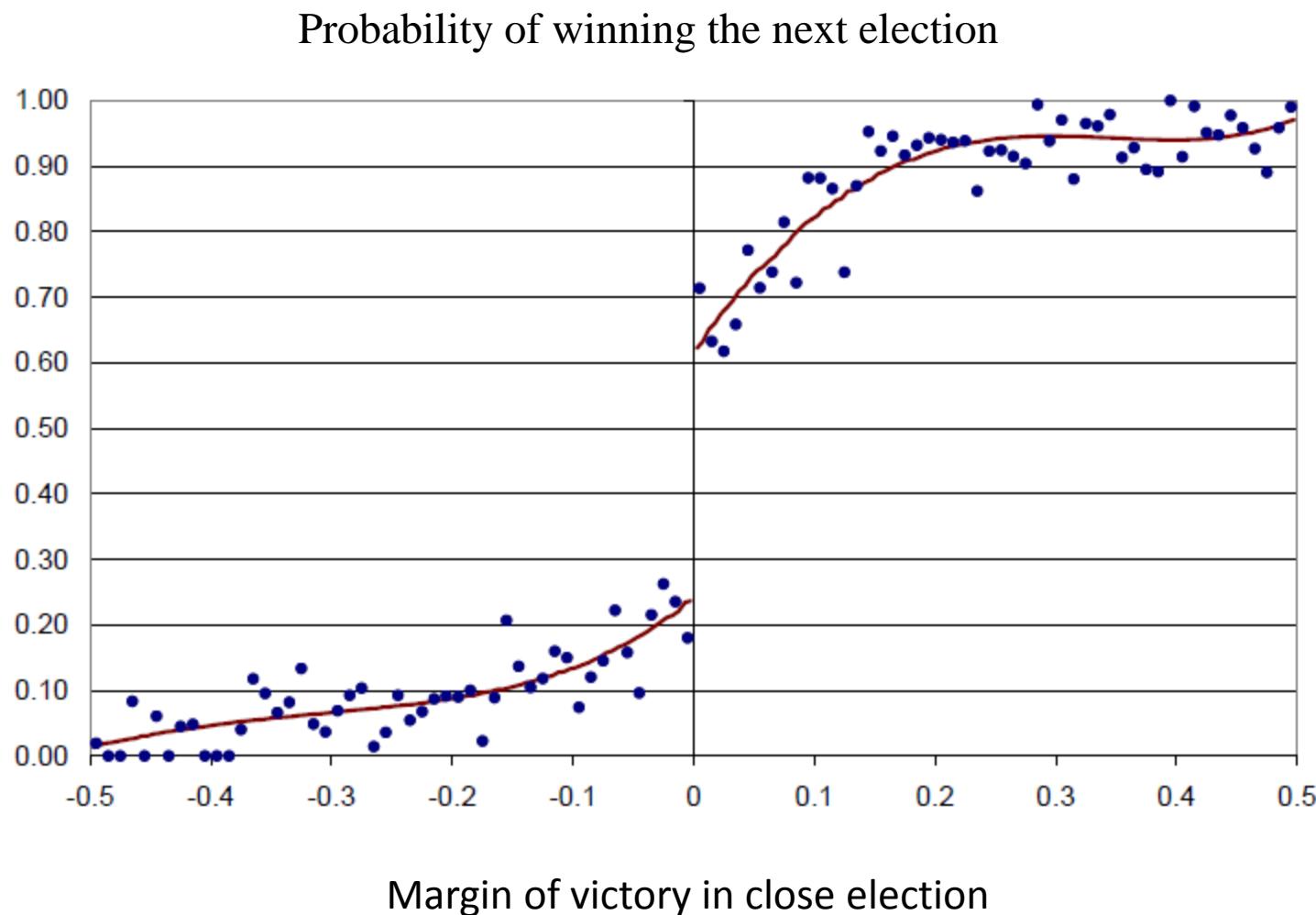
Regression Discontinuity

Share of votes in the next election



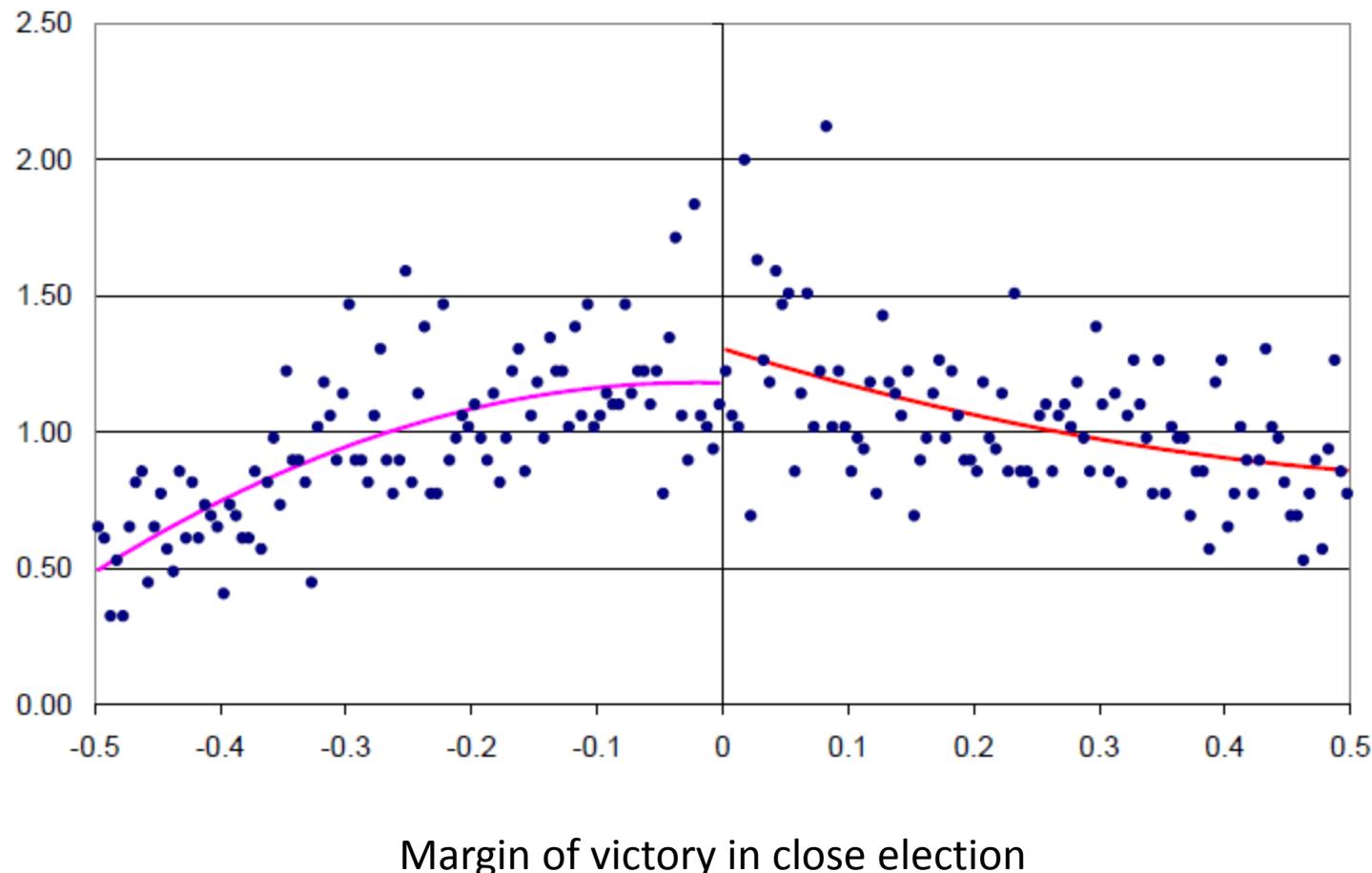
Margin of victory in close election

Regression Discontinuity



Regression Discontinuity

Density of the running variable



Regression Discontinuity

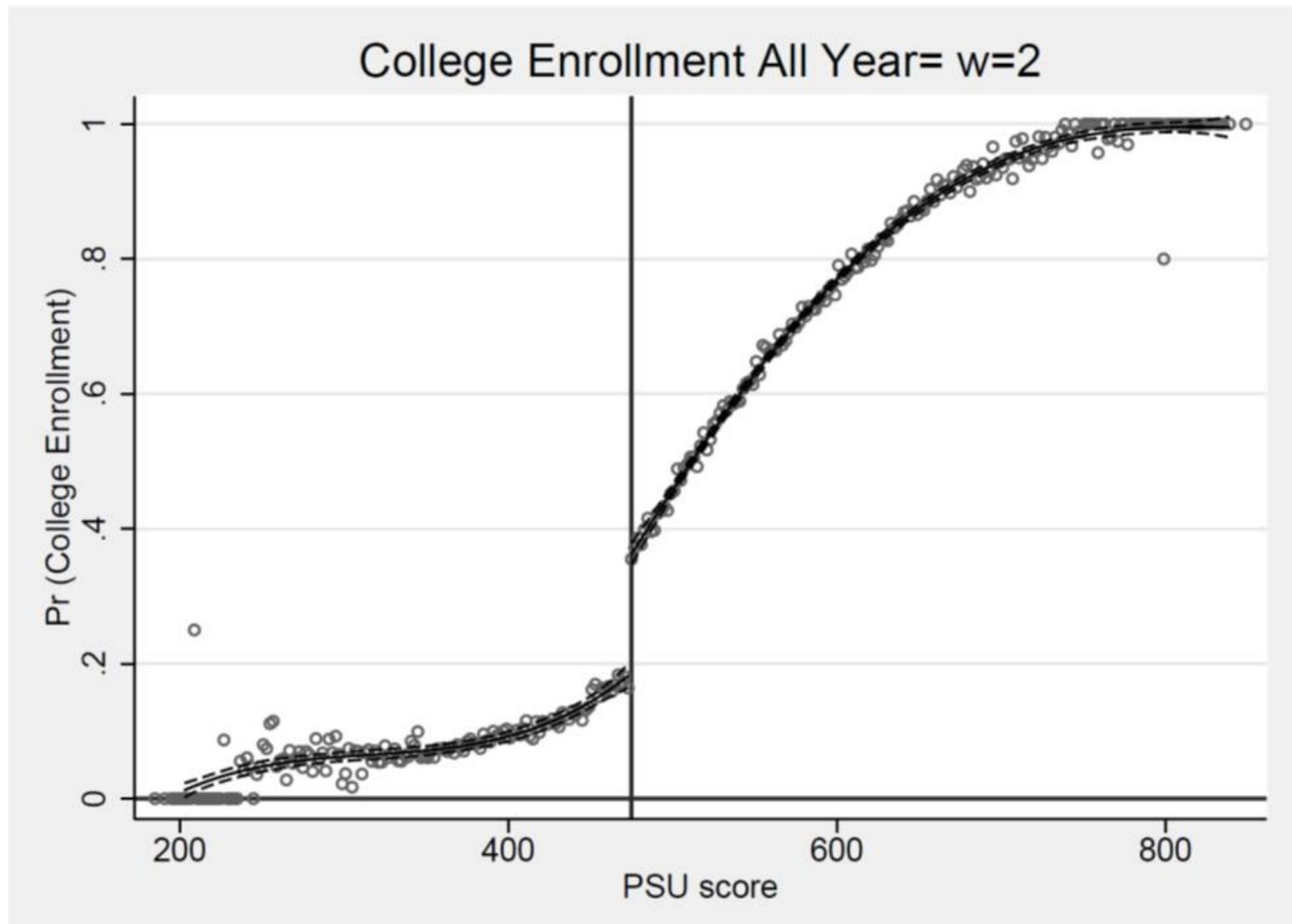
Table 2a: RD estimates of the effect of winning the previous election on the share of votes in the next election

Bandwidth:	1.00	0.50	0.25	0.15	0.10	0.05	0.04	0.03	0.02	0.01
Polynomial of order:										
Zero	0.347 (0.003) [0.000]	0.257 (0.004) [0.000]	0.179 (0.004) [0.000]	0.143 (0.005) [0.000]	0.125 (0.006) [0.003]	0.096 (0.009) [0.047]	0.080 (0.011) [0.778]	0.073 (0.012) [0.821]	0.077 (0.014) [0.687]	0.088 (0.015)
One	0.118 (0.006) [0.000]	0.090 (0.007) [0.332]	0.082 (0.008) [0.423]	0.077 (0.011) [0.216]	0.061 (0.013) [0.543]	0.049 (0.019) [0.168]	0.067 (0.022) [0.436]	0.079 (0.026) [0.254]	0.098 (0.029) [0.935]	0.096 (0.028)
Two	0.052 (0.008) [0.000]	0.082 (0.010) [0.335]	0.069 (0.013) [0.371]	0.050 (0.016) [0.385]	0.057 (0.020) [0.458]	0.100 (0.029) [0.650]	0.101 (0.033) [0.682]	0.119 (0.038) [0.272]	0.088 (0.044) [0.943]	0.098 (0.045)
Three	0.111 (0.011) [0.001]	0.068 (0.013) [0.335]	0.057 (0.017) [0.524]	0.061 (0.022) [0.421]	0.072 (0.028) [0.354]	0.112 (0.037) [0.603]	0.119 (0.043) [0.453]	0.092 (0.052) [0.324]	0.108 (0.062) [0.915]	0.082 (0.063)
Four	0.077 (0.013) [0.014]	0.066 (0.017) [0.325]	0.048 (0.022) [0.385]	0.074 (0.027) [0.425]	0.103 (0.033) [0.327]	0.106 (0.048) [0.560]	0.088 (0.056) [0.497]	0.049 (0.067) [0.044]	0.055 (0.079) [0.947]	0.077 (0.063)
Optimal order of the polynomial	6	3	1	2	1	2	0	0	0	0
Observations	6558	4900	2763	1765	1209	610	483	355	231	106

Notes: Standard errors in parentheses. P-values from the goodness-of-fit test in square brackets. The goodness-of-fit test is obtained by jointly testing the significance of a set of bin dummies included as additional regressors in the model. The bin width used to construct the bin dummies is .01. The optimal order of the polynomial is chosen using Akaike's criterion (penalized cross-validation)

Regression Discontinuity

Chilean fin. aid is determined by Prueba de Selección Universitaria (PSU) score:



Lesson 2

Fundamental Concepts

Outline

Previous Lesson:

1. Stages of econometric analysis
2. Recurring examples

This Lesson:

1. Introductory Example
2. Statistical Framework
3. Properties of estimators
4. Hypothesis testing (introduction)

Next Lesson:

1. Types of data
2. Data handling

Introductory Example

Example:

- Compare the average age at death of women and men.
- Our goal is to estimate the sign and magnitude of the difference and determine if it is statistically significant.

Comparing means:

- Involves many of the same steps and requires many of the same assumptions as more complicated econometric analyses.
- In this simple context we will state several important definitions, discuss the desirable properties of estimators, and introduce hypothesis testing.

A Motivating Example

Table 1.1 Average age at death for the EU15 countries (2002)

	<i>Women</i>	<i>Men</i>
Austria	81.2	75.4
Belgium	81.4	75.1
Denmark	79.2	74.5
Finland	81.5	74.6
France	83.0	75.5
Germany	80.8	74.8
Greece	80.7	75.4
Ireland	78.5	73.0
Italy	82.9	76.7
Luxembourg	81.3	74.9
Netherlands	80.6	75.5
Portugal	79.4	72.4
Spain	82.9	75.6
Sweden	82.1	77.5
UK	79.7	75.0
Mean	81.0	75.1
Standard deviation	1.3886616	1.2391241

A Motivating Example

Theory: Women live longer than men.

Econometric model: Find the difference between the means.

$$\bar{Y}_w - \bar{Y}_m$$

$$\bar{Y}_w = \frac{1}{n} \sum_i Y_{wi} = \frac{1}{15} \sum_i Y_{wi} = 81.0$$

$$\bar{Y}_m = \frac{1}{n} \sum_i Y_{mi} = \frac{1}{15} \sum_i Y_{mi} = 75.1$$

A Motivating Example

Variance: measure of dispersion of age at death for males and females [note the typo in the book].

$$S_j^2 = \frac{1}{15} \sum_{i=1}^n (Y_{ji} - \bar{Y}_j)^2 \quad j = w, m$$

The standard deviation is the square root of the variance S_j .

t – statistic: method of testing for statistically significant differences.

$$t = \frac{\bar{Y}_w - \bar{Y}_m}{\sqrt{\frac{S_w^2}{n_w} + \frac{S_m^2}{n_m}}} = \frac{\bar{Y}_w - \bar{Y}_m}{\sqrt{\frac{S_w^2}{15} + \frac{S_m^2}{15}}} = \frac{81 - 75.1}{\sqrt{\frac{1.389^2}{15} + \frac{1.24^2}{15}}} = 12.27$$

This is very large t-statistic (typically anything greater than 1.96 indicates significance at the 5 percent level), so we conclude that women live longer than men.

Statistical Framework

This simple comparison of means introduces several questions:

- 1) Is this data set a good one for making claims about the world population (i.e. is it a good sample)?
- 2) Does the difference in means have desirable properties for an estimator?
- 3) How are we able to assign probabilities to hypothesis tests?

Samples

We are typically interested in some population of individuals, but we can only get data for a subset (or sample) of them.

Population:

Data for all individuals in a group that we wish to study:

- all countries in the world
- all UC Santa Cruz students
- all companies in California

Sample:

Data selected from the population by a defined procedure.

- European countries
- 100 students at the UCSC library
- companies in CA with at least 50 employees

Samples

In order to make accurate statements about the population, we need our sample to be representative of the population.

Random sample: members of the population are selected to be part of the sample at random. Each member of the population has an equal probability of being in the sample.

Violations of random sampling:

- *Selection bias* – those selected to be in the sample are not representative of the population. This is typically due to the data collection method.
- *Response bias* – we are unable to collect data for those selected to be in the sample, and the missing individuals are not random. This is common in survey data.

Samples

Example:

Suppose we want to know how many hours UCSC students study each day. Evaluate each of the following sampling methods in terms of *selection bias* and *non-response bias*:

1. Survey 100 students outside of the campus library.
2. Survey 20 students at each of the college dining halls.
3. Pull student ids from the registrar, randomly select 100 of them, and send them a survey by email.

Any other ideas?

Samples

Assumption: data are drawn at random.

Implications:

- observations are *independent* of each other (uncorrelated)
- observations have *identical distribution* (expected value, variance)
- i.i.d. – independent and identically distributed

Do you think the data on age at death is i.i.d. for the world population?

Properties of Estimators

Consider a random variable Y_i :

Population mean and sample estimator:

$$E(Y) = \mu_Y \quad \hat{\mu}_Y = \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$$

Population variance and sample estimator (note Bessel's Correction):

$$\text{var}(Y) = \sigma_Y^2 \quad \hat{\sigma}_Y^2 = s^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$$

Population standard deviation and sample estimator:

$$s.d.(Y) = \sigma_Y \quad \hat{\sigma}_Y = s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2}$$

Properties of Estimators

unbiased – the expected value is equal to the true population value.

$$E(\bar{Y}) = \mu_Y$$

consistent – the expected value converges to the true population value as the sample becomes large.

$$\lim_{n \rightarrow \infty} \bar{Y} = \mu_Y$$

efficiency – minimizes some loss function. For example, minimizes the variance of the estimate.

If there are multiple estimators that are unbiased or consistent, we can use efficiency to choose between them.

Properties of Estimators

Let's see how a sample mean performs as an estimator:

$$\begin{aligned} E(\bar{Y}) &= E\left(\frac{1}{n} \sum_{i=1}^n Y_i\right) \\ &= \frac{1}{n} \sum_{i=1}^n E(Y_i) \\ &= \frac{1}{n} \sum_{i=1}^n \mu_Y && \text{since random} \\ &= \left(\frac{1}{n} n\right) \mu_Y = \mu_Y \end{aligned}$$

So, the population mean is an unbiased estimator of the population mean.

Note that Y_i is also an unbiased estimator of the population mean since $E(Y_i)=\mu_Y$. So what makes the sample mean better?

Properties of Estimators

Let's estimate the variance of the mean:

$$\begin{aligned} \text{var}(\bar{Y}) &= E(\bar{Y} - \mu_Y)^2 \\ &= E\left(\frac{1}{n} \sum_{i=1}^n Y_i - \mu_Y\right)^2 \\ &= E\left(\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n (Y_i - \mu_Y)(Y_j - \mu_Y)\right) \\ &= E\left(\frac{1}{n^2} \sum_{i=1}^n (Y_i - \mu_Y)^2 + \frac{1}{n^2} \sum_{i=1}^n \sum_{j \neq i}^n (Y_i - \mu_Y)(Y_j - \mu_Y)\right) \\ &= \frac{1}{n^2} \left(\sum_{i=1}^n \text{var}(Y_i) + \sum_{i=1}^n \sum_{j \neq i}^n \text{cov}(Y_i, Y_j) \right) \\ &= \frac{\sigma_Y^2}{n} \quad (\text{cov}=0 \text{ because independent}) \end{aligned}$$

Whereas the $\text{var}(Y_i) = \sigma_Y^2$ by definition, so the sample mean has smaller variance.

Properties of Estimators

$$var(\bar{Y}) = \frac{\sigma_Y^2}{n}$$

So, the variance is smaller when:

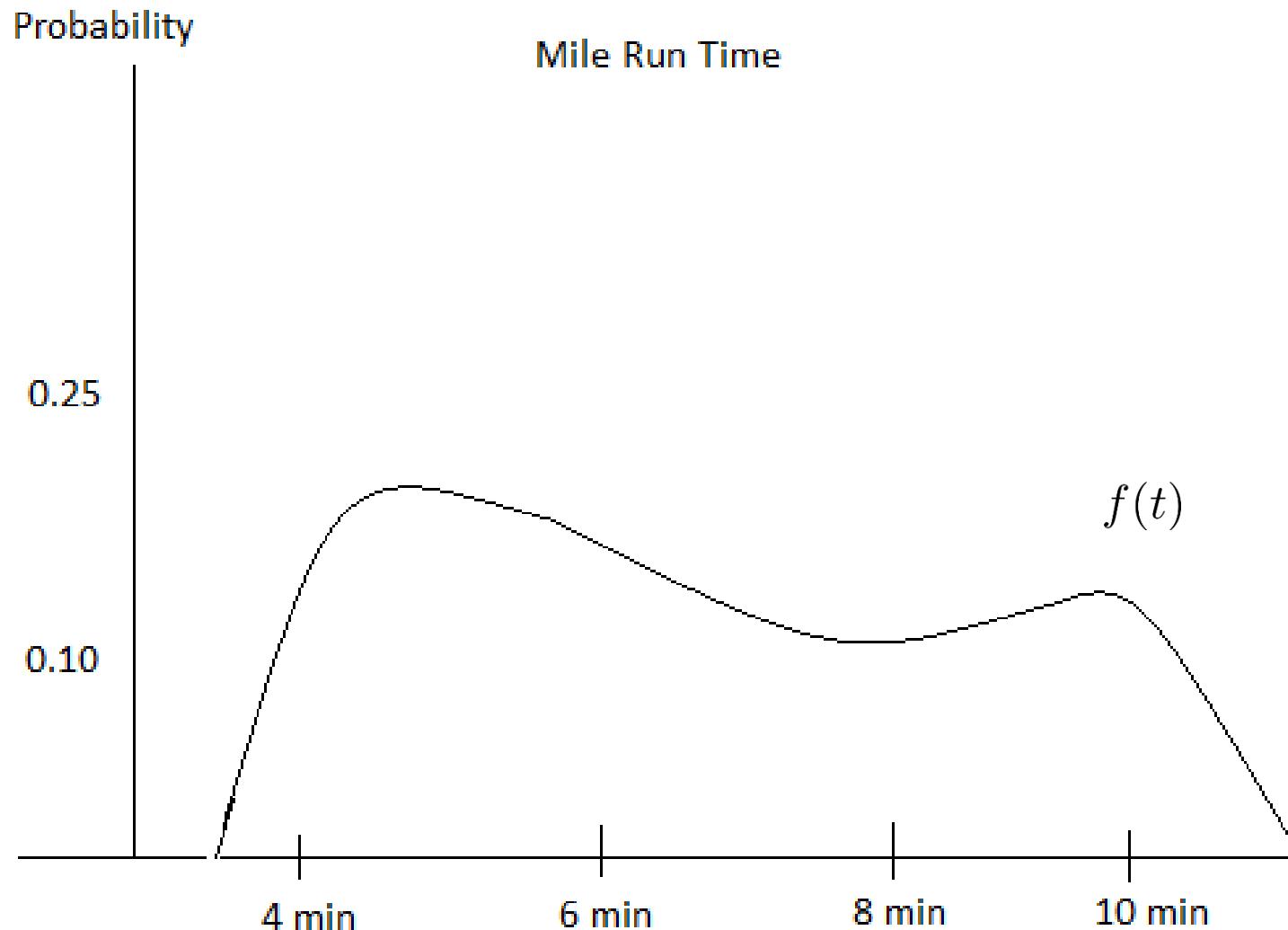
- the sample is larger (more data is better)
- the population variance σ_Y^2 is smaller

Recall, the standard deviation is just the square root of the variance:

$$s.d.(\bar{Y}) = \frac{\sigma_Y}{\sqrt{n}}$$

Distributions

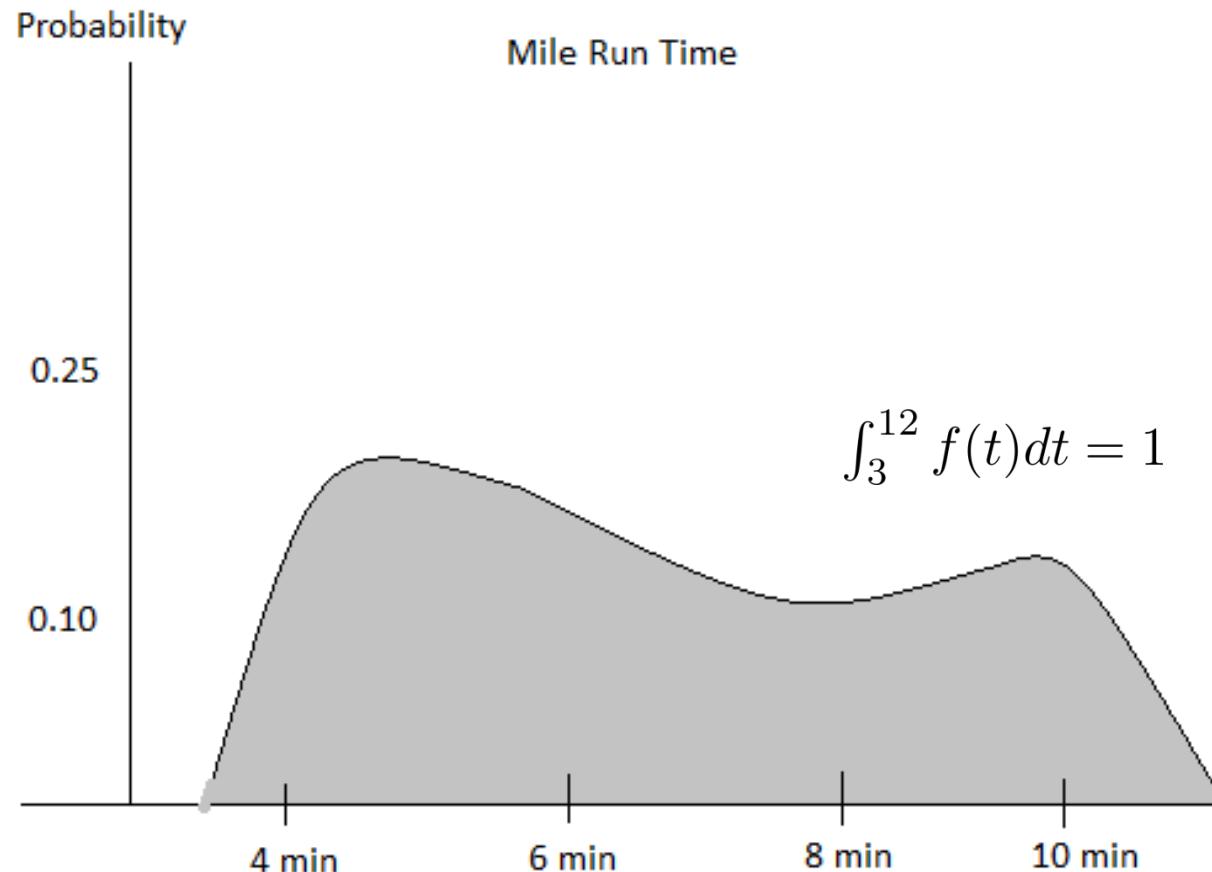
The probability distribution is given by a density function: $f(x)$



Distributions

If we integrate to get the area under the entire distribution, we will get 1.
[This is like adding up the probabilities.]

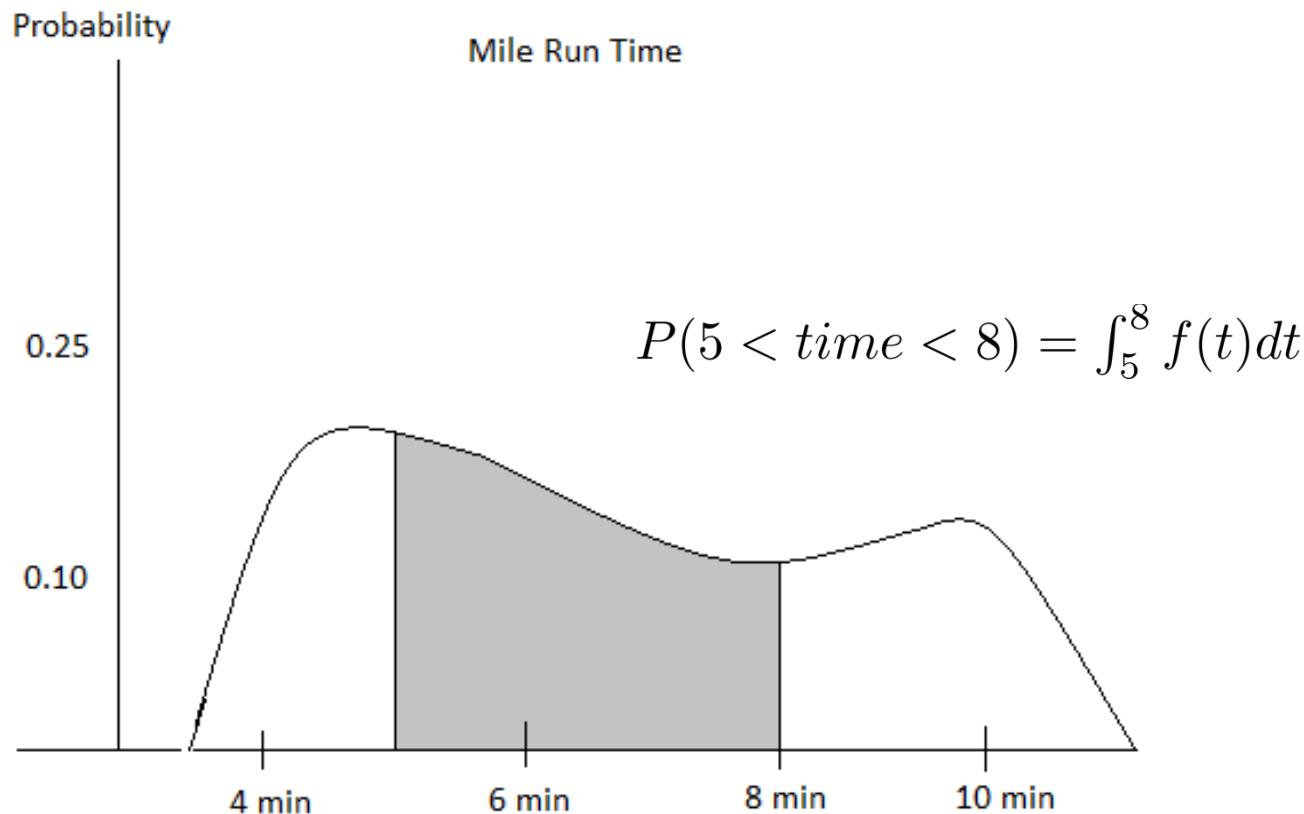
$$\int_{-\infty}^{\infty} f(x)dx = 1$$



Distributions

The probability that a random draw X falls between c and d is found by integrating over the density function from c to d .

$$P(c < X < d) = \int_c^d f(x)dx$$



Hypothesis Testing

State a null hypothesis, H_0 , about the true value of the estimate:

- the true value is equal to 0
- the true value is equal to some constant

$$H_0 : \tilde{Y} = c$$

State an alternative hypothesis, H_1 :

- the true value is smaller (one-sided)
- the true value is larger (one-sided)
- the true value is either larger or smaller (two-sided test)

$$H_1 : \tilde{Y} \neq c$$

- We want to calculate the probability that the observed estimate could have occurred by chance if the null hypothesis is true. This requires that we have an idea of the distribution of estimator.

Hypothesis Testing

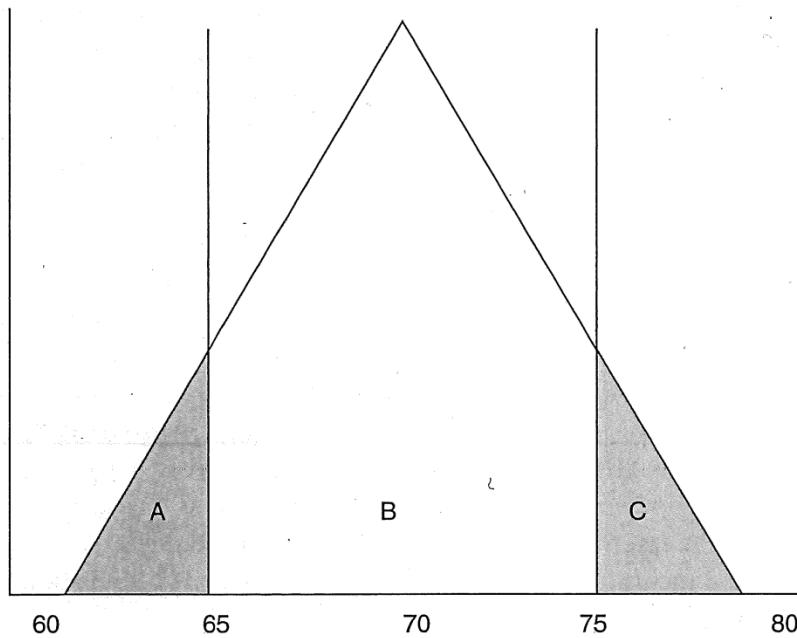


Figure 1.1 A possible distribution for life expectancy

Suppose we want to test if the average age at death for a man is equal to 70 against the alternative that it is not. We know the sample average is 75.1.

Assume that the distribution of means is the triangle above. If 70 is the true mean, then the probability of being more than 5 years off is equal to the shaded region: A+C.

Hypothesis Testing

The problem:

- we typically do not observe the distribution
- the distribution may be very complicated

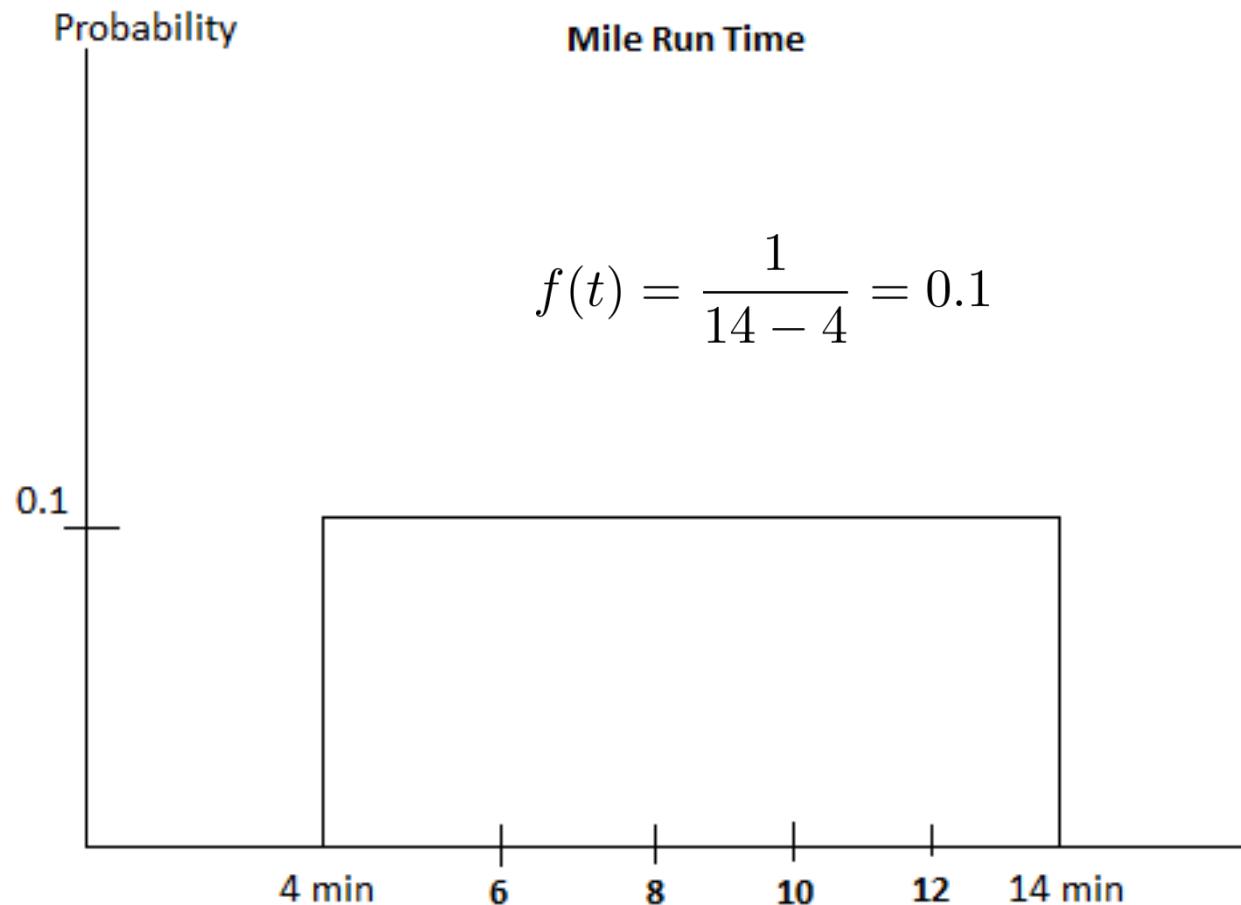
Two notable exceptions:

- the uniform distribution
 - occurs in many discrete settings
- the standard normal distribution
 - normal distributions occur naturally
 - means are normally distributed (Central Limit Thm)

Uniform distribution

A uniform distribution that has equal probability between a and b has a very simple density function:

$$f(x) = \frac{1}{b - a} \quad \text{for } x \text{ in } (a, b)$$



Uniform distribution

If we want to know the probability of X falling between c and d , we integrate over this range:

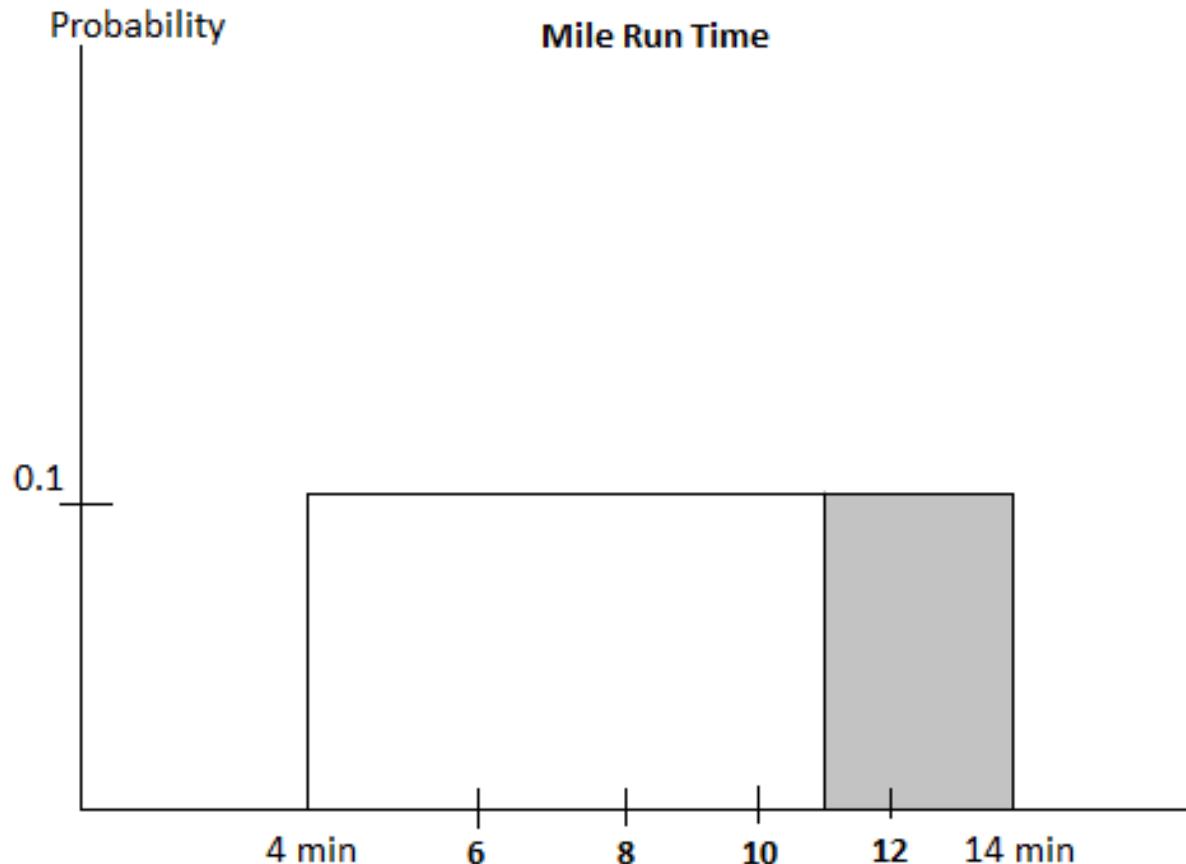
$$\begin{aligned} P(c < X < d) &= \int_c^d \frac{1}{b-a} dx \\ &= \frac{1}{b-a} \int_c^d dx \\ &= \left(\frac{1}{b-a} \right) x \Big|_c^d \\ &= \frac{1}{b-a} [d - c] \end{aligned}$$

This is just the density times the range of values (i.e. calculating the area of a rectangle).

Uniform distribution

If we want to know the probability of X falling between c and d , we integrate over this range:

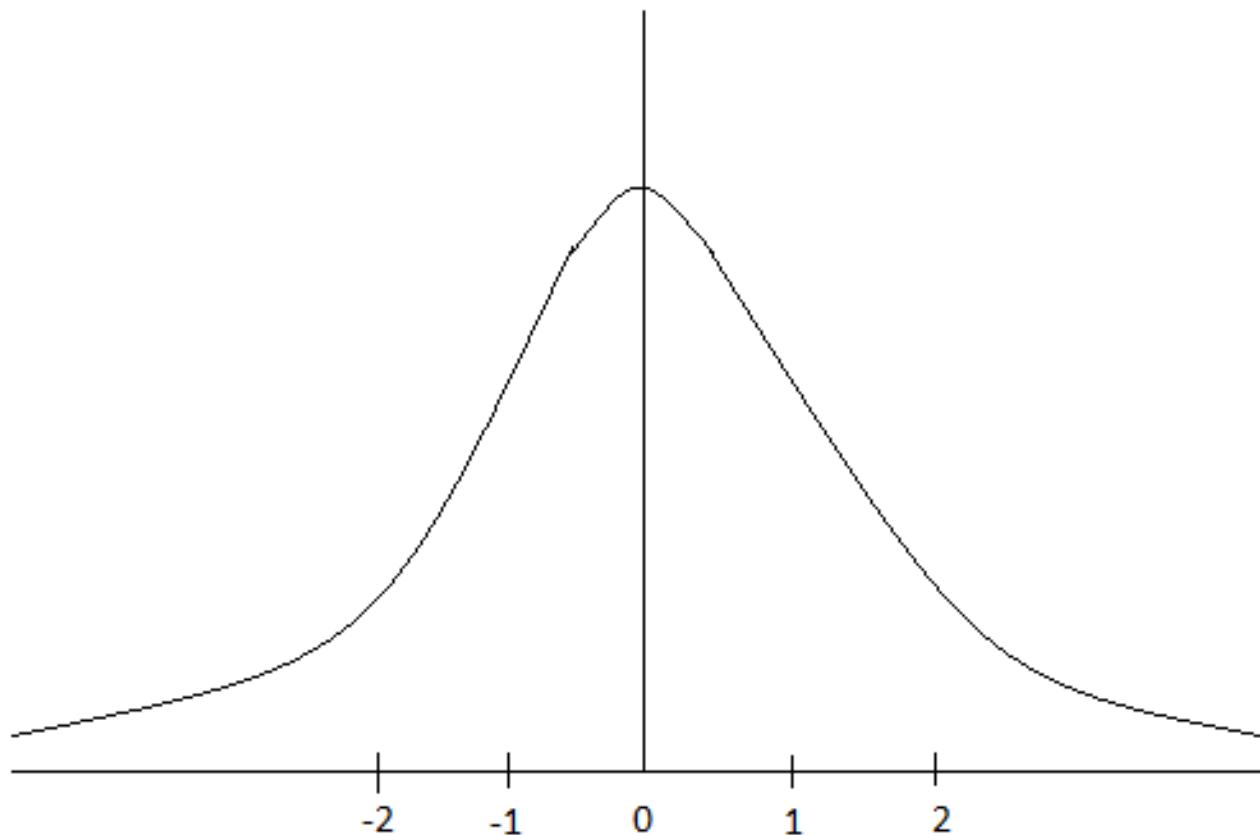
$$P(X > 11) = 0.1[14 - 11] = 0.1[3] = 0.3$$



Standard normal distribution

The standard normal distribution has a mean of 0 and standard deviation of 1.

$$Y \sim N(0, 1)$$



Standard normal distribution

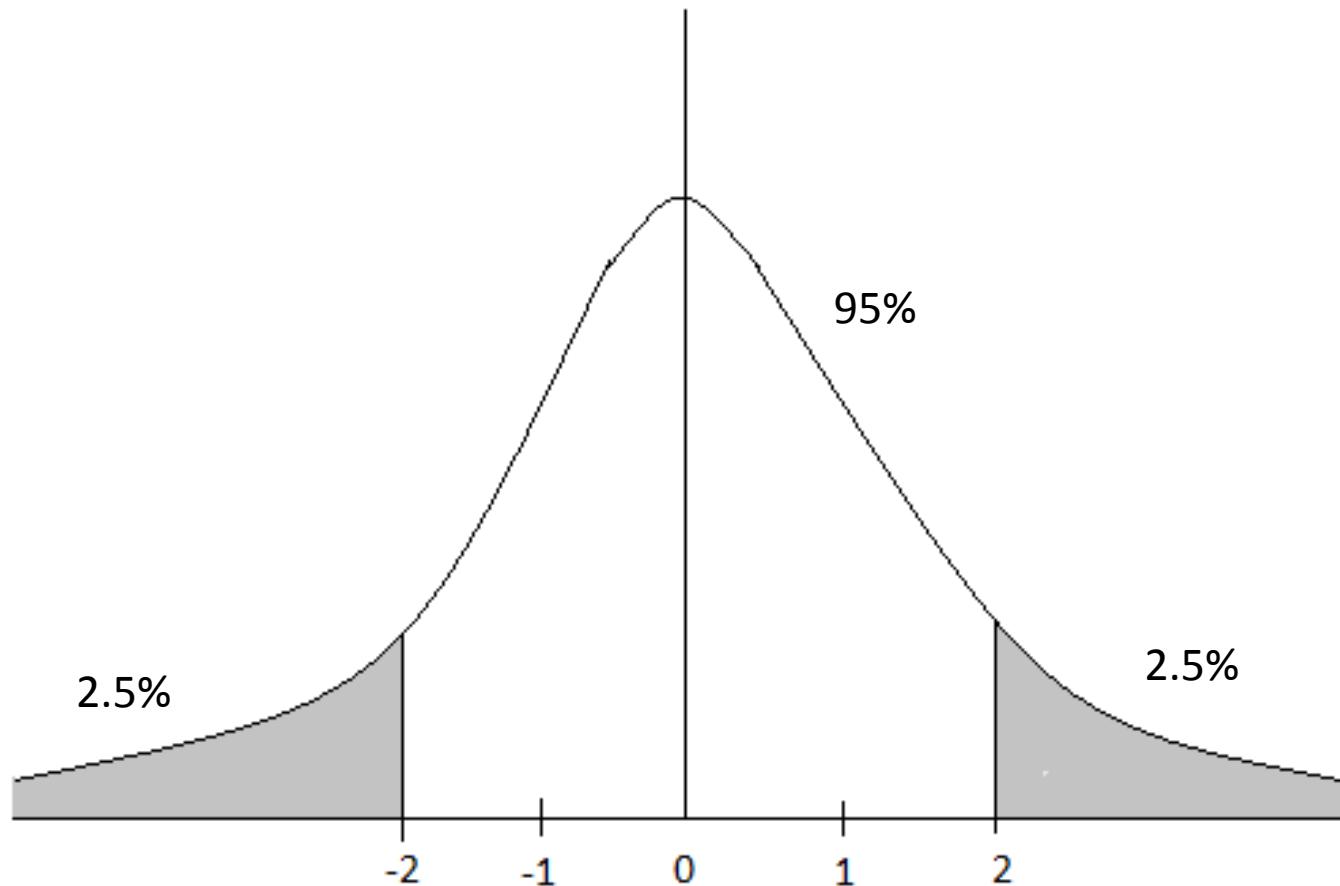
The standard normal distribution has a complicated density function that is difficult to integrate.

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

Fortunately, we will be able to find all of the integral values we need for the standard normal distribution using a **t-table**.

Standard normal distribution

If we want to reject at the 95% confidence level, then we want there to be a 2.5% of having a t-statistic this large and a 2.5% chance of there being a test statistic this small.



Standard normal distribution

Many distributions are normal, but they are typically not standard normal (with mean of 0 and standard deviation of 1).

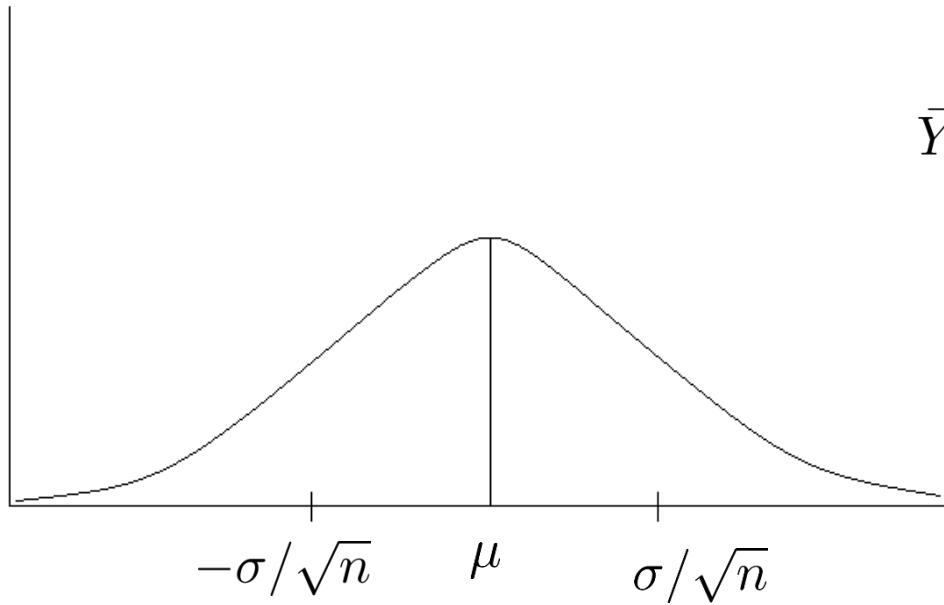
However, it is easy to convert a normal distribution to the standard normal distribution. This is called standardizing

$$Y \sim N(\mu, \sigma)$$

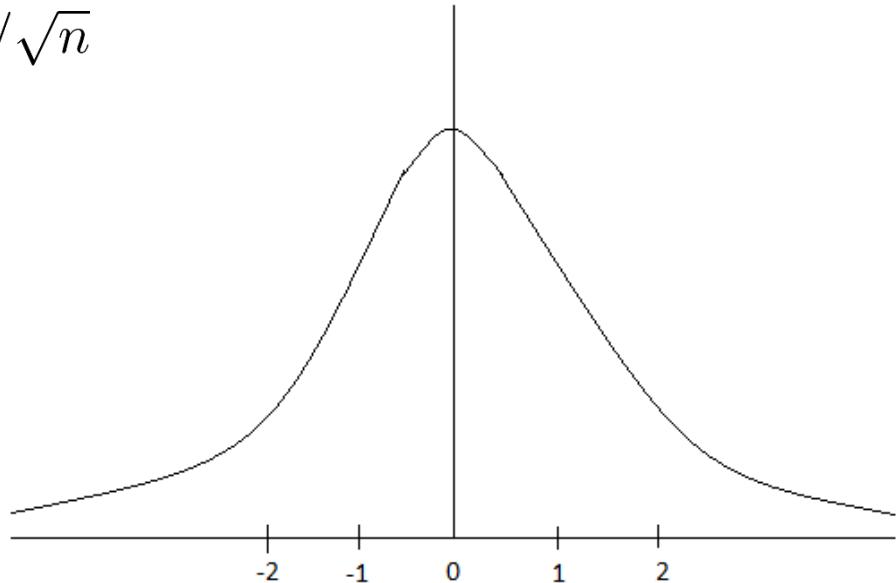
1. Subtract the mean (makes the mean 0)
2. Divide by the standard deviation (makes the std deviation 1)

$$\frac{\bar{Y} - \mu}{\sqrt{\sigma^2/n}} \sim N(0, 1)$$

Standard normal distribution



$$\bar{Y} \sim N(\mu, \sigma/\sqrt{n})$$



$$\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Student's t-distribution

We do not know the population variance σ^2 , so we need to estimate it.

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$$

The Student's t-distribution is standard normal. Thus, when we calculate the t statistic, we know the probability of getting a t-stat of this size.

$$t = \frac{\bar{Y} - c}{\sqrt{s^2/n}}$$

In our example, the mean age at death for men is very far from 70, resulting in a large t-statistic (which occurs with very low probability).

$$t = \frac{75.1 - 70}{\sqrt{s^2/15}} = 14.2$$

Student's t-distribution

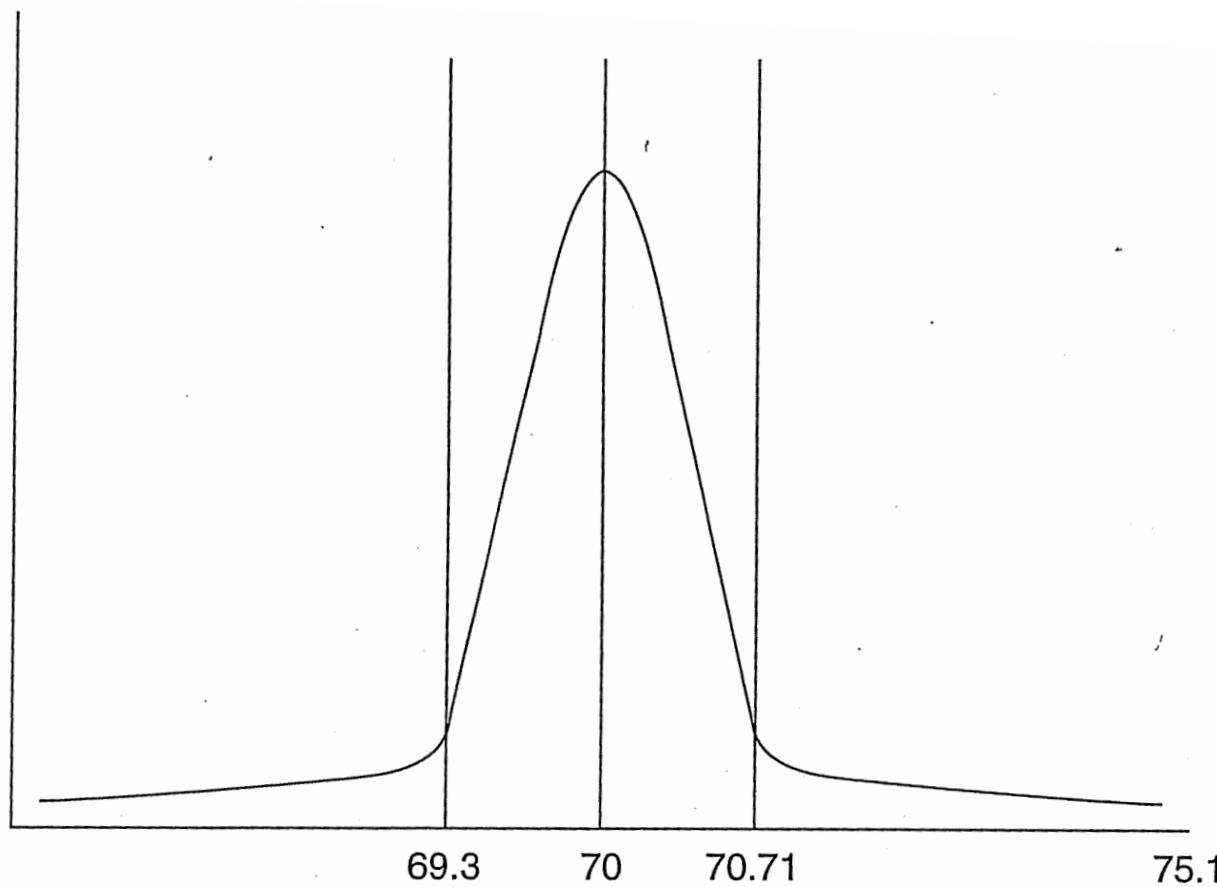


Figure 1.2 A normal distribution for life expectancy around the null

Student's t-distribution

Null Hypothesis:

$$H_0: \bar{x} = c$$

Steps:

1. Compute the "t-statistic" – the number of standard deviations from c.

$$\frac{\bar{x} - c}{\hat{\sigma}_{\bar{x}}} \sim t_{n-1}$$

2. Look at the "t-table" to see what the probability is of being this many standard deviations away from c when there are n-1 degrees of freedom.
3. If the probability is very small of being this far away, then reject the null.

[Important: use the standard dev of the mean, not the standard dev of X]

Student's t-distribution

We are specifically interested in testing hypotheses about:

1. Sample means
2. Regression coefficients (\hat{B}_1)

Luckily for us, sample means and regression coefficients both have normal distributions:

1. Sample means, \bar{x} , are normally distributed around the true population mean, μ_x , according to the Central Limit Theorem.
2. Regression coefficients, \hat{B}_1 , are normally distributed around their true population values, B_1 , if the regression errors are normally distributed.

Confidence Intervals

p-value: The actual probability of observing a t-statistic.

confidence interval: the range around the estimate for which we are confident that the true population value falls, for a given confidence level.

For example, for a 95% confidence interval:

$$CI_{95\%} = \{\bar{Y} - 1.96 * s.e., \bar{Y} + 1.96 * s.e.\}$$

In our example of men's age at death:

$$CI_{95\%} = \{\bar{Y} - 0.71, \bar{Y} + 0.71\} = \{74.39, 75.81\}$$

Confidence Intervals

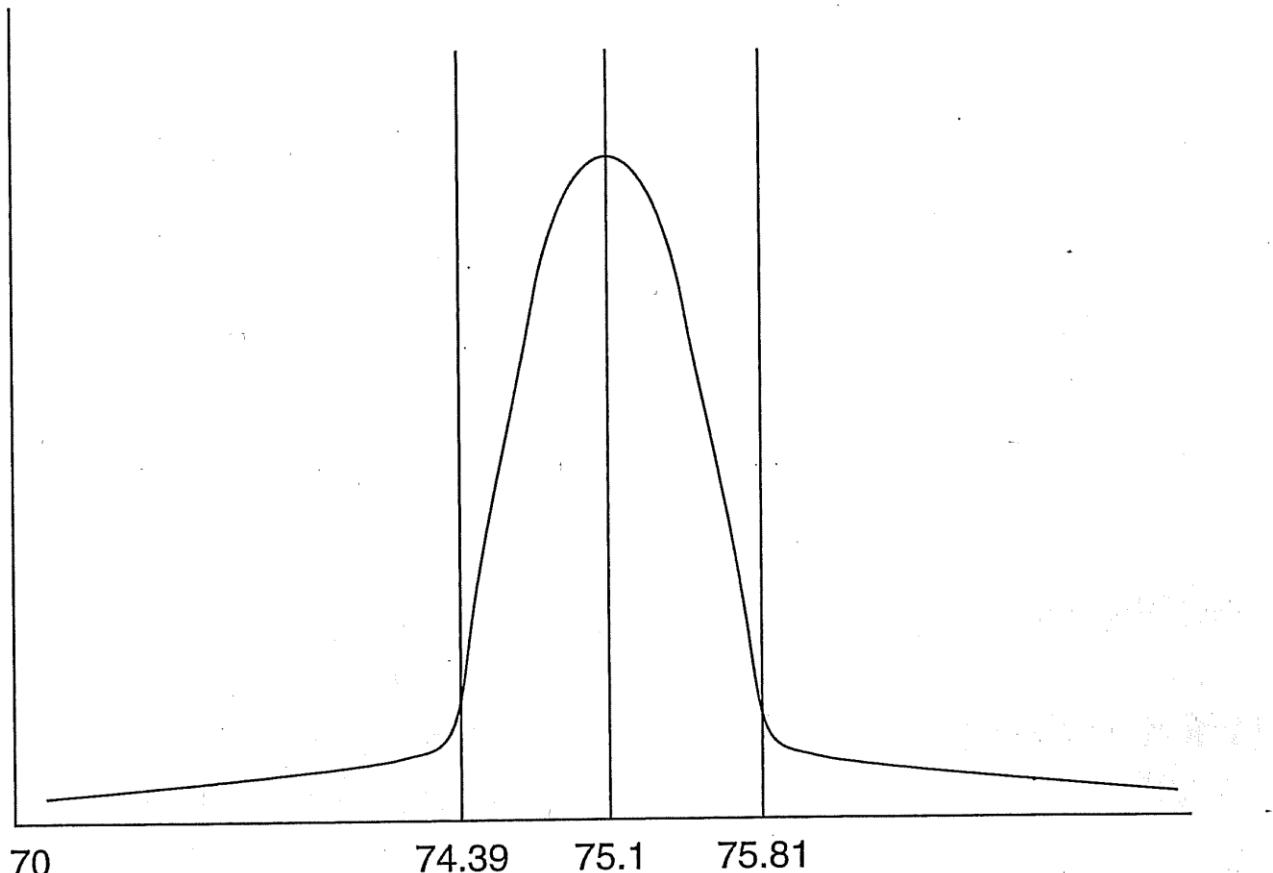


Figure 1.3 A 95% confidence interval around the estimated mean

Lesson 20

Regression Kink

Outline

Previous Lesson

1. Regression Discontinuity

This Lesson

1. Regression Kink
2. Summary of course

Identification

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

□ Endogeneity (sources of bias): 3 causes

1. Omitted variables

Any X_3 that is correlated with the variable of interest and Y

Note that selection bias is a frequent cause

2. Reverse causality

When the variable of interest causes Y

3. Measurement error

When X is measured with error.

Identification

Empirical Methods we have seen:

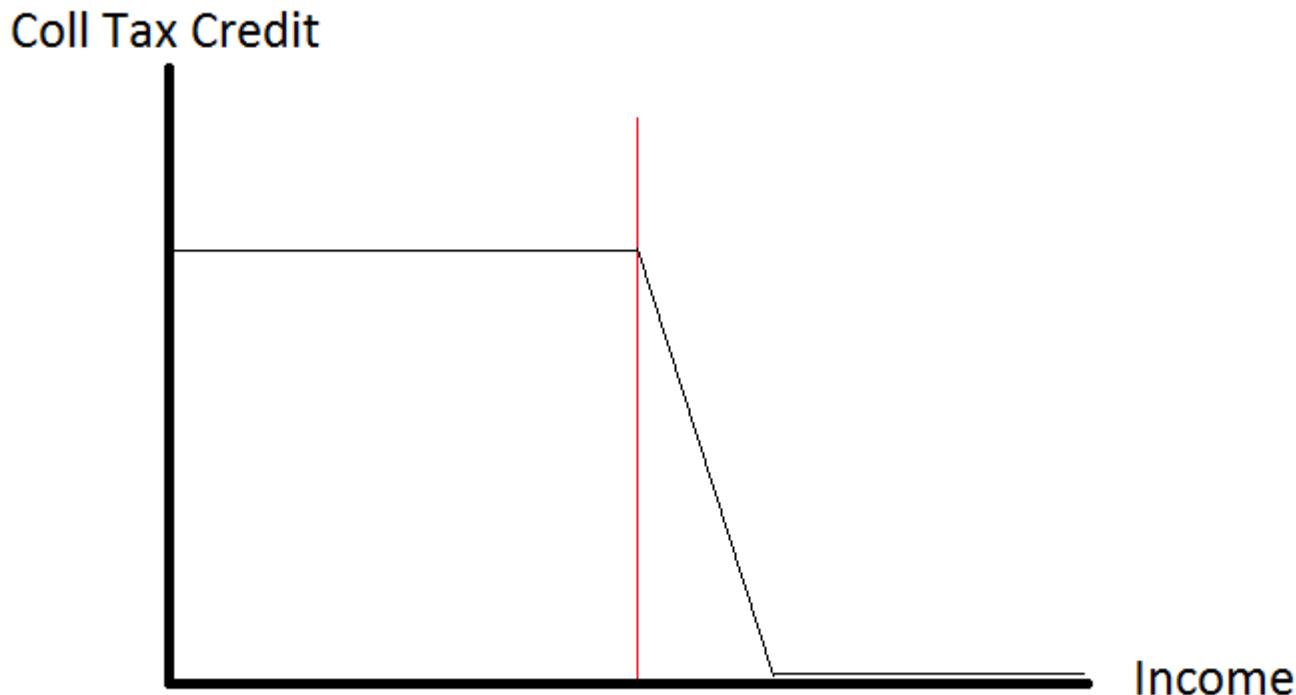
1. Experiment (diff-in-diff)
2. Natural experiment (diff-in-diff)
 - Choose control using matching
3. Instrumental variables
4. Regression Discontinuity
5. Today: Regression kink

Regression Kink

How regression kink works:

There is a sharp change in the slope of the treatment. This is almost always caused by a policy formula (e.g. tax rate, benefit formula...)

For example, eligibility for college tax credits phases out at a specific household income level.



Regression Kink

Consider a tax credit that covers 100% of tuition expenses up to \$5,000. Suppose eligibility for the tax credit phases out for household income ranges from \$80,000 to \$100,000. This means that for each \$1,000 of income, their maximum tax credit decreases by \$250.

Consider a household that makes \$79,000. What is the cost of attending a university that charges \$5,000 per year?

Consider a household that makes \$80,000. What is the cost of attending a university that charges \$5,000 per year?

Consider a household that makes \$81,000. What is the cost of attending a university that charges \$5,000 per year?

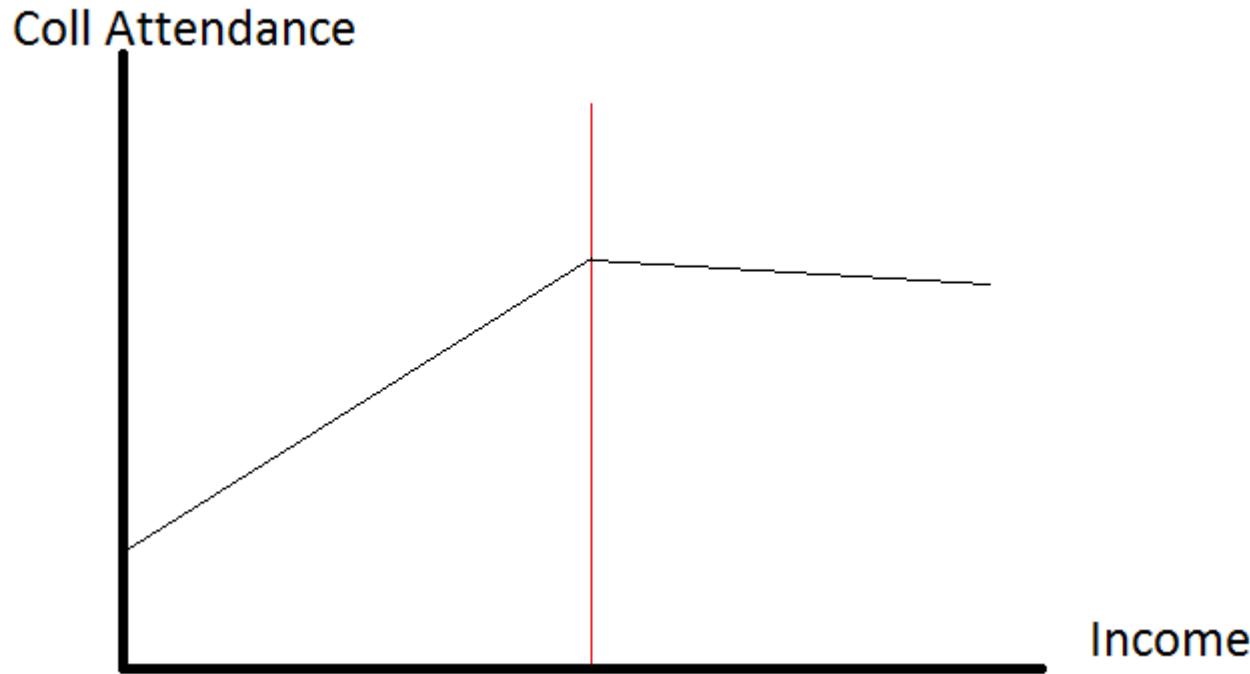
So, the cost of college increases starting at \$80,000.

Regression Kink

We now examine whether there is a corresponding change in the slope of the outcome of interest at the same spot.

For example, we look at the change in the slope of college attendance.

Specifically, we are looking to see if the increase in cost slows the rate of increase in college enrollment relative to household income.



Regression Kink

The math of regression kink:

1. Compute the change in the slope of the treatment variable at the kink:

Formula below kink: $\frac{\partial TaxCredit}{\partial Income} = 0$

Formula above kink: $\frac{\partial TaxCredit}{\partial Income} = -0.25$

So, the change in slope if -0.25

This can also be estimated using a regression (on those spending $\geq 5k$):

$$taxcredit_i = \beta_0 + \beta_1 * Income_i + \beta_2 * Income_i * D_{Inc>80k} + \epsilon_i$$

What is β_1 and what is the expected value?

What is β_2 and what is the expected value?

Regression Kink

The math of regression kink:

2. Compute the change in the slope of the outcome variable at the kink.
This must be done with a regression:

$$attend_i = \alpha_0 + \alpha_1 * Income_i + \alpha_2 * Income_i * D_{Inc>80k} + \eta_i$$

If all we want to know is whether tax credits affect college attendance, then we can just examine whether α_2 is significant.

What sign do we expect this coefficient to have?

Regression Kink

However, suppose that we are interested in the following:

$$attend_i = \gamma_0 + \gamma_1 * taxcredit_i + \mu_i$$

Then we need to scale the change in attendance that occurs at the kink by the change in tax credits that occurs at the kink (from the formula or from the regression):

$$\gamma_1 = \frac{\alpha_2}{\beta_2}$$

So, in our example, $\beta_2 = -0.25$ and we might get something like $\alpha_2 = -0.0001$.

This means that $\gamma_1 = 0.0004$. That is, each \$1 of tax credit increases college attendance by .04 percent. So, \$100 dollars of tax credit increases enrollment by 4 percent.

Formally:

$$\tau = \frac{\lim_{v \rightarrow 0^+} \frac{\partial E[Y|V=v]}{\partial v} - \lim_{v \rightarrow 0^-} \frac{\partial E[Y|V=v]}{\partial v}}{\lim_{v \rightarrow 0^+} \frac{\partial b(v)}{\partial v} - \lim_{v \rightarrow 0^-} \frac{\partial b(v)}{\partial v}}.$$

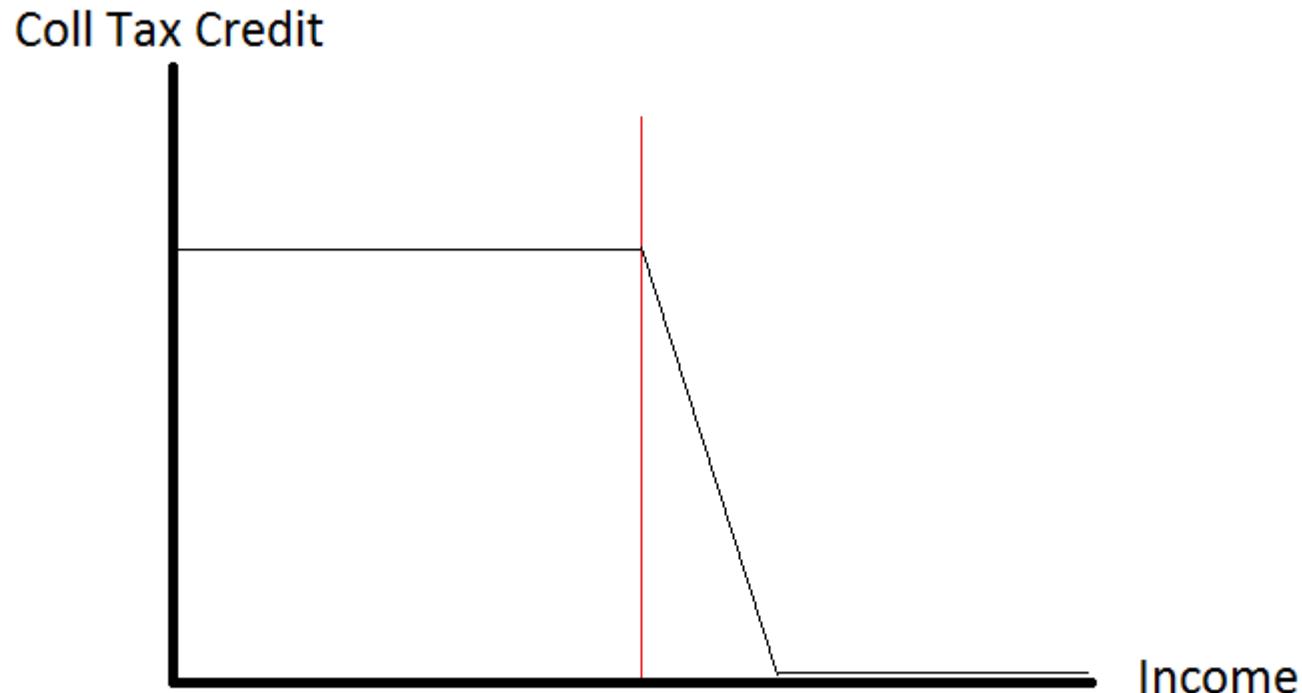
Regression Kink

The same issues and checks as RD:

- Show results for different bandwidths.
- Show results for different polynomials.
[Even though the first order coefficient is the only one we care about.]
- Make sure there is no change in slope of observables at the kink
- Make sure there is no change in density of observations around kink

Regression Kink

Have we exploited everything we can in this context?



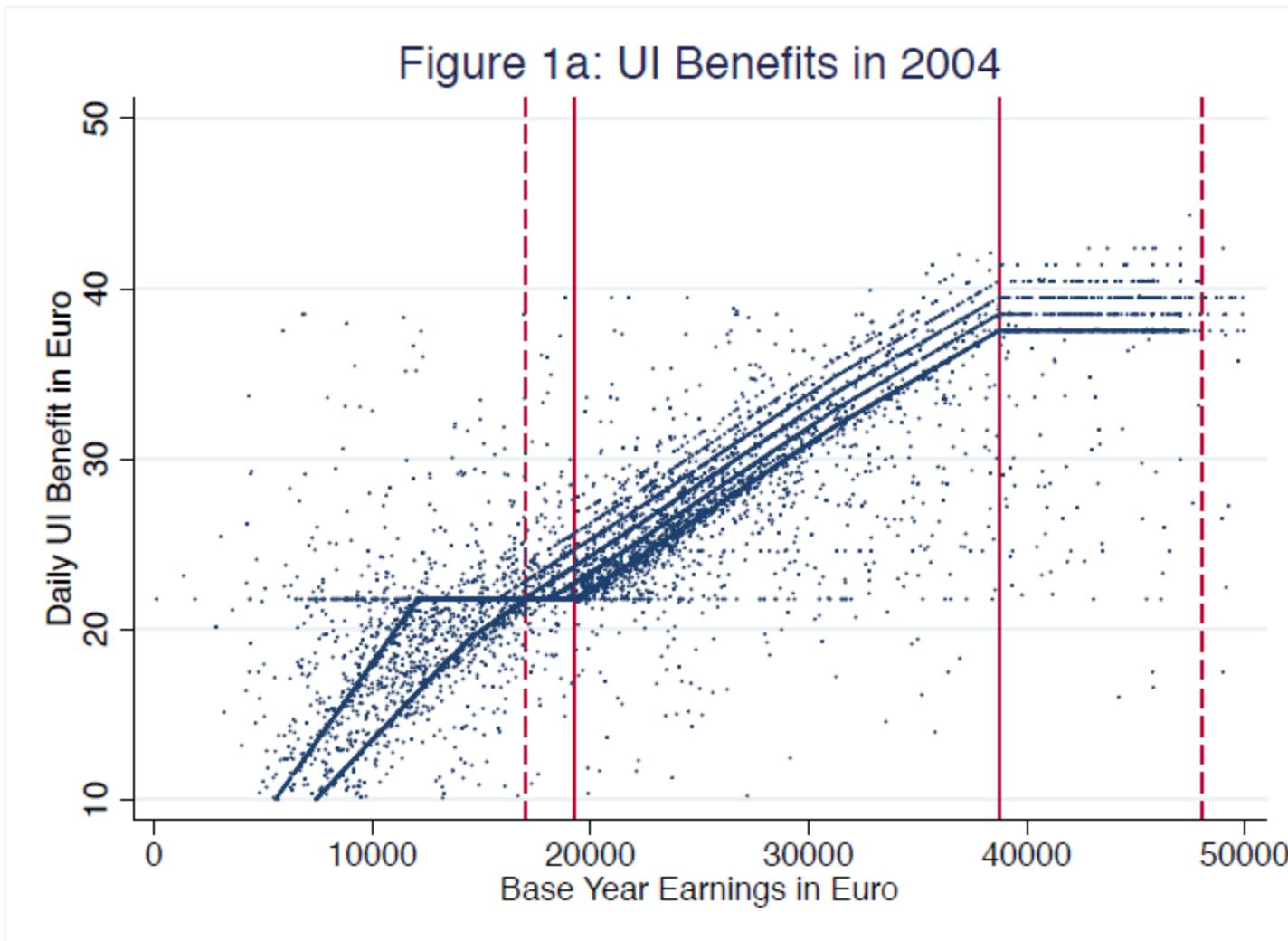
Regression Kink

- Unemployment insurance:
 - Paid by employers for every worker.
 - Workers are compensated when they get laid off.
 - Benefits based on a formula that uses prior income.
 - Typically a fraction of earnings up to a max level.
- Primary concern:
 - People won't look for a job if they get benefits
- What we want to estimate:

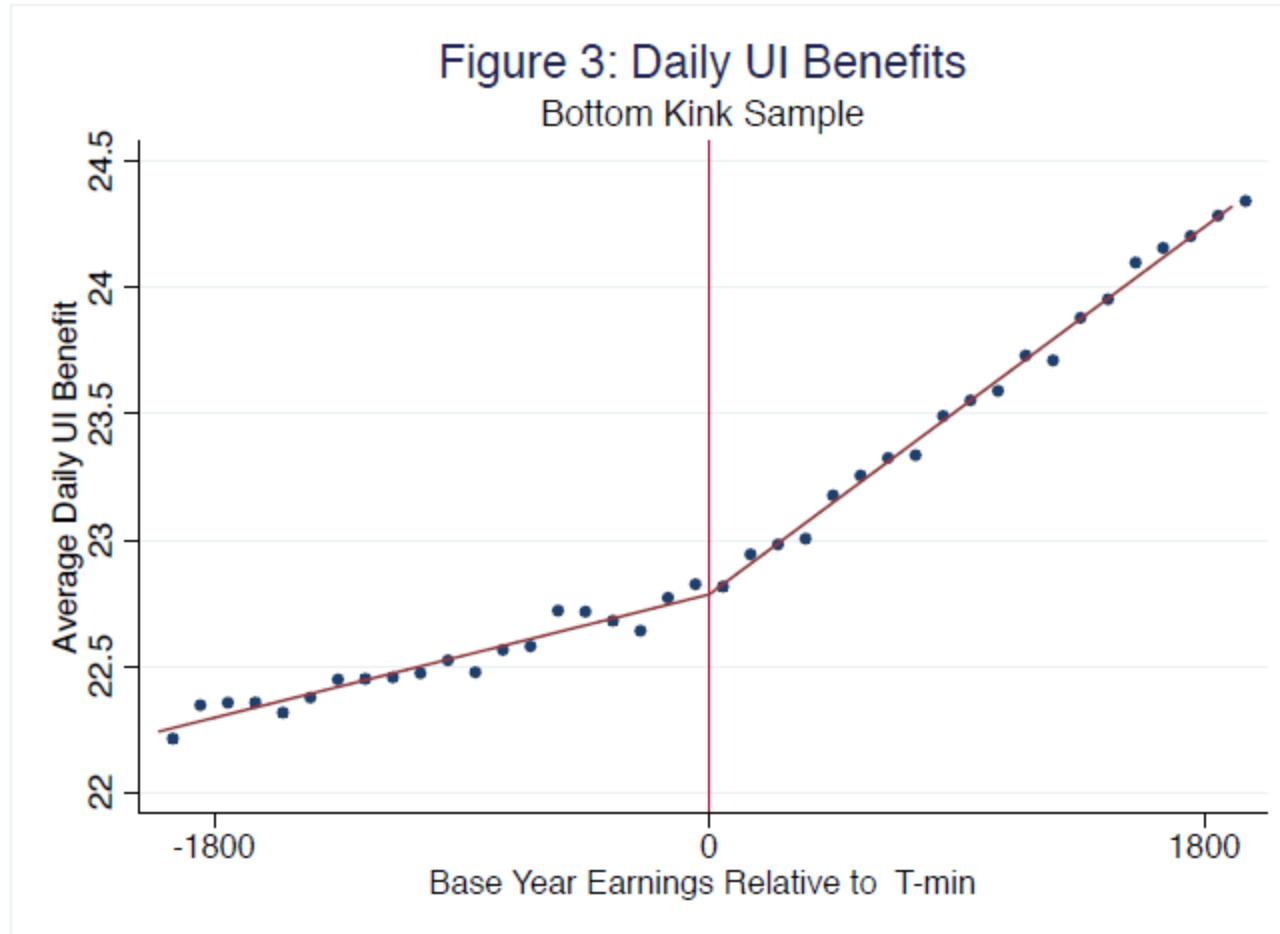
$$\ln(\text{weeksjobless}_i) = \beta_0 + \beta_1 \ln(\text{UIbenefit}_i) + \epsilon_i$$

So, the goal is basically to estimate the moral hazard of unemployment insurance benefits.

Regression Kink: Example



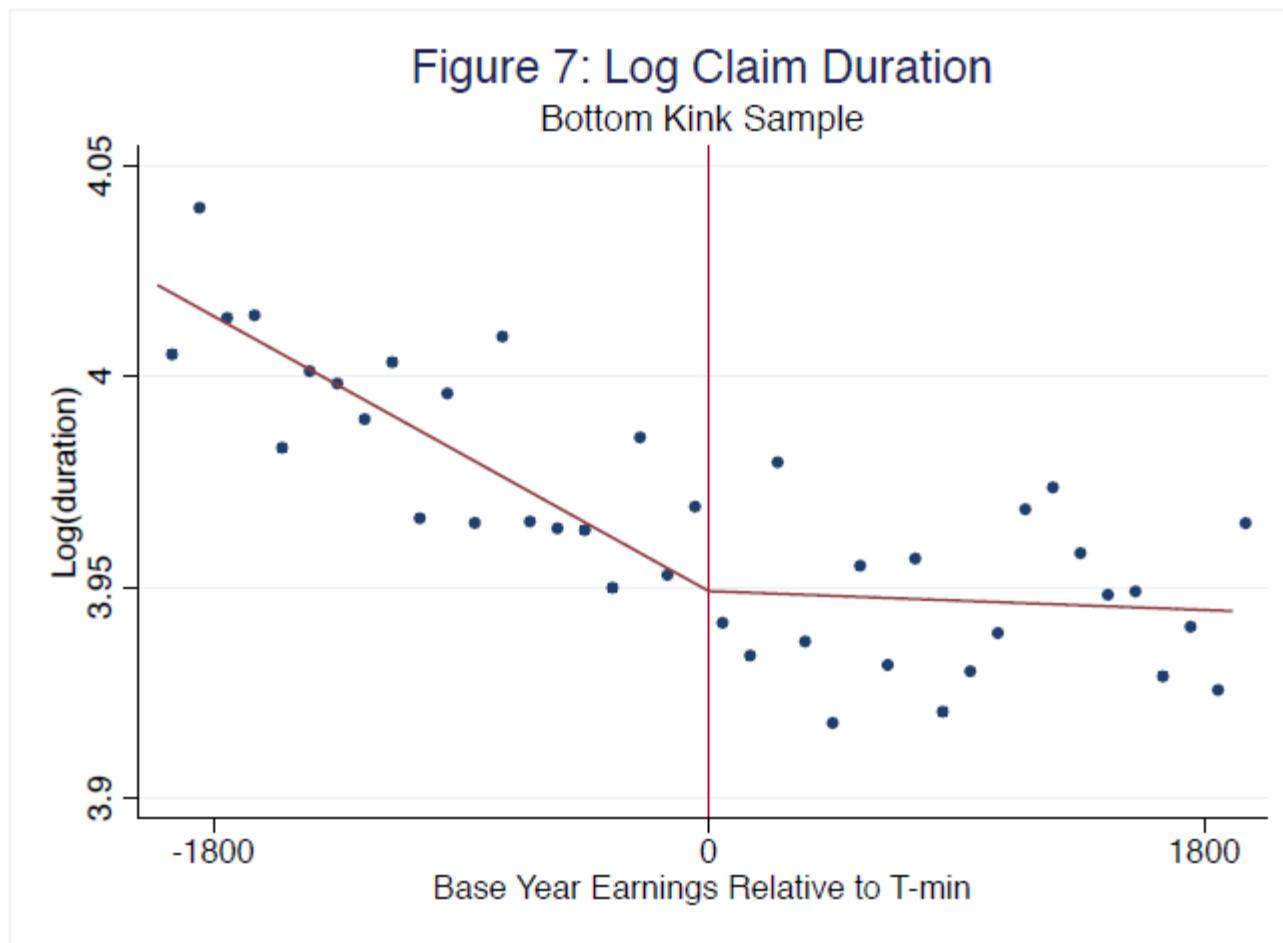
Regression Kink: Example



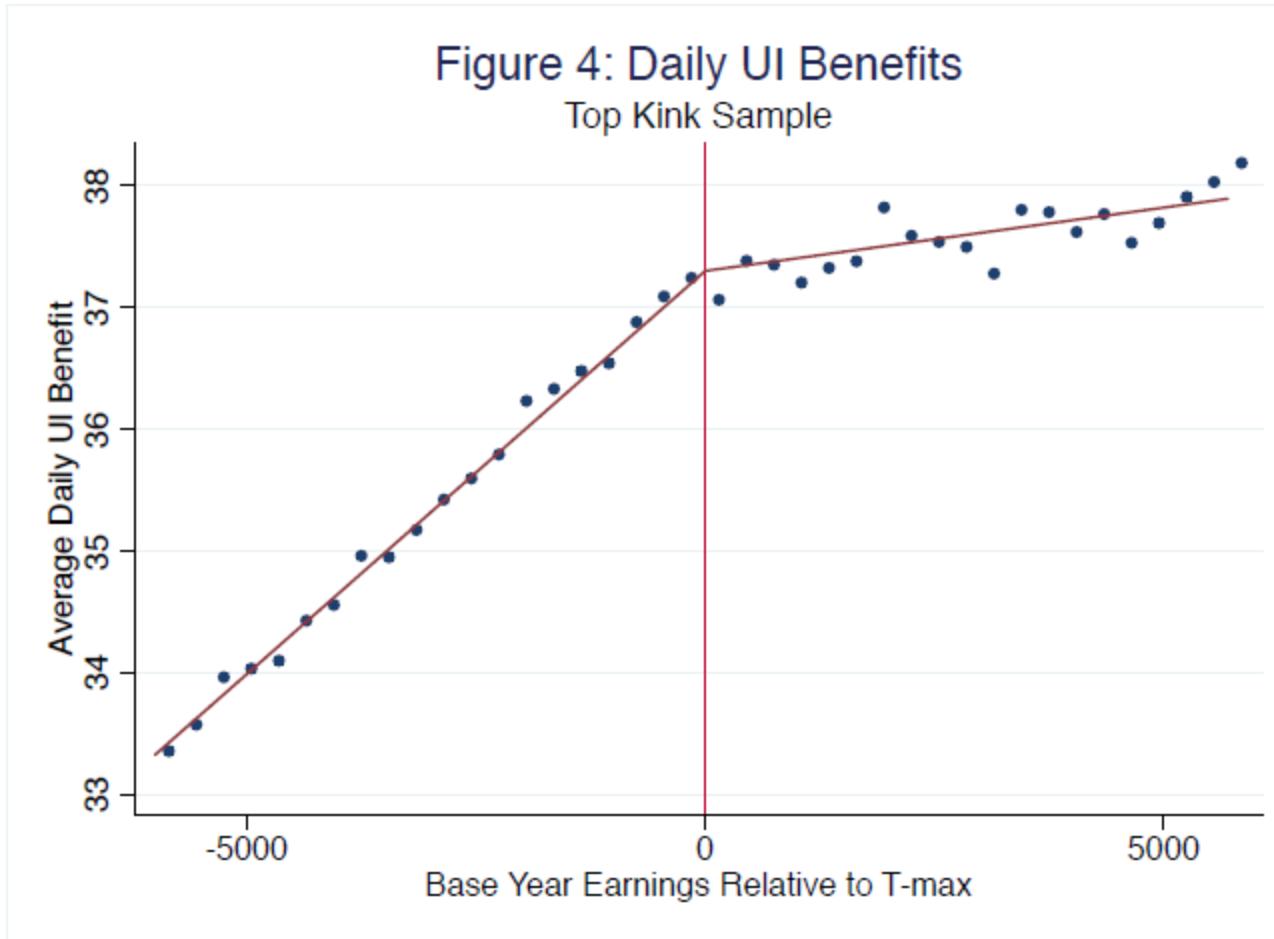
Regression Kink: Example



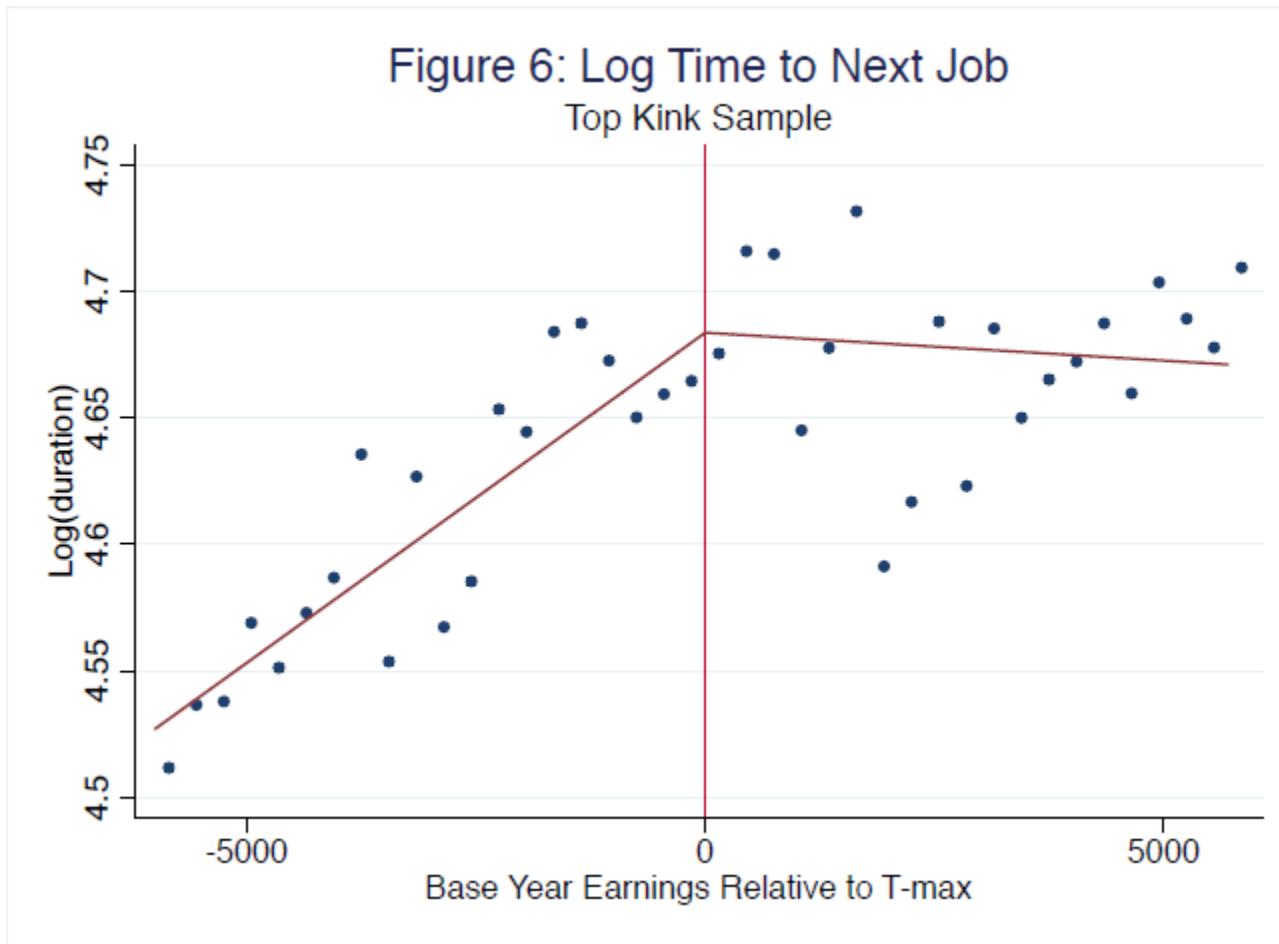
Regression Kink: Example



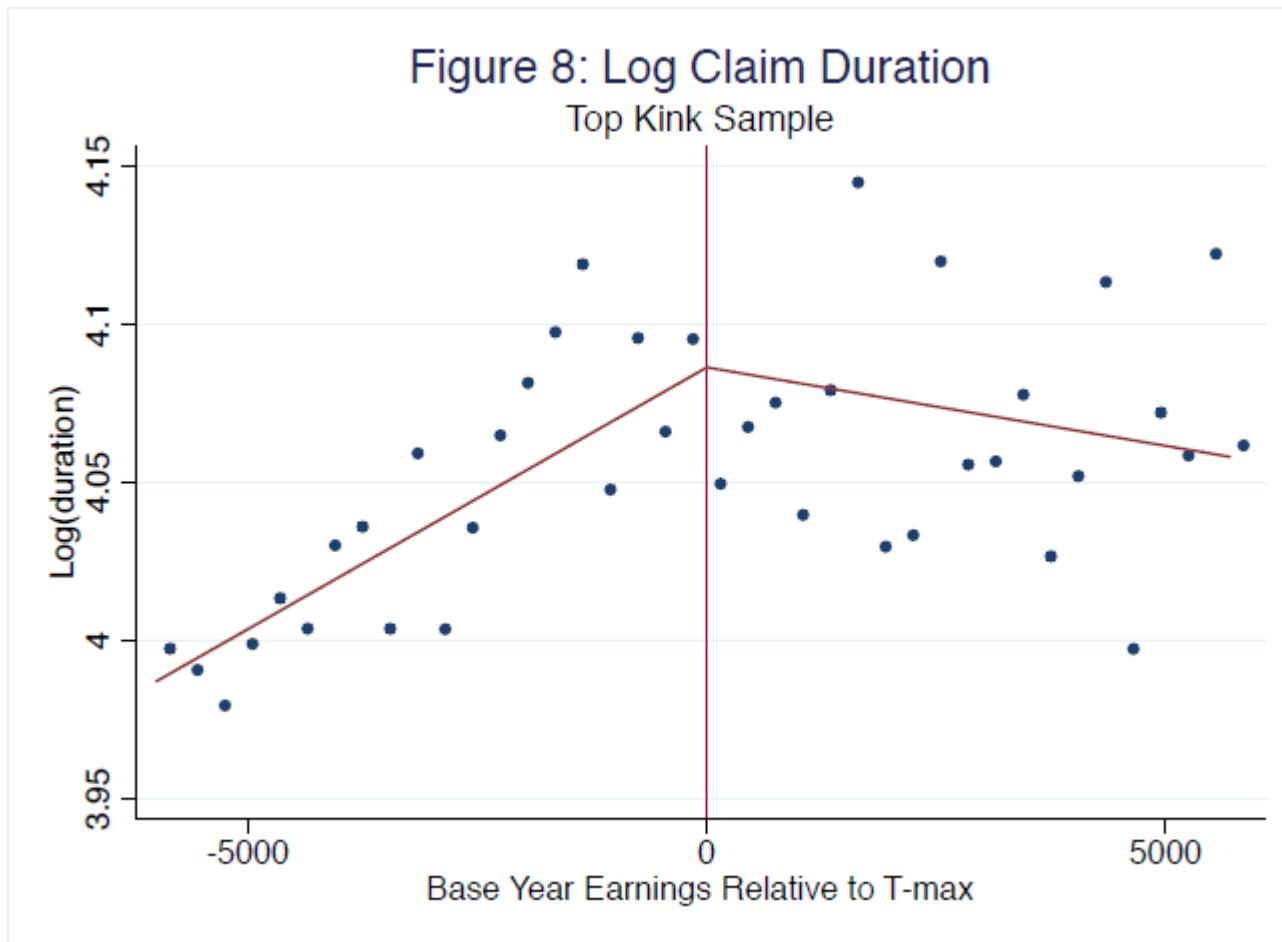
Regression Kink: Example



Regression Kink: Example



Regression Kink: Example



Regression Kink: Example

Table 2: Reduced Form Estimates of Kink Effects in Benefits and Durations

	Local Linear Models		Local Quadratic Models	
	FG Bandwidth (1)	Estimated Kink (2)	FG Bandwidth (3)	Estimated Kink (4)
<u>A. Bottom Kink:</u>				
Log daily UI benefit	2,133	0.0222 (0.0010)	4,564	0.0192 (0.0025)
Log time to next job	2,615	0.0375 (0.0093)	4,328	0.0598 (0.0280)
Log claim duration	2,651	0.0269 (0.0085)	4,564	0.0541 (0.0254)

Regression Kink: Example

Table 2: Reduced Form Estimates of Kink Effects in Benefits and Durations

	Local Linear Models		Local Quadratic Models	
	FG Bandwidth (1)	Estimated Kink (2)	FG Bandwidth (3)	Estimated Kink (4)
B. Top Kink:				
Log daily UI benefit	7,064	-0.0154 (0.0006)	6,577	-0.0166 (0.0027)
Log time to next job	4,148	-0.0396 (0.0100)	7,521	-0.0577 (0.0191)
Log claim duration	9,067	-0.0221 (0.0038)	9,355	-0.0363 (0.0151)

Regression Kink: Example

Table 4: Estimated Structural Coefficients from Fuzzy Regression Kink Design

	Local Linear Models		Local Quadratic Models	
	FG Bandwidth (1)	Estimated Elasticity (2)	FGI Bandwidth (3)	Estimated Elasticity (4)
<u>A. Bottom Kink:</u>				
Log time to next job	2,615	1.726 (0.440)	4,328	3.024 (1.501)
Log claim duration	2,651	1.250 (0.406)	4,564	2.816 (1.401)
<u>B. Top Kink:</u>				
Log time to next job	4,148	2.643 (0.715)	7,521	3.497 (1.278)
Log claim duration	9,067	1.312 (0.228)	9,355	2.500 (1.103)

Figure 2a: Density in Bottom Kink Sample

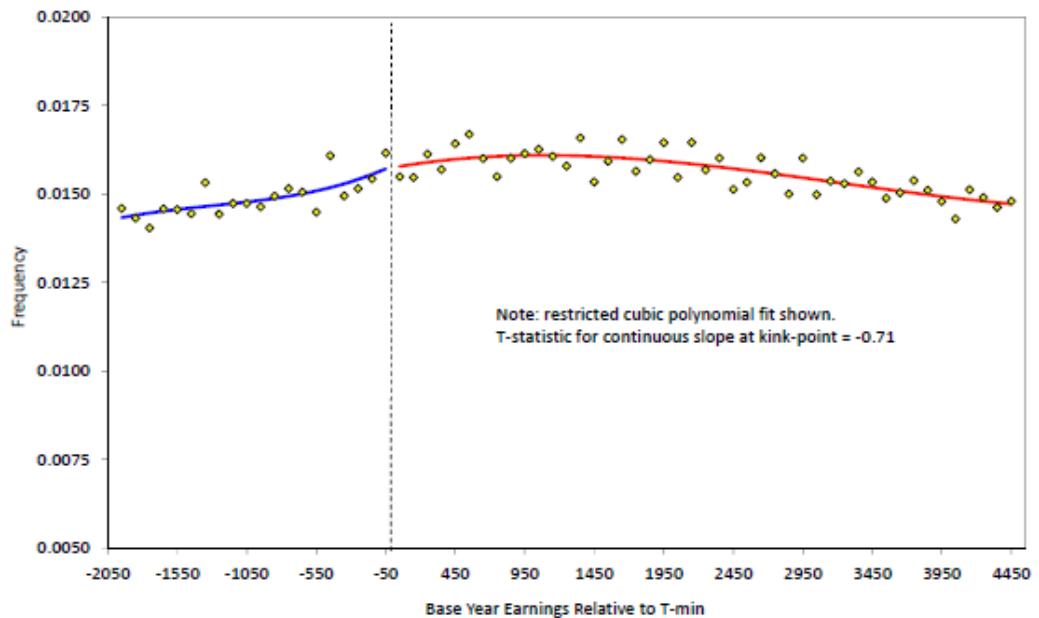
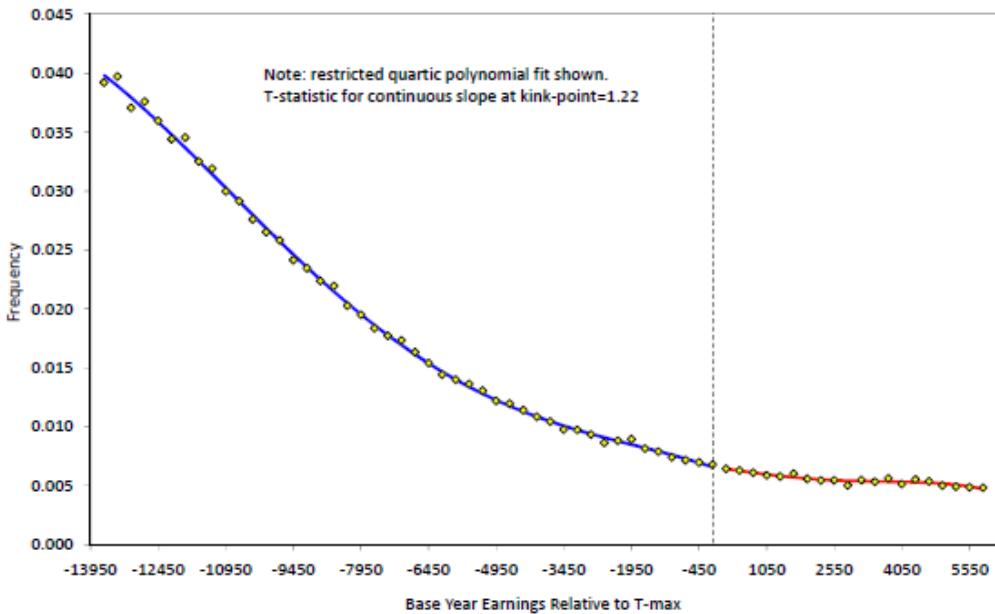


Figure 2b: Density in Top Kink Sample



Permutation Test

- Problem:
 - RK is highly sensitive to curvature.
 - Curvature can make there appear to be a change in slope at the kink when there is not
 - One solution would be to correctly account for that curvature
- Proposed solution:
 - Perform placebo tests in the vicinity where there is no kink
 - Determine the range of estimates
 - Use these for inference
- For example, run 100 placebo tests. If result at true kink is in largest or smallest 5, then 90 percent significant.

Mean Reversion

In statistics, regression toward the mean is the phenomenon that if a variable is extreme on its first measurement, it will tend to be closer to the average on its second measurement. Why is this an issue?

Example 1

Consider a case in which low performing students on an exam are given a treatment such as tutoring.

The students are then retested to estimate the effect of the tutoring.

Example 2

Surgeons are evaluated on how many patients they have who die. An organization decides to grade surgeons and an economist wishes to test if surgeons improve in response.

The surgeons who receive and “F” are evaluated the next year.

How do we fix this issue?

Survivorship Bias

Survivorship bias, or survival bias, is the logical error of concentrating on the people or things that "survived" some process and inadvertently overlooking those that did not because of their lack of visibility.

Example

Suppose you wish to evaluate the performance of actively managed mutual funds against the S&P 500 index. You collect data on 100 funds and look at their 10 year returns.

How do we fix this issue?

Lesson 3

Structure of Data

Outline

Previous Lesson:

1. Samples
2. Properties of estimators
3. Hypothesis testing

This Lesson

1. Types of data
2. Indices and base dates
3. Data transformations

The text goes into more detail about graphing. I will leave this to you to read and for the lab class.

Next Lesson:

1. Ordinary least squares regression

Cross-sectional data

Cross-sectional data: This is data for many individuals (e.g. people, firms, countries) at a single point in time. Typically there will be one observation (row) per individual.

$$Y_i \quad \text{for } i = 1, 2, 3, \dots, N$$

Time-series data: This is data for a single individual (e.g. person, firm, country) over time (e.g. days, months, years). Typically there will be several rows for one individual.

$$Y_t \quad \text{for } t = 1, 2, 3, \dots, T$$

Panel data: This is data for many individuals (e.g. people, firms, companies) over time (e.g. days, months, years). Typically there will be multiple rows for each individual.

$$Y_{it} \quad \text{for } i = 1, 2, 3, \dots, N \text{ and } t = 1, 2, 3, \dots, T$$

[balanced panel: same t's for every individual]

Cross-sectional data

name	year	wage	educ	tenure
Amanda	2003	3.10	11	0
Jacob	2003	3.20	12	2
Nicole	2003	3.00	11	0
William	2003	6.00	8	28
Jennifer	2003	5.30	12	2
Courtney	2003	8.80	16	8
Olivia	2003	11.00	18	7
Patrick	2003	5.00	12	3
Carlos	2003	3.60	12	4
Angelica	2003	18.00	17	21

What do we observe in the data about wages, education and tenure?

Are some observations more informative than others?

Time-series data

name	year	wage	educ	tenure
Amanda	1998	4.50	12	2
Amanda	1999	5.00	12	3
Amanda	2000	5.00	13	4
Amanda	2001	5.50	14	5
Amanda	2002	6.00	15	6
Amanda	2003	12.00	16	0
Amanda	2004	13.00	16	1
Amanda	2005	14.00	16	2
Amanda	2006	14.00	16	3
Amanda	2007	12.00	16	0

What do we observe in the data about wages, education and tenure?

Panel data

name	year	wage	educ	tenure
Amanda	2000	5.00	13	4
Amanda	2001	5.50	14	5
Amanda	2002	6.00	15	6
Amanda	2003	12.00	16	0
Amanda	2004	13.00	16	1
Jacob	2000	8.00	12	4
Jacob	2001	8.30	12	5
Jacob	2002	8.50	12	6
Jacob	2003	9.00	12	7
Jacob	2004	9.00	12	8

What do we observe in the data about wages, education and tenure?

Types of Data

Cross-sectional data:

- wages and education across people
- crime rate and number of police across cities
- GDP levels across countries

Time-series data:

- wages and education for one individual over time
- crime rate and number of police for one city over time
- GDP levels for one country over time

Panel data:

- wages and education for multiple people over time
- crime rate and number of police for multiple cities over time
- GDP levels for multiple countries over time

Structure of Data

Time-series and cross-sectional data can be represented as T x 1 and N x 1 vectors respectively:

$$Y_t^{\text{ARGENTINA}} = \begin{pmatrix} Y_{1990} \\ Y_{1991} \\ Y_{1992} \\ \vdots \\ Y_{2012} \end{pmatrix}; \quad Y_i^{1990} = \begin{pmatrix} Y_{\text{ARGENTINA}} \\ Y_{\text{BRAZIL}} \\ Y_{\text{URUGUAY}} \\ \vdots \\ Y_{\text{VENEZUELA}} \end{pmatrix}$$

And panel data can be represented as a T x N matrix:

$$Y_{it} = \begin{pmatrix} Y_{\text{ARG}, 1990} & Y_{\text{BRA}, 1990} & \dots & Y_{\text{VEN}, 1990} \\ Y_{\text{ARG}, 1991} & Y_{\text{BRA}, 1991} & \dots & Y_{\text{VEN}, 1991} \\ \vdots & \vdots & & \vdots \\ Y_{\text{ARG}, 2012} & Y_{\text{BRA}, 2012} & \dots & Y_{\text{VEN}, 2012} \end{pmatrix}$$

Transforming data: indices

Indices are a very common type of data. However, they will sometimes have different base years.

In the example below, the standardized price index use 1990 as a base year and use the ratio during the overlapping year to convert the 1985 base year index.

Year	<i>Price index (1985 base year)</i>	<i>Price index (1990 base year)</i>	<i>Standardized price index (1990 base year)</i>
1985	100		45.9
1986	132		60.6
1987	196		89.9
1988	213		97.7
1989	258		118.3
1990	218	100	100
1991		85	85
1992		62	62

Transforming data: natural logs

Transforming data to natural logs results in data that changes in percents (not to be confused with percentage points):

$$\frac{d(\ln x)}{dx} = \frac{1}{x}$$

And thus:

$$d(\ln x) = \frac{dx}{x} = \text{percent change in } x$$

It is very likely that things in the real world change in percents rather than in absolute amounts:

- country GDP
- house prices
- return to a year of education

Transforming data: natural logs

Natural logs can also transform otherwise intractable functional forms:

$$Y = AL^\alpha K^\beta e^u$$

$$\ln(Y) = \ln(A) + \alpha \ln(L) + \beta \ln(K) + u$$

The result is linear and can be used easily in regression analysis (next lesson).

Transforming data: differencing

We can remove a linear time trend by taking the first difference:

$$\Delta Y_t = Y_t - Y_{t-1}$$

This would turn a linear time trend into a constant.

If there is still a time trend, we could take a second difference:

$$\Delta^2 Y_t = \Delta(Y_t - Y_{t-1}) = \Delta Y_t - \Delta Y_{t-1} = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2})$$

This differences out the change in the slope of the time trend.

Transforming data: growth rates

Growth rate (option 1):

$$\text{growth rate of } Y_t = \frac{Y_t - Y_{t-1}}{Y_{t-1}}$$

Growth rate (option 2):

$$\text{growth rate of } Y_t = \ln\left(\frac{Y_t}{Y_{t-1}}\right) = \ln(Y_t) - \ln(Y_{t-1})$$

Lesson 4

Ordinary Least Squares: Univariate

Outline

Previous Lesson

1. Types of data
2. Data transformations

This Lesson:

1. Deriving ordinary least squares (OLS)
2. Assumptions of OLS
3. Properties of OLS
4. Goodness of Fit
5. Interpreting coefficients

Next Lesson:

1. Multivariate regression in matrix form

Regression

Table 3.2 Data for simple regression example

<i>Consumption Y</i>	<i>Disposable income X</i>
72.30	100
91.65	120
135.20	200
94.60	130
163.50	240
100.00	114
86.50	126
142.36	213
120.00	156
112.56	167
132.30	189
149.80	214
115.30	188
132.20	197
149.50	206
100.25	142
79.60	112
90.20	134
116.50	169
126.00	170

Regression

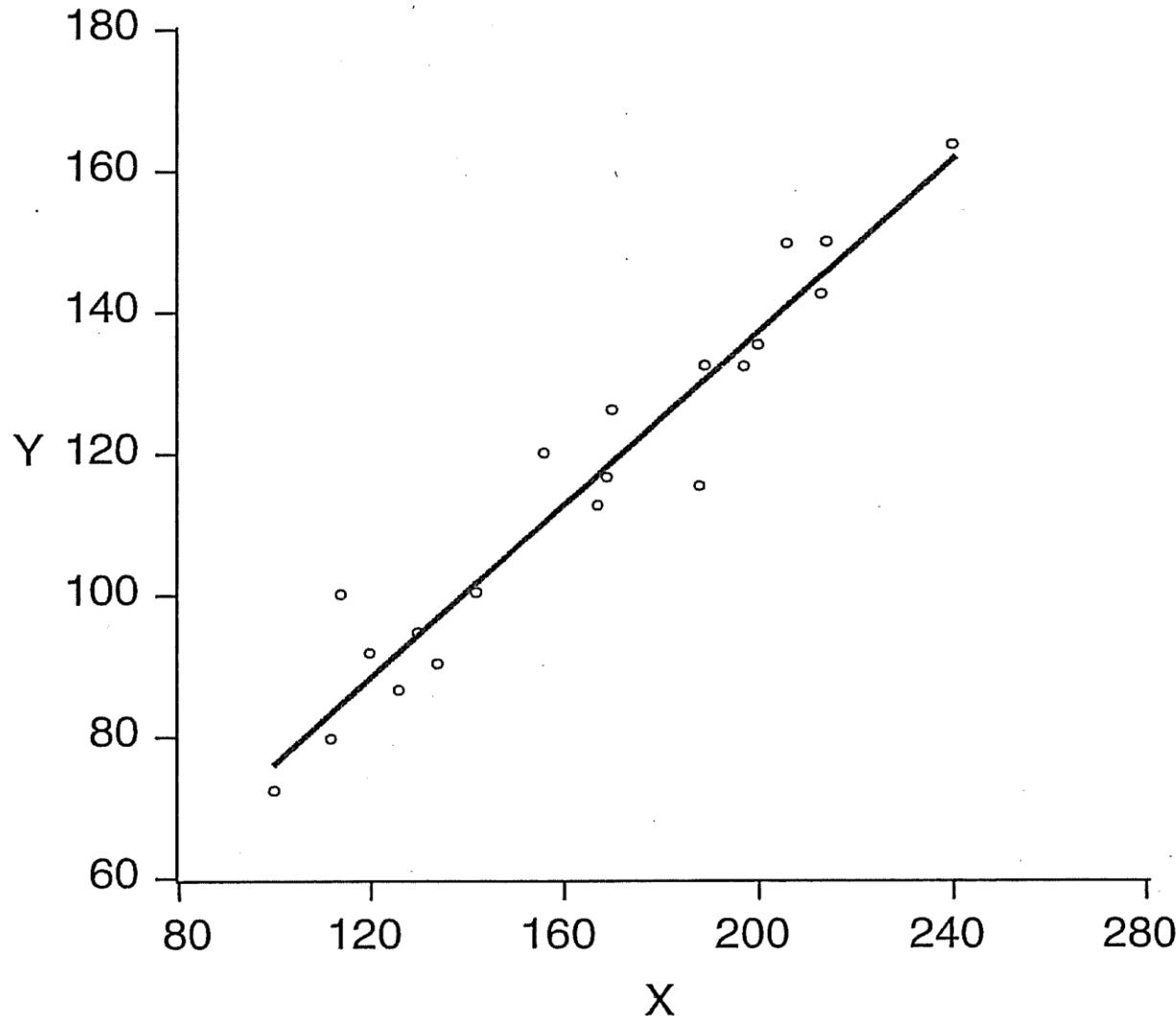


Figure 3.2 Scatter plot

Estimating OLS

Population regression equation:

$$Y_i = a + \beta X_i + u_i$$

Terminology: “The regression of Y on X”

Y = outcome variable

X = explanatory variable

a = intercept

B = coefficient on X (also parameter, slope)

u = error

Note: Generally the subscript “i” is used for cross-sectional data and the subscript “t” is used for time-series data.

Estimating OLS

$$Y_i = a + \beta X_i + u_i$$

Why do we need an error term? [Why doesn't our regression line fit the data perfectly?]

Omitted variables – we failed to include all relevant explanatory variables.

Misspecification – the functional form may not actually be perfectly linear.

Measurement error – we may have an incorrect measure of Y or X.

Aggregate variables – we might not be using individual level measures of X.

Estimating OLS

$$Y_i = a + \beta X_i + u_i$$

We do not observe this population regression equation. We want to estimate a sample regression equation:

$$\hat{Y}_i = \hat{a} + \hat{\beta} X_i$$

We want to estimate the coefficients in such a way that we minimize the distance between the estimated value \hat{Y}_i and the true value Y_i . That is, we want to minimize the errors \hat{u}_i in some way.

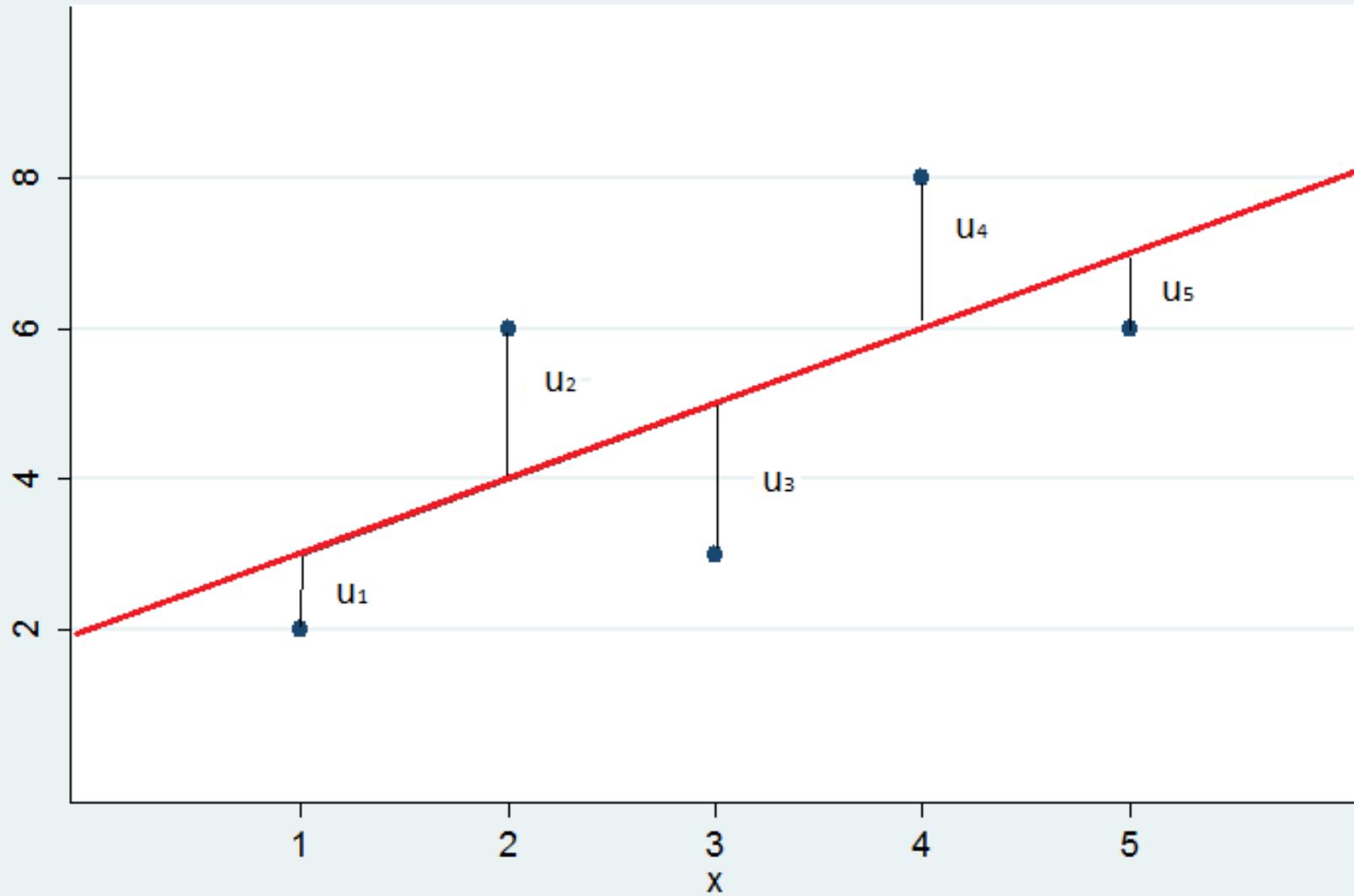
$$\hat{u}_i = Y_i - \hat{Y}_i = Y_i - \hat{a} - \hat{\beta} X_i$$

We do this by minimizing the sum of the squared errors (rather than, for example, minimizing the sum of the absolute values of the errors):

- by squaring we place more weight on outliers
- squaring is mathematically tractable
- other desirable properties

Deriving OLS

Residuals



Deriving OLS

We want to minimize the Sum of the Squared Residuals (RSS):

$$\begin{aligned} RSS &= \hat{u}_1^2 + \hat{u}_2^2 + \dots + \hat{u}_n^2 \\ &= \sum_{i=1}^n \hat{u}_i^2 \\ &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \\ &= \sum_{i=1}^n (Y_i - \hat{a} - \hat{\beta} X_i)^2 \end{aligned}$$

Deriving OLS

$$\text{Min} \sum_{i=1}^n (Y_i - \hat{a} - \hat{\beta} X_i)^2$$

Take first order conditions (set partial derivatives = 0):

$$\frac{\partial RSS}{\partial \hat{a}} = -2 \sum_{i=1}^n (Y_i - \hat{a} - \hat{\beta} X_i) = 0 \quad (1)$$

$$\frac{\partial RSS}{\partial \hat{\beta}} = -2 \sum_{i=1}^n X_i (Y_i - \hat{a} - \hat{\beta} X_i) = 0 \quad (2)$$

We have two equations, (1) and (2), and two unknowns: \hat{a} and $\hat{\beta}$. Note that we used the chain rule for the derivative.

Deriving OLS

If we divide (1) by -2 and then by n, we get:

$$\begin{aligned}(1) \Rightarrow & \sum_{i=1}^n (Y_i - \hat{a} - \hat{\beta}X_i) = 0 \\ \Rightarrow & \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{a} - \hat{\beta}X_i) = 0 \\ \Rightarrow & \bar{Y} - \hat{a} - \hat{\beta}\bar{X} = 0 \\ \Rightarrow & \hat{a} = \bar{Y} - \hat{\beta}\bar{X} \quad (\text{This is half of the solution})\end{aligned}$$

Deriving OLS

If we divide (2) by -2 and then plug in for \hat{a} using (1), we get:

$$\begin{aligned}(2) \Rightarrow & \sum_{i=1}^n X_i(Y_i - \hat{a} - \hat{\beta}X_i) = 0 \\ \Rightarrow & \sum_{i=1}^n X_i(Y_i - (\bar{Y} - \hat{\beta}\bar{X}) - \hat{\beta}X_i) = 0 \\ \Rightarrow & \sum_{i=1}^n X_i((Y_i - \bar{Y}) + \hat{\beta}(\bar{X} - X_i)) = 0 \\ \Rightarrow & \sum_{i=1}^n X_i(Y_i - \bar{Y}) + \hat{\beta} \sum_{i=1}^n X_i(\bar{X} - X_i) = 0\end{aligned}$$

Deriving OLS

Solve for $\hat{\beta}_1$:

$$\hat{\beta} = \frac{\sum_{i=1}^n X_i(Y_i - \bar{Y})}{\sum_{i=1}^n X_i(X_i - \bar{X})}$$

$$\hat{\beta} = \frac{\sum_{i=1}^n X_i Y_i - \bar{Y} \sum_{i=1}^n X_i}{\sum_{i=1}^n X_i^2 - \bar{X} \sum_{i=1}^n X_i}$$

$$\hat{\beta} = \frac{\sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y}}{\sum_{i=1}^n X_i^2 - n \bar{X}^2}$$

$$\hat{\beta} = \frac{\sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y} - n \bar{X} \bar{Y} + n \bar{X} \bar{Y}}{\sum_{i=1}^n X_i X_i - n \bar{X}^2 - n \bar{X}^2 + n \bar{X}^2}$$

(to set up the next step)

Deriving OLS

Solve for \hat{B}_1 :

$$\hat{\beta} = \frac{\sum_{i=1}^n (X_i Y_i - \bar{X} \bar{Y})}{\sum_{i=1}^n (X_i X_i - \bar{X} \bar{X})}$$

(work backwards if helpful)

$$\hat{\beta} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$\hat{\beta} = \frac{Cov(X, Y)}{Var(X)}$$

Estimating OLS

$$\hat{Y}_i = \hat{a} + \hat{\beta}X_i$$

So, we can estimate the regression coefficients as follows:

$$\hat{\beta} = \frac{Cov(X, Y)}{Var(X)}$$

$$\hat{a} = \bar{Y} - \hat{\beta}\bar{X}$$

Practice Estimating OLS

We have employment data for 5 people. Y is the hourly wage a person earns and X is the number of years of work experience they have.

Y	X
2	1
6	2
3	3
8	4
6	5

Make sure you can use the equations we derived on the previous slides to estimate the regression equation:

$$\hat{\beta}_1 = \frac{Cov(X, Y)}{Var(X)} = \frac{2.5}{2.5} = 1.0$$

$$\hat{\beta}_0 = \bar{Y} - \beta_1 \bar{X} = 5 - 1 * 3 = 2.0$$

$$\hat{Y} = 2.00 + 1.00 * X$$

Interpreting OLS

The relationship we want to estimate:

$$Cons_i = a + \beta * Income_i + u_i$$

Estimated regression equation:

$$\widehat{Cons}_i = \hat{a} + \hat{\beta} * Income_i$$

The interpretation of $\hat{\beta}_1$ as a derivative.

$$\frac{\partial \widehat{Cons}}{\partial Income} = \hat{\beta}$$

So, the coefficient estimate tells us the expected change in consumption for each 1 dollar increase in income.

Prediction using OLS

$$\widehat{Cons}_i = 8,000 + .75 * Income_i$$

We can use the results of a regression to:

- Predict consumption given income.
- Predict how much consumption will change for a given change in income.
- Determine if income is a statistically significant determinant of consumption (we would need standard errors for the estimate).

Regression

OLS requires several assumptions:

(some are only needed for hypothesis testing)

1. *Linearity* – population data can be described using an equation that is linear in the parameters: $Y=B_0 + B_1X_1 + \dots + B_NX_N + u$.
2. *No exact multicollinearity* – there are no linear relationships between the Xs. Otherwise we can not determine which X is causing Y to change.
3. *Intercept is catchall* – $E(u)=0$.
4. *Exogeneity* – $E(u|X)$ [equivalently the $\text{cov}(X,u)=0$]
5. *Homoskedasticity* – the variance of the error term is constant
 $\text{var}(u|X)=\sigma^2$.
6. *Normality* – the error term is normally distributed: $u \sim N(0, \sigma^2)$
7. *Serial correlation* – the errors are not correlated with each other.

Regression

Table 3.1 The assumptions of the CLRM

<i>Assumption</i>	<i>Mathematical expression</i>	<i>Violation may imply</i>	<i>Chapter</i>
(1) Linearity of the model	$Y_t = \alpha + \beta X_t + u_t$	Wrong regressors Non-linearity Changing parameters	8
(2) X is variable	$Var(X)$ is not 0	Errors in variables	8
(3) X is non-stochastic and fixed in repeated samples	$Cov(X_s, u_t) = 0$ for all s and $t = 1, 2, \dots, n$	Autoregression	10
(4) Expected value of disturbance is zero	$E(u_t) = 0$	Biased intercept	—
(5) Homoskedasticity	$Var(u_t) = \sigma^2 = \text{constant}$	Heteroskedasticity	6
(6) Serial independence	$Cov(u_t, u_s) = 0$ for all $t \neq s$	Autocorrelation	7
(7) Normality of disturbance	$u_t \sim N(\mu, \sigma^2)$	Outliers	8
(8) No linear relationships	$\sum_{t=1}^T (\delta_i X_{it} + \delta_j X_{jt}) \neq 0 \quad i \neq j$	Multicollinearity	5

Properties of the OLS estimator

Properties of the OLS estimator:

1. Unbiased
2. Efficient

Properties of the OLS estimator

Unbiasedness:

$$\begin{aligned} \text{Cov}(X, Y) &= \text{Cov}(X, a + \beta X + u) \\ &= \text{Cov}(X, a) + \text{Cov}(X, \beta X) + \text{Cov}(X, u) \\ &= \beta \text{Var}(X) + \text{Cov}(X, u) \end{aligned}$$

$$\begin{aligned} \hat{\beta} &= \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \frac{\beta \text{Var}(X) + \text{Cov}(X, u)}{\text{Var}(X)} \\ &= \beta + \frac{\text{Cov}(X, u)}{\text{Var}(X)} \end{aligned}$$

So, $\hat{\beta}$ is equal to the population β and a random component that depends on the covariance of X and u .

Properties of the OLS estimator

Unbiasedness continued...:

$$\begin{aligned} E(\hat{\beta}) &= E[\beta + \frac{Cov(X, u)}{Var(X)}] \\ &= \beta + \frac{1}{Var(X)} E[Cov(X, u)] \\ &= \beta + \frac{1}{Var(X)} E\left[\frac{1}{n} \sum (X_t - \bar{X})(u_t - \bar{u})\right] \\ &= \beta + \frac{1}{Var(X)} \frac{1}{n} \sum (X_t - \bar{X}) E(u_t - \bar{u}) \end{aligned}$$

Because $E(u_t) = 0$. Thus we have:

$$E(\hat{\beta}) = \beta$$

Properties of the OLS estimator

Unbiasedness continued...:

$$\begin{aligned} E(\hat{a}) &= E(\bar{Y}) - E(\hat{\beta})\bar{X} \\ &= a + \beta\bar{X} - E(\hat{\beta})\bar{X} && \text{Since } E(Y)=a+BX \\ &= a + \beta\bar{X} - \beta\bar{X} \\ &= a \end{aligned}$$

Thus the OLS estimators are unbiased.

Properties of the OLS estimator

BLUE - We want to show that OLS is the **B**est (most efficient) **L**inear **U**nbiased **E**stimator.

This essentially boils down to solving a minimization problem for the variance of the estimator (best) while requiring that the estimator to be linear in Y and unbiased.

The steps:

1. Define a generic linear estimator
2. Force the estimator to be unbiased – this results in 2 constraints
3. Now examine the variance of the generic estimator
4. Minimize this variance subject to the 2 constraints
5. The result is the OLS estimator.

Properties of the OLS estimator

Note that there are two definitions of “linear” in this context:

1. Population data can be described using an equation that is linear in the parameters: $Y = B_0 + B_1 X_1 + \dots + B_N X_N + u$.
2. The OLS coefficient estimate \hat{B} can be written as an expression that is linear in Y (see next slide).

$$\check{\beta} = \delta_1 Y_1 + \dots + \delta_n Y_n = \sum \delta_i Y_i$$

The L in BLUE refers to this latter type of linearity and thus the generic linear estimator must have the form shown above.

Properties of the OLS estimator

Define a generic linear (in Y) estimator:

$$\check{\beta} = \delta_1 Y_1 + \dots + \delta_n Y_n = \sum \delta_i Y_i$$

Impose unbiasedness:

$$\begin{aligned} E(\check{\beta}) &= E\left(\sum \delta_i Y_i\right) = \sum \delta_i E(Y_i) \\ &= \sum \delta_i E(a + \beta X_i) \\ &= a \sum \delta_i + \beta \sum \delta_i X_i \end{aligned}$$

In order for this to be equal to β , we need the following two constraints to hold:

$$\sum \delta_i = 0 \text{ and } \sum \delta_i X_i = 1$$

Now we have two constraints that must be satisfied to be linear and unbiased.

Properties of the OLS estimator

Now minimize the variance of the estimator: This requires that we do some math first.

$$\begin{aligned}Var(\check{\beta}) &= E[\check{\beta} - E(\check{\beta})]^2 \\&= E\left[\sum \delta_i Y_i - E\left(\sum \delta_i Y_i\right)\right]^2 \\&= E\left[\sum \delta_i Y_i - \sum \delta_i E(Y_i)\right]^2 \\&= E\left[\sum \delta_i (Y_i - E(Y_i))\right]^2 \\&= E\left[\sum \delta_i (a + \beta X_i + u_i - (a + \beta X_i))\right]^2 \\&= E\left[\sum \delta_i u_i\right]^2\end{aligned}$$

Properties of the OLS estimator

Now minimize the variance of the estimator: This requires that we do some math first.

$$\begin{aligned}Var(\check{\beta}) &= E(\delta_1^2 u_1^2 + \dots + \delta_N^2 u_N^2 \\&\quad + 2\delta_1\delta_2 u_1 u_2 + 2\delta_1\delta_3 u_1 u_3 + \dots) \\&= \delta_1^2 E(u_1^2) + \dots + \delta_N^2 E(u_N^2) \\&\quad + 2\delta_1\delta_2 E(u_1 u_2) + 2\delta_1\delta_3 E(u_1 u_3) + \dots \\&= \sum \delta_i^2 \sigma^2\end{aligned}$$

We use the fact that the error are uncorrelated so that $E(u_1, u_2) = 0$, etc.

Properties of the OLS estimator

Minimize the variance: subject to the 2 constraints

$$\text{Min } \sum \delta_i^2 \sigma^2 \quad \text{s.t. } \sum \delta_i = 0 \text{ and } \sum \delta_i X_i = 1$$

The solution to this minimization problem (using Legrangians):

$$\delta_i = \frac{X_i - \bar{X}}{\sum(X_i - \bar{X})^2}$$

Now we plug this into our generic expression for a linear estimator:

$$\begin{aligned}\check{\beta} &= \sum \delta_i Y_i = \frac{\sum(X_i - \bar{X})Y_i}{\sum(X_i - \bar{X})^2} = \frac{\sum X_i Y_i - \bar{X} \sum Y_i}{\sum(X_i - \bar{X})^2} \\ &= \frac{\sum X_i Y_i - N \bar{X} \bar{Y}}{\sum(X_i - \bar{X})^2} \\ &= \hat{\beta}\end{aligned}$$

The last step can be seen from the algebra used to derive OLS.

Properties of the OLS estimator

So, of all the possible linear estimators that are unbiased, OLS has the smallest variance.

[under the assumptions we made]

Another thing we have done is produce an equation for the standard errors of the estimate of B:

$$\begin{aligned}Var(\hat{\beta}) &= \sum \delta_i^2 \sigma^2 \\&= \sum \left(\frac{X_i - \bar{X}}{\sum(X_i - \bar{X})^2} \right)^2 \sigma^2 \\&= \sigma^2 \frac{1}{\sum(X_i - \bar{X})^2}\end{aligned}$$

Properties of the OLS estimator

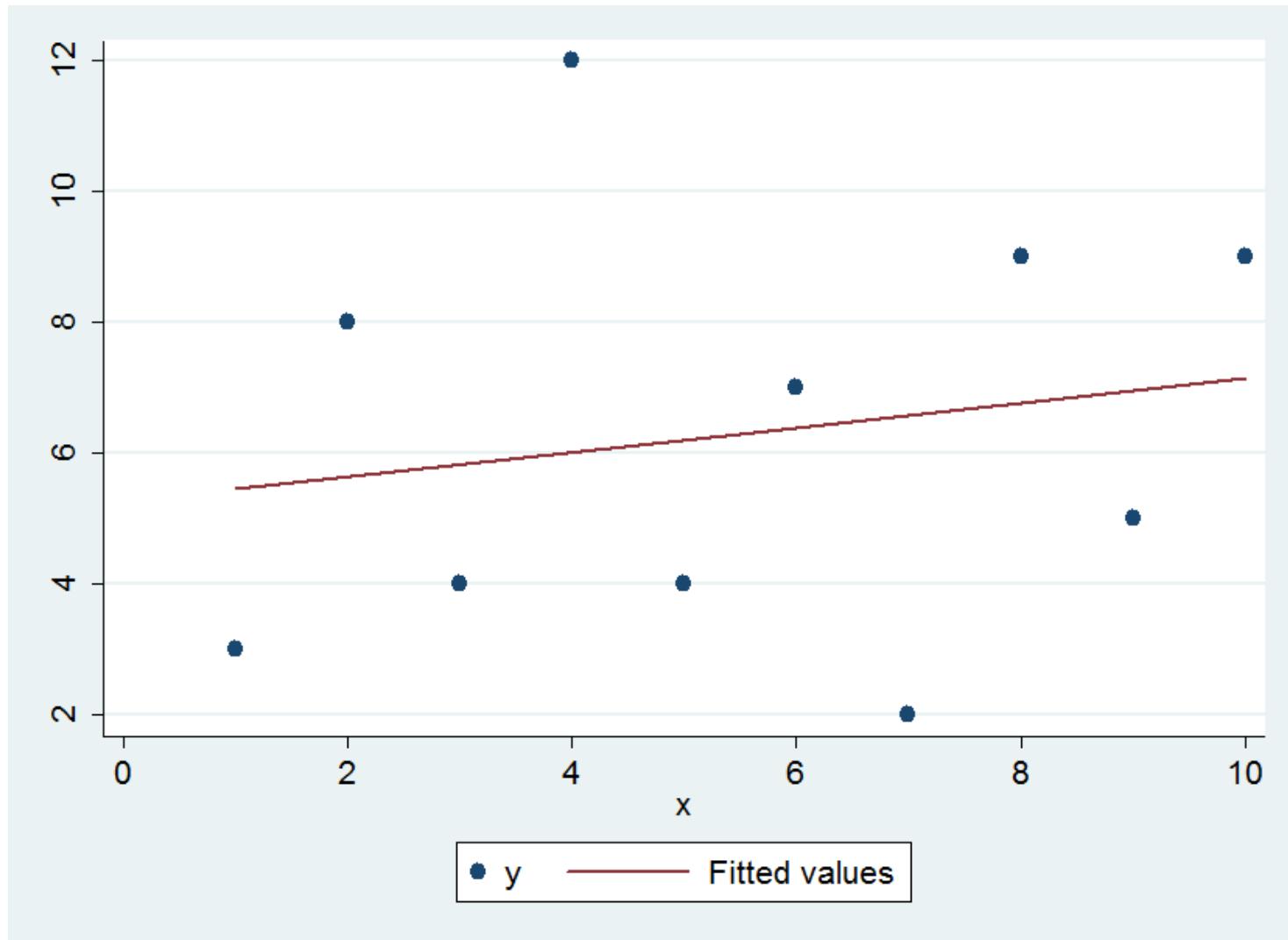
You might think that requiring the estimator to be linear in Y is kind of odd and restrictive (it is).

This requirement can be relaxed, and the OLS estimate is still best, if you impose further assumptions (specifically, that the error terms u_i are normally distributed).

In this case, the OLS estimate is BUE (Best Unbiased Estimator), but we won't be proving this in class.

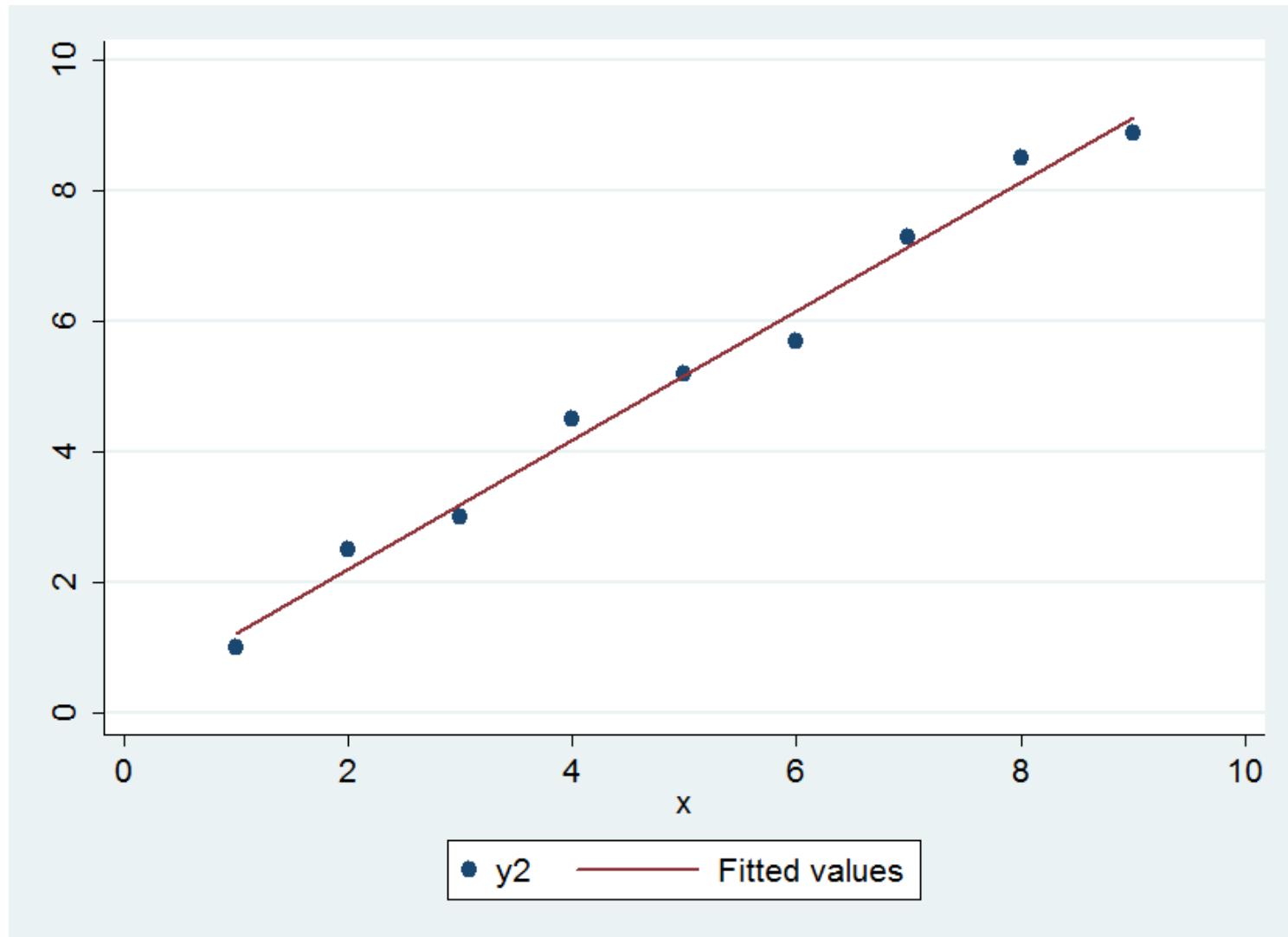
Goodness of Fit

Sometimes a regression line matches the data very well and sometimes it does not.



Goodness of Fit

Sometimes a regression line matches the data very well and sometimes it does not.



Goodness of Fit

RSS – measure the difference between the true Y and the predicted value \hat{Y} .

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

SST – measures the difference between the true Y and the mean \bar{Y} .

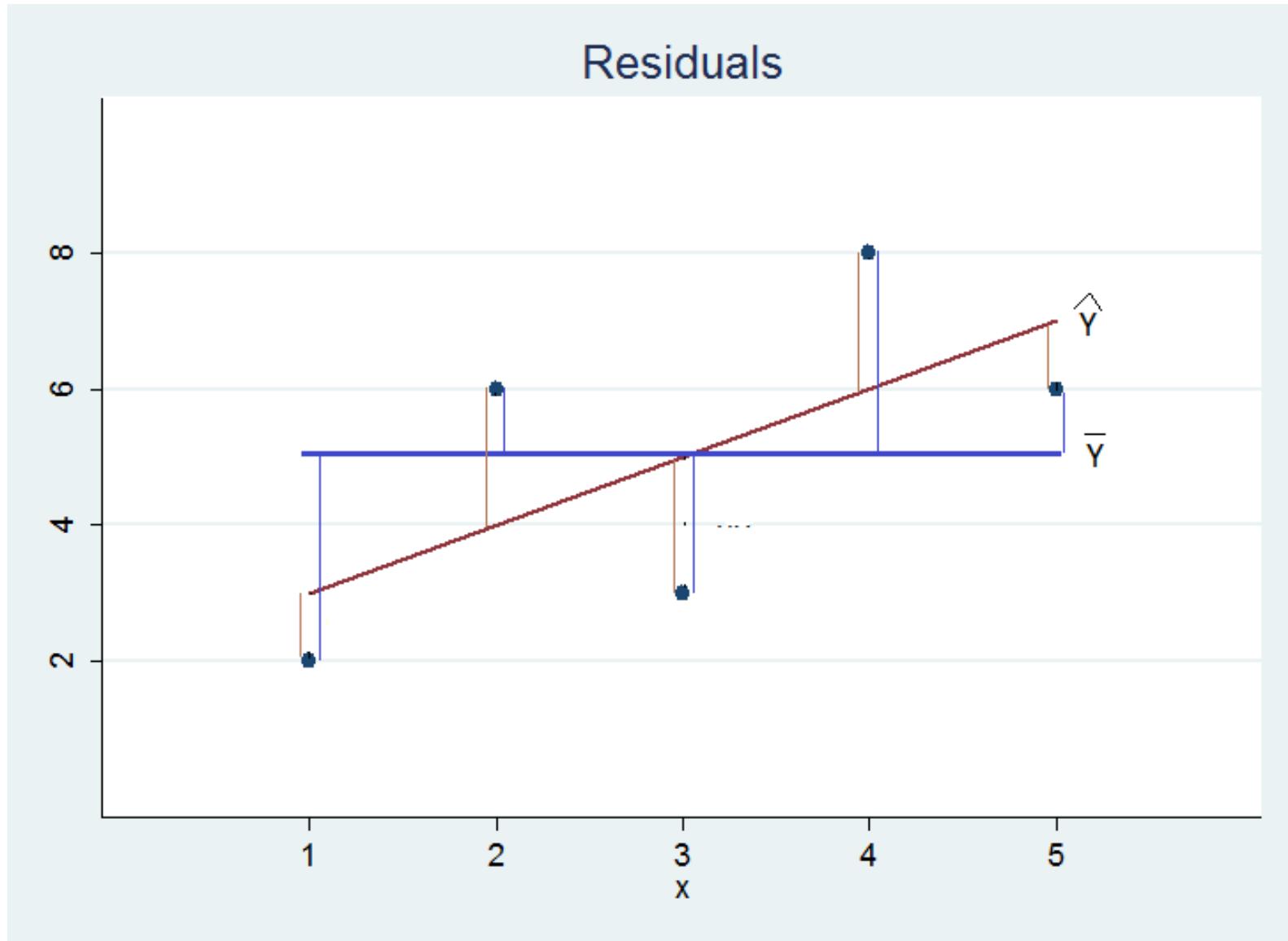
$$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

R-Squared – measures how much better the regression line predicts the true values of y than just the mean.

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

Note that $RSS \leq SST$, so the R-squared should fall between 0 and 1. A large value (close to 1) implies that the regression fits the data well.

Goodness of Fit



Goodness of fit

Shortcomings of the R-Squared:

- The R – squared says almost nothing about causality:

Spurious regression and trends – If X and Y follow a shared path (which is frequently the case in time-series analysis), then the R-squared will be large, even if X does not cause Y.

Correlation does not imply causality – There may be some third variable Z, or several, that is causing X and Y to move together. The variables are omitted variables.

- The R-squared has limited use for comparing models

Functional form – The value is strongly determined by the choice of functional form.

Adding variables – Adding more variables always increases the value.

STATA

use filename.dta

regress Y X

Lesson 5

Ordinary Least Squares: Multivariate

Outline

Previous Lesson:

1. Deriving ordinary least squares (OLS)
2. OLS is BLUE
3. Goodness of Fit

This Lesson:

1. Interpreting multivariate regressions
2. Deriving multivariate regression
3. Matrix form
4. Multivariate regression is BLUE

Next Lesson:

1. Hypothesis testing with estimated coefficients
2. Comparing models

Multivariate: Conditionality

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

Variables:

- Y is the outcome variable,
- X_1 and X_2 are explanatory variables

Coefficients:

- B_0 is the intercept
- B_1 is the change in Y with respect to X_1 , holding X_2 fixed
- B_2 is the change in Y with respect to X_2 , holding X_1 fixed

Multivariate: Conditionality

□ Interpreting a multivariate regression:

$$\widehat{wage} = 10.50 + 1.50 * educ + 0.75 * exper$$

- These regression results suggest that, holding experience fixed, each year of education increases wages by \$1.50.
- These regression results suggest that, holding education fixed, each year of experience increases wages by \$0.75.

□ Regressions can have many explanatory variables: The coefficients B_1, B_2, \dots are not only interesting in their own right, but become more meaningful when the other factors are fixed.

$$wage_i = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 male + u_i$$

Multivariate: Conditionality

Suppose we regress log consumption on log income and a person's sex.

$$\ln(\text{consumption}_i) = 7.238 + 0.65\ln(\text{income}_i) + 0.14\text{male}_i + u_i$$

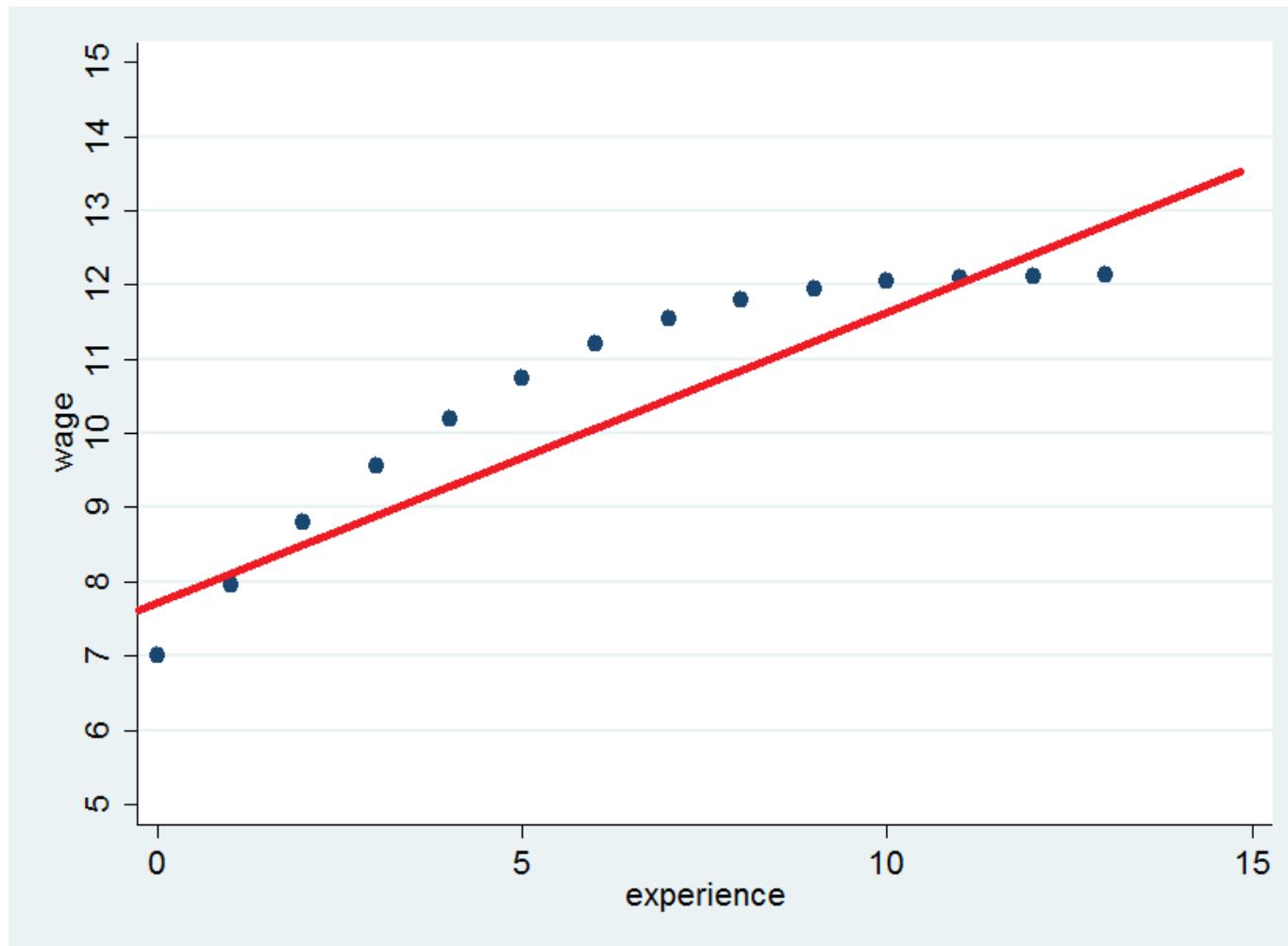
How do we interpret B_2 ?

How do we interpret B_3 ?

What other variables might we add?

Multivariate: Curvature

Consider the following graph that compares wages to experience. It has a curved shape that does not fit well with a straight line.



Multivariate: Curvature

- **Curved regression lines:** By adding higher order polynomials we allow Y to change non-linearly in X.

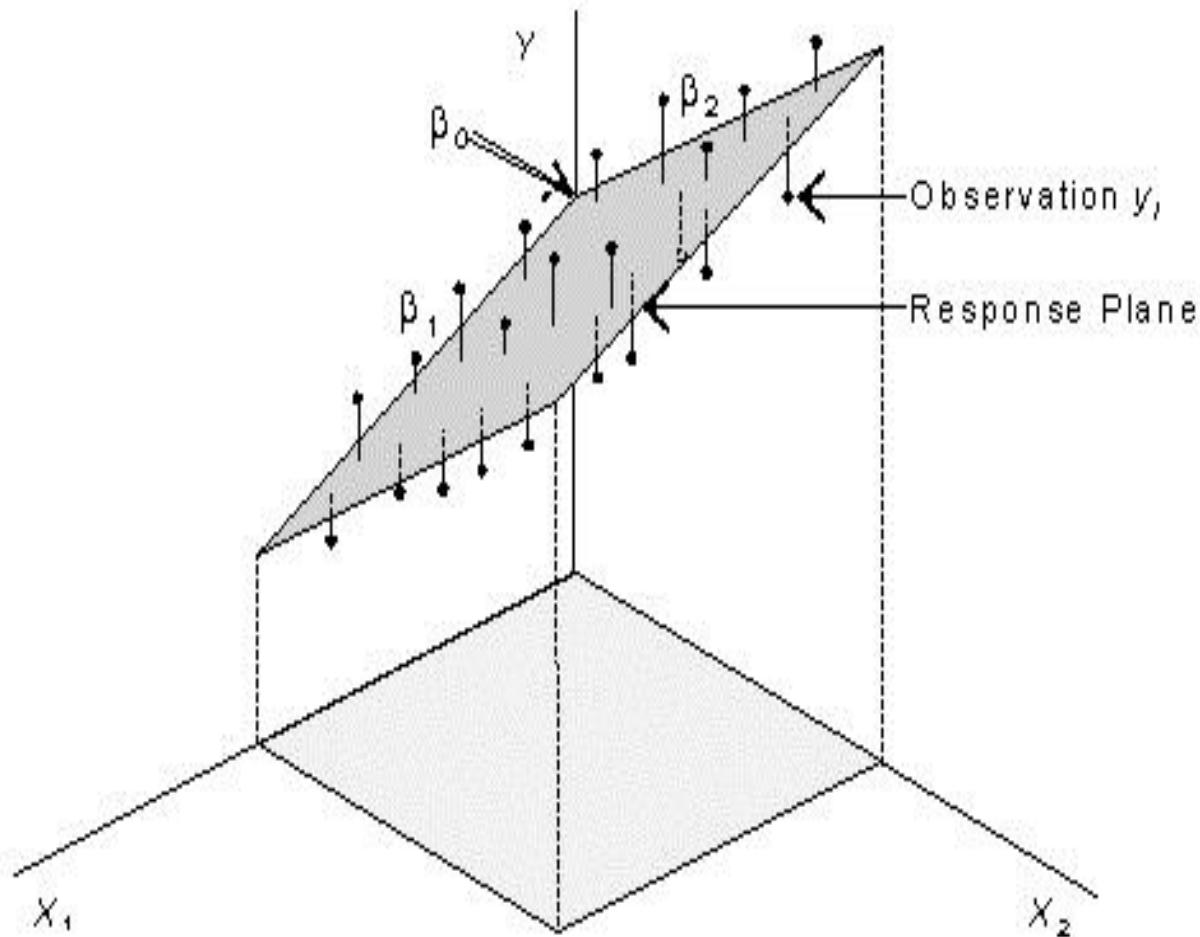
$$wage_i = \beta_0 + \beta_1 * exper + \beta_2 * exper^2 + u_i$$

- **The slope varies with the level:**

$$\widehat{wage}_i = \hat{\beta}_0 + \hat{\beta}_1 * exper + \hat{\beta}_2 * exper^2$$

$$\frac{\partial(\widehat{wage})}{\partial(exper)} = \hat{\beta}_1 + 2\hat{\beta}_2 exper$$

Multivariate: 2 explanatory variables



Multivariate: 2 explanatory variables

Minimize the sum of the squared errors:

$$\text{Min } \sum_{i=1}^N (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{i2} - \hat{\beta}_3 X_{i3})^2$$

Take the first order derivatives with respect to B_1 , B_2 , and B_3 :

$$\frac{\partial RSS}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^N (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{i2} - \hat{\beta}_3 X_{i3}) = 0 \quad (1)$$

$$\frac{\partial RSS}{\partial \hat{\beta}_2} = -2 \sum_{i=1}^N X_{i2} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{i2} - \hat{\beta}_3 X_{i3}) = 0 \quad (2)$$

$$\frac{\partial RSS}{\partial \hat{\beta}_3} = -2 \sum_{i=1}^N X_{i3} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{i2} - \hat{\beta}_3 X_{i3}) = 0 \quad (3)$$

We have 3 equations and 3 unknowns.

Multivariate: 2 explanatory variables

$$\begin{aligned}(1) \Rightarrow \sum Y_i &= \sum \hat{\beta}_1 + \sum \hat{\beta}_2 X_{i2} + \sum \hat{\beta}_3 X_{i3} \\ \sum Y_i &= N\hat{\beta}_1 + \hat{\beta}_2 \sum X_{i1} + \hat{\beta}_3 \sum X_{i3} \\ \bar{Y} &= \hat{\beta}_1 + \hat{\beta}_2 \bar{X}_2 + \hat{\beta}_3 \bar{X}_3 \\ \hat{\beta}_1 &= \bar{Y} - \hat{\beta}_2 \bar{X}_2 - \hat{\beta}_3 \bar{X}_3\end{aligned}$$

After a fair bit of manipulation, we can derive the following estimates:

$$\hat{\beta}_2 = \frac{Cov(X_2, Y)Var(X_3) - Cov(X_3, Y)Cov(X_2, X_3)}{Var(X_2)Var(X_3) - [Cov(X_2)(X_3)]^2}$$

$$\hat{\beta}_3 = \frac{Cov(X_3, Y)Var(X_2) - Cov(X_2, Y)Cov(X_3, X_2)}{Var(X_3)Var(X_2) - [Cov(X_3)(X_2)]^2}$$

As you might be able to guess, this is both tedious to derive and tedious to compute with data.

Multivariate: k explanatory variables

Minimize the sum of the squared errors:

$$\text{Min } \sum_{i=1}^N (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{i2} - \hat{\beta}_3 X_{i3} - \dots - \hat{\beta}_k X_{ik})^2$$

Take the first order derivatives with respect to B_0, B_1, \dots, B_k :

$$\frac{\partial RSS}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^N (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{i2} - \dots - \hat{\beta}_k X_{ik}) = 0 \quad (1)$$

$$\frac{\partial RSS}{\partial \hat{\beta}_2} = -2 \sum_{i=1}^N X_{i2} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{i2} - \dots - \hat{\beta}_k X_{ik}) = 0 \quad (2)$$

⋮

⋮

$$\frac{\partial RSS}{\partial \hat{\beta}_k} = -2 \sum_{i=1}^N X_{ik} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{i2} - \dots - \hat{\beta}_k X_{ik}) = 0$$

Multivariate: k explanatory variables

Population regression equation:

$$Y_i = \beta_1 + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + u_i$$

The problem:

This math is very tedious.

The expressions are too long to be practical.

The solution:

Use computer software to do the estimation.

Use matrix algebra and notation for ease of expression.

Note: often times **bold** font is used to denote a vector or matrix

Multivariate: Matrices

Review of matrix operations and properties (in the 2 x 2 case):

Addition/subtraction:

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} \\ a_{21} + b_{21} & a_{22} + b_{22} \end{bmatrix}$$

Two matrices must have the same dimensions in order to be added and subtracted.

Multiplication:

$$\begin{aligned}\mathbf{AB} &= \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \\ &= \begin{bmatrix} a_{11} * b_{11} + a_{12} * b_{21} & a_{11} * b_{12} + a_{12} * b_{22} \\ a_{21} * b_{11} + a_{22} * b_{21} & a_{21} * b_{12} + a_{22} * b_{22} \end{bmatrix}\end{aligned}$$

A matrix with dimensions m x n can be multiplied with another matrix with dimensions n x k.

Multivariate: Matrices

Review of matrix operations and properties:

Transpose:

$$\mathbf{A}' = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}' = \begin{bmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{bmatrix}'$$

Inverse:

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Only square matrices are invertible. The inverse is found by setting up a partitioned matrix and then converting the left side into the identity using row operations (addition and subtraction of other rows, division or multiplication of the row):

$$\left[\begin{array}{cc|cc} a_{11} & a_{12} & 1 & 0 \\ a_{21} & a_{22} & 0 & 1 \end{array} \right]$$

Multivariate: Matrices

Some properties of matrix operations:

$$1. \quad \mathbf{ABC} = \mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$$

$$2. \quad (\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$$

$$2. \quad (\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'$$

$$3. \quad \mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$$

$$4. \quad (\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$$

$$5. \quad \mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

$$6. \quad (\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$$

Multivariate: Matrices

Derivatives:

$$\mathbf{A} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$

$$\frac{\partial \mathbf{a}}{\partial \boldsymbol{\beta}} = \begin{bmatrix} \frac{\partial a}{\partial \beta_1} & \frac{\partial a}{\partial \beta_2} & \dots & \frac{\partial a}{\partial \beta_k} \end{bmatrix}$$

$$\frac{\partial \mathbf{A}}{\partial \boldsymbol{\beta}} = \begin{bmatrix} \frac{\partial a_1}{\partial \beta_1} & \frac{\partial a_1}{\partial \beta_2} & \dots & \frac{\partial a_1}{\partial \beta_k} \\ \frac{\partial a_2}{\partial \beta_1} & \frac{\partial a_2}{\partial \beta_2} & \dots & \frac{\partial a_2}{\partial \beta_k} \\ \vdots & \vdots & & \\ \frac{\partial a_N}{\partial \beta_1} & \frac{\partial a_N}{\partial \beta_2} & \dots & \frac{\partial a_N}{\partial \beta_k} \end{bmatrix}$$

Multivariate: k explanatory variables

$$Y_i = \beta_1 + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + u_i$$

We have k explanatory variables and k coefficients. So B is a k x 1 vector and X_i is a k x 1 vector.

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_k \end{bmatrix} \quad X_i = \begin{bmatrix} 1 \\ X_{i2} \\ X_{i3} \\ \vdots \\ X_{ik} \end{bmatrix}$$

Multivariate: k explanatory variables

We have k explanatory variables and k coefficients. The expression for each individual can be expressed in terms of vectors:

$$\begin{aligned} Y_i &= \mathbf{X}'_i \boldsymbol{\beta} + u_i \\ &= [1 \quad X_{i2} \quad \cdots \quad X_{ik}] \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + u_i \\ &= \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \dots + \beta_k X_{ik} + u_i \\ &= \sum_{j=1}^k \beta_j X_{ij} + u_i \end{aligned}$$

Multivariate: k explanatory variables

We can expand this and have a vector of individuals, with one row for each individual:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} \mathbf{X}'_1 \boldsymbol{\beta} \\ \mathbf{X}'_2 \boldsymbol{\beta} \\ \vdots \\ \mathbf{X}'_N \boldsymbol{\beta} \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix}$$

Where the \mathbf{X}_i is $1 \times k$ and $\boldsymbol{\beta}$ is $k \times 1$.

We can express the vector of vectors as a single matrix. Each row has the values of \mathbf{X} for one individual:

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}'_1 \\ \mathbf{X}'_2 \\ \vdots \\ \mathbf{X}'_N \end{bmatrix} = \begin{bmatrix} 1 & X_{12} & X_{13} & \cdots & X_{1k} \\ 1 & X_{22} & X_{23} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{N2} & X_{N3} & \cdots & X_{Nk} \end{bmatrix}$$

Multivariate: k explanatory variables

Thus we can express the set of population relationships for every individual as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

$$= \begin{bmatrix} 1 & X_{12} & X_{13} & \cdots & X_{1k} \\ 1 & X_{22} & X_{23} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{N2} & X_{N3} & \cdots & X_{Nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix}$$

So, \mathbf{Y} is an $N \times 1$ vector of outcome values, \mathbf{X} is an $N \times k$ matrix of explanatory variables, \mathbf{B} is an $k \times 1$ vector of coefficients, and \mathbf{u} is an $N \times 1$ vector of error terms.

A good exercise is to verify each of the matrix steps, or at least to check to make sure that the dimensionality is valid in each step.

Multivariate: k explanatory variables

Our goal is to get an estimate of B:

$$\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \rightarrow \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$$

Our estimated regression equation will look like this:

$$\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_k \end{bmatrix} = \begin{bmatrix} 1 & X_{12} & X_{13} & \cdots & X_{1k} \\ 1 & X_{22} & X_{23} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{N2} & X_{N3} & \cdots & X_{Nk} \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} + \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_N \end{bmatrix}$$

Multivariate: k explanatory variables

In order to estimate the B coefficients, we need to minimize the sum of the squared errors:

$$\hat{\mathbf{u}}' \hat{\mathbf{u}} = [\hat{u}_1 \quad \hat{u}_2 \quad \cdots \quad \hat{u}_N] \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_N \end{bmatrix} = \sum_{i=1}^N \hat{u}_i^2$$

The sum of the squared errors can be expressed in matrix form (check dimensions):

$$\begin{aligned}\hat{\mathbf{u}}' \hat{\mathbf{u}} &= (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= (\mathbf{Y}' - \hat{\boldsymbol{\beta}}'\mathbf{X}')(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} && \text{The 2nd and 3rd terms are identical.} \\ &= \mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} && \text{Each term is scalar.}\end{aligned}$$

Multivariate: k explanatory variables

Just as in the univariate case we take the derivative with respect to the Bs and set it equal to 0:

$$RSS = \hat{\mathbf{u}}' \hat{\mathbf{u}} = \mathbf{Y}' \mathbf{Y} - 2\mathbf{Y}' \mathbf{X} \hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}' \mathbf{X}' \mathbf{X} \hat{\boldsymbol{\beta}}$$

$$\frac{\partial RSS}{\partial \hat{\boldsymbol{\beta}}} = -2\mathbf{X}' \mathbf{Y} + 2\mathbf{X}' \mathbf{X} \hat{\boldsymbol{\beta}} = 0 \quad (\text{write out long form and verify})$$

$$\Rightarrow \mathbf{X}' \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X}' \mathbf{Y}$$

$$\Rightarrow \hat{\boldsymbol{\beta}} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{Y}$$

This is certainly much easier to express than without matrix algebra. However, these steps contain a lot of math that is not shown.

We are now going to show that these OLS estimates are BLUE.

Multivariate: Assumptions

1. **Linearity:** The model is correct and intrinsically linear. This means that it is linear in the parameters (coefficients B) or can be made linear.

$$Y_i = \beta_1 + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + u_i$$

Example: $Y_i = \beta_1 + \beta_2 \left(\frac{1}{X_{i2}} \right) + u_i$

Example:

$$\begin{aligned} Q_i &= AK_i^{\beta_2} L_i^{\beta_3} u_i \\ \Rightarrow \ln(Q_i) &= \ln(A) + \beta_2 \ln(K_i) + \beta_3 \ln(L_i) + u_i \end{aligned}$$

Multivariate: Assumptions

2. **No exact multicollinearity:** None of the explanatory variables X_j is a linear combination of the others.

Example: Suppose X_3 is a linear function of X_2

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + u_i$$

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3(\alpha_1 + \alpha_2 X_{i2}) + u_i$$

$$Y_i = \beta_1 + \beta_3\alpha_1 + (\beta_2 + \beta_3\alpha_2)X_{i2} + u_i$$

$$Y_i = \beta_1^* + \beta_2^* X_{i2} + v_i$$

$$\beta_1^* = \beta_1 + \beta_3\alpha_1$$

$$\beta_2^* = \beta_2 + \beta_3\alpha_2$$

So we can't estimate B_1 , B_2 and B_3 . These coefficients are not identified. This is also called the full rank assumption (none of the rows of the matrix can be eliminated by a linear combination of the others).

Multivariate: Assumptions

3. Intercept is a catchall: The expected value of the error term is 0.

$$E(\mathbf{u}) = 0$$

This is by design. Suppose that it were not true:

$$Y_i = \beta_1^* + \beta_2 X_{i2} + \beta_3 X_{i3} + u_i^*$$

And:

$$E(u_i^*) = \mu$$

Then:

$$Y_i = \beta_1^* + \mu + \beta_2 X_{i2} + \beta_3 X_{i3} + u_i^* - \mu$$

$$Y_i = \beta_1 + \mu + \beta_2 X_{i2} + \beta_3 X_{i3} + u_i$$

Now we have:

$$E(u_i) = 0$$

Multivariate: Assumptions

4. **Exogeneity**: The expected value of the error term is 0 conditional on the Xs. In other words, the explanatory variables are not correlated with the error term. If they are, the B's will be biased.

$$E(\mathbf{u}|\mathbf{X}) = \mathbf{0}$$

This is a very strong assumption. We will spend much of the last part of the class discussing how to address **endogeneity**, which is a violation of this assumption.

Empirical methods such as instrumental variables, difference-in-differences, and regression discontinuity are used to correct for this.

5. **Homoscedasticity**: spherical errors. A violation of this implies that OLS is not efficient (it is still unbiased).

$$Var(\mathbf{u}|\mathbf{X}) = \sigma^2 \mathbf{I}$$

Multivariate: BLUE – Unbiased

The estimator is unbiased: $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$

We know that: $\mathbf{Y} = \mathbf{X}\beta + \mathbf{u}$

So,

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{u}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \\ &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\end{aligned}$$

We take the expected value:

$$\begin{aligned}E(\hat{\beta}) &= E(\beta) + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{u}) \\ &= \beta\end{aligned}$$

Multivariate: BLUE – Efficient

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$$

We need to find an expression for the variance of the estimator:

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}) &= E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'] \\ &= E\{[\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} - \boldsymbol{\beta}][\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} - \boldsymbol{\beta}]'\} \\ &= E\{[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}][(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}]'\} \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{u}\mathbf{u}')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

Multivariate: BLUE – Efficient

Variance continued:

$$\begin{aligned}Var(\hat{\beta}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2 I \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\&= \sigma^2 I(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

Multivariate: BLUE – Efficient

Consider any other linear estimator (just as we did for the univariate case):

$$\begin{aligned}\check{\beta} &= [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{Z}]\mathbf{Y} && \text{(where Z is any k by N matrix)} \\ &= [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{Z}](\mathbf{X}\beta + \mathbf{u}) \\ &= \beta + \mathbf{Z}\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} + \mathbf{Z}\mathbf{u}\end{aligned}$$

Thus we will satisfy:

$$E(\check{\beta}) = \beta$$

Only if:

$$\mathbf{Z}\mathbf{X} = \mathbf{0}$$

Multivariate: BLUE – Efficient

So, now we have:

$$\begin{aligned}\check{\beta} &= \beta + \mathbf{Z}\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} + \mathbf{Z}\mathbf{u} \\ &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} + \mathbf{Z}\mathbf{u}\end{aligned}$$

And thus:

$$\check{\beta} - \beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} + \mathbf{Z}\mathbf{u}$$

Now consider the variance-covariance matrix:

$$\begin{aligned}E[(\check{\beta} - \beta)(\check{\beta} - \beta)'] &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} + \mathbf{Z}\mathbf{u}] [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} + \mathbf{Z}\mathbf{u}]' \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} + \mathbf{Z}\mathbf{u}] [\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} + \mathbf{u}'\mathbf{Z}] \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} + \sigma^2\mathbf{Z}\mathbf{Z}'\end{aligned}$$

Note that the middle terms disappear because $\mathbf{Z}\mathbf{X} = 0$. Because $\mathbf{Z}\mathbf{Z}'$ is always weakly positive, any non-zero matrix will increase variance-covariance matrix, so $\hat{\mathbf{B}}$ is the best of the linear estimators.

Multivariate: Measures of Fit

RRS – difference between the true Y and the predicted value \hat{Y} .

$$RSS = \sum_{i=1}^N (Y_i - \hat{Y})^2$$

SST – difference between the true Y and the mean \bar{Y} .

$$SST = \sum_{i=1}^N (Y_i - \bar{Y})^2$$

R-Squared – measures how much better the regression line predicts the true values of y than just the mean.

$$R^2 = 1 - \frac{SSR}{SST}$$

Note: Adding additional variables always improves the R-Squared, which is one of the reason that economists don't like it.

Multivariate: Interpretation

Cobb-Douglas:

$$\begin{aligned} Q_i &= AK_i^{\beta_2} L_i^{\beta_3} u_i \\ \Rightarrow \ln(Q_i) &= \ln(A) + \beta_2 \ln(K_i) + \beta_3 \ln(L_i) + u_i \end{aligned}$$

How do we interpret B_2 ?

How do we interpret B_3 ?

Multivariate: STATA

regress crime police

Source	SS	df	MS	Number of obs	=	97
Model	10663878.3	1	10663878.3	F(1, 95)	=	104.24
Residual	9719011.76	95	102305.387	Prob > F	=	0.0000
Total	20382890	96	212321.771	R-squared	=	0.5232
				Adj R-squared	=	0.5182
				Root MSE	=	319.85

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
crime						
police	21.32293	2.088519	10.21	0.000	17.17669	25.46916
_cons	-42.5565	53.72957	-0.79	0.430	-149.2232	64.11019

What does the estimated equation look like?

What variables have we omitted?

Multivariate: STATA

regress crime police enroll

Source	SS	df	MS	Number of obs	=	97
Model	14903894.4	2	7451947.22	F(2, 94)	=	127.85
Residual	5478995.6	94	58287.1873	Prob > F	=	0.0000
Total	20382890	96	212321.771	R-squared	=	0.7312

crime	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
police	7.570163	2.255043	3.36	0.001	3.092723	12.0476
enroll	.0244432	.0028659	8.53	0.000	.0187529	.0301335
_cons	-153.6529	42.59608	-3.61	0.000	-238.2285	-69.07743

What is the new regression equation?

Multivariate: Stata

$$\widehat{crime} = -42.46 + 21.32 * police$$

$$\widehat{crime} = -153.67 + 7.57 * police + 0.24 * enroll$$

How did adding enrollment affect the coefficient on police? Why?

How did adding enrollment affect the estimated fit of the model?

If the college reduces its police force by 10, but admits an additional 1,000 students, how do we predict the number of crimes will change?

Lesson 6

Hypothesis Testing: Regression Coefficients

Outline

Previous Lesson:

1. Interpreting multivariate regressions
2. Deriving multivariate regression
3. Multivariate regression is BLUE

This Lesson:

1. Hypothesis testing with regression coefficients
2. Comparing coefficients
3. Joint hypothesis tests
4. Comparing models

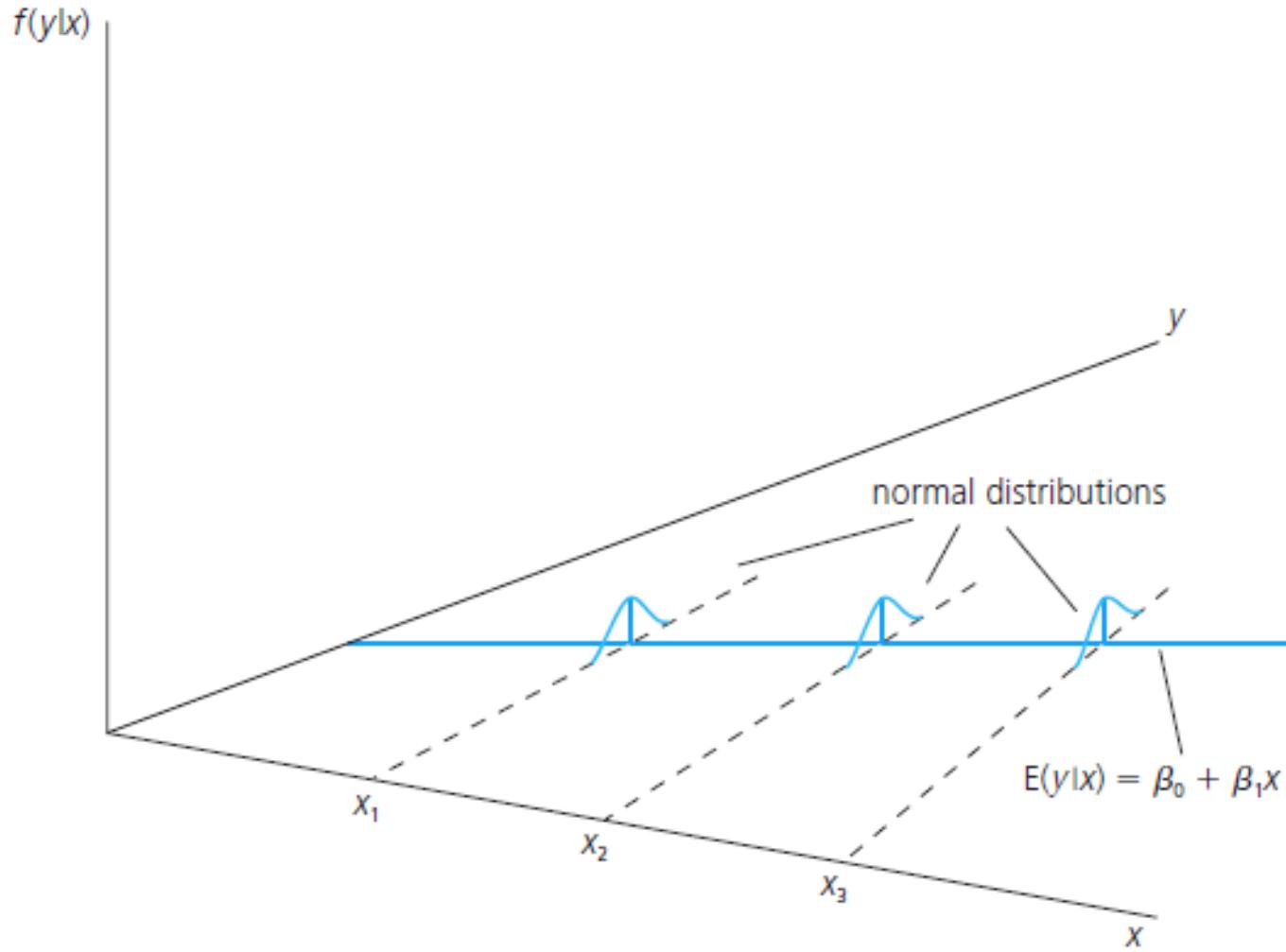
Text: Chap 3 (p.45-48), Chap 4 (p.74-88)

Next Lesson:

1. Dummy variables
2. Multicollinearity

t-distribution

The homoskedastic normal distribution with a single explanatory variable.



t-statistics

Regression coefficients are normally distributed around the true parameter:

$$\hat{\beta}_j \sim (\beta_j, Var(\hat{\beta}_j))$$

The t-statistic is found by converting to standard normal:

$$\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{n-k}$$

- β_j is the true value of the regression coefficient.
- $\hat{\beta}_j$ is the estimated value of the regression coefficient.
- $se(\hat{\beta}_j)$ is the standard error of the estimated regression coefficient.
- t_{n-k} is the t-distribution with $n-k$ degrees of freedom, where n is the number of observations in the data, k is the number of slope parameters (including the intercept).

Regression Coefficients

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + u$$

Null hypothesis:

$$H_0 : \beta_j = 0$$

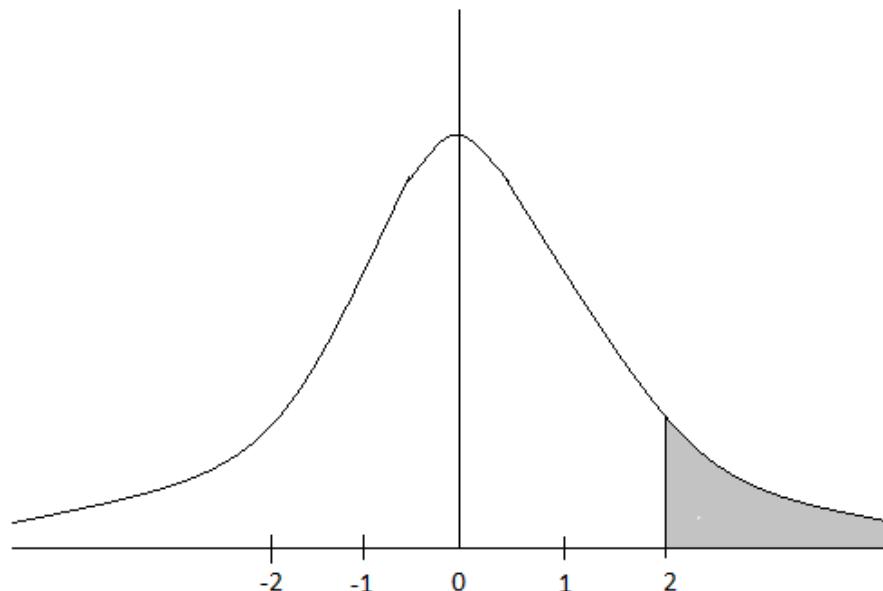
- This hypothesis states that the explanatory variable X_j has no effect on the outcome variable Y .
- We will either reject the null hypothesis at some level of confidence level (typically 1, 5, or 10 %), or we will not reject. We will reject if $\hat{\beta}_j$ is far away from 0.

Regression Coefficients

Null hypothesis: $H_0 : \beta_j = 0$

One-sided test: $H_A : \beta_j > 0$

- The alternative hypothesis is that B_j is larger than 0.
- In this one sided test we will reject the null hypothesis only if B_j falls too far to the right of 0.



Regression Coefficients

Suppose we estimate the effects of company profits and revenue on CEO salary for 60 companies and get the following results:

$$\widehat{CEO\text{Salary}} = 432,751 + 0.258 * Profits + 0.021 * Revenue$$
$$(72,000) \quad (0.150) \quad (0.005)$$

Test the hypothesis that profits is not a determinant of the CEO's salary versus the alternative that it has a positive effect at the 5% and 10% levels.

Null hypothesis: $H_0 : \beta_2 = 0$

Alternative hypothesis: $H_A : \beta_2 > 0$

T-statistic: $t(\hat{\beta}_2 | \beta_2 = 0) = \frac{\hat{\beta}_2 - 0}{se(\hat{\beta}_2)} = \frac{0.258}{0.150} = 1.72 \sim t_{57}$

Reject or not reject at 5% confidence level:

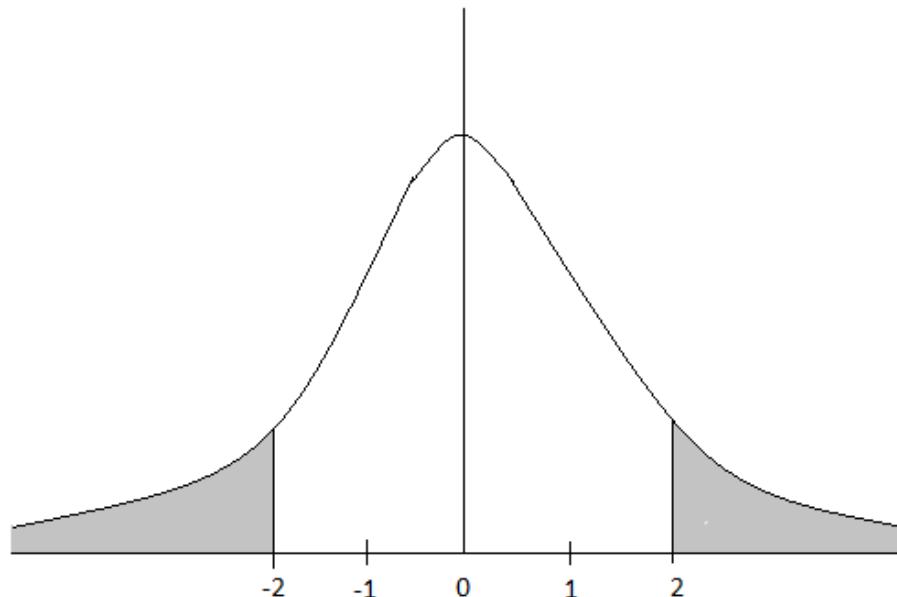
Reject or not reject at 10% confidence level:

Regression Coefficients

Null hypothesis: $H_0 : \beta_j = 0$

Two-sided test: $H_A : \beta_j \neq 0$

- The alternative hypothesis is that B_j is not equal to 0, so it could either be larger or smaller.
- In this two-sided test we will reject the null hypothesis only if B_j falls too far to the right of 0 or too far to the left (so we take the absolute value of the t-statistic).



Regression Coefficients

$$\widehat{CEO\text{Salary}} = 432,751 + 0.258 * Profits + 0.021 * Revenue$$
$$(72,000) \quad (0.150) \quad (0.005)$$

Test the hypothesis that a firm's revenue is not a determinant of the CEO's salary versus the alternative that it does have an effect at the 5% and 1% levels.

Null hypothesis: $H_0 : \beta_3 = 0$

Alternative hypothesis: $H_A : \beta_3 \neq 0$

T-statistic:

$$t(\hat{\beta}_3 | \beta_3 = 0) = \frac{\hat{\beta}_3 - 0}{se(\hat{\beta}_3)} = \frac{0.021}{0.005} = 4.2 \sim t_{57}$$

Reject or not reject at 5% confidence level:

Reject or not reject at 1% confidence level:

Regression Coefficients

Steps:

1. State the null and alternative hypothesis
2. Compute the t-statistic
3. Find the critical t-value with $n-k$ degrees of freedom for the 1 or 2 sided test and the desired confidence level.
4. Reject the null if the t-stat is larger in magnitude than the critical t-value.
Not reject if the t-stat is smaller in magnitude than the critical t-value.
[for the two-sided test]

Intuition:

We are constructing a standard normal distribution around the hypothesized value of the coefficient (often 0). If the estimated value of the coefficient is in the thin part of the tails of this distribution, then the hypothesized value is probably not correct.

Confidence Intervals

Confidence interval - a range in which we are confident that the true coefficient must fall given the estimated coefficient.

The range is determined by three factors:

- 1) the estimated coefficient,
- 2) the standard error of the estimate,
- 3) the critical t-value needed for our desired confidence level.

It is given by the following expression:

$$\hat{\beta}_j - t_{crit} * se(\hat{\beta}_j) \leq \hat{\beta}_j \leq \hat{\beta}_j + t_{crit} * se(\hat{\beta}_j)$$

Where t is the critical value associated with the desired level of confidence for our interval.

Confidence Interval

$$\widehat{CEO\text{Salary}} = 432,751 + 0.258 * Profits + 0.021 * Revenue$$
$$(72,000) \quad (0.150) \quad (0.005)$$

What is the 95 percent confidence interval for the estimated effect of company profits on CEO salary. Recall that there were 60 companies included in the regression.

$$t_{57}(95\%) \approx 2.00$$

$$0.258 - 2.00 * 0.150 \leq \beta_j \leq 0.258 + 2.00 * 0.150$$

$$-0.042 \leq \beta_j \leq 0.558$$

So the true value of B_j falls in the range [-0.042, 0.558] with probability of 95 percent.

Stata

```
regress savings income age
```

Source	SS	df	MS	Number of obs	=	100
Model	69993876.4	2	34996938.2	F(2, 97)	=	3.40
Residual	998273721	97	10291481.7	Prob > F	=	0.0374
Total	1.0683e+09	99	10790581.8	R-squared	=	0.0655
				Adj R-squared	=	0.0463
				Root MSE	=	3208

sav	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
inc	.1555624	.0596697	2.61	0.011	.0371346 .2739903
age	-26.72823	45.03278	-0.59	0.554	-116.1058 62.64938
_cons	1072.28	1726.415	0.62	0.536	-2354.176 4498.736

Note that magnitude and statistical significance are very different things.

Research

Academics are lazy and use stars to help indicate significance:

Promotion Effects: Separate Polynomials

	(1) Separate Linear	(2) Separate Quadratic	(3) Separate Cubic
Long-Run Effects			
Future Tournaments/Yr	1.857** (0.736)	0.946 (0.937)	1.013 (1.124)
AIC Model Rank	1	2	3
Future Earnings/Yr			
	191,948*** (42,228)	152,446*** (53,796)	139,288** (64,520)
AIC Model Rank	1	2	3

* 10 percent significance

** 5 percent significance

*** 1 percent significance

Regression Coefficients

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + u$$

□ Statistical significance:

- depends on the magnitude of B_j and the std errors
- it is a statistical fact
- increases as you add observations (reduces std errors)

□ Economic significance:

- only relates to the magnitude of B_j
- subjective assessment of whether the explanatory variable is an important determinant of Y
- may be statistically insignificant due to small sample size

Joint tests

Suppose we want to test if a group of explanatory variables (e.g. X_2 and X_3) has an effect on the outcome variable.

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + u$$

Null hypothesis: $H_0 : \beta_3 = 0, \beta_4 = 0$

Alternative hypothesis: $H_A : H_0$ is not true

Unfortunately, you can't simply do this by testing each coefficient separately. [How would we know if the combined effect is significant at the 95%?]

Joint tests

Here is how we handle this:

Run the unrestricted model. Then compute the RSS_{ur} :

$$SSR_{ur} = \sum_{i=1}^N (Y_i - \hat{Y}_{ur})^2$$

Run the restricted model. Then compute the RSS_r :

$$SSR_r = \sum_{i=1}^N (Y_i - \hat{Y}_r)^2$$

The difference in these values gives us a measure of whether X_3 and X_4 are important determinants of Y . [Note that $\text{SSR}_r \geq \text{SSR}_{\text{ur}}$ always.]

Joint tests

So, we have measures of how well the regression line fits with and without X_3 and X_4 . We use this to construct the F-statistic (which performs the same function as the t-statistic). Then we compare to the critical F-values.

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n - k)} \sim F_{q,n-k}$$

q = the number of restrictions in numerator (two in our example)

$n-k$ = the degrees of freedom denominator ($n-4$ in our example)

The F-statistic is larger when the omitted variables matter.

The F-statistic gets smaller when we omit a lot of variables
(it is measuring the average benefit per variable).

Joint tests

$$\ln(\widehat{\text{salary}}) = 11.19 + .069\text{years} + 0.002\text{bavg} + 0.015\text{HmRuns}$$
$$(.290) \quad (.0121) \quad (0.001) \quad (0.016)$$

We want to test if batting average and home runs are jointly predictive of salary at the 95% confidence interval.

$$H_0 : \beta_3 = 0, \beta_4 = 0$$

$$H_A : H_0 \text{ is not true}$$

In STATA:

```
test bavg HmRuns
```

Joint tests

Find the SSR for the restricted and unrestricted model

$$\ln(\widehat{\text{salary}}) = 11.19 + .069\text{years} + 0.002\text{bavg} + 0.015\text{HmRuns}$$

$$N=350, \text{SSR}_{ur} = 190.2$$

$$\ln(\widehat{\text{salary}}) = 11.22 + .0713\text{years}$$

$$N=350, \text{SSR}_r = 198.3$$

Compute the F-Statistic:

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n - k - 1)} = \frac{(198.3 - 190.2)/2}{190.2/(350 - 3 - 1)} = 7.39$$

Using the F-table, we see that we can reject the null hypothesis at the 5% level.

Testing Linear Restrictions

Sometimes we want to impose a restriction on the coefficients.

$$Q = AL^\alpha K^\beta u$$

$$Q = A(cL)^\alpha (cK)^\beta u$$

$$Q = Ac^{\alpha+\beta} L^\alpha K^\beta u$$

Constant returns to scale means that $\alpha+\beta=1$. We might be interested in imposing this restriction on the model:

1. We may have a reason to believe it is true.
2. We may want to test if it is true.

Testing Linear Restrictions

We impose the linear restriction and estimate the restricted model.
Specifically, we force $\alpha+\beta=1$ (equivalently, $\alpha=1-\beta$):

$$\begin{aligned} Q &= AL^\alpha K^\beta u \\ \Rightarrow \ln(Q) &= \ln(A) + \alpha \ln(L) + \beta \ln(K) + u \\ \Rightarrow \ln(Q) &= \ln(A) + (1 - \beta) \ln(L) + \beta \ln(K) + u \\ \Rightarrow \ln(Q) - \ln(L) &= \ln(A) + \beta [\ln(K) - \ln(L)] + u \\ \Rightarrow Q^* &= \ln(A) + \beta K^* + u \end{aligned}$$

We estimate the restricted regression of Q^* on K^* to get β .

1. The restricted regression can be used to get estimates:
 - we have estimated β
 - we can estimate $\alpha = 1 - \beta$
2. We can use the restricted regression to test if the restriction is valid.

Testing Linear Restrictions

To test if the linear restrictions are valid using an F-Test:

1. The null hypothesis is that the restriction is valid: $\alpha + \beta = 1$
2. Estimate the unrestricted and restricted model:

$$\text{Unrestricted : } \ln(Q) = \ln(A) + \alpha \ln(L) + \beta \ln(K) + u$$

$$\text{Restricted : } Q^* = \ln(A) + \beta K^* + u$$

3. Calculate the F-statistic:

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n - k)} \sim F_{q, n-k}$$

4. Compare the F-statistic to the critical value for the appropriate number of restrictions and number of observations.

Comparing coefficients

Suppose we want to test if two coefficients in a regression are equal.

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

Null hypothesis: $H_0 : \beta_2 = \beta_3$

Alternative hypothesis: $H_A : \beta_2 \neq \beta_3$

We can rewrite the null hypothesis like this:

$$H_0 : \beta_2 - \beta_3 = 0$$

One option is to treat this as a linear restriction and test using the F-test approach we just examined. The restrict model would just be:

$$Y = \beta_1 + \beta_2(X_2 + X_3) + u$$

Comparing coefficients

There are other options:

1. STATA can compute this after you have run the unrestricted regression:

```
test inc = age
```

2. Manipulate the regression equation to give us what we want

Comparing coefficients

We define a new coefficient that we are interested in:

$$\theta = \beta_2 - \beta_3$$

This can be written as:

$$\beta_2 = \theta + \beta_3$$

Plug this into our regression and regroup terms:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

$$Y = \beta_1 + (\theta + \beta_3) X_2 + \beta_3 X_3 + u$$

$$Y = \beta_1 + \theta X_2 + \beta_3 (X_2 + X_3) + u$$

So, we can estimate Θ and its standard error by regressing Y on X_2 and $X_2 + X_3$. Then we can easily perform hypothesis tests on it.

Comparing coefficients

$$wage = \beta_0 + \beta_1 YrsPrim + \beta_2 YrsColl + u$$

We want to test if the effect of a year of primary (or secondary) school on wages is the same as a year of college on wages.

We can make a new variable that is the total years of any kind of education, including college.

$$YrsEduc = YrsPrim + YrsColl$$

$$wage = \beta_0 + \beta_1 YrsEduc + \theta YrsColl + u$$

The coefficient on years of education is the average return per year of primary education. The coefficient on years of college is the amount that the return to a year of college differs from this amount.

Comparing coefficients

$$\widehat{wage} = 5.00 + 1.25YrsPrim + 2.30YrsColl$$

(1.25) (0.35) (1.02)

$$\widehat{wage} = 5.00 + 1.25YrsEduc + 1.05YrsColl$$

(1.25) (0.35) (0.85)

- How much is a year of college estimated to increase wages in the first regression?
- How much is a year of college estimated to increase wages in the second regression?
- Is there a statistically significant difference in the returns to a year of primary education and a year of college education at the 95% confidence level?

Comparing coefficients

Example 2: Suppose we want to know if male and female employee have a different effect on company profits:

$$\text{Profits} = \beta_1 + \beta_2 \text{MaleEmployees} + \beta_3 \text{FemaleEmployees} + u$$

$$\text{Profits} = \beta_1 + \beta_2 \text{Employees} + \theta \text{FemaleEmployees} + u$$

Example 3: Suppose we want to know if people are more likely to spend money if they get it as a gift relative to earned income:

$$\text{Expenditure} = \beta_1 + \beta_2 \text{EarnedIncome} + \beta_3 \text{Gifts} + u$$

$$\text{Expenditure} = \beta_1 + \beta_2 \text{TotIncome} + \theta \text{Gifts} + u$$

Comparing Models

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \dots + \beta_k X_{ik} + u_i$$

Should we add or delete an explanatory variable X_k :

- Single variable*: check the t-statistic for the estimate of the coefficient B_k .
- Multiple variables*: conduct an F-test of joint significance of the variables.

These tests are great for nested models. That is, when one of the models is different because it has some coefficients set to 0.

For models that are fundamentally different, this does not work. For example, models with completely different sets of explanatory variables or interactions between variables.

Comparing Models

- There are many alternative measures for comparing model fit. All of them are a bit ad-hoc.
- The measures consider average RSS (i.e. RSS per observation), so smaller values are better.
- The measures penalize using more explanatory variables, so there is a fit/variables tradeoff.

Akaike Information Criterion

$$AIC = \left(\frac{RSS}{N}\right)e^{2k/N}$$

Schwarz Bayesian Criterion

$$SBC = \left(\frac{RSS}{N}\right)e^{k/N}$$

Finite Prediction Error

$$FPE = \left(\frac{RSS}{N}\right)\left(\frac{n+k}{n-k}\right)$$

Hannan and Quin Criterion

$$HQC = \left(\frac{RSS}{N}\right)(\ln N)^{2k/n}$$

Lesson 7

Multicollinearity & Dummy Variables

Outline

Previous Lesson

1. Hypothesis testing with regression coefficients
2. Comparing coefficients & joint hypothesis tests
3. Comparing models

[We have covered all of Chapters 1-4]

This Lesson

1. Multicollinearity [Chapter 5]
2. Dummy Variables [Chapter 9]
3. Interaction Terms

Next Lesson:

1. Heteroscedasticity

Dummy Variables

Dummy variables – (aka binary variables) take on the value 0 or 1, where 1 typically indicates “Yes” and 0 indicates “No”.

Examples:

Gender: male = 1 if male, 0 if female
female = 1 if female, 0 if male

Race: Asian, black, hispanic, native American, white

Time: year1990, year1991, year1992,...
quarter1, quarter2, quarter3, quarter4
Jan, Feb, Mar,...
Mon, Tue, Wed,...

Groups: countries, states, companies

Perfect Multicollinearity

Perfect multicollinearity – occurs when an explanatory variable is a linear function of the other explanatory variables (i.e. the explanatory variables are linearly dependent).

Note: not having multicollinearity between the X variables is one of the core assumptions needed to use OLS.

Example:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

$$X'_2 = [1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6]$$

$$X'_3 = [2 \quad 4 \quad 6 \quad 8 \quad 10 \quad 12]$$

Perfect Multicollinearity

Formally, two variables X_2 and X_3 are linearly dependent if the following expression is satisfied with two non-zero coefficients:

$$\delta_2 X_2 + \delta_3 X_3 = 0$$

More generally, there is perfect multicollinearity in a regression if the following expression can be satisfied by two or more non-zero coefficients (where X_1 is the intercept = 1):

$$\delta_1 X_1 + \delta_2 X_2 + \delta_3 X_3 + \cdots + \delta_n X_n = 0$$

In our example: $X'_2 = [1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6]$

$$X'_3 = [2 \quad 4 \quad 6 \quad 8 \quad 10 \quad 12]$$

So, $\delta_2 = 2$ and $\delta_3 = -1$ would work (there are many options).

Perfect Multicollinearity

Dummy variable trap – the most common source of perfect multicollinearity is when you add dummy variables to a regression.

Gender in a regression:

$$Y = \beta_1 + \beta_2 male + \beta_3 female + u$$

This regression has a multicollinearity problem if the following is satisfied:

$$\delta_1 + \delta_2 male + \delta_3 female = 0$$

This is satisfied by $\delta_1 = 1$, $\delta_2 = -1$, and $\delta_3 = -1$ (and many others).

So, if we include the intercept, male, and female in the regression, then we will have perfect multicollinearity. Intuitively, adding “female” does not add any explanatory power if “male” is included (and vice-versa)

This is true any time you include the dummy variables for all possible (non-overlapping) groups.

Perfect Multicollinearity

Why is multicollinearity a problem?

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

$$Y = \beta_1 + \beta_2 X_2 + \beta_3(\alpha_1 + \alpha_2 X_2) + u$$

$$Y = \beta_1 + \beta_3 \alpha_1 + (\beta_2 + \beta_3 \alpha_2) X_2 + u$$

$$Y = \beta_1^* + \beta_2^* X_2 + v$$

Estimating the last equation will give us only two coefficient estimates, which is not sufficient to uniquely determine B_1 , B_2 , and B_3 :

$$\hat{\beta}_1^* = \hat{\beta}_1 + \hat{\beta}_3 \alpha_1$$

$$\hat{\beta}_2^* = \hat{\beta}_2 + \hat{\beta}_3 \alpha_2$$

Specifically:

$$\hat{\beta}_1 = \hat{\beta}_1^* - \hat{\beta}_3 \alpha_1$$

$$\hat{\beta}_2 = \hat{\beta}_2^* - \hat{\beta}_3 \alpha_2$$

So for any value of \hat{B}_3 , we can find a \hat{B}_1 and \hat{B}_2 that will work. Thus we have an infinite number of solutions.

Perfect Multicollinearity

In matrix form, this means that one of the rows of \mathbf{X} is a linear combination of the other rows. This means that \mathbf{X} is singular. It also means that $\mathbf{X}'\mathbf{X}$ is **singular** (i.e. it is degenerate or has a determinant = 0). That is, it is impossible to find the inverse because it does not exist.

Note:
$$\mathbf{A}^{-1} = \frac{1}{|A|} * [adj A]$$

Recall the OLS regression:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{|\mathbf{X}'\mathbf{X}|} [adj \mathbf{X}'\mathbf{X}]$$

Since the determinant of $\mathbf{X}'\mathbf{X}$ is 0, then $(\mathbf{X}'\mathbf{X})^{-1}$ is undefined and we can't compute the OLS estimate of \mathbf{B} .

Perfect Multicollinearity

Dummy variable trap – avoiding this problem when coding involves either excluding one of the dummy variables yourself or letting the program exclude one of the variables.

Let STATA drop one:

```
reg Y male female  
reg Y asian black hispanic nativeamer white
```

Drop one yourself:

```
reg Y female  
reg Y asian black hispanic nativeamer
```

There are also commands that generate dummy variables for you (e.g. if you have a variable “state” that takes on 50 values). You can then tell STATA which of the states to omit the dummy for.

Perfect Multicollinearity

Dummy variable trap – the interpretation of the remaining dummy variables is relative to the omitted group (i.e. the omitted group is the intercept)

Gender: We have data for male and female students:

$$wage = 12.75 - 1.75female$$

Race: Suppose we have data for asian, black, hispanic, native american, and white students:

$$wage = 11.50 + 1.42asian + 0.34black - 0.72hispanic + 0.11natamer$$

Time: Suppose we have data for years 2000-2005.

$$wage = 8.25 + 0.22Yr01 + 0.37Yr02 + 0.52Yr03 + 0.94Yr04 + 0.70Yr05$$

Imperfect Multicollinearity

If an econometric model has a perfect multicollinearity problem, then it means that the econometrician is careless.

However, there are many cases where we want to estimate the effect of a variable that is highly correlated with another variable. This may cause a multicollinearity problem.

Suppose we are interested in estimating the relationship between average household spending, average house prices, and average income.

Example 1: We may have time series data for the US from 1900 to 2000.

$$spending_{year} = \beta_1 + \beta_2 houseprices_{year} + \beta_3 income_{year} + u_{year}$$

Example 2: We may have cross-sectional data on US states for a single year.

$$spending_{state} = \beta_1 + \beta_2 houseprices_{state} + \beta_3 income_{state} + u_{state}$$

Imperfect Multicollinearity

Why this is a problem?

Recall the equation for the variance: $Var(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$

When there is a high level of multicollinearity, the determinant of a matrix will be close to 0: $|\mathbf{X}'\mathbf{X}| \approx 0$

And thus: $(\mathbf{X}'\mathbf{X})^{-1}$ is large

Which means that the variance will be large. This means that the estimates of B will be estimated with very large standard errors. Thus:

1. Estimated coefficients may be imprecise (and may even have the wrong sign).
2. Estimates may be sensitive to alternative specifications or to dropping a few observations.
3. Unlikely that we will have statistically significant estimates.

Imperfect Multicollinearity

The variance of an OLS estimate can be written as follows (we will not derive this):

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{\sum(X_j - \bar{X}_j)^2} \frac{1}{(1 - R_j^2)}$$

The Variance Inflation Factor is the second term:

$$VIF_j = \frac{1}{(1 - R_j^2)}$$

Where R_j^2 is the R-squared from a regression of X_j on all the other X 's, called the auxiliary regression. The variance of the coefficient on X_j depends on how much of its variation is explained by other explanatory variables.

As the R^2 increases, the variance blows up.

In the case of perfect multicollinearity, obviously the R^2 would be 1 and thus the denominator would be 0.

Imperfect Multicollinearity

R_j^2	VIF_j
0	1
0.5	2
0.8	5
0.9	10
0.95	20
0.975	40
0.99	100
0.995	200
0.999	1000

How much is too much collinearity? This is a subjective. Book says VIF=10.

Note that adding observations (larger N) reduces std. errors. Thus you might be able to have a lot of collinearity and be OK with a very large number of observations.

In time series you often have limited observations and collinearity, so this problem is most common in that context.

Detecting Multicollinearity

Methods of detecting problematic multicollinearity:

1. Find the correlation of two variables. High correlation could indicate a problem:
 - $\text{corr } X_2 \ X_3$
2. Run the auxiliary regression of one of the X_j on all of the other Xs and see what the R-squared is. A high value could suggest a problem.
 - $\text{reg } X_1 \ X_2 \ X_3$
3. Look for coefficients that change dramatically or, generally, use common sense when examining explanatory variables and look for coefficients with std errors that seem too large given the size of the data set.

Detecting Multicollinearity

Consider data for the UK that includes imports, GDP, the consumer price index, and the producer price index.

Table 5.6 Correlation matrix

	<i>IMP</i>	<i>GDP</i>	<i>CPI</i>	<i>PPI</i>
<i>IMP</i>	1.000000	0.979713	0.916331	0.883530
<i>GDP</i>	0.979713	1.000000	0.910961	0.899851
<i>CPI</i>	0.916331	0.910961	1.000000	0.981983
<i>PPI</i>	0.883530	0.899851	0.981983	1.000000

Note that the CPI and PPI are very highly correlated.

Detecting Multicollinearity

Table 5.7 First model regression results (including only *CPI*)

Dependent variable: LOG(IMP)

Method: least squares

Date: 02/17/04 *Time:* 02:16

Sample: 1990:1 1998:2

Included observations: 34

Variable	Coefficient	Std. error	t-statistic	Prob.
<i>C</i>	0.631870	0.344368	1.834867	0.0761
LOG(GDP)	1.926936	0.168856	11.41172	0.0000
LOG(CPI)	0.274276	0.137400	1.996179	0.0548
<i>R-squared</i>	0.966057	Mean dependent var		10.81363
Adjusted <i>R-squared</i>	0.963867	S.D. dependent var		0.138427
S.E. of regression	0.026313	Akaike info criterion		-4.353390
Sum squared resid	0.021464	Schwarz criterion		-4.218711
Log likelihood	77.00763	F-statistic		441.1430
Durbin-Watson stat	0.475694	Prob(F-statistic)		0.000000

Detecting Multicollinearity

Table 5.9 Third model regression results (including only *PPI*)

Dependent variable: LOG(IMP)

Method: least squares

Date: 02/17/04 Time: 02:22

Sample: 1990:1 1998:2

Included observations: 34

Variable	Coefficient	Std. error	t-statistic	Prob.
<i>C</i>	0.685704	0.370644	1.850031	0.0739
<i>LOG(GDP)</i>	2.093849	0.172585	12.13228	0.0000
<i>LOG(PPI)</i>	0.119566	0.136062	0.878764	0.3863
<i>R-squared</i>	0.962625	Mean dependent var		10.81363
Adjusted <i>R-squared</i>	0.960213	S.D. dependent var		0.138427
S.E. of regression	0.027612	Akaike info criterion		-4.257071
Sum squared resid	0.023634	Schwarz criterion		-4.122392
Log likelihood	75.37021	<i>F</i> -statistic		399.2113
Durbin-Watson stat	0.448237	Prob(<i>F</i> -statistic)		0.000000

Detecting Multicollinearity

Table 5.8 Second model regression results (including both *CPI* and *PPI*)

Dependent variable: LOG(IMP)

Method: least squares

Date: 02/17/04 Time: 02:19

Sample: 1990:1 1998:2

Included observations: 34

Variable	Coefficient	Std. error	t-statistic	Prob.
<i>C</i>	0.213906	0.358425	0.596795	0.5551
<i>LOG(GDP)</i>	1.969713	0.156800	12.56198	0.0000
<i>LOG(CPI)</i>	1.025473	0.323427	3.170645	0.0035
<i>LOG(PPI)</i>	-0.770644	0.305218	-2.524894	0.0171
<i>R-squared</i>	0.972006	Mean dependent var		10.81363
Adjusted <i>R-squared</i>	0.969206	S.D. dependent var		0.138427
S.E. of regression	0.024291	Akaike info criterion		-4.487253
Sum squared resid	0.017702	Schwarz criterion		-4.307682
Log likelihood	80.28331	F-statistic		347.2135
Durbin-Watson stat	0.608648	Prob(F-statistic)		0.000000

What happens to the coefficients on CPI and PPI?

What happens to their standard errors?

Note: we appear to be splitting a 0.

Dummy Variables

Dummy variables (often called fixed effects) shift the intercept:

$$wage = \beta_1 + \beta_2 educ + \beta_3 female + u$$

$$GDPgr = \beta_1 + \beta_2 prcman + \beta_3 EU + u$$

What are they good for:

1. Test to see if there are differences between groups (pos or neg).
2. May produce more valid coefficients on other variables (e.g. B_2)

Dummy Variables

Figure 9.1

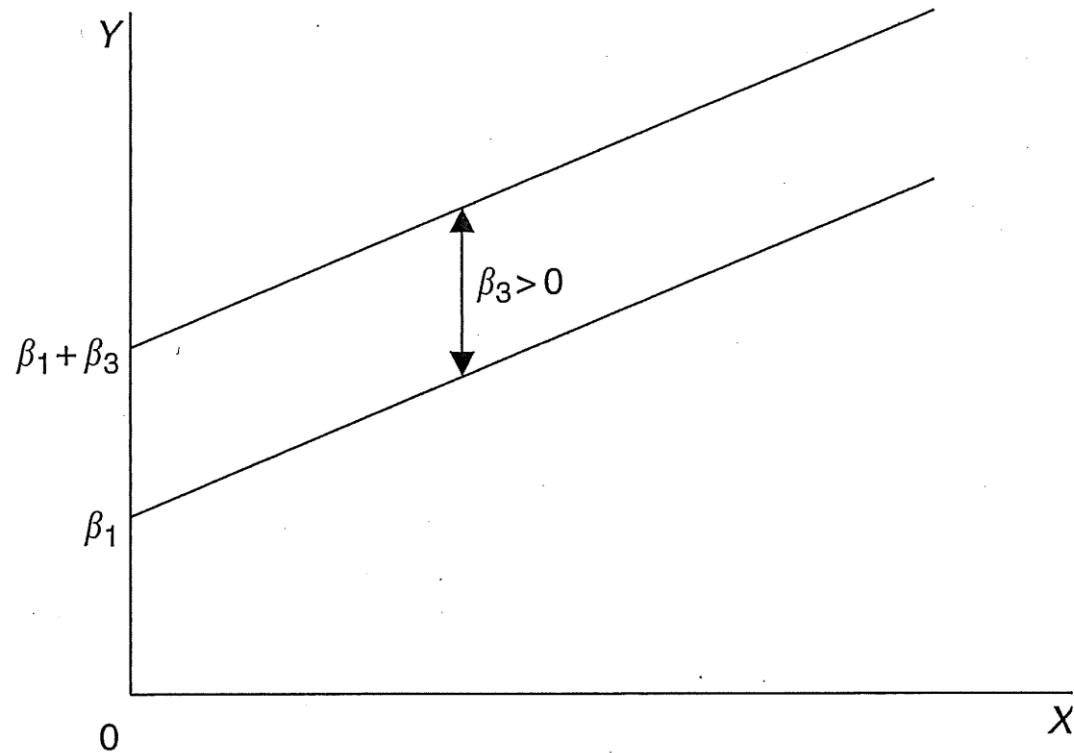


Figure 9.1 The effect of a dummy variable on the constant of the regression line

Dummy Variables

Figure 9.2

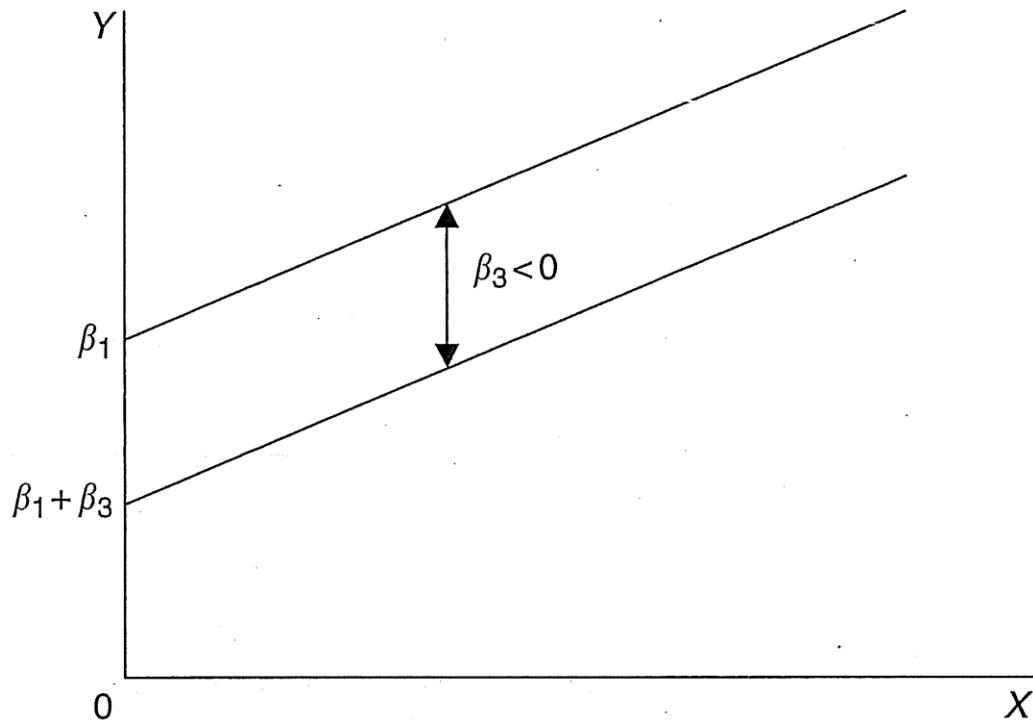
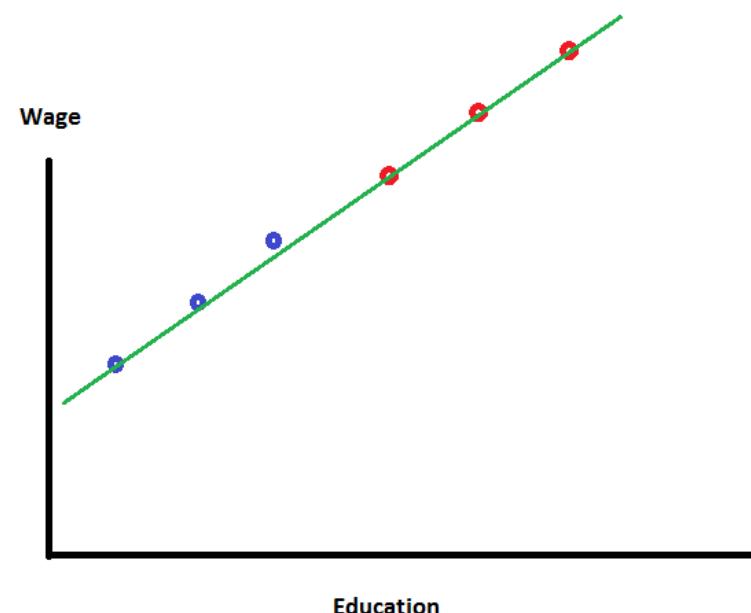
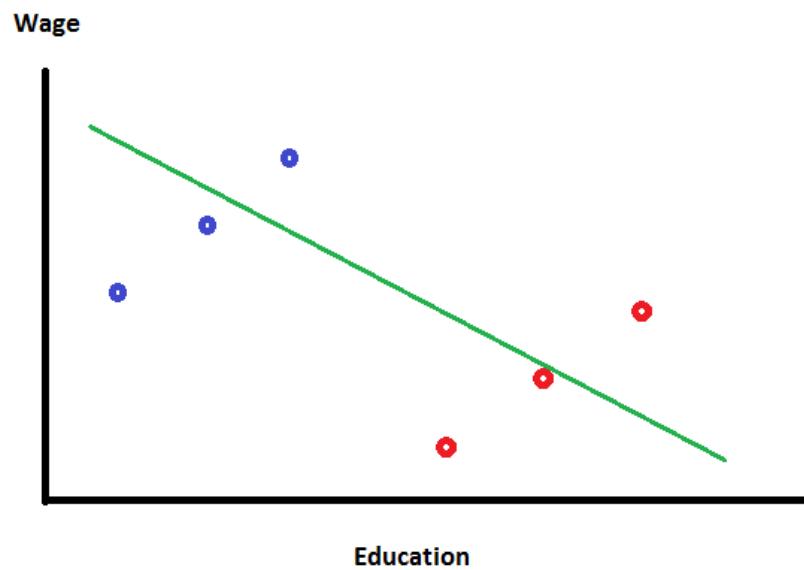


Figure 9.2 The effect of a dummy variable on the constant of the regression line

Fixed Effects: Cross-Section

Here is what the data might look like. Men are in blue and women are in red.



Without the fixed effect, we could actually get the wrong sign.

Dummy Variables

Interaction terms – Including the product of two explanatory variables as an additional explanatory variable.

In the case where one of the interacted variables is a dummy, these are also called: dummy slope effects, heterogeneous effects.

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 D + \beta_4 D * X_2 + u$$

In this regression, B_2 represents the effect of X_2 on Y for a person with $D=0$, while B_2+B_4 represents the effect of X_2 on Y for a person with $D=1$. That is, B_4 is the difference in the effect for people with $D=1$ relative to $D=0$.

Mathematically:

$$\frac{\partial Y}{\partial X_2} = \beta_2 \text{ if } D=0$$

$$\frac{\partial Y}{\partial X_2} = \beta_2 + \beta_4 \text{ if } D=1$$

Dummy Variables

Interaction terms – You must always include the interacted terms separately in the regression. These are called the main effects.

Failure to do this will result in biased coefficients (the main effects are omitted variables, which we will discuss more later).

$$wage = \beta_1 + \beta_2 educ + \beta_3 female + \beta_4 female * educ + u$$

What is the interpretation of B_4 ?

$$GDPgr = \beta_1 + \beta_2 prcman + \beta_3 EU + \beta_4 EU * prcman + u$$

What is the interpretation of B_4 ?

Dummy Variables

Figure 9.5

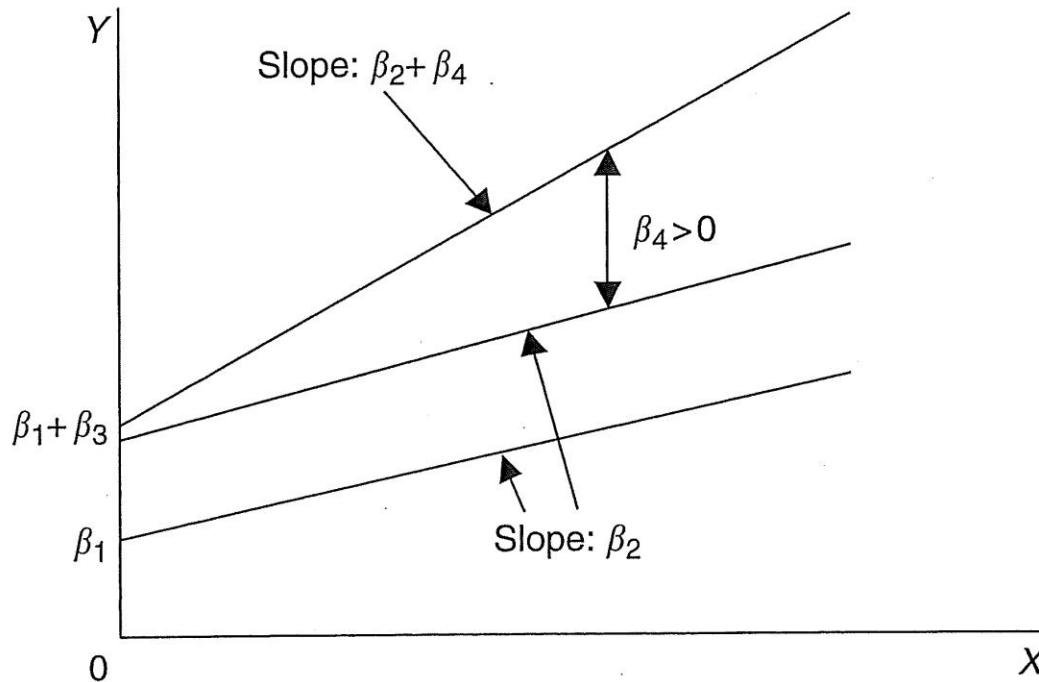


Figure 9.5 The combined effect of a dummy variable on the constant and the slope of the regression line

Dummy Variables

- We can use multiple categories of dummies in one regression.
Consider the following specifications of the determinants of wages:

$$wage = \beta_1 + \beta_2 fem + \beta_3 asian + \beta_4 black + \beta_5 hisp + \beta_6 exp + u$$

What is the predicted wage of an asian male?

What is the predicted wage of an asian female?

- We can let the effect of experience differ by gender and race:

$$\begin{aligned} wage = & \beta_1 + \beta_2 fem + \beta_3 asian + \beta_4 black + \beta_5 hisp + \beta_6 exp \\ & + \beta_7 fem * exp + \beta_8 asian * exp + \beta_9 black * exp + \beta_{10} hisp * exp + u \end{aligned}$$

What is the predicted wage of a white male?

What is the predicted wage of a hispanic female?

What is the effect of one more year of experience for a hispanic female?

Dummy Variables

Figure 9.6

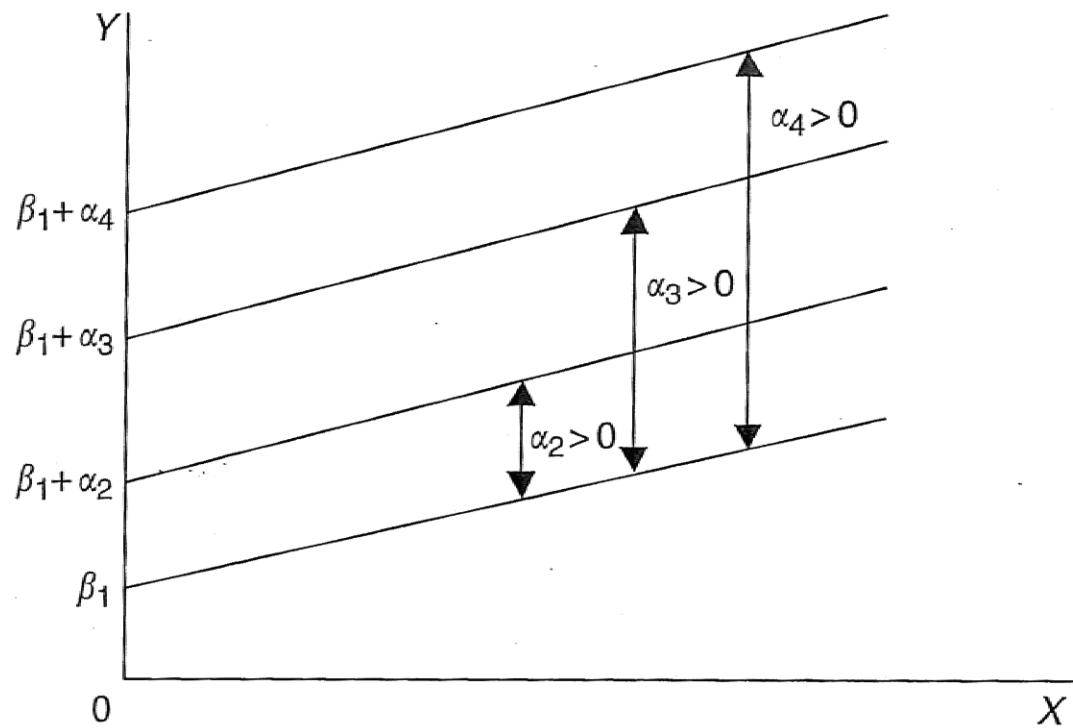


Figure 9.6 The effect of a dummy variable on the constant of the regression line

Dummy Variables

Dummies can overlap in a category. Suppose we assign a value of 1 to each level of education a person completed:

$$wage = \beta_1 + \beta_2 compHS + \beta_3 compBA + \beta_4 compMA + u$$

What is the wage of someone who completed a BA?

What is the wage of someone who completed an MA?

In this specification, β_4 is the additional wage effect of completing an MA (above the return of a HS degree and a BA).

Dummy Variables

Often times you will want to turn a single variable into many dummy variables in the data:

Race = 1, 2, 3, 4, 5

State = “AK”, “AL”,

Manually:

```
gen alaska=0;
```

```
replace alaska=1 if state ==“AK”;
```

STATA shortcuts:

```
reg wage exp i.state
```

```
areg wage exp, absorb(state)
```

```
xtset panelvar
```

```
xtreg
```

Lesson 8

Heteroskedasticity

Outline

Previous Lesson

1. Multicollinearity [Chapter 5]
2. Dummy Variables [Chapter 9]
3. Interaction Terms

This Lesson:

1. Heteroscedasticity: Detection
2. Heteroscedasticity: Correction

Next Lesson:

1. Autocorrelation

Heteroskedasticity

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

Homoskedasticity – When all of the errors have the same variance.

$$\text{var}(u_i | \mathbf{X}) = \sigma^2$$

$$\text{var}(\mathbf{u} | \mathbf{X}) = \sigma^2 \mathbf{I}$$

Heteroskedasticity – When the errors have different variances. More testably, when the variance of the errors is correlated with the explanatory variables.

$$\text{var}(u_i | \mathbf{X}) = \sigma_i^2$$

$$\text{var}(\mathbf{u} | \mathbf{X}) \neq \sigma^2 \mathbf{I}$$

Heteroskedasticity

Recall that homoskedasticity was crucial for deriving the variance of \hat{B} and thus for two important results:

1. Proving that OLS was the best (BLUE)
2. Inference (aka hypothesis testing such as t-tests...)

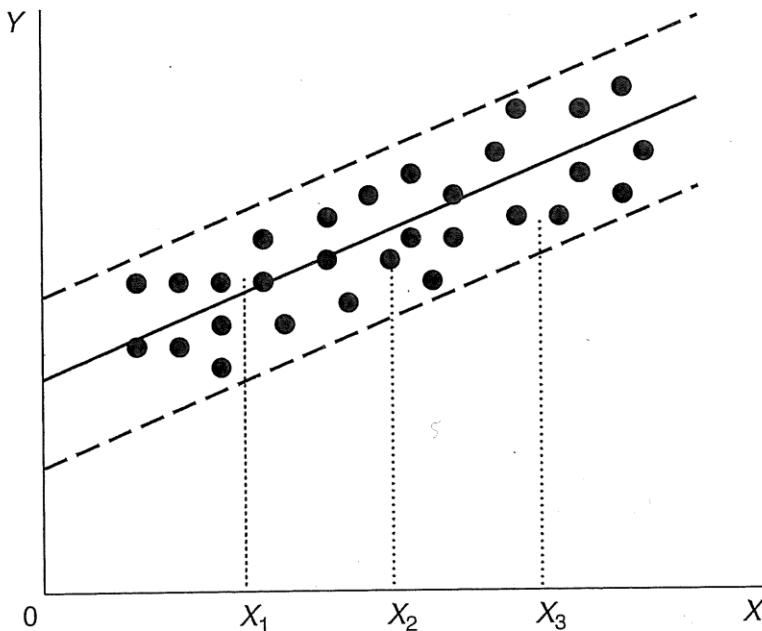


Figure 6.1 Data with a constant variance

Heteroskedasticity

Heteroskedasticity exists when the variance of the errors is larger or smaller for larger values of the explanatory variable.

That is, that the prediction of Y is worse for large or small values of X.

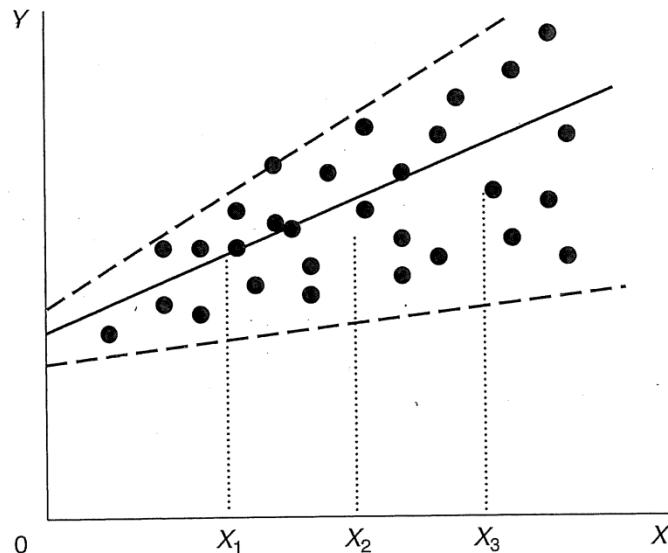


Figure 6.2 An example of heteroskedasticity with increasing variance

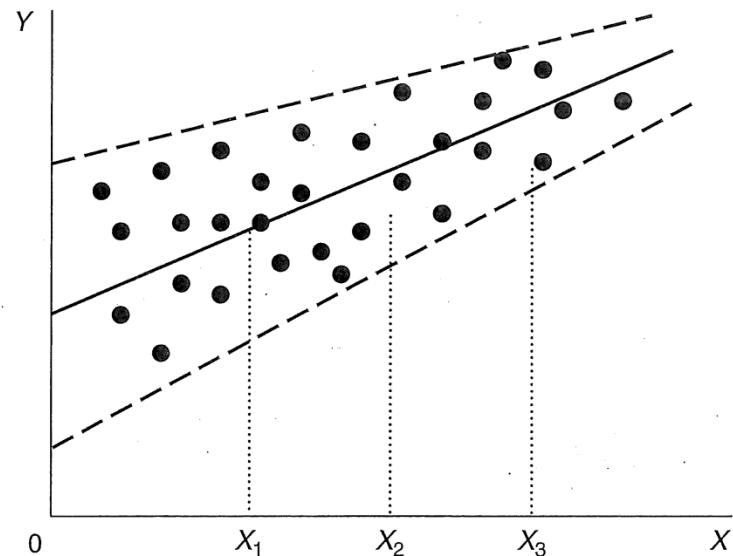


Figure 6.3 An example of heteroskedasticity with falling variance

Heteroskedasticity

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_k X_{ki} + u_i$$

If there is heteroskedasticity, what are the implications?

1. OLS is still unbiased. The $E(u|X)$ is still 0, so the estimates \hat{B} are still centered around the true B .
2. However, OLS is no longer the most efficient estimator of B . An alternative approach can produce B 's with smaller variance.
3. OLS estimates of variance are wrong:

$$\text{var}(\hat{\beta}) \neq \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

Thus inference (aka hypothesis testing) is wrong because the t-stats (and other stats) are wrong. The usual variance is usually too small (optimistic).

Heteroskedasticity

The variance of the estimators under homoskedasticity:

$$var(\hat{\beta}) = \frac{\sum(X_i - \bar{X})^2 \sigma^2}{[\sum(X_i - \bar{X})^2]^2} = \frac{\sigma^2}{\sum(X_i - \bar{X})^2}$$

The variance of the estimators under heteroskedasticity:

$$var(\hat{\beta}) = \frac{\sum(X_i - \bar{X})^2 \sigma_i^2}{[\sum(X_i - \bar{X})^2]^2}$$

Thus heteroskedasticity can produce standard errors that are either larger or smaller:

- larger if larger errors occur with X's further from the mean***
- smaller if larger errors occur with X's closer to the mean

Heteroskedasticity

Consider the variance-covariance matrix for the errors:

$$E(\mathbf{u}\mathbf{u}') = \begin{bmatrix} E[u_1u_1] & E[u_1u_2] & E[u_1u_3] & \cdots & E[u_1u_N] \\ E[u_2u_1] & E[u_2u_2] & E[u_2u_3] & \cdots & E[u_2u_N] \\ E[u_3u_1] & E[u_3u_2] & E[u_3u_3] & \cdots & E[u_3u_N] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ E[u_Nu_1] & E[u_Nu_2] & E[u_Nu_3] & \cdots & E[u_Nu_N] \end{bmatrix}$$

$$= \begin{bmatrix} var[u_1] & cov[u_1u_2] & cov[u_1u_3] & \cdots & cov[u_1u_N] \\ cov[u_2u_1] & var[u_2] & cov[u_2u_3] & \cdots & cov[u_2u_N] \\ cov[u_3u_1] & cov[u_3u_2] & var[u_3] & \cdots & cov[u_3u_N] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ cov[u_Nu_1] & cov[u_Nu_2] & cov[u_Nu_3] & \cdots & var[u_N] \end{bmatrix}$$

The values on the horizontal represent the variances and the values on the off-diagonals represent the covariances.

Heteroskedasticity

In the case of homoskedasticity we have the same expected variance for all errors [and expected covariance of 0 across errors – i.e. no autocorrelation] :

$$E(\mathbf{u}\mathbf{u}') = \begin{bmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}$$

In the case of heteroskedasticity, we have different variances for the errors.

$$E(\mathbf{u}\mathbf{u}') = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma_3^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma_N^2 \end{bmatrix} = \sigma_i^2 \mathbf{I} = \boldsymbol{\Omega}$$

Heteroskedasticity

Now consider the variances for the OLS coefficients under heteroskedasticity:

$$\begin{aligned} \text{var}(\hat{\beta}) &= E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] \\ &= E\{[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}][(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}]'\} \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{u}\mathbf{u}')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \boxed{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Omega\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}} \end{aligned}$$

This may be very different from the collapsed form that occurs with homoskedasticity:

$$\text{var}(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

Heteroskedasticity

Reality check #1:

Our primary concern is that our OLS standard errors are wrong. Thus we may be claiming that estimates are statistically significant when they are not.

Reality check #2:

There is almost always heteroskedasticity in reality. This can be due to model misspecification (including X when you need X and X^2 , including fixed effects, etc.).

At the end of the lesson, we will discuss the two most common ways of correcting standard errors in STATA (that will be Reality check #3).

Heteroskedasticity

How do we detect heteroskedasticity:

Compute the prediction errors:

$$\hat{u}_i = Y_i - \hat{Y}_i$$

1. Informal visual inspection of the errors:
 - squared predicted errors increase with X
 - squared predicted errors decrease with X
 - squared predicted errors exhibit any relationship with any X

Why do we square or take the absolute value for inspection?

1. Conduct a formal test:
 - Breusch-Pagan LM test
 - close alternatives: Glesjer, Harvey-Godfrey, and Park tests
 - White's test

Heteroskedasticity

Visual inspection – Under homoskedasticity we expect to see that the variance of the errors does not vary with the magnitude of X or Y. In the graph below, note two things:

- magnitude of the predicted squared errors is constant (flat)
- the spread of the predicted squared errors is constant

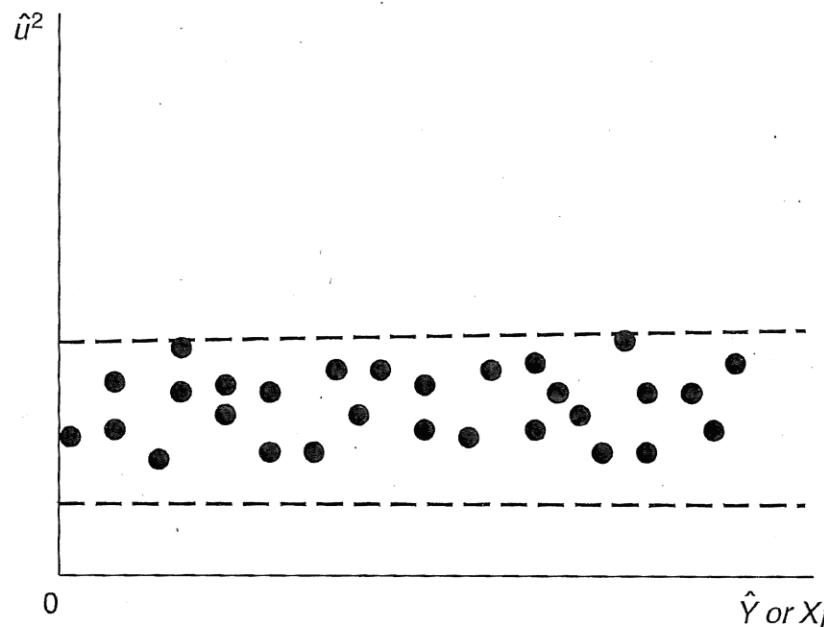


Figure 6.5 A ‘healthy’ distribution of squared residuals

Heteroskedasticity

In the graph on the left, the magnitude of the squares of the prediction errors are increasing in X. In the graph on the right, the squares of the prediction errors are increasing in X and they are becoming more spread out.

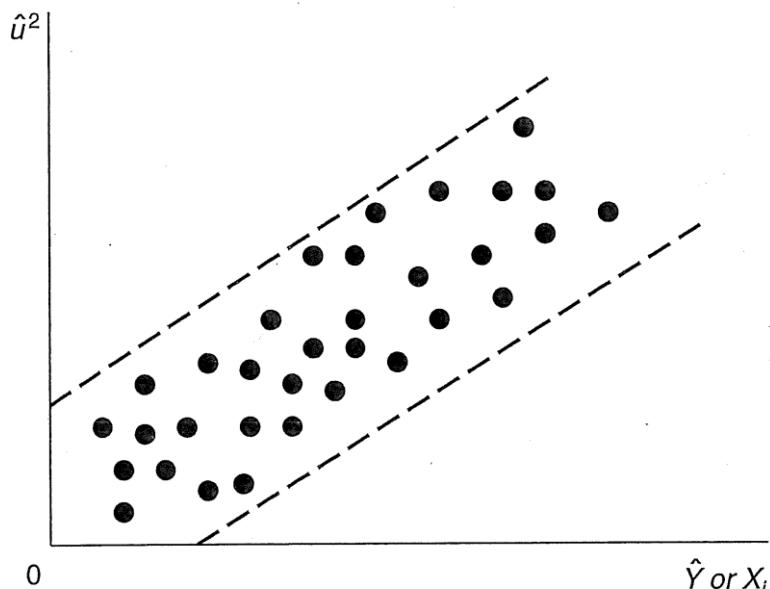


Figure 6.7 Another indication of heteroskedasticity

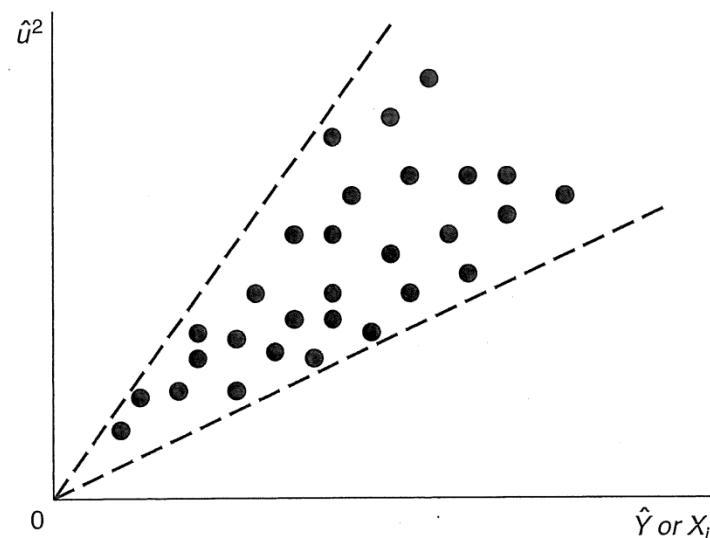


Figure 6.6 An indication of the presence of heteroskedasticity

Heteroskedasticity

Heteroskedasticity can take on a more complicated relationship relative to X. On the left, the squared errors are smaller for more extreme values of X, while on the right, the magnitude of the squared errors increase exponentially with X.

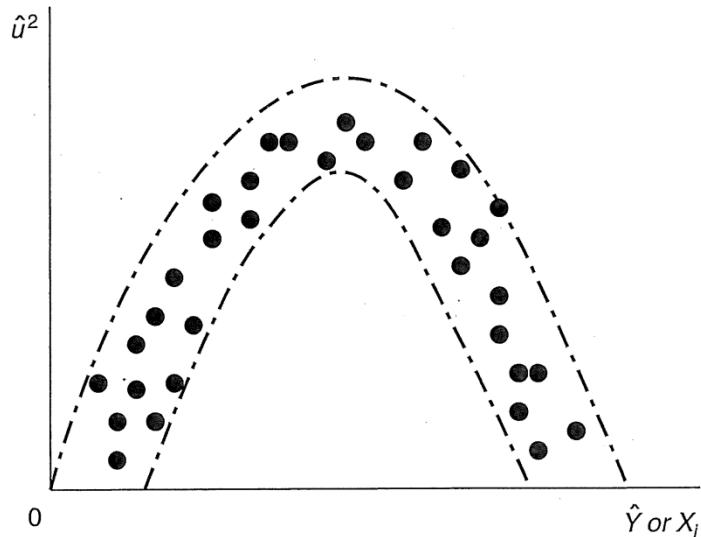


Figure 6.8 A non-linear relationship leading to heteroskedasticity

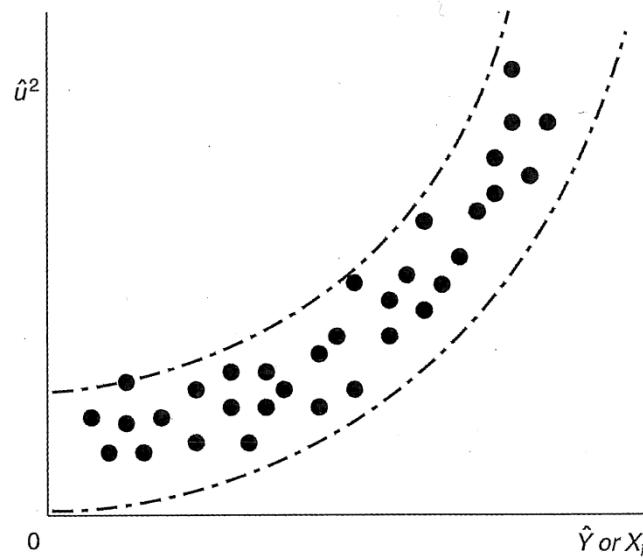


Figure 6.9 Another form of non-linear heteroskedasticity

Heteroskedasticity

Breusch-Pagan LM Test steps - the goal is to test if the errors are correlated with the Xs.

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_k X_{ki} + u_i$$

1. Estimate the regression above and get the predicted error \hat{u}_i
2. Run an auxiliary regression of the square of these estimated prediction errors on all of the explanatory variables to see if the magnitude of the errors is correlated with the X's:

$$\hat{u}_i^2 = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \dots + \alpha_k X_{ki} + v_i$$

3. Test if the alpha coefficients are jointly = 0. Many options for this. One is by computing the LM statistic and comparing to the critical Chi-squared value for k degrees of freedom:

$$LM = N * R^2 \quad \text{vs.} \quad \text{Chi-Squared with } k \text{ freedom}$$

Heteroskedasticity

Breusch-Pagan LM Test in STATA:

```
regress Y X1 X2 X3 ... XK
predict ui, residual
generate ui_squared=ui^2
regress ui_squared X1 X2 X3 ... XK
```

Use the resulting R-squared and the sample size to compute the LM statistic and compare to Chi-Squared.

```
test X1 X2 ... Xk
```

Note again that the last regression is trying to determine if the magnitude of the errors is related to the X's. That is, do the errors get smaller and larger as the X's get larger or smaller.

```
regress Y X1 X2 X3 ... XK
estat hettest
```

Heteroskedasticity

Alternatives: The following tests are identical to the B-P test except that they run different versions of the auxilliary regression in step 2. The idea being that they might do a better job of capturing a relationship between the errors and the X's.

Glesjer LM test:

$$|\hat{u}_i| = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \dots + \alpha_k X_{ki} + v_i$$

Harvey-Godfrey LM test:

$$\ln(\hat{u}_i^2) = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \dots + \alpha_k X_{ki} + v_i$$

Park LM test:

$$\ln(\hat{u}_i^2) = \alpha_1 + \alpha_2 \ln(X_{2i}) + \alpha_3 \ln(X_{3i}) + \dots + \alpha_k \ln(X_{ki}) + v_i$$

What does each step 2 assume about the relationship between the errors and the explanatory variables in each of these tests?

Heteroskedasticity

White's Test: (2 variable case for ease of expression)

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

1. Estimate the regression above and get the predicted error \hat{u}_i
2. Run an auxiliary regression of the square of these estimated prediction errors on all of the explanatory variables, their squares, and their interaction terms. This looks for more complicated relationships between the X's and the error terms:

$$\hat{u}_i^2 = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 X_{2i}^2 + \alpha_5 X_{3i}^2 + \alpha_6 X_{2i}X_{3i} + v_i$$

3. Test if the alpha coefficients are jointly = 0 by computing the LM statistic and comparing to the critical Chi-Squared value:

$$LM = N * R^2$$

Heteroskedasticity

White's Test in STATA:

```
reg Y X2 X3
```

```
predict ui, residual
```

```
gen uisq=ui^2
```

```
gen X2sq=X2^2
```

```
gen X3sq=X3^2
```

```
gen X2xX3=X2*X3
```

```
reg uisq X2 X3 X2sq X3sq X2xX3
```

Use the resulting R-squared to compute the LM statistic.

Heteroskedasticity

Consider the following regression of wages on sex, years of schooling, and experience.

Source	SS	df	MS	Number of obs	=	3294
Model	4666.31659	3	1555.43886	F(3, 3290)	=	167.63
Residual	30527.8705	3290	9.27898798	Prob > F	=	0.0000
Total	35194.187	3293	10.6875758	R-squared	=	0.1326

wage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
school	.6387977	.0327958	19.48	0.000	.5744954	.7031
male	1.344369	.1076759	12.49	0.000	1.13325	1.555487
exper	.1248255	.0237628	5.25	0.000	.0782342	.1714167
_cons	-3.380018	.4649765	-7.27	0.000	-4.291691	-2.468346

Heteroskedasticity

After computing the squared residuals, these are regressed on the X's. The results suggest that the errors have larger variance for males and for people with higher levels of education. Does this make sense?

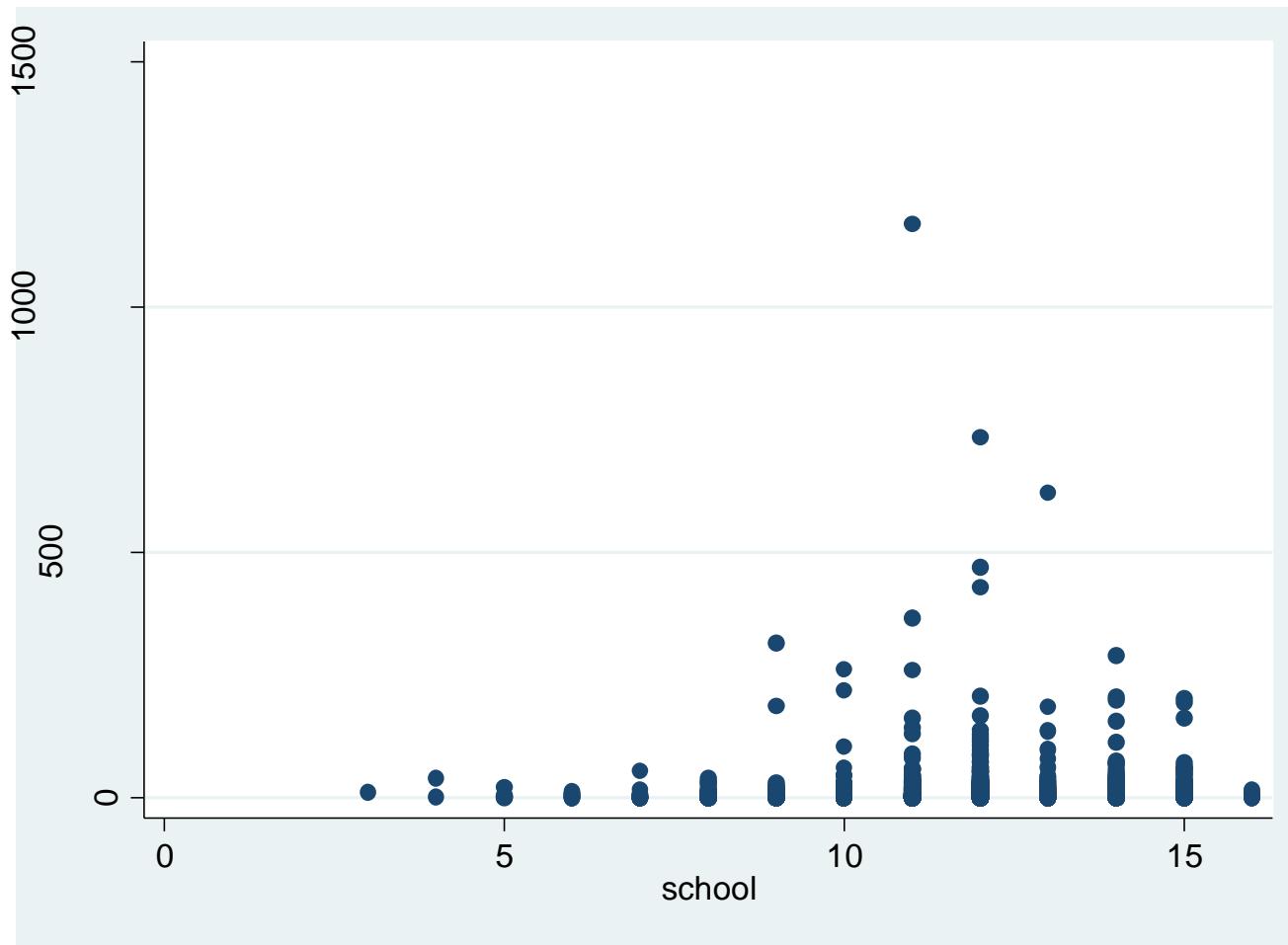
Source	SS	df	MS	Number of obs	=	3294
Model	27208.7699	3	9069.58997	F(3, 3290)	=	7.56
Residual	3948538.67	3290	1200.16373	Prob > F	=	0.0000
Total	3975747.44	3293	1207.33296	R-squared	=	0.0068

ui_squared	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
male	4.135426	1.224584	3.38	0.001	1.734403	6.536448
school	1.197484	.3729828	3.21	0.001	.4661817	1.928785
exper	-.3955445	.2702507	-1.46	0.143	-.9254211	.1343321
_cons	-3.64377	5.288116	-0.69	0.491	-14.0121	6.724561

The LM statistic = $3,294 * .0068$ which we compare to the Chi-Squared with 3 degrees of freedom.

Heteroskedasticity

Here is what the graph of the squared residual looks like when compared to schooling level. Is there evidence of heteroskedasticity?



Heteroskedasticity

How do we address heteroskedasticity:

1. Estimate with OLS (which is unbiased but is not efficient). Use the correct version of variance matrix for the estimated B's in order to be able to conduct valid hypothesis tests.
2. Use a new estimation method that accounts for the heteroskedasticity. Specifically, derive a new least squared error estimator that is more efficient. This method is called Generalized Least Squares (GLS). However, in order to do this, we need to know the form of the heteroskedasticity.

In practice, most researchers do #1 by using commands in STATA or another program that compute the corrected standard errors. To implement #2, the heteroskedasticity needs to be observable. For example, if males have higher variance than females. In many cases, the results won't be that different between the two since OLS is still unbiased.

Heteroskedasticity

Weighted least squares: A special case of GLS occurs when there is heteroskedasticity but there is no autocorrelation (i.e. when the errors differ in variance but are not correlated across observations).

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_k X_{ki} + u_i$$

And suppose that we have some idea of what the heteroskedasticity is for each observation (or group of observations). That is, we can write:

$$\text{var}(u_i) = \sigma^2 h_i$$

We call h_i the weight. For example, if we believe (or have determined) that the variance of the errors are twice as large for men as for women, then $h_i=2$ for all men and $h_i=1$ for all women.

We could, of course, have a different h_i for every observation i .

Heteroskedasticity

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_k X_{ki} + u_i$$

We then divide each observation by the square root of the relevant weight h_i to get the weighted least squares regression:

$$\frac{Y_i}{\sqrt{h_i}} = \beta_1 \frac{1}{\sqrt{h_i}} + \beta_2 \frac{X_{2i}}{\sqrt{h_i}} + \beta_3 \frac{X_{3i}}{\sqrt{h_i}} + \dots + \beta_k \frac{X_{ki}}{\sqrt{h_i}} + \frac{u_i}{\sqrt{h_i}}$$

Now, note that the variance of the error term is:

$$\text{var}\left(\frac{u_i}{\sqrt{h_i}}\right) = \frac{\sigma^2 h_i}{h_i} = \sigma^2$$

So our new, weighted regression is homoskedastic. This means that OLS is now BLUE!

Heteroskedasticity

To review:

- If we know the nature of the heteroskedasticity, then we can reweight the Y's and X's for each observation and then run OLS.
- This is called weighted least squares. The basic idea is that we give less weight to observations with higher error variance (i.e. men in the example).
- The procedure will produce efficient estimates (unlike just running OLS on the original data) and the correct standard errors for hypothesis testing.

Heteroskedasticity

So what's the problem?:

- We don't know the weights h_i .
- One method of trying to get at these weights is called feasible weighted least squares (Feasible WLS):
 1. Run the regular OLS regression and get the \hat{u}_i .
 2. Regress the squared \hat{u}_i 's against the X's in an auxilliary regression— just like all of the tests for heteroskedasticity.
 3. Use the coefficients from this auxilliary regression, get a predicted squared residual, and call this the weight h_i .

Basically, we are estimating the expected variance of the error using the values of the X's.

Heteroskedasticity

These weighted procedures are a special case of GLS, which also allows for correlation across observations:

$$E(\mathbf{u}\mathbf{u}') = \Omega = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1N} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \cdots & \sigma_{2N} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \cdots & \sigma_{3N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{N1} & \sigma_{N2} & \sigma_{N3} & \cdots & \sigma_N^2 \end{bmatrix}$$

We are not going to derive the optimal estimator in this general case, but the result is:

$$\text{Min } (Y - X\beta)' \Omega^{-1} (Y - X\beta)$$

$$\hat{\beta} = (X'\Omega^{-1}X)^{-1} X'\Omega^{-1} Y$$

$$var(\hat{\beta}) = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \Omega \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1}$$

[Note that these collapse to the usual case if $\Omega = I$.]

Heteroskedasticity

Reality check #3

In applied work, most researchers simply estimate OLS and then use one of two commands for making sure that their errors are adjusted for heteroskedasticity:

1. regress Y X1 X2...Xk, robust
2. regress Y X1 X2...Xk, vce(cluster group)

Let's discuss each of these.

Heteroskedasticity

Robust standard errors: The idea is that the squared errors from the prediction for each observation are used to compute the variances of the OLS estimators of B:

Specifically, instead of (where sigma is the average predicted error):

$$var(\hat{\beta}) = \hat{\sigma}^2 (\mathbf{X}'\mathbf{X})^{-1}$$

STATA computes:

$$var(\hat{\beta}) = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \hat{\Omega} \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1}$$

Where:

$$\hat{\Omega} = \begin{bmatrix} \hat{u}_1^2 & 0 & 0 & \cdots & 0 \\ 0 & \hat{u}_2^2 & 0 & \cdots & 0 \\ 0 & 0 & \hat{u}_3^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \hat{u}_N^2 \end{bmatrix}$$

Heteroskedasticity

Clustered standard errors: The idea is that each group, or cluster, is treated as a unit. That is, we assume that individuals in the same cluster group have correlated errors.

Groups could be:

- People who work at the same company
- People who attend the same school
- Houses sold in the same state

Often there are fixed effects for these groups in the regression.

Why might their errors be correlated?

- Share similar shocks.
- Have similar unobserved characteristics.

Lesson 9

Autocorrelation

Outline

Previous Lesson

1. Heteroscedasticity: Detection
2. Heteroscedasticity: Correction

This Lesson:

1. Autocorrelation: Detection
2. Autocorrelation: Correction

Next Lesson:

1. Misspecification: Functional Form
2. Misspecification: Measurement Error
3. Misspecification: Omitted Variables

Autocorrelation

Heteroskedasticity is a violation of the assumption of that all errors have the same variance (i.e. homeskedasticity):

$$\text{var}(u_i) \neq \sigma^2$$

Autocorrelation is a violation of the assumption that errors are not correlated with each other:

$$\text{cov}(u_t, u_s) \neq 0$$

That is, the error terms for observations are correlated with the error terms for other observations. It is easy to think of autocorrelation happening in one, primary way: error terms for observations adjacent in time are correlated (time-series).

In other words, the errors are correlated over time.

Autocorrelation

Causes of autocorrelation:

□ Omitted variables example:

- Y_t depends on X_{2t} and X_{3t}
- We omit X_{3t} in our regression so its effect is in the error u_t
- If X_{3t} depends on X_{3t-1} and X_{3t-2} ... then u_t will be correlated with u_{t-1} , u_{t-2}

Why is X_{3t} omitted?

□ Misspecification of the model example:

- Y_t depends on X_{2t}^2 (or any other non-linear relationship)
- We include X_{2t} linearly in our regression.
- The error term will depend on X_{2t} .
- If X_{2t} increases/decreases over time, then u_t 's will be correlated over time.

Why is the model misspecified?

□ Measurement error:

- Suppose Y_t or X_t is measured with error.
- If the measurement error persists then the u_t 's will be correlated over time.

Why might measurement errors persist over time?

Autocorrelation

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \dots + \beta_k X_{kt} + u_t$$

First order autocorrelation:

$$u_t = \rho u_{t-1} + \epsilon_t$$

The constant ρ is the first order autocorrelation coefficient. It ranges from -1 to 1 (if it were larger then the errors would be blowing up over time). The newly defined error, ϵ_t , is the component of the error term that is i.i.d.

- A positive value indicates that when u_{t-1} is positive u_t is more likely to also be positive. (positive residuals follow positive residuals) – **MOST COMMON**
- A negative value indicates that when u_{t-1} is positive u_t is more likely to be negative. (negative residuals follow positive residuals)

Higher order autocorrelations:

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \dots + \rho_p u_{t-p} + \epsilon_t$$

Autocorrelation

Left: Graph of error over time (note: not the square of the error)

Right: Graph of adjacent errors and the prior period error

What might Y vs Time look like?

Example: consumer spending – some unobserved income shock (positive or negative) causes an error in year t and may continue to have an effect in year $t+1$.

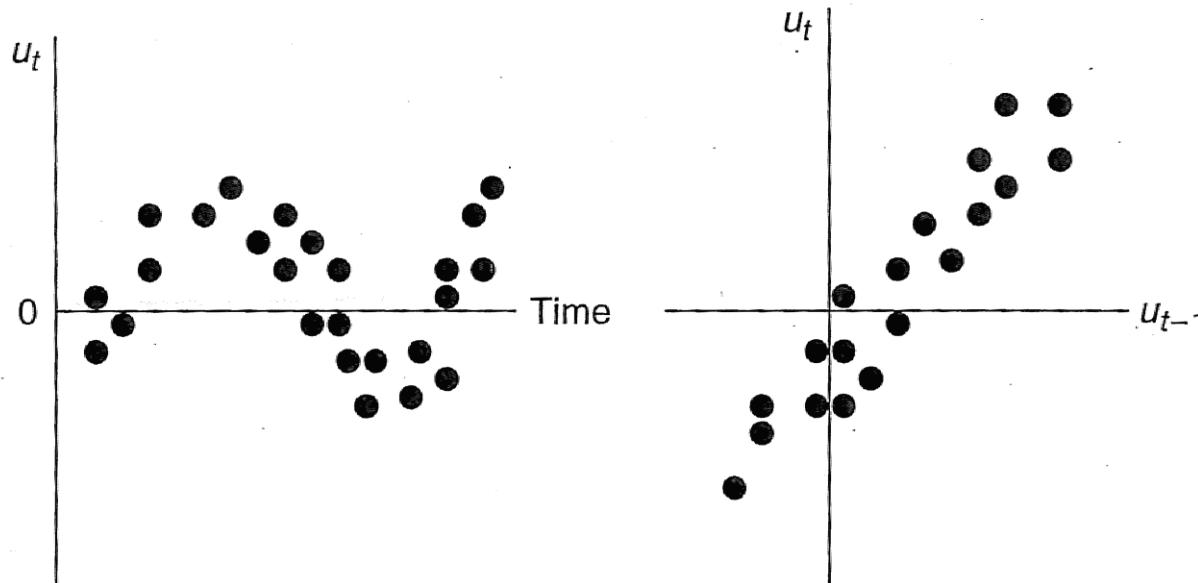


Figure 7.1 Positive serial correlation

Autocorrelation

Left: Graph of error over time.

Right: Graph of adjacent errors and the prior period error

What might Y vs Time look like?

Example: Length of time a doctor spends with patients. If she has a positive shock with one visit (positive residual) then she may try to speed up the next to get on schedule (negative residual).

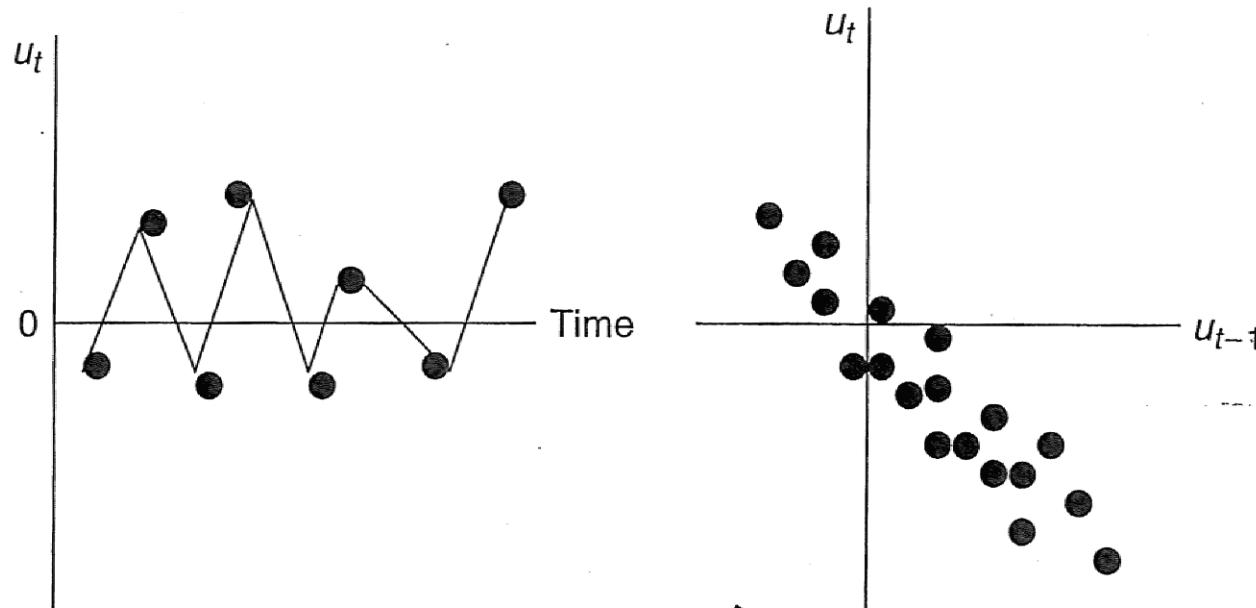


Figure 7.2 Negative serial correlation

Autocorrelation

Consequences of autocorrelation for OLS:

1. OLS is still unbiased. As with heteroskedasticity, $E(u|X)=0$.
2. OLS is no longer BLUE (i.e. not efficient). There is an alternative estimator of B that has smaller variance.
3. The variance of the estimated B 's is incorrect since $\text{var}(u) \neq \sigma^2 I$.

$$\text{var}(\hat{\beta}) \neq \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$$

This will result in invalid inference. Typically the OLS s.e.'s will be too small and result in t-statistics that are too large.

Autocorrelation

Derive variance-covariance matrix:

$$u_t = \rho u_{t-1} + \epsilon_t$$

But we also know (since the autocorrelation holds for t-1, t-2, ...):

$$u_{t-1} = \rho u_{t-2} + \epsilon_{t-1}$$

So we have:

$$\begin{aligned} u_t &= \epsilon_t + \rho[\rho u_{t-2} + \epsilon_{t-1}] \\ u_t &= \epsilon_t + \rho\epsilon_{t-1} + \rho^2[\rho u_{t-3} + \epsilon_{t-2}] \\ u_t &= \epsilon_t + \rho\epsilon_{t-1} + \rho^2\epsilon_{t-2} + \rho^3[\rho u_{t-4} + \epsilon_{t-3}] \\ &\vdots & \vdots \\ u_t &= \epsilon_t + \rho\epsilon_{t-1} + \rho^2\epsilon_{t-2} + \rho^3\epsilon_{t-3} + \dots \end{aligned}$$

Autocorrelation

So we can compute the variance of the error using the iid errors:

$$u_t = \epsilon_t + \rho\epsilon_{t-1} + \rho^2\epsilon_{t-2} + \rho^3\epsilon_{t-3} + \dots$$

$$\begin{aligned} E(u_t, u_t) &= \sigma_\epsilon^2 + \rho^2\sigma_\epsilon^2 + \rho^4\sigma_\epsilon^2 + \rho^6\sigma_\epsilon^2 + \dots \\ &= \frac{\sigma_\epsilon^2}{1 - \rho^2} \end{aligned}$$

And the covariances of the errors:

$$E(u_t, u_{t-1}) = \rho\sigma_u^2$$

$$E(u_t, u_{t-2}) = \rho^2\sigma_u^2$$

$$E(u_t, u_{t-3}) = \rho^3\sigma_u^2$$

Autocorrelation

Thus our variance-covariance matrix for the errors has the following form:

$$E(\mathbf{u}\mathbf{u}') = \begin{bmatrix} \text{var}[u_1] & \text{cov}[u_1 u_2] & \text{cov}[u_1 u_3] & \cdots & \text{cov}[u_1 u_N] \\ \text{cov}[u_2 u_1] & \text{var}[u_2] & \text{cov}[u_2 u_3] & \cdots & \text{cov}[u_2 u_N] \\ \text{cov}[u_3 u_1] & \text{cov}[u_3 u_2] & \text{var}[u_3] & \cdots & \text{cov}[u_3 u_N] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \text{cov}[u_N u_1] & \text{cov}[u_N u_2] & \text{cov}[u_N u_3] & \cdots & \text{var}[u_N] \end{bmatrix}$$

$$E(\mathbf{u}\mathbf{u}') = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{N-1} \\ \rho & 1 & \rho & \cdots & \rho^{N-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{N-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{N-1} & \rho^{N-2} & \rho^{N-3} & \cdots & 1 \end{bmatrix} = \boldsymbol{\Omega}_2$$

Autocorrelation

And the Variance of $\hat{\beta}$ is as follows:

$$\begin{aligned}Var(\hat{\beta}) &= E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] \\&= E\{[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}][(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}]'\} \\&= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \\&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{u}\mathbf{u}')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Omega_2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

Autocorrelation

How do we detect autocorrelation?:

Compute the prediction errors:

$$\hat{u}_i = Y_i - \hat{Y}_i$$

1. Informal visual inspection of the errors:
 - predicted errors positively correlated over time
 - predicted errors negatively correlated over time
2. Conduct a formal test:
 - Durbin-Watson test
 - Breusch-Godfrey test

Autocorrelation

Visual inspection: consumption as a function of disposable income and price

```
reg lcons ldisp lprice  
predict res01, residual  
twoway (line res01 time)
```

```
gen res01_1=L1.res01  
twoway (scatter res01_1 res01)
```

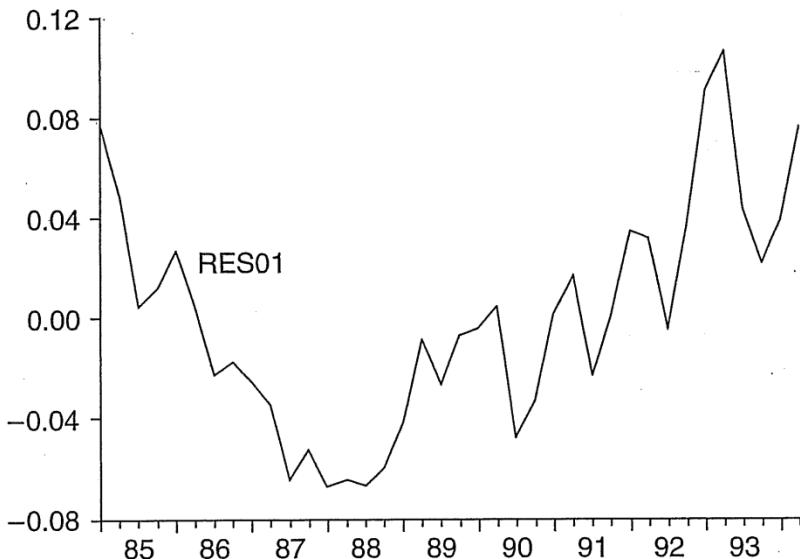


Figure 7.3 Residuals plot from computer example

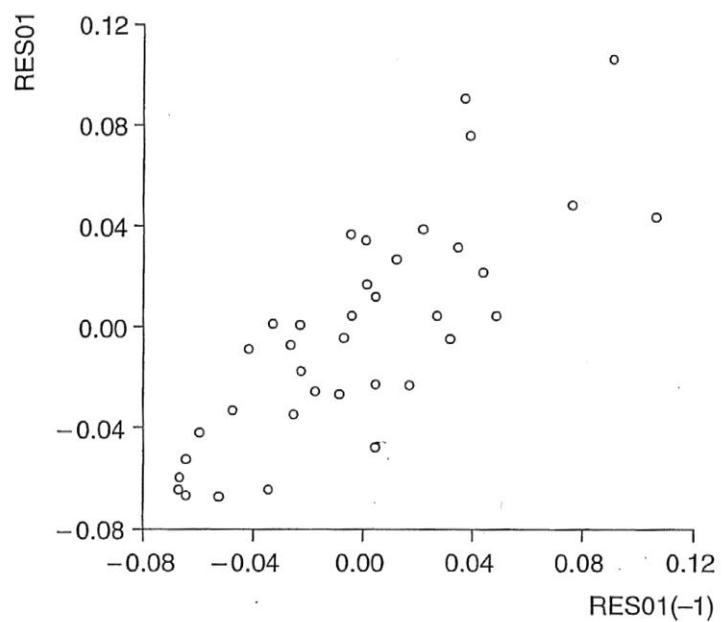


Figure 7.4 Residuals scatter plot from computer example

Autocorrelation

Breusch – Godfrey Test:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \dots + \beta_k X_{kt} + u_t$$

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \dots + \rho_p u_{t-p} + \epsilon_t$$

We are interested in determining if the lagged errors partially determine the current time period error:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \dots + \beta_k X_{kt} + \rho_1 u_{t-1} + \rho_2 u_{t-2} + \dots + \rho_p u_{t-p} + \epsilon_t$$

Specifically are the ρ 's jointly significant.

In the case where we suspect first-order autocorrelation:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \dots + \beta_k X_{kt} + \rho_1 u_{t-1} + \epsilon_t$$

Autocorrelation

B-G Test Steps:

1. Estimate the regression above and get the predicted error \hat{u}_t
2. Run a regression of the prediction errors on the explanatory variables and the lagged prediction errors for prior periods:

$$\hat{u}_t = \alpha_1 + \alpha_2 X_{2t} + \dots + \alpha_k X_{kt} + \alpha_{k+1} \hat{u}_{t-1} + \dots + \alpha_{k+p} \hat{u}_{t-p} + \epsilon_t$$

3. Test if the alpha coefficients on the lagged errors are jointly = 0. Compute the LM statistic and comparing to the critical Chi-squared value for k degrees of freedom:

$$LM = (N-p) * R^2 \quad \text{vs.} \quad \text{Chi-Squared with } p \text{ freedom}$$

STATA (two methods): all steps or shortcut - estat bgodfrey, lags(1)

Autocorrelation

Durbin-Watson Test:

1. Run OLS on the model and get the residuals $\hat{u}_t, \hat{u}_{t-1}, \dots$
2. Calculate the DW statistic:

$$d = \frac{\sum_{t=2}^N (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^N \hat{u}_t^2}$$

Note:

small value indicates close errors (i.e. positive correlation)
large value indicates negative correlation.
with iid expect value of about 2

3. Conduct a hypothesis test for positive correlation:

$$H_O: \rho=0$$

$$H_A: \rho>0$$

Look up the critical values which depend on k (in book and on eCommons).
If $d < d_L$ then reject that there is no positive autocorrelation.
Analogous test for negative correlation (see p. 157 of text).

Autocorrelation

Durbin-Watson Rule of Thumb:

Note that ρ is the coefficient of the error regression (cov/var where we know the mean of u is 0):

$$\hat{\rho} = \frac{\sum_{t=2}^N \hat{u}_t \hat{u}_{t-1}}{\sum_{t=1}^N \hat{u}_t^2}$$

You can show that the D-W statistic reduces:

$$d = \frac{\sum_{t=2}^N \hat{u}_t^2 + \sum_{t=2}^N \hat{u}_{t-1}^2 - 2 \sum_{t=2}^N \hat{u}_t \hat{u}_{t-1}}{\sum_{t=1}^N \hat{u}_t^2}$$

$$d \approx 2 - \frac{2 \sum_{t=2}^N \hat{u}_t \hat{u}_{t-1}}{\sum_{t=1}^N \hat{u}_t^2} = 2 - 2\rho = 2(1 - \rho)$$

So, what is the D-W value when:

1. We have no autocorrelation of errors?
2. We have perfect positive autocorrelation of errors?
3. We have perfect negative autocorrelation of errors?

Autocorrelation

Table 7.2 The DW test

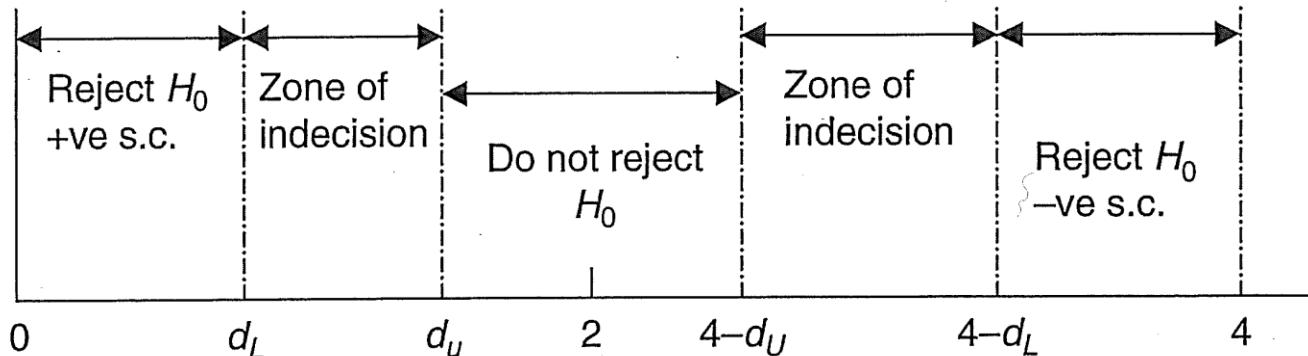
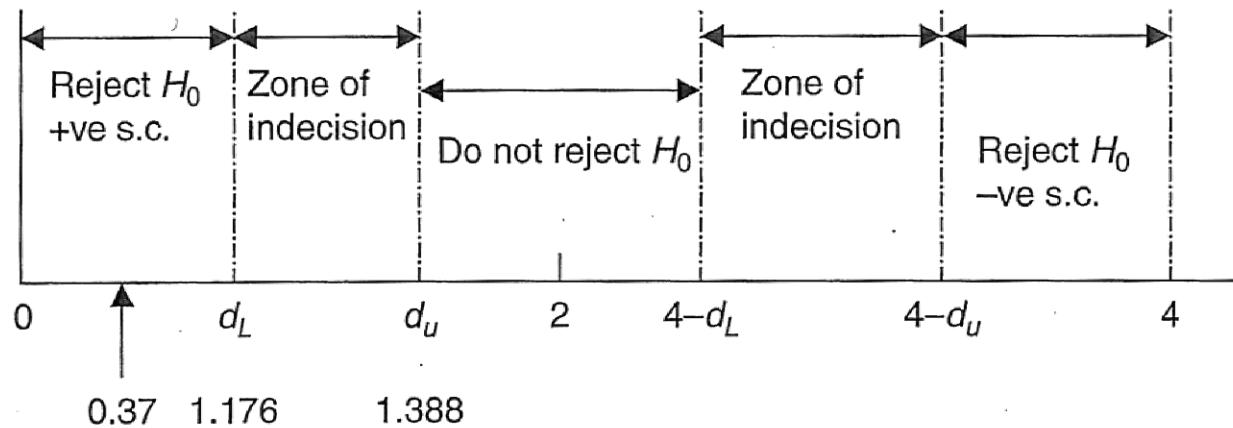


Table 7.3 An example of the DW test



Autocorrelation

Resolving autocorrelation when ρ known:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \dots + \beta_k X_{kt} + u_t$$

$$u_t = \rho u_{t-1} + \epsilon_t$$

Note that:

$$Y_{t-1} = \beta_1 + \beta_2 X_{2t-1} + \beta_3 X_{3t-1} + \dots + \beta_k X_{kt-1} + u_{t-1}$$

$$\rho Y_{t-1} = \rho \beta_1 + \rho \beta_2 X_{2t-1} + \rho \beta_3 X_{3t-1} + \dots + \rho \beta_k X_{kt-1} + \rho u_{t-1}$$

Now we construct:

$$Y_t - \rho Y_{t-1} = \beta_1(1-\rho) + \beta_2(X_{2t} - \rho X_{2t-1}) + \dots + \beta_k(X_{kt} - \rho X_{kt-1}) + (u_t - \rho u_{t-1})$$

$$Y_t^* = \beta_1^* + \beta_2 X_{2t}^* + \beta_3 X_{3t}^* + \dots + \beta_k X_{kt}^* + \epsilon_t$$

Verify that the error term is correct: $u_t = \rho u_{t-1} + \epsilon_t$

Most important is to note that the new regression is homoskedastic. So OLS on the transformed data is now BLUE.

Autocorrelation

Cochrane-Orcutt procedure: resolving autocorrelation: ρ unknown.

Example: suspect first order autocorrelation only:

1. Run the OLS regression to get an estimated u .
2. Regress u on one lag of u . The resulting coefficient is the estimate of ρ .
3. Compute the adjust $Y_t - \rho Y_{t-1}$ and $X_t - \rho X_{t-1}$ and run the new regression which is homoskedastic.

The book discusses doing multiple iterations of this procedure, but I don't think it adds any value to intuition.

STATA:

1. Manually
2. Shortcuts: `prais Y X1 X2, corc`

Breusch-Godfrey, DW test, Cochrane-Orcutt

1. [10 points] Properties of Estimators. You are interested in estimating the mean wage in Santa Cruz. The true average wage of the population is w which has variance σ_w^2 . You have data for two individual's wages (w_1 and w_2) that were drawn at random from the population. You are considering three possible estimators of the mean based on different weightings of your two data points:

$$\text{Estimator 1: } 0.4*w_1+0.6*w_2$$

$$\text{Estimator 2: } 0.3*w_1+0.7*w_2$$

$$\text{Estimator 3: } 0.3*w_1+0.3*w_2$$

Evaluate these three estimators in terms of being: a) unbiased; b) efficient. Show your work.

Estimator 1:

$$a) E(0.4w_1+0.6w_2) = 0.4E(w_1) + 0.6E(w_2) = 0.4w + 0.6w = \boxed{w} \checkmark$$

$$b) \text{Var}(0.4w_1+0.6w_2) = 0.16\text{Var}(w_1) + 0.36\text{Var}(w_2) = \\ 0.16\sigma_w^2 + 0.36\sigma_w^2 = \boxed{0.52\sigma_w^2}$$

Estimator 2:

$$a) E(0.3w_1+0.7w_2) = \boxed{w} \checkmark$$

$$b) \text{Var}(0.3w_1+0.7w_2) = \boxed{0.58\sigma_w^2}$$

Estimator 3:

$$a) E(0.3w_1+0.3w_2) = \boxed{0.6w} \text{ biased}$$

$$b) \text{Var}(0.3w_1+0.3w_2) = \boxed{0.18\sigma_w^2}$$

Based on the analysis above, which estimator is the best choice (1, 2, or 3)?

Estimator 1 - unbiased and has lower variance than 2.

2. [10 points] Ordinary Least Squares. You are interested in estimating the effect of the minimum wage on the unemployment rate. Your plan is to estimate this using cross-state differences in minimum wage levels, so you collect the data in the table below.

state	minwage	unemp	percmmanuf
california	10	11	11
oregon	7	4	30
washington	11	8	14
nevada	4	5	4

Use the data above to estimate the regression coefficients B_0 and B_1 in the regression below and write the resulting regression equation. Show the equations you are using for B_0 and B_1 .

$$unemp_i = \beta_0 + \beta_1 minwage + u_i$$

$$\hat{\beta}_1 = \frac{Cov(minwage, unemp)}{Var(minwage)}$$

$$\hat{\beta}_0 = \bar{unemp} - \hat{\beta}_1 \bar{minwage}$$

$$(x) \bar{minwage} = 8$$

$$(y) \bar{unemp} = 7$$

$$\begin{aligned} Var(minwage) &= \frac{1}{3} [(10-8)^2 + (7-8)^2 + (11-8)^2 + (4-8)^2] \\ &= \frac{1}{3} [4 + 1 + 9 + 16] = 10 \end{aligned}$$

$$\begin{aligned} Cov(minwage, unemp) &= \frac{1}{3} [(10-8)(11-7) + (7-8)(4-7) + (11-8)(8-7) + (4-8)(5-7)] \\ &= \frac{1}{3} [8 + 3 + 3 + 8] = \frac{1}{3} [22] = \frac{22}{3} \text{ or } 7\frac{1}{3} \end{aligned}$$

$$\hat{\beta}_1 = \frac{22/3}{10} = 0.73$$

$$\hat{\beta}_0 = 7 - 0.73(8) = 7 - 5.86 = 1.14$$

$$\hat{Unemp}_i = 1.14 + 0.73 minwage_i$$

3a. [6 points] Hypothesis Testing. The following regression examines the relationship between hours worked and years of education, union status (binary), and whether or not an individual works in finance (binary). The regression is based on 13,200 individuals.

$$\ln(\text{hours}_i) = 3.224 + 0.041\text{educ}_i + 0.185\text{fin}_i - 0.110\text{union}_i$$

(0.885) (0.019) (0.045) (0.040)

- i. Interpret the coefficient on *finance* in a sentence. Be precise.

Working in finance is associated with working 18.5 percent more hours (holding education & union status fixed).

- ii. Test if the effect of *education* is statistically significant at the 95% confidence level?

$$H_0: \beta_2 = 0 \quad t = \frac{0.041}{0.019} \approx 2.15 \quad t_{\text{crit}}^{95} = 1.96$$

$$|2.15| > 1.96 \Rightarrow \text{reject null} \Rightarrow \boxed{\text{is statistically significant}}$$

- iii. Find the 99% confidence interval for the effect of *finance* on hours worked.

$$t_{\text{crit}}^{99} = 2.576 \quad 0.045 + 2.576 = 0.1159$$

$$0.185 - 0.1159 < \beta_3 < 0.185 + 0.1159 \Rightarrow \boxed{[0.069, 0.301]}$$

3b. [4 points] You wish to test the restriction that $\alpha = 0.5\beta$ in the Cobb-Douglas production function below. Write out the restricted and unrestricted regressions you would estimate to test this restriction.

$$Y_i = AL_i^\alpha K_i^\beta u_i$$

unrestricted:

$$\boxed{\ln(Y) = \ln(A) + \alpha \ln(L) + \beta \ln(K) + \ln(u)}$$

restricted: $\alpha = 0.5\beta$

$$\begin{aligned} \ln(Y) &= \ln(A) + 0.5\beta \ln(L) + \beta \ln(K) + \ln(u) \\ \Rightarrow \boxed{\ln(Y) &= \ln(A) + \beta (\ln(K) + 0.5 \ln(L)) + \ln(u)} \end{aligned}$$

4. [10 points] Omitted variables. Shown below is the true regression equation with included variables X^* and omitted variables X^o , and the estimated regression equation with only the omitted variables. Matrix X^* has k_1 variables and matrix X^o has $k - k_1$ variables.

$$Y = X^* \beta^* + X^o \beta^o + u \quad Y = X^* b_R^* + u$$

- a) Write the expression for estimating b_R^* in matrix form and then derive the expression for omitted variable bias.

$$\begin{aligned} b_R^* &= (X^{*'} X^*)^{-1} X^{*'} Y \\ &= (X^{*'} X^*)^{-1} X^{*'} (X^* \beta^* + X^o \beta^o + u) \\ &= \beta^* + (X^{*'} X^*)^{-1} X^{*'} X^o \beta^o + (X^{*'} X^*)^{-1} X^{*'} u \end{aligned}$$

$$E(b_R^* | X) = \beta^* + (X^{*'} X^*)^{-1} X^{*'} X^o \beta^o$$

$$\text{omitted variable bias} = \boxed{(X^{*'} X^*)^{-1} X^{*'} X^o \beta^o}$$

- b) Explain how the expression for omitted variable bias can be thought of as depending on an auxiliary regression.

The expression $\underbrace{(X^{*'} X^*)^{-1} X^{*'} X^o}_{(X' X)^{-1} X' Y}$ has the

exact same form as a regression of X^o on X^* , where X^o is like Y on X^* , X is X .

5a. [5 points] The following regression shows the effect of education, working in finance, and union membership on the natural log of hours worked.

$$\ln(\widehat{\text{hours}}_i) = 3.224 + 0.041 \text{educ}_i + 0.185 \text{fin}_i - 0.110 \text{union}_i$$

The relationship between union status and the other explanatory variables can be written as follows:

$$\widehat{\text{union}}_i = 0.26 + 0.10 \text{educ}_i - 0.30 \text{fin}_i$$

Derive the coefficients on education and finance in a regression that omits union status.

$$\begin{aligned} \ln(\widehat{\text{hours}}) &= 3.224 + 0.041 \text{educ} + 0.185 \text{fin} + (-.110) \text{union} \\ \Rightarrow \ln(\widehat{\text{hours}}) &= 3.224 + 0.041 \text{educ} + 0.185 \text{fin} + (-.110)(.26 + .1 \text{ed} - .3 \text{fin}) \\ \Rightarrow \ln(\widehat{\text{hrs}}) &= 3.224 + 0.041 \text{educ} + 0.185 \text{fin} - 0.0286 - 0.0110 \text{ed} + 0.0337 \\ \Rightarrow \boxed{\ln(\widehat{\text{hrs}}) &= 3.1954 + 0.03 \text{educ} + 0.218 \text{fin}} \\ &\quad \underline{0.03} \quad \text{and} \quad \underline{0.218} \end{aligned}$$

5b. [5 points] You want to regress Y on X. However, the variables X and Y are both measured with random error:

$$Y^* = B_0 + B_1 X^* + u \quad X^* = X + e \quad Y^* = Y + \eta$$

Derive the expression for B_1 in the univariate regression above. Show your work.

$$\begin{aligned} \hat{B}_1 &= \frac{\text{cov}(x^*, y^*)}{\text{var}(x^*)} = \frac{\text{cov}(x + e, y + \eta)}{\text{var}(x + e)} \\ &= \frac{\text{cov}(x, y) + \text{cov}(e, y) + \text{cov}(x, \eta) + \text{cov}(e, \eta)}{\text{var}(x) + \text{var}(e)} \\ &= \frac{\text{cov}(x, y) + 0 + 0 + 0}{\text{var}(x) + \text{var}(e)} \\ &= \boxed{\frac{\text{cov}(x, y)}{\text{var}(x) + \text{var}(e)}} \end{aligned}$$

6. [10 points] Stata. You have the data set “hrs_worked.dta” that includes each employees hours (*hours*), years of education (*educ*), and whether they are female (*fem*). There is also a column for industry (*industry*), which takes on one of three values: “energy”, “tech”, and “transport”.

- a) Write Stata code to estimate the effect of education and female on hours worked.
 - b) Manually test for heteroskedasticity using a White Test (without using shortcuts).

Use hrs_worked, dta

- | reg hours educ fem
- | predict w_i , residual
- | gen $w_i \text{sq} = w_i * w_i$
- | gen educ sq = educ * educ
- | gen fem sq = fem * fem
- | gen educXfem = educ * fem
- | reg $w_i \text{sq}$ educ fem educsq femsq educXfem

- c) Write code to test if education has differential effects on hours worked across industries. You need to make new variables (no shortcuts). Your regression should not suffer from perfect multicollinearity.

$$\begin{cases} \text{gen. educ} \times \text{energy} = \text{educ} * \text{energy} \\ \text{gen. educ} \times \text{tech} = \text{educ} * \text{tech} \\ \text{reg hours } \boxed{\text{educ tech energy}} \quad \boxed{\text{educ} \times \text{energy} \text{ educ} \times \text{tech}} \\ \text{Main} \qquad \qquad \qquad \text{interacted} \end{cases}$$

Note: omit one group.

7a. [5 points] You estimate the determinants of imports to a country and get the following:

$$\ln(\widehat{\text{imports}}_t) = 23.22 + 0.25\ln(\text{avginc}_t) + 0.88\ln(\text{pop}_t) \quad \text{RSS}=245 \quad \text{SST}=1,205$$

$$\ln(\widehat{\text{avginc}}_t) = 10.77 + 0.72\ln(\text{pop}_t) \quad \text{RSS}=35 \quad \text{SST}=2,022$$

Use the expression below and the results above to decide whether or not multicollinearity is likely to be an issue for the coefficient on avginc in the first regression. Provide empirical evidence and explain.

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\sum(X_j - \bar{X}_j)^2} \frac{1}{(1-R_j^2)}$$

$\frac{1}{1-R_j^2}$ is the variance inflation factor (VIF)

$$R_j^2 = 1 - \frac{35}{2,022} = 0.983$$

$$\text{So, VIF} = \frac{1}{1-0.983} = 58.8 \Rightarrow \begin{array}{l} \text{the standard errors are} \\ \text{inflated by a multiple of nearly 60.} \end{array}$$

This strongly suggests that we have a multicollinearity problem

7b. [5 points] Cochrane Orcutt

- i. What is the purpose of the Cochrane-Orcutt procedure (i.e. what problem does it fix)?

The Cochrane-Orcutt procedure does two things when there is autocorrelation:

1. produces a more efficient estimator
2. results in the correct standard errors.

- ii. List the steps (in words) needed to execute the Cochrane Orcutt procedure.

1. Run the desired regression
2. Get the predicted error for each period u_t
3. Regress current error on lagged error to get ρ (rho)
4. Adjust Y and X: $\hat{Y}_t = Y_t - \rho Y_{t-1}$, $\hat{X}_t = X_t - \rho X_{t-1}$
5. Regress new \hat{Y}^* on \hat{X}^* .

