Metrics Honors Review

Peter Tu

petertu@fas.harvard.edu

Harvard University

Department of Economics

24 March 2016

Logistics

- Exam Date: **Wednesday, April 6** from **3-6pm** Emerson 105
  - The exam covers material from 1010a, 1010b, and 1123 (not 1011a, 1011b, or 1126).
- Econometrics Office Hours in Littauer
  - Wednesday 3/30: 2 - 4pm, 3rd Floor Lounge
  - Friday 4/1: 2 - 4pm, M-17
  - Monday 4/3: 9 - 11am, Rm 219
- http://economics.harvard.edu/pages/honors has previous exams, review session videos, and slides

Table of Contents

Section 1

Ordinary Least Squares (OLS)

Ordinary Least Squares

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

Assumptions:

- Observations (X_{1i}, X_{2i}, Y_i) are independent and identically distributed (**iid**)
- No perfect multicollinearity of Xs
- Linear form is correctly-specified
- Conditional Mean Independence of X

If these assumptions hold, then OLS estimates **unbiased**, **consistent**, and asymptotically-normal coefficients

```
regress y x1 x2, robust
```

Perfect Multicollinearity

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

Regressors are **perfectly multicollinear** if one can be expressed as a linear function of the others:

i.e. if for all i, Xs are perfectly correlated. Ex:

- $X_{1i} = X_{2i}$
- $X_{3i} = X_{2i} - 2X_{1i}$
- ...

This is especially common if we include an intercept & fail to omit a dummy term (**dummy variable trap**). Ex:

$$Y_i = \beta_0 + \beta_1 \text{MALE}_i + \beta_2 \text{FEMALE}_i + \cdots + u_i$$

Perfect Multicollinearity – Intuition

- Multicollinearity is a measure of how much variation is *lacking* in your dataset. Generally, the more variation the better
  - **Ex:** Suppose you want to estimate the effect of graduating Harvard on future life outcomes, but everyone in your dataset graduated Harvard
- Now suppose you have data on Harvard & Yale students
  - **Ex:** Suppose you want to estimate the effect of graduating Harvard and the effect of graduating Harvard Econ, but all the Harvard students in your dataset are Econ students
  - Then you suffer **perfect multicollinearity** and cannot separately estimate β_{Harvard} and $\beta_{\text{Harvard Ec}}$

The Error Term u_i

The error term u_i is **unobserved** and typically the culprit behind our econometric woes

u_i contains all the stuff related to Y_i but isn't explicitly in the regression

$$\text{WAGE}_i = \beta_0 + \beta_1 \text{EDUC}_i + u_i$$

In this case, u_i includes the effect of:

The Error Term u_i

The error term u_i is **unobserved** and typically the culprit behind our econometric woes

u_i contains all the stuff related to Y_i but isn't explicitly in the regression

$$\text{WAGE}_i = \beta_0 + \beta_1 \text{EDUC}_i + u_i$$

In this case, u_i includes the effect of:

- age
- age^2
- past work experience
- health
- ... (all things that affect wage other than education)

The Error Term u_i

Suppose we control for age explicitly:

$$\text{WAGE}_i = \beta_0 + \beta_1 \text{EDUC}_i + \beta_2 \text{AGE}_i + u_i$$

Then u_i includes the effect of...

- ~~age~~
- age^2
- past work experience
- health
- ... (all things that affect wage other than age & educ)

Homoskedastic v. Heteroskedastic Errors

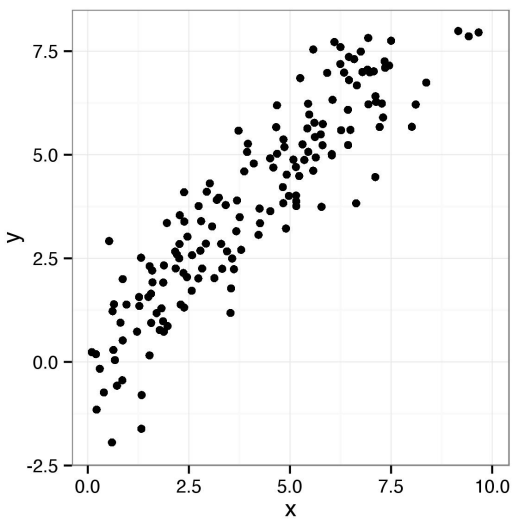
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + u_i$$

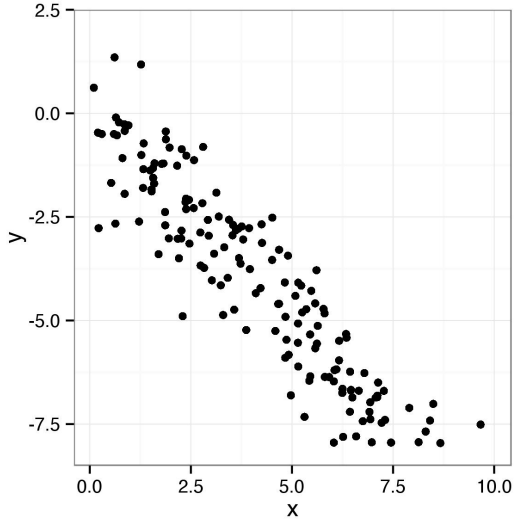
Homoskedastic or Heteroskedastic is an assumption about the pattern of errors u_i

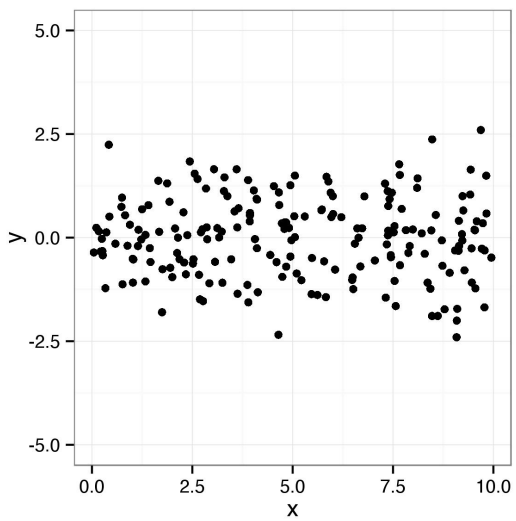
- **Homoskedastic**: $\text{Var}(u|X)$ is constant for all X
- **Heteroskedastic**: $\text{Var}(u|X)$ can vary with X

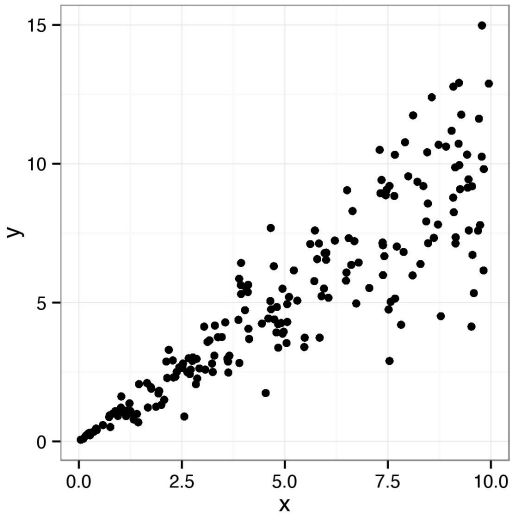
Homoskedasticity is a strong assumption that we basically never have enough evidence to make, because u_i is unobserved

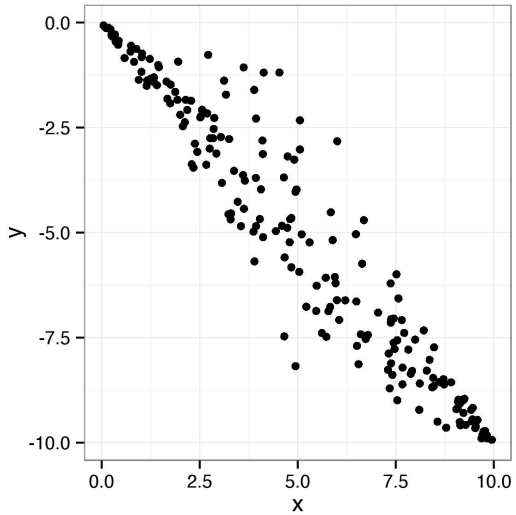
Homoskedastic Errors: Var($u|X$) constant for all X

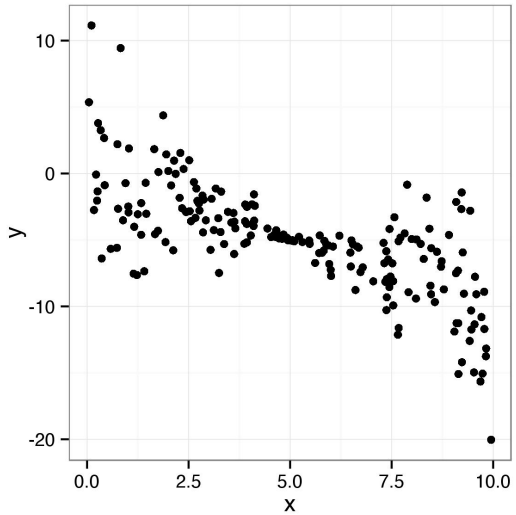












Heteroskedastic Errors: Var($u|X$) may change with X

Homoskedastic v. Heteroskedastic Errors

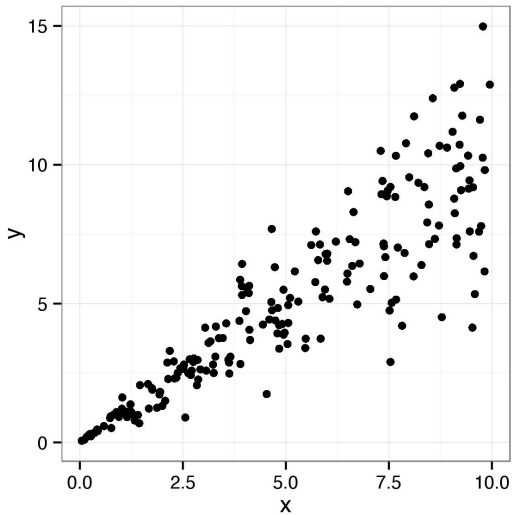
- The problem is that error u is always unobserved
- Fortunately, if we allow for heteroskedasticity, standard error estimates will be right, even if the errors are homoskedastic
- **NEVER** assume homoskedasticity
  - In `STATA`, use the "`robust`" command

```
regress y x1 x2, robust
```

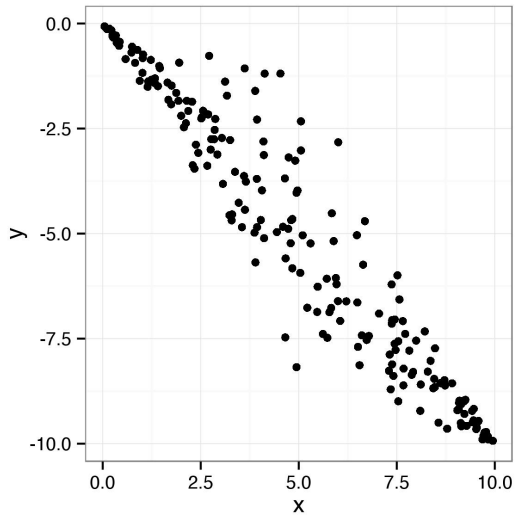
Heteroskedasticity Intuition

Heteroskedasticity implies that X is better at predicting Y for some values than others

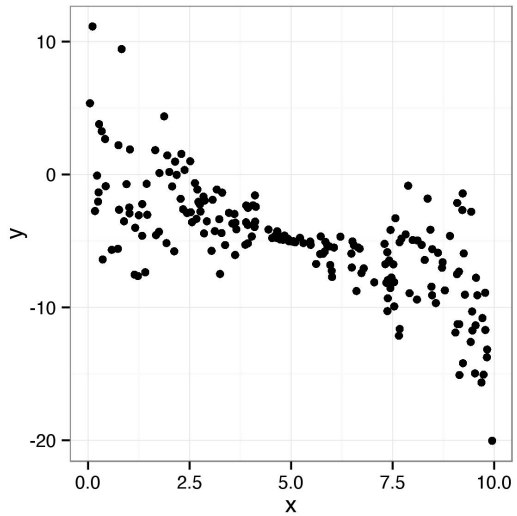
In these examples, X is a better predictor of Y for:



... low X



... extreme X



... middling X

Hypothesis Testing – Single Equality

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

Suppose we wanted to test if β_1 is statistically different from a constant C, where C is usually 0:

Null Hypothesis	H_0 :	$\beta_1 = C$
Alternative Hypothesis	H_a :	$\beta_1 \neq C$

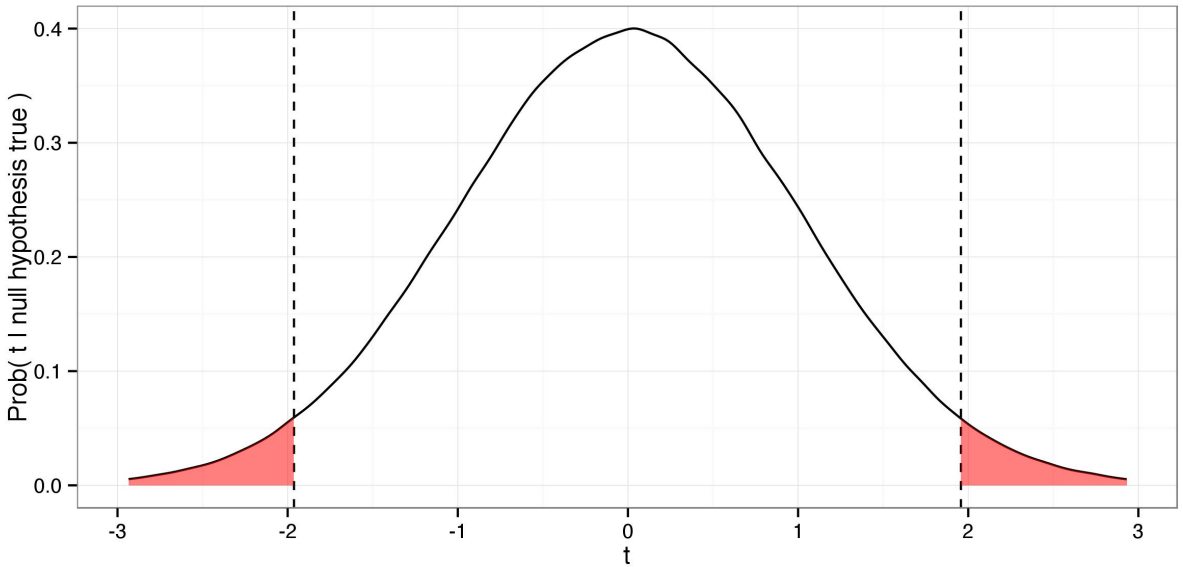
We calculate a t-statistic using our estimate $\widehat{\beta}_1$ and its standard error:

$$t = \frac{\widehat{\beta}_1 - C}{\text{se}(\widehat{\beta}_1)}$$

Single Hypothesis Testing – One Equality

$$t = \frac{\widehat{\beta}_1 - C}{\text{se}(\widehat{\beta}_1)}$$

For a 95% two-sided confidence test, we reject the null hypothesis when $t \geq 1.96$ or $t \leq -1.96$:



Make sure you also understand how to construct 95% confidence intervals!

Joint Hypothesis Testing – Multiple Equalities

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

Suppose we wanted to test **multiple** coefficients are different from 0

$$H_0: \quad \beta_1 = \beta_2 = \beta_3 = 0$$

H_a : **At least** one of $\beta_1, \beta_2,$ or β_3 is nonzero

Now we have to use a *F*-**test**, which is like a multiple *t*-test that takes into account the correlation between $\widehat{\beta}_1, \widehat{\beta}_2$, and $\widehat{\beta}_3$

Note: If we reject H_0, we cannot say which coefficient(s) is/are non-zero, **only that at least one is non-zero**

Single Hypothesis Testing

$$H_0: \quad \beta_1 = 0$$
$$H_a: \quad \beta_1 \neq 0$$

Use a **t-test**

Joint Hypothesis Testing

$$H_0: \quad \beta_1 = \beta_2 = \beta_3 = 0$$
$$H_a: \quad \textbf{At least} \text{ one of } \beta_1, \beta_2, \text{ or } \beta_3 \text{ is nonzero}$$

Use a **F-test**

Testing a Linear Combination of Coefficients

Suppose we wanted to test $\beta_1 = \beta_2$:

$$H_0: \quad \beta_1 - \beta_2 = 0$$
$$H_a: \quad \beta_1 - \beta_2 \neq 0$$

Which test?

Testing a Linear Combination of Coefficients

Suppose we wanted to test $\beta_1 = \beta_2$:

$$H_0: \quad \beta_1 - \beta_2 = 0$$
$$H_a: \quad \beta_1 - \beta_2 \neq 0$$

Which test? Single equality, so **t-test**!

$$t = \widehat{\beta}_1 - \widehat{\beta}_2$$
$$\text{Var}(\widehat{\beta_1} - \widehat{\beta_2}) = \text{Var}(\widehat{\beta_1}) + \text{Var}(\widehat{\beta_1}) - 2\text{Cov}(\widehat{\beta}_1, \widehat{\beta}_2)$$

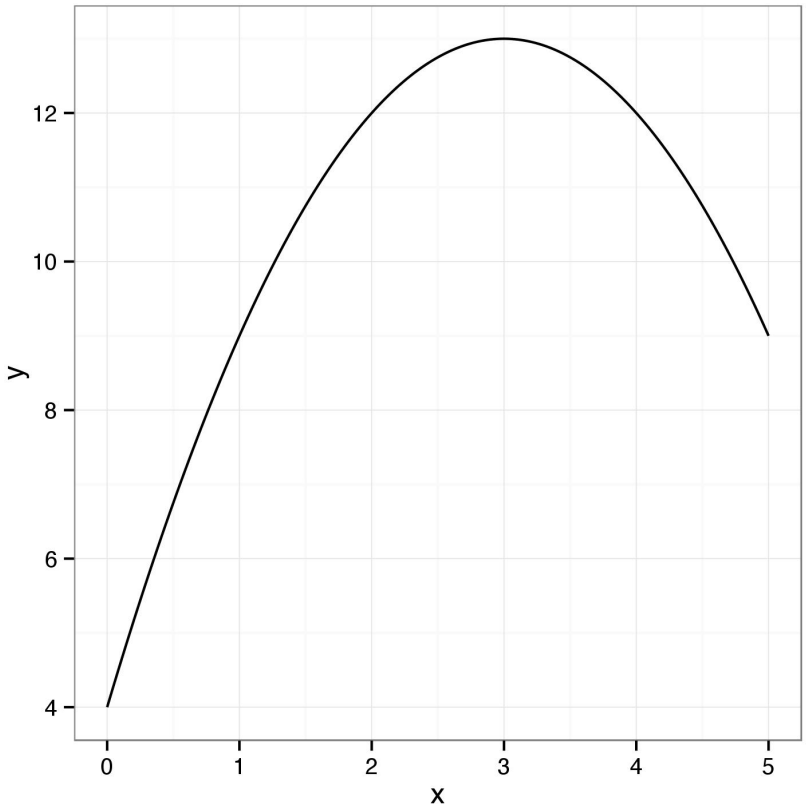
From the variance formula:

$$\text{Var}(A \pm B) = \text{Var}(A) + \text{Var}(B) \pm 2\text{Cov}(A, B)$$

Polynomial regressions

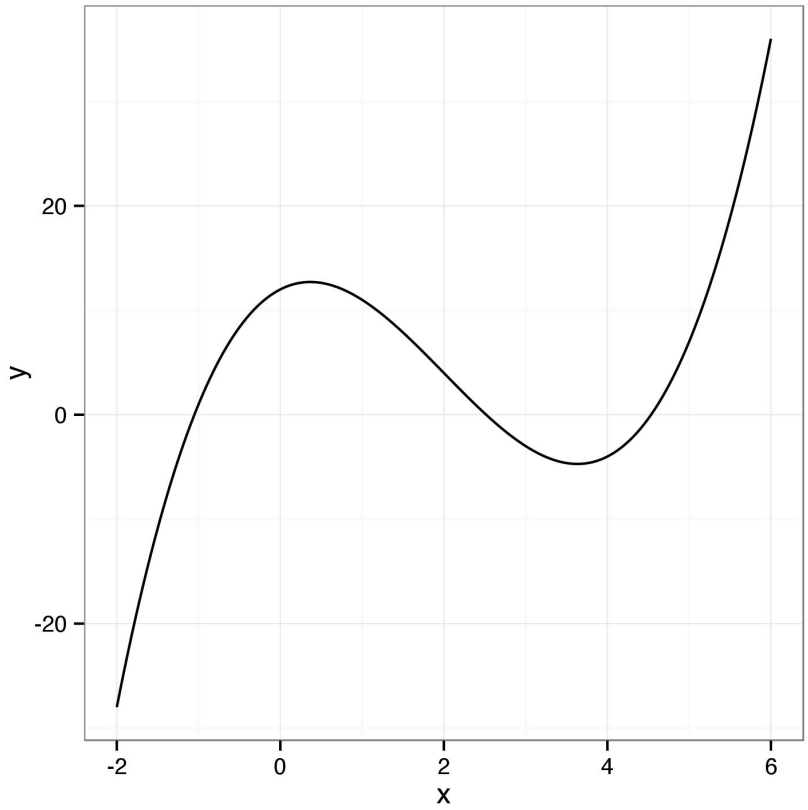
Quadratic

$$Y = 4 + 6X - X^2$$



Cubic

$$Y = 12 + 4X - 6X^2 + X^3$$



Examples of Regressions with Polynomials

Regressing with polynomials useful whenever Y and X do not have a linear relationship

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$$

- Diminishing marginal returns [$\beta_2 < 0$]
  - Kitchen output $\sim$ number of chefs
  - Total cost $\sim$ quantity
- Increasing marginal returns [$\beta_2 > 0$]
  - Cell-phone carrier demand $\sim$ number of antennas
  - Most natural monopolies

Testing a Regression with Polynomials

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + u_i$$

Suppose we wanted to conduct the following hypothesis test:

$$\begin{aligned} H_0: &\quad Y \text{ has a linear relationship with } X \\ H_a: &\quad Y \text{ is non-linear with } X \end{aligned}$$

Mathematically:

$$\begin{aligned} H_0: &\quad \beta_2 = \beta_3 = 0 \\ H_a: &\quad \text{Either } \beta_2 \neq 0, \beta_3 \neq 0, \text{ or both} \end{aligned}$$

Testing **multiple** equalities, so have to use an **F-test**

Interpreting Coefficients without Polynomials

What is the average effect of changing X from $X = x$ to $X = x + \Delta x$, holding all else fixed?

Without non-linearities, this was easy:

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

The average effect is $\widehat{\beta}_1 \times \Delta x$ with a standard error of:

$$se(\widehat{\Delta Y}) = \sqrt{\text{Var}(\widehat{\beta}_1 \times \Delta x)} = \Delta x \times se(\widehat{\beta}_1)$$

We did this all the time with $\Delta x = 1$

Interpreting Coefficients with Polynomials

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$$

What is the average effect of changing X from $X = x$ to $X = x + \Delta x$, holding all else fixed?

$$\text{BEFORE:} \quad Y_{\text{before}} = \beta_0 + \beta_1 x + \beta_2 x^2 + u$$
$$\text{AFTER:} \quad Y_{\text{after}} = \beta_0 + \beta_1 (x + \Delta x) + \beta_2 (x + \Delta x)^2 + u$$

On average, the effect of Δx is:

$$\mathbb{E}[Y_{\text{after}} - Y_{\text{before}}] = \beta_1 \Delta x + \beta_2 \left[(x + \Delta x)^2 - x^2 \right]$$

Notice that the effect of changing Δx depends on the initial x!

Interpreting Coefficients with Nonlinearities

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$$

With nonlinearities, the effect of changing x by Δx depends on the initial x

Intuition:

- What is the effect of one additional hour of study on your grade?
  - ... depends on how much you've already studied
- What is the effect of one additional campaign ad on election?
  - ... depends on how much was already spent

log-Regressions

CASES:

1. **Linear-Log**

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

A 1% increase in X is associated a $(0.01 \times \beta_1)$ increase in Y

2. **Log-Linear**

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

A unit increase in X is associated with a $(100 \times \beta_1 \%)$ change in Y

3. **Log-Log**

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

A 1% increase in X is associated with a $(\beta\%)$ change in Y
Ex: price elasticities of demand and supply

Interaction Terms

Interaction terms are the product of two or more variables.

Ex:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + u_i$$

Interaction terms allow for heterogenous treatment effects by group

Interaction Terms – Binary Case

$$\text{WAGE}_i = \beta_0 + \beta_1 F_i + \beta_2 B_i + \beta_3 (F_i \times B_i) + u_i$$

- $F_i = 1$ if female; 0 otherwise
- $B_i = 1$ if black; 0 otherwise

How would we test for any wage discrimination?

$$H_0: \quad \beta_1 = \beta_2 = \beta_3 = 0$$
$$H_a: \quad \text{At least one of } \beta_1, \beta_2, \beta_3 \text{ is non-zero}$$

According to this model, what is the source of discrimination if $\beta_1 \neq 0$? $\beta_2 \neq 0$? $\beta_3 \neq 0$?

Interaction Terms – Hybrid Case

$$\text{WAGE}_i = \beta_0 + \beta_1 EDU_i + \beta_2 F_i + \beta_3 (\text{EDU}_i \times F_i) + u_i$$

The average wage when $EDU = 0$ is...

- β_0 for males
- $\beta_0 + \beta_2$ for females

Dummy variables allow for **different intercepts** across groups

The effect of one additional year of education is...

- β_1 for males
- $\beta_1 + \beta_3$ for females

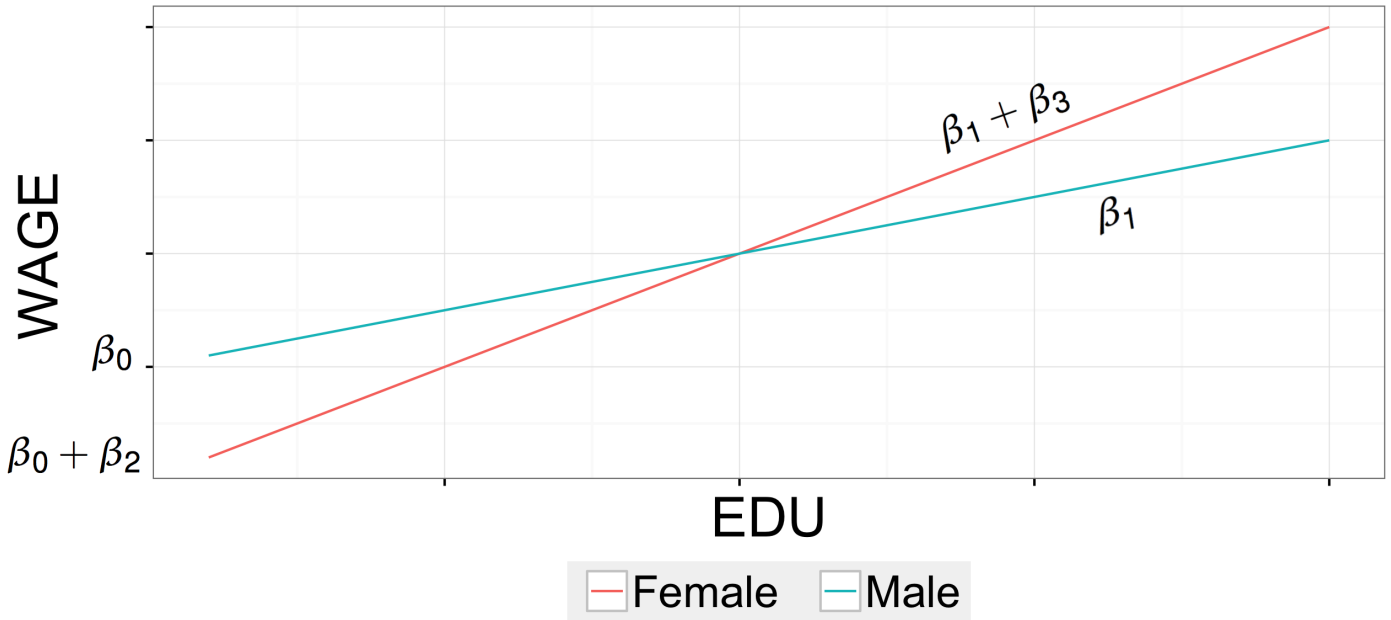
Interaction terms allow for **different slopes** across groups

Interaction Terms – Hybrid Case

$$\text{WAGE}_i = \beta_0 + \beta_1 EDU_i + \beta_2 F_i + \beta_3 (\text{EDU}_i \times F_i) + u_i$$

How would we test for any wage discrimination? $H_0 : \beta_2 = \beta_3 = 0$

How would we test for any effect of education? $H_0 : \beta_1 = \beta_1 + \beta_3 = 0$



Combining It All

Polynomials, logs, interactions, and control variables:

$$\ln Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 D_i + \beta_4 (X_i \times D_i) + \cdots$$
$$\beta_5 (X_i^2 \times D_i) + \beta_6 W_{1i} + \beta_1 W_{2i} + u_i$$

Interpreting Coefficients

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \beta_3 W_{2i} + u_i$$

> "All else equal, a unit increase in X is **associated** with a β change in Y on average"

But economists care about **causality**

When can we claim causality?

> "All else equal, a unit increase in X **causes** a β change in Y on average"

Causality in OLS requires **conditional mean independence** of X

Conditional Mean Independence

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \beta_3 W_{2i} + u_i$$

Conditional Mean Independence of X requires:

$$\mathbb{E}[u|X, W_1, W_2] = \mathbb{E}[u|W_1, W_2]$$

Intuition:

- CMI implies that those with high X are not (unobservably) different from those with low X
- CMI implies X is **as-if randomly assigned**. This parallels a randomized experiment, so we can make statements about **causality**

Endogeneity and Exogeneity

Regressors X that satisfy *conditional mean independence* are called

exogeneous

$\Rightarrow$ Exogenous Xs are **as-if** randomly assigned

Regressors X that fail *conditional mean independence* are called

endogenous

$\Rightarrow$ OLS with endogenous regressors yields biased coefficients that cannot be interpreted causally

Omitted Variable Bias

One of the most common violations of CMI is **omitted variable bias**

OVB occurs when we fail to control for a variable in our regression

Suppose we ran:

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

Instead of:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i$$

Omitted Variable Bias (OVB)

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

Conditions for OVB?

Omitted Variable Bias

OVB arises if a variable W is omitted from the regression and

1. W is a determinant of Y
   - W lies in u_i
2. W is correlated with X
   - $\text{corr}(W, X) \neq 0$

Omitted Variable Bias Example

What is the effect of $X =$ **Student-Teacher Ratio** on $Y =$ average district test scores?

$$\text{Avg Test Scores}_i = \beta_0 + \beta_1 \left(\frac{\text{\# Students}}{\text{\# Teachers}} \right)_i + u_i$$

We estimate the model above and produce

$$\widehat{\beta}_1 = -2.28^{**}$$

If $\widehat{\beta}_1$ were *unbiased*, then we would claim a unit increase in the student-teacher ratio **causes** average test scores to fall by 2.28.

Omitted Variable Bias Example

$$\text{Avg Test Scores}_i = 698.9 - 2.28\left(\frac{\text{\# Students}}{\text{\# Teachers}}\right)_i + u_i$$

But we should doubt the causality claim here. $\widehat{\beta}$ is likely biased due to OVB. Notice that districts with higher incomes likely have higher test grades (**Condition 1**) and lower ratios (**Condition 2**).

One proxy for district income is $X_2 = \%$ of the district who are English learners. Including X_2 in the model:

$$\text{Avg Test Scores}_i = \beta_0 + \beta_1\left(\frac{\text{\# Students}}{\text{\# Teachers}}\right)_i + \beta_2 X_2 + u_i$$

Now we get a *different* causal impact:

$$\widehat{\beta}_1 = -1.10^*$$

Omitted Variable Bias Example

Dependent variable: average test score in the district.

Regressor	(1)	(2)	(3)	(4)	(5)
Student–teacher ratio (X_1)	−2.28** (0.52)	−1.10* (0.43)	−1.00** (0.27)	−1.31** (0.34)	−1.01** (0.27)
Percent English learners (X_2)		−0.650** (0.031)	−0.122** (0.033)	−0.488** (0.030)	−0.130** (0.036)
Percent eligible for subsidized lunch (X_3)			−0.547** (0.024)		−0.529** (0.038)
Percent on public income assistance (X_4)				−0.790** (0.068)	0.048 (0.059)
Intercept	698.9** (10.4)	686.0** (8.7)	700.2** (5.6)	698.0** (6.9)	700.4** (5.5)

- Omitting X_2 raised major OVB, but after including X_2, other controls don't seem to matter much (see (3)-(5))
- The difference in $\widehat{\beta}_1$ between (1) and (2) implies that % ESL must affect average test scores (condition 1) and be correlated with Student-Teacher ratio (condition 2)

Omitted Variable Bias

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

The **direction of the bias** depends on the direction of the two OVB conditions, i.e how W, X, and Y are correlated.

	$\text{Corr}(W, Y) > 0$	$\text{Corr}(W, Y) < 0$
$\text{Corr}(W, X) > 0$	+	-
$\text{Corr}(W, X) < 0$	-	+

When the two correlations go in the same correlation, the bias is positive. When opposite, the bias is negative.

Omitted Variable Bias – Example 1

How does the student-teacher ratio affect test scores?

$$\text{Avg Test Scores}_i = \beta_0 + \beta_1 \left(\frac{\text{\# Students}}{\text{\# Teachers}} \right)_i + u_i$$

Omitted Variable Bias?

$W =$ % English as a Second Language learners

$$\text{Corr}\left(W, \frac{\text{\# Students}}{\text{\# Teachers}} \right) > 0$$

$$\text{Corr}(W, \text{Avg Test Scores}) < 0$$

Hence, there is negative OVB if we neglect to control for % ESL

Omitted Variable Bias – Example 2

Did stimulus spending reduce local unemployment during the Great Recession?

$$\begin{pmatrix} \text{DISTRICT} \\ \text{UNEMPLOYMENT} \end{pmatrix}_i = \beta_0 + \beta_1 \times \begin{pmatrix} \text{LOCAL STIMULUS} \\ \text{SPENDING} \end{pmatrix}_i + u_i$$

Omitted Variable Bias? $W =$ previous unemployment in the area

$$\text{Corr}(W, \text{District Unemployment}) > 0$$

$$\text{Corr}(W, \text{Stimulus Spending}) > 0$$

Hence, there is **positive** OVB if we fail to control for initial unemployment

The OVB Formula

When OVB exists,

$$\widehat{\beta_1} = \beta_1 + \left(\frac{\sigma_u}{\sigma_X}\right)\rho_{Xu}$$

where

$$\begin{aligned}\rho_{Xu} = \text{corr}(X, u) &= \text{corr}(X, \gamma W + e) \\ &= \text{corr}(X, \gamma W) + \text{corr}(X, e) \\ &= \gamma \text{corr}(X, W) + 0\end{aligned}$$

where γ and e come from:

$$Y_i = \beta_0 + \beta_1 X_i + \gamma W_i + e_i$$

I recommend relying on the previous 2x2 table for finding OVB direction

Deriving the OVB Bias Formula

Suppose we *naively* believe the model to be $Y_i = \beta_0 + \beta_1 X_i + u_i$. But we've omitted W, so the true model is:

$$Y_i = \beta_0 + \beta_1 X_i + \gamma W_i + e_i$$

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{\text{Cov}(X_i, Y_i)}{\text{Var}(X_i)} \\
&= \frac{\text{Cov}(X_i, \beta_0 + \beta_1 X_i + \gamma W_i + e_i)}{\text{Var}(X_i)} \\
&= \frac{0 + \text{Cov}(X_i, \beta_1 X_i) + \text{Cov}(X_i, \gamma W_i) + 0}{\text{Var}(X_i)} \\
\widehat{\beta}_1 &= \beta_1 + \gamma \frac{\text{Cov}(X_i, W_i)}{\text{Var}(X_i)}
\end{aligned}$$

Rearranging by $\text{Corr}(X, u) = \frac{\text{Cov}(X,u)}{\sigma_X \sigma_u}$ yields the OVB equation

Fixing OVB

Fixing some causes of OVB is straightforward – we just control for variables by including them in our regression

However, usually implausible to control for *all* potential omitted variables

Other common strategies for mitigating OVB include

- Fixed effects and panel data
- Instrument Variable (IV) regression

Omitted Variable Bias Examples

$$\text{Income Inequality} = \beta_0 + \beta_1 \times \text{Segregation} + U$$

$$\text{Winning \%} = \beta_0 + \beta_1 \times \text{NBA Team Payroll} + U$$

$$\text{Recidivism} = \beta_0 + \beta_1 \times \text{Length of Prison Sentence} + U$$

$$\text{GDP growth} = \beta_0 + \beta_1 \times \text{Civil Conflict} + U$$

$$\text{Health} = \beta_0 + \beta_1 \times \text{Smoking} + U$$

$$\text{Health care expenditures} = \beta_0 + \beta_1 \times \text{Amount of Health Insurance} + U$$

$$\text{Gov't Corruption} = \beta_0 + \beta_1 \times \text{Foreign Aid} + U$$

Internal Validity

Internal validity is a measure of how well our estimates capture what we intended to study (or how unbiased our causal estimates are)

- Suppose we wanted to study wanted to study the impact of student-teacher ratio on education outcomes?
  - **Internal Validity**: Do we have an unbiased estimate of the true causal effect?

Threats to Internal Validity

1. Omitted variable bias
2. Wrong functional form
   - Are we assuming linearity when really the relationship between Y and X is nonlinear?
3. Errors-in-variables bias
   - Measurement errors in X biases $\widehat{\beta}$ toward 0
4. Sample selection bias
   - Is the sample representative of the population?
5. Simultaneous causality bias
   - Is Y "causing" X?
6. "Wrong" standard errors
   - Homoskedasticity *v.* heteroskedasticity
   - Are our observations iid or autocorrelated?

Assessing Internal Validity of a Regression

When assessing the internal validity of a regression:

- No real-world study is 100% internally valid
  - Do **not** write "Yes, it is internally valid."
- Write intelligently
  - What two conditions would potential omitted variables have to satisfy?
  - Why might there be measurement error?
  - ...
- Lastly, assess whether you think these threats to internal validity are large or small
  - i.e. Is your $\widehat{\beta}$ estimate very biased or only slightly biased?
  - Which direction is the bias? Why? There could be multiple OVBs acting in different directions

External Validity

External validity measures our ability to **extrapolate** conclusions from our study to *outside* its own setting.

- Does our study of California generalize to Massachusetts?
- Can we apply the results of a study from 1990 to today?
- Does our pilot experiment on 1,000 students scale to an entire country?
  - Ex: more spending on primary school in the Perry Pre-School Project

No study can be externally valid to *all* other settings. Pick a (important) setting and in a sentence, explain why the study's results may not extrapolate

Section 2

Binary Dependent Variables

Binary Dependent Variables

Previously, we discussed regression for continuous Y, but sometimes Y is binary (0 or 1)

Examples of binary Y:

- Harvard admitted or denied
- Employed or unemployed
- War or no war
- Election victory or loss
- Mortgage application approved or rejected

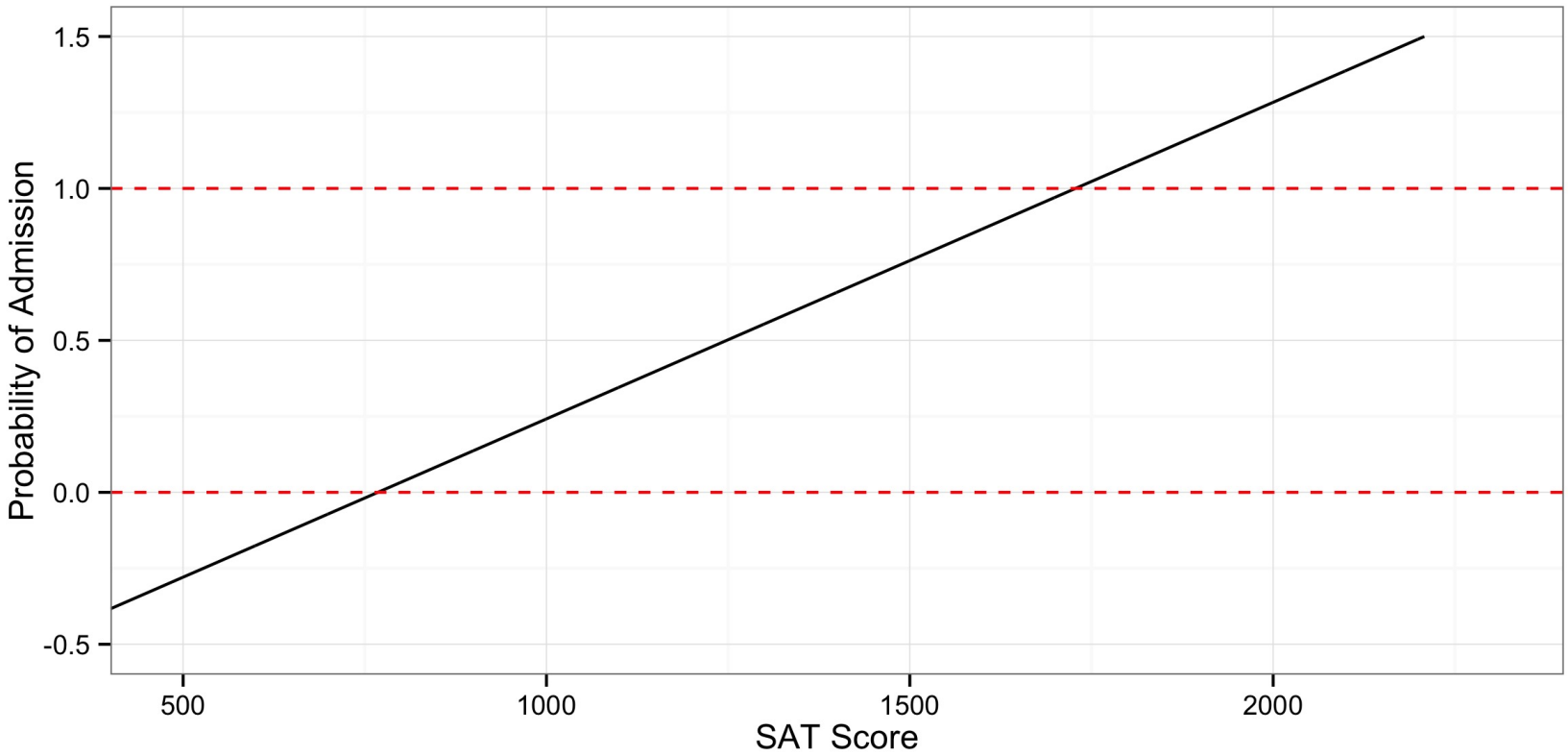
When Y is binary, predicted Y is the probability that $Y = 1$:

$$\widehat{Y}_i = \Pr(Y_i = 1)$$

Binary Dependent Variables

OLS is generally problematic when Y is binary, because

- it generates probabilities $\Pr(Y = 1)$ greater than 1 or less than 0
- it assumes changes in X have a constant effect on $\Pr(Y = 1)$



Probit

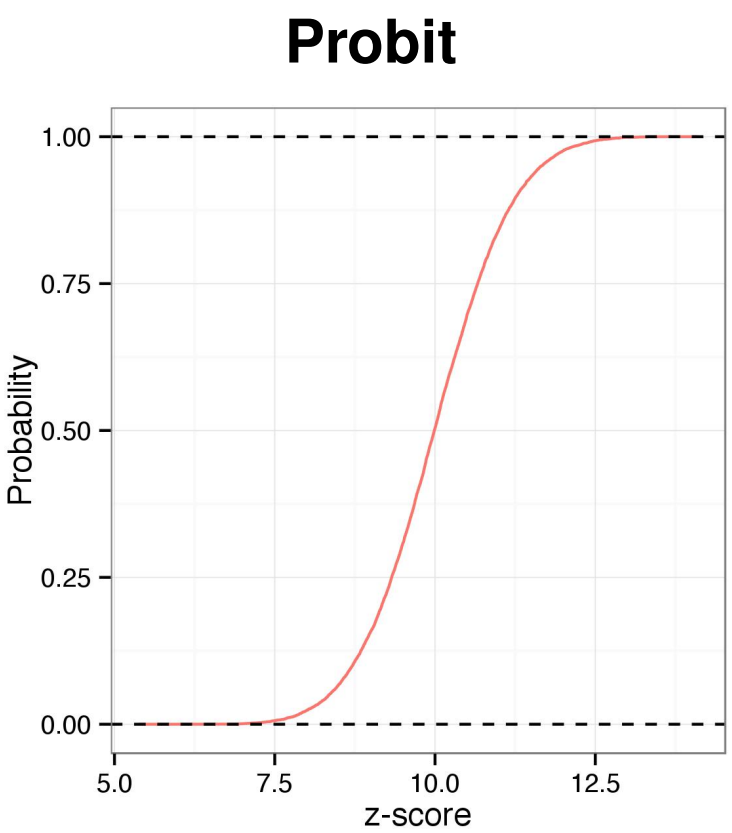
Instead, we put a non-linear wrapper:

$$Y_i = \Phi\left(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i\right)$$

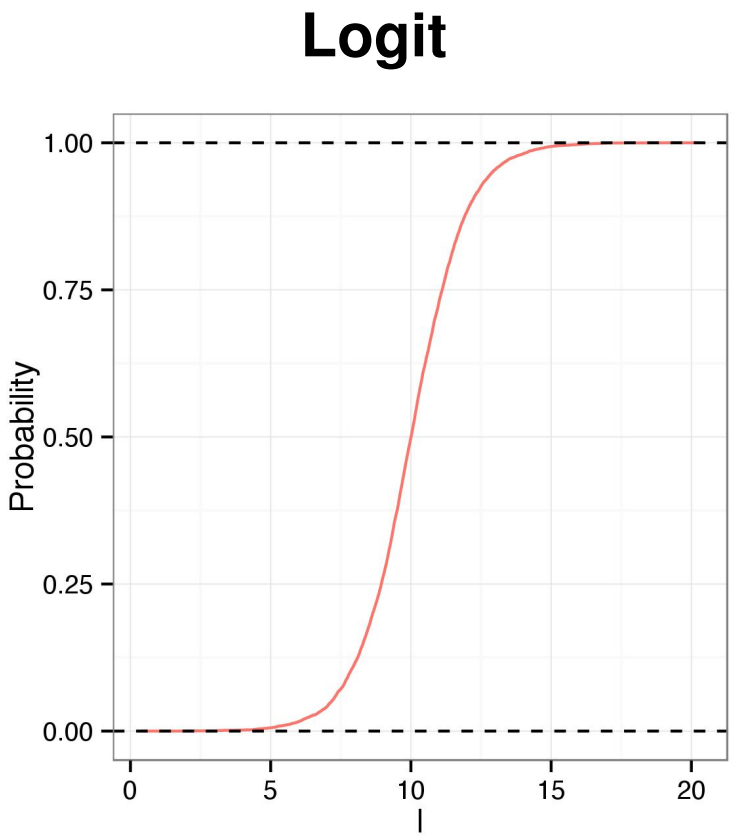
Using **Probit**:

- $\Phi(\cdot)$ is the **normal cumulative density function**
- z-score $= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}$
- So β_1 is the effect of an additional X_1 on the **z-score** (*not on the actual probability* $\Pr(Y_i = 1)$)

Logit & Probit



Normal CDF $\Phi(\cdot)$



Logistic CDF $F(\cdot)$

- Probit and logit nearly identical; just use Probit

Estimating Probit & Logit Model

Both Probit and Logit are usually estimated using **Maximum Likelihood Estimation**

What are the coefficients $\beta_1, \beta_2, \ldots, \beta_j$ that produce expected probabilities $\widehat{Y}_i = \Pr(Y_i = 1)$ most consistent with the data we observe $\{Y_1, Y_2, \ldots, Y_n\}$?

Interpreting Probit Coefficients

$$\Pr(Y_i = 1) = \Phi\left(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}\right)$$

Probit is **non-linear**, so the effect of $\Delta X_{1i} = 1$ is not just β_1

- The effect on probability $\Pr(Y_i = 1)$ depends on initial $\Pr(Y_i = 1)$

Intuitively, the impact of another club presidency on admission to Harvard depends on the previous Pr(Admission)

- If initially $\Pr(Y_i = 1) \approx 0$ or $\Pr(Y_i = 1) \approx 1$, likely ΔX_1 will has no effect
- If initially $\Pr(Y_i = 1) \approx 0.5$, then ΔX_1 could have a big effect if β_1 large

Section 3

Panel Data

Panel Data

Panel Data means we observe the same group of entities over time.

- $N =$ number of **entities**
- $T =$ number of **time periods**

Previously, we studied **cross-section** data, which was a snapshot of entities at just one period of time ($T = 1$).

Panel Data

EXAMPLES:

- Alcohol-related fatalities by state over time
  - What is the effect of a beer tax on alcohol-related fatalities?
- Terrorist fatalities by country over time
  - What is the effect of repressing political freedom on terrorism?
- GPA by student over 1st grade, 2nd grade, 3rd grade, ...
  - What is the effect of a good teacher on test scores?
- Crop output by farmer over time
  - Does access to microfinance help farmers in developing countries?
- Local unemployment by city by month
  - Did stimulus spending improve local labor market conditions during the Great Recession?

Example

Recall: Does a lower student-teacher ratio improve test scores?

$$\text{Avg Test Scores}_{it} = \beta_0 + \beta_1 \left(\frac{\text{Students}}{\text{Teachers}} \right)_{it} + u_{it}$$

Ton of OVB problems:

- district income
- percent english-learners
- quantity and quality extracurricular activities at school
- teacher quality
- access to resources
- ...

Example

Recall: Does a lower student-teacher ratio improve test scores?

$$\text{Avg Test Scores}_{it} = \beta_0 + \beta_1 \left(\frac{\text{Students}}{\text{Teachers}} \right)_{it} + u_{it}$$

Ton of OVB problems, and unrealistic to control for all of them

Instead, we can have a school dummy variable, i.e. a school **fixed effect**

$$\text{Avg Test Scores}_{it} = \alpha_i + \beta_1 \left(\frac{\text{Students}}{\text{Teachers}} \right)_{it} + u_{it}$$

α_i is a school-specific intercept (it's just like a dummy variable equal to 1 only for school i)

Advantages of Panel Data

Panel Data enables us better control for **entity-specific, time-invariant** effects.

because we observe how the *same* entity responds to different X's

Back to the test score example:

With entity fixed effects, we are only relying on variation **within** the **same** school over time

i.e. comparing School #56 (2002) v. School #56 (2004) if its student-teacher ratio changed.

More credible than comparing School #56 to School #89 in totally different states

Advantages of Panel Data

Can panel data solve all omitted variable bias?

No. There's no such silver bullet (other than perfectly random assignment in an experiment setting). OVB could still arise from omitting factors that change over time.

In our example, what if some schools have successfully recruited better teachers than others during the sample period?

Entity Fixed Effects

$$Y_{it} = \alpha_i + \beta_1 X_{1,it} + \beta_2 X_{2,it} + u_{it}$$

Entity Fixed Effects α_i allow each entity i to have a different intercept

- Using entity FE controls for any factors that vary across states but are constant over time
  - **e.g.** geography, environmental factors, anything relatively static across time
- Entity FE are like per-entity dummy variables

Entity FE means we are using only **within-entity** variation for identification

Time Fixed Effects

$$Y_{it} = \alpha_i + \gamma_t + \beta_1 X_{1,it} + \beta_2 X_{2,it} + u_{it}$$

Time Fixed Effects γ_t control for factors constant across entities but not across time

Time FE are basically dummy variables for time. **Ex:**

$$Y_{it} = \alpha_i + \mathbf{D}_{2013} + \mathbf{D}_{2012} + \cdots + \mathbf{D}_{2014} + \beta_1 X_{1,it} + \beta_2 X_{2,it} + u_{it}$$

Estimating Standard Errors in Panel Data

Typically we assume that observations are **independent**, so u_1 is independent from u_2, u_3, ...

With **panel data**, we observe the same entity over time, so u_{it} and u_{it+1} may be *correlated*

- Unobservables are likely to linger more than a single period

For the same entity over time, we allow the errors to be **serially correlated** or **autocorrelated** (i.e. correlated with itself)

Allowing autocorrelation within entities is a much weaker assumption than independence. Assuming independence across u_{it} produces wrong standard error estimates!

Clustering Standard Errors by Entity

To account for this autocorrelation, we **cluster** standard errors by entity

```
xtreg y x1 x2, fe vce(cluster entity)
```

This assumes that observations across different entities are still independent, but observations within the same entity (i.e. cluster) may be correlated

Clustered Standard Errors Example

Back to our example:

$$\text{Avg Test Scores}_{it} = \alpha_i + \beta_1 \left(\frac{\text{Students}}{\text{Teachers}} \right)_{it} + u_{it}$$

In this case, we have panel data on schools over time,

- Schools are our entity; we should cluster errors by school

```
xtreg testscore ratio, fe vce(cluster school)
```

This approach generates correct standard errors so long as different schools give independent observations (even though observations for the same school over time are autocorrelated)

Section 4

Instrumental Variables

Instrumental Variables

Instrumental Variables (IV) are useful for estimating models

- with simultaneous causality or
- with omitted variable bias
  - IV especially useful when we cannot plausibly control for all omitted variables

More generally, IV useful whenever conditional mean independence of X fails

Instrumental Variables

What is the **causal** relationship between X and Y?

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

Suppose our model suffers from omitted variable bias and simultaneous casuality, so CMI fails.

Hence, OLS produces a biased estimate of $\widehat{\beta}_1$ that we cannot interpret causally

Suppose we have another variable Z that is a valid instrument, then we can recover a $\widehat{\beta}_{IV}$ with a causal interpretation

IV Conditions

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + u_i$$

Suppose OLS yields a biased estimate $\widehat{\beta}_1$ because conditional mean independence fails:

$$\mathbb{E}[u_i | X_i, W_{1i}] \neq \mathbb{E}[u_i | W_{1i}]$$

Conditions for IV

Z is a valid instrumental variable for X if:

1. **Relevance:** Z is related to X:

$$\text{Corr}(Z, X) \neq 0$$

2. **Exogeneity:** Controlling for Ws, the only effect that Z has on Y goes through X:

$$\text{Corr}(Z, u) = 0$$

Conditions for IV: Intuition

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

Z is an instrumental variable for X in this model if:

Condition 1: **Relevance**

Z is related to X $\qquad$ $\text{Corr}(Z, X) \neq 0$

Condition 2: **Exogeneity** of Z

$$\text{Corr}(Z, u) = 0$$

Two ways of saying the **Exogeneity** condition:

- Z is **as-if randomly** assigned
- The only effect of Z on Y goes through X

Both C1 and C2 must hold after controlling for Ws

IV Intuition

- **Problem:** $\widehat{\beta}_1$ is biased, because X fails conditional mean independence, because X is **not** as-if randomly assigned
- Our goal is to find the variation in X is that is as-if randomly assigned, so we can estimate causality
- We have an instrument Z and assuming Z satisfies Condition 2: Exogeneity, Z is as-if randomly assigned
- Furthermore, assuming Condition 1: Relevance is also satisfied, Z is related to X, so there must be some part of X that is **as-if randomly assigned**

IV uses the as-if random assignment of X induced by Z to estimate $\widehat{\beta}_{IV}$

IV Conditions – Graphically

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

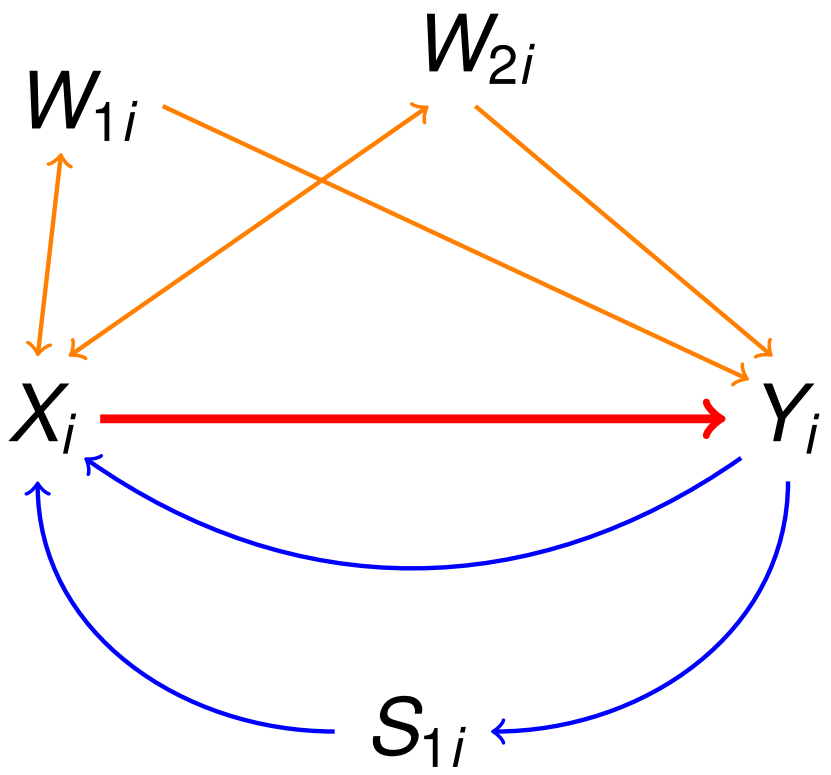
Suppose we are investigating the effect of $X \longrightarrow Y$ but we suffer OVB and simultaneous causality:

- Omitted variables W_{1i} and W_{2i}:
  - $W_1 \longrightarrow Y$ and $W_1 \longleftrightarrow X$
  - $W_2 \longrightarrow Y$ and $W_2 \longleftrightarrow X$
- Simultaneous causality
  - $Y \longrightarrow X$
  - $Y \longrightarrow S_1 \longrightarrow X$

Suppose we have an instrument Z

Instrumental Variables

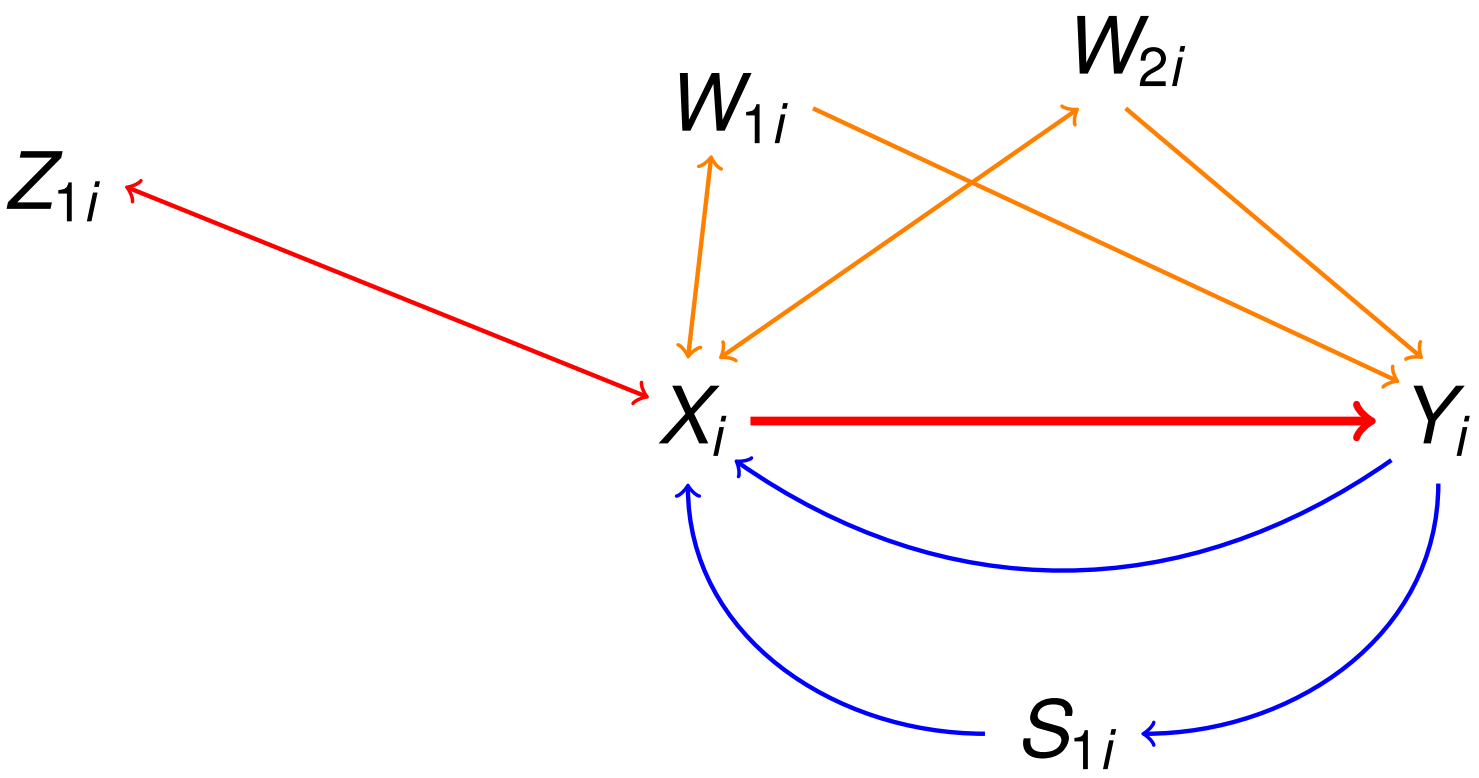
$$Y_i = \beta_0 + \beta_1 X_i + u_i$$



Because of OVB and simultaneous causality: $\mathbb{E}[u|X] \neq 0$

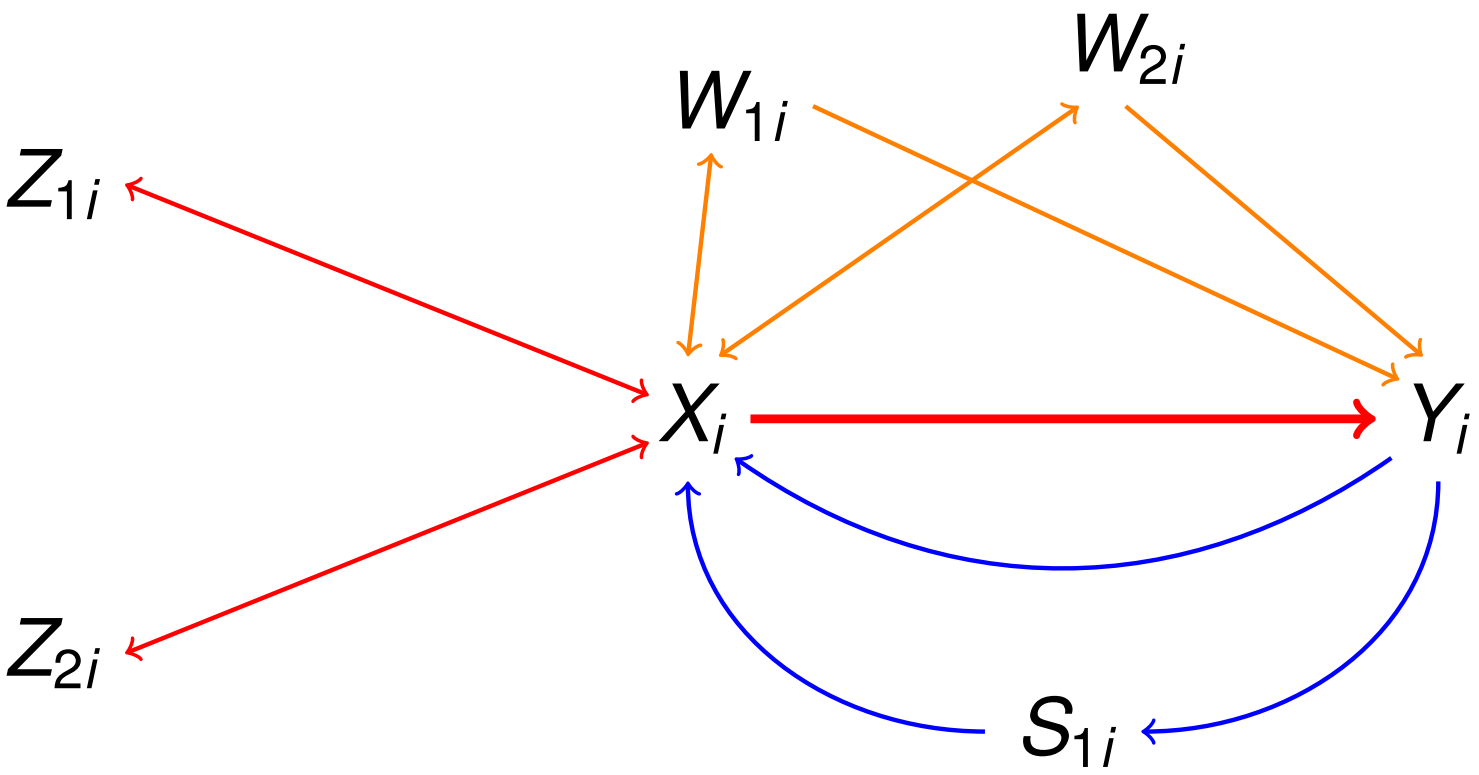
Instrumental Variables

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



Instrumental Variables

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



There can be multiple instruments Z_1 and Z_2 for the same X

IV

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

What is NOT allowed (by Condition 2: Exogeneity) –

- $Z_i \longleftrightarrow Y_i$
- $Z_i \longleftrightarrow W_{1i}$ and $Z \longleftrightarrow W_{2i}$
- $Z_i \longleftrightarrow S_{1i}$
- ...

Suppose we also include control variable W_{1i}

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + u_i$$

What is NOT allowed (by Condition 2: Exogeneity) –

- $Z_i \longleftrightarrow Y_i$
- $Z_i \longleftrightarrow W_{2i}$
- $Z_i \longleftrightarrow S_{1i}$
- ...

Examples of IVs

Z	X	Y
How does prenatal health affect long-run development?		
Pregnancy during Ramadan	Prenatal health	Adult health & income
What effect does serving in the military have on wages?		
Military draft lottery #	Military service	Income
What is the effect of riots on community development?		
Rainfall during month of MLK assassination	Number of riots	Long-run property values

Each of these examples requires some control variables Ws for the exogeneity condition to hold. In general, arguing the exogeneity condition can be very difficult.

Testing the Validity of Instruments

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \beta_3 W_{2i} + u_i$$

Conditions for IV

Z is an instrumental variable for X in this model if:

1. **Relevance:** $\text{Corr}(Z, X) \neq 0$
2. **Exogeneity:** $\text{Corr}(Z, u) = 0$

- Testing **Condition 1** is straightforward, since we have data on both Z and X
- Testing **Condition 2** is trickier, because we never observe u. In fact, we can only test Condition 2 when we have more instruments Zs than endogenous Xs

Testing Condition 1: Relevance

Condition 1: **Relevance**

Z must be related to X. i.e. $\text{Corr}(Z, X) \neq 0$

We need the relationship between X and Z to be meaningfully "large"

How to check?

Run **first-stage** regression with OLS (if we have multiple instruments, include all of them)

$$X_i = \alpha_0 + \alpha_1 Z_{1i} + \alpha_2 Z_{2i} + \alpha_3 W_{1i} + \alpha_4 W_{2i} + \cdots + v_i$$

Check that the F-test on all the coefficients on the instruments α_1, α_2

- If $\widehat{F} > 10$, we claim that Z is a **strong instrument**
- If $\widehat{F} \leq 10$, we have a weak instruments problem

Testing Condition 2: Exogeneity

J-test for overidentifying restrictions:

H_0 : Both Z_1 and Z_2 satisfy the exogeneity condition

H_a : Either $Z_1, Z_2,$ or both are invalid instruments

```
ivregress y w1 w2 (x = z1 z2), robust
estat overid
display "J-test = " r(score) " p-value = " r(p_score)
```

If the p-value < 0.05, then we reject the null hypothesis that all our instruments are valid

But just like in the F-test case, rejecting the test alone does not reveal which instrument is invalid, only that at least one fails the exogeneity condition

Two-Stage Least Squares

IV regression is typically estimated using **two-stage least squares**

First Stage: Regress X on Z and W

$$X_i = \alpha_0 + \alpha_1 Z_i + \alpha_2 W_i + v_i$$

Second Stage: Regress Y on $\widehat{X}$ and W

$$Y_i = \beta_0 + \beta_1 \widehat{X}_i + \beta_2 W_i + u_i$$

```
ivregress 2sls y w (x = z), robust
```

Intuition: If the instrument Z satisfies the two IV conditions:

- The first stage of 2SLS isolates the as-if random parts of X
- $\widehat{X}$ satisfies conditional mean independence

Local Average Treatment Effect (LATE)

If CMI is satisfied, $\widehat{\beta}_{\text{OLS}}$ identifies the **average treatment effect**

IV estimate $\widehat{\beta}_{\text{IV}}$ identifies the ***local* average treatment effect** (LATE)

INTUITION:

LATE is the weighted-average treatment effect for entities affected by the instrument, weighted more heavily toward those most affected by the instrument Z

The word *local* indicates the LATE is the average for this affected group known as **compliers**. Compliers are those affected by the instrument (i.e. they **complied** with Z)

More LATE Intuition

Recall that IV works by using the as-if random variation in X induced by the instrument Z

- Possible that for some entities (*non-compliers*), their X will not change according to Z
  - Hence, there is no variation in X induced by Z, so IV ignores these "non-compliers"
- Instead, IV estimates the treatment effect for entities whose X is related to Z (this is the **LATE**)

More LATE

Suppose we are estimating the causal effect of X on Y:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{2i} + \beta_3 W_{3i} + u_i$$

We have two valid instruments Z_1 and Z_2.

- We just use Z_1 and run 2SLS to estimate $\widehat{\beta}_{2SLS}$
- We just use Z_2 and run 2SLS to estimate $\tilde{\beta}_{2SLS}$

Should we expect our estimates to equal?

$$\widehat{\beta}_{2SLS} \stackrel{?}{=} \tilde{\beta}_{2SLS}$$

More LATE

Suppose we are estimating the causal effect of X on Y:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{2i} + \beta_3 W_{3i} + u_i$$

We have two valid instruments Z_1 and Z_2.

- We just use Z_1 and run 2SLS to estimate $\widehat{\beta}_{2SLS}$
- We just use Z_2 and run 2SLS to estimate $\tilde{\beta}_{2SLS}$

Should we expect our estimates to equal?

$$\widehat{\beta}_{2SLS} \stackrel{?}{=} \tilde{\beta}_{2SLS}$$

No. Different complier groups may respond to the different instruments Z_1 and Z_2. So each instrument may have a different LATE

ATE v. LATE

LATE $=$ **ATE** if any of the following is true

- no heterogeneity in treatment effects
  - **i.e.** the effect of X_i on Y_i is the same for all i
- no heterogeneity in first-stage responses to the instrument Z
  - **i.e.** the effect of Z_i on X_i is the same for all i; everyone's X_i complies identically to the instrument Z_i
- no correlation between treatment effect (X_i on Y_i) and first-stage effect (Z_i on X_i)
  - **i.e.** compliers are on average the same as non-compliers

ATE v. LATE

Which do we care about: ATE or LATE?

Depends on the context.

- if proposed policy is to give everyone the treatment, then ATE
- if proposed policy only affects a subset, then maybe LATE is more appropriate

Internal Validity with IV

If the IV is valid, then instrument variable regression takes care of:

- Omitted Variable Bias
- Simultaneous Causality Bias
- Errors-in-variables (or measurement error)

Thus, internal validity in an IV regression is mostly about assessing the two IV conditions:

- Relevance of Z
- Exogeneity of Z

Section 5

Forecasting

Forecasting

With forecasting, forget causality. It's all about **prediction**.

How can we use past Y to predict future Y?

i.e. What can $(Y_{t-4}, Y_{t-3}, \ldots, Y_t)$ tell us about Y_{t+1} or Y_{t+n}?

Examples of Y_t

- GDP
- Oil prices
- Stock market indices
- Exchanges rates
- ...

Forecasting Vocabulary

- **Lag**: Y_{t-p} is the p^{th} lag of Y_t
- **Autocovariance** – covariance of a variable with a lag of itself

$$\text{Cov}(Y_t, Y_{t-j}) \qquad \text{“}j^{th}\text{ autocovariance”}$$

- **Autocorrelation** – correlation of a variable with a lag of itself

$$\text{Corr}(Y_t, Y_{t-j}) \qquad \text{“}j^{th}\text{ autocorrelation”}$$

- Both autocovariance and autocorrelation measure how Y_{t-j} is related to Y_t

Stationarity

For the past to be useful for predicting the future, the process must be **stationary**

Let Y_t be a process that evolves over time: $(Y_{t_0}, Y_{t_0+1}, \ldots, Y_T)$

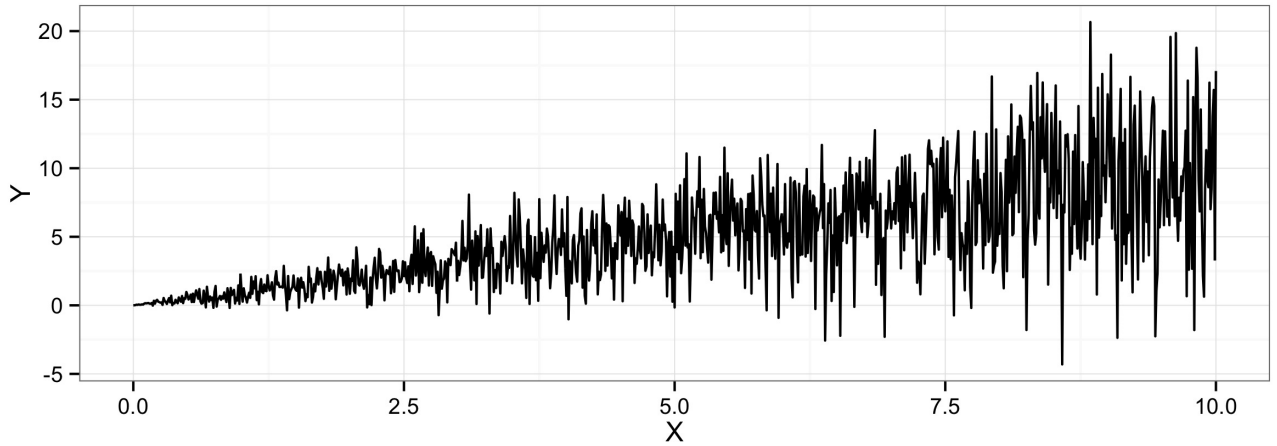
Y_t is **stationary** if all three are true:

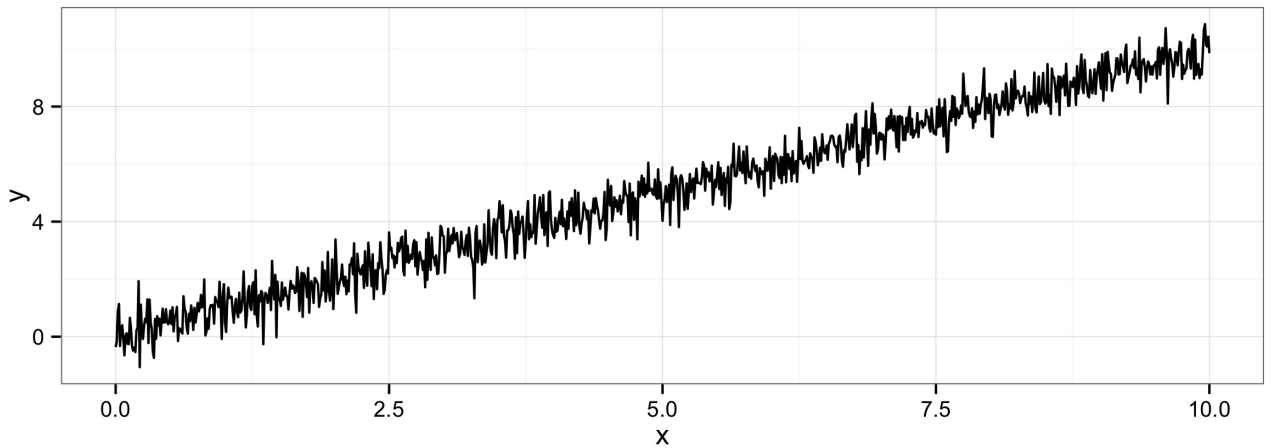
- $\mathbb{E}[Y_t]$ is constant over time
- $\text{Var}(Y_t)$ is constant over time
- Autocorrelation $\text{Corr}(Y_t, Y_{t-j})$ depends only on j and not t

i.e. the behavior of Y_t isn't fundamentally changing over time

Examples of Non-Stationarity

Non-stationary. $\text{Var}(Y_t)$ is increasing over time



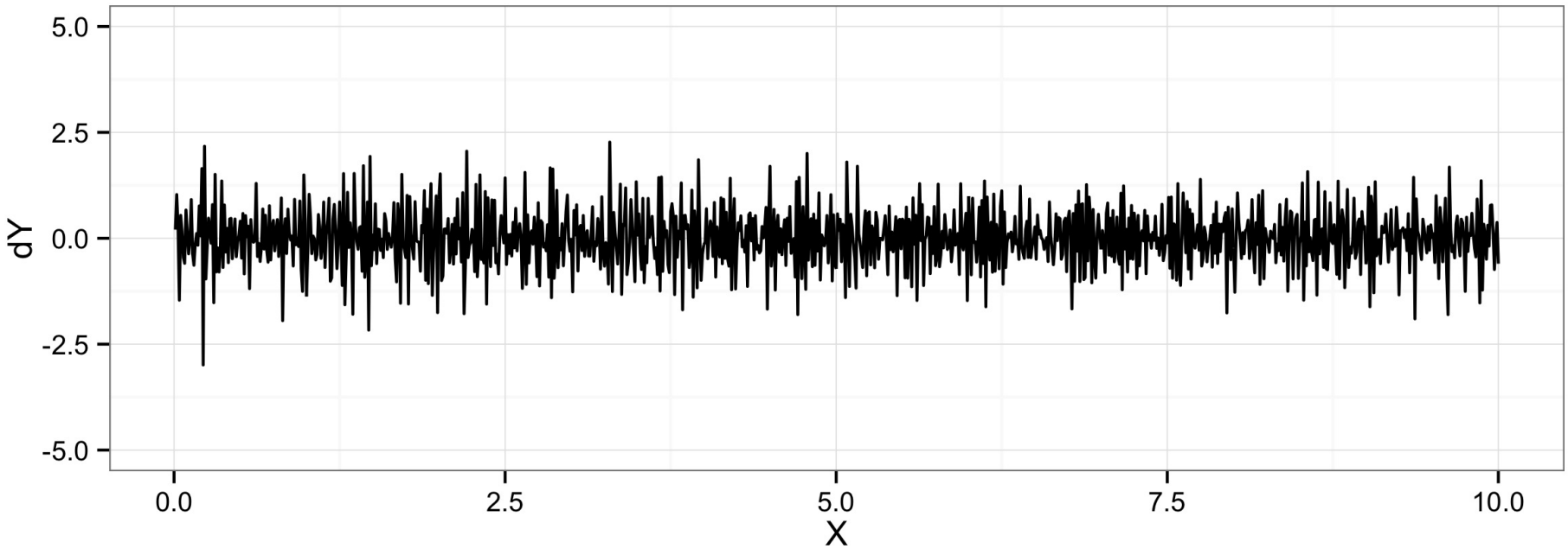


Non-stationary. $\mathbb{E}[Y_t]$ is increasing over time

First Differences

Even when Y_t is non-stationary, first-differences might be stationary!

$$\Delta Y_t = Y_t - Y_{t-1}$$



For example, GDP is not stationary but GDP growth is

Forecasting Models

Assuming Y_t is stationary:

AR(p): Autoregressive Model of Order p:

Regression of Y_t on p lags of Y_t

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \cdots + \beta_p Y_{t-p} + u_t$$

Ex: **AR(1)**: $Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t$

Ex: **AR(4)**: $Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \beta_4 Y_{t-4} + u_t$

ADL(p,q): Autoregressive Distributed Lag Model

Regression of Y_t on p lags of Y_t and q lags of X_t

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \cdots + \beta_p Y_{t-p} + \delta_1 X_{t-1} + \cdots + \delta_q X_{t-q} + u_t$$

Models

Suppose you were trying to predict $Y_t =$ US GDP Growth

- Past ΔGDP is likely a good predictor
  - Using just past Y means you're applying an **AR** model
- However, you might also want to test gas prices, housing demand, inflation, European GDP as additional predictors
  - Now this becomes a **ADL** model

Model Selection: Choosing number of lags p

How many lags p should the model include?

$$\textbf{AR}(p): \quad Y_t = \beta_0 + \beta_1 Y_{t-1} + \cdots + \beta_p Y_{t-p} + u_t$$

We choose p by minimizing the **information criterion**:

Information Criterion

information criterion is a measure of how much *information* in our dataset is **not captured** by our model

Intuitive that we want to choose the model (i.e. choose p) with the smallest **IC**

Minimizing Information Criterion IC(p)

Choose p to minimize **Bayes' information criterion**, BIC(p)

$$\min_{0 \leq p \leq p^{\max}} \text{BIC}(p) = \min_{0 \leq p \leq p^{\max}} \ln\left(\frac{\text{SSR}(p)}{T}\right) + (p+1)\left(\frac{\ln T}{T}\right)$$

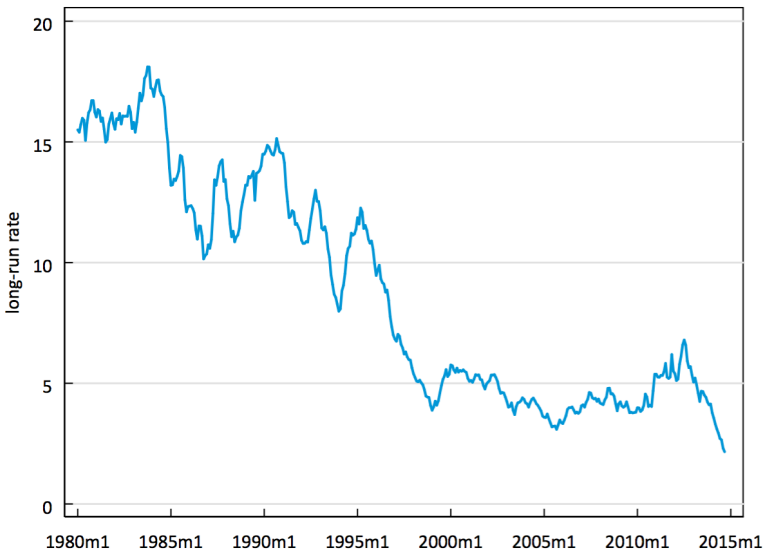
- SSR(p) is the sum of squared residuals when number of lags is p
  - SSR(p) is the variation in Y_t not captured by the model
- $\left(\frac{\ln T}{T}\right)$ is a "penalty" factor associated with increasing p by 1
  - Need this penalty term because SSR(p) is always decreasing in p
- BIC trades off the *decrease* in bias from including important lags against the *increase* in variance from including irrelevant lags

Testing Stationarity

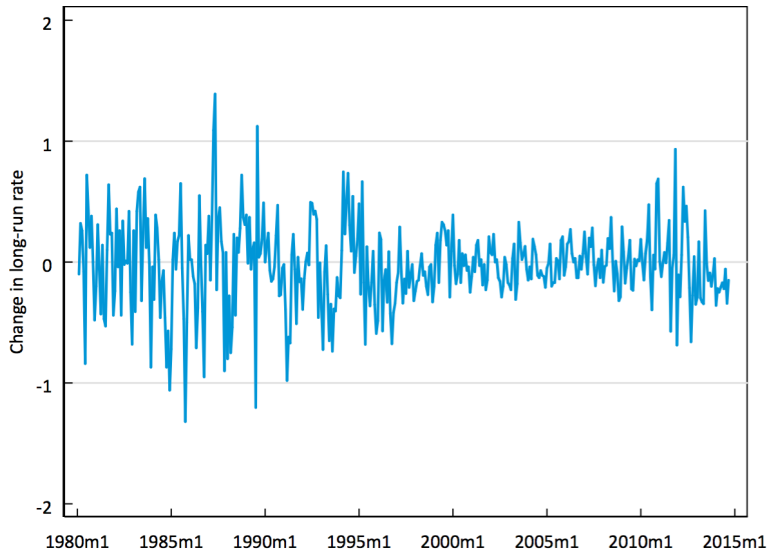
- Often difficult to confirm Y_t is stationary
- Breaks (or **structural breaks**) imply non-stationarity because the underlying relationships between Y_t and Y_{t-j} have changed

Ex: Was there a break? If so, when did the break occur?

Levels Y_t

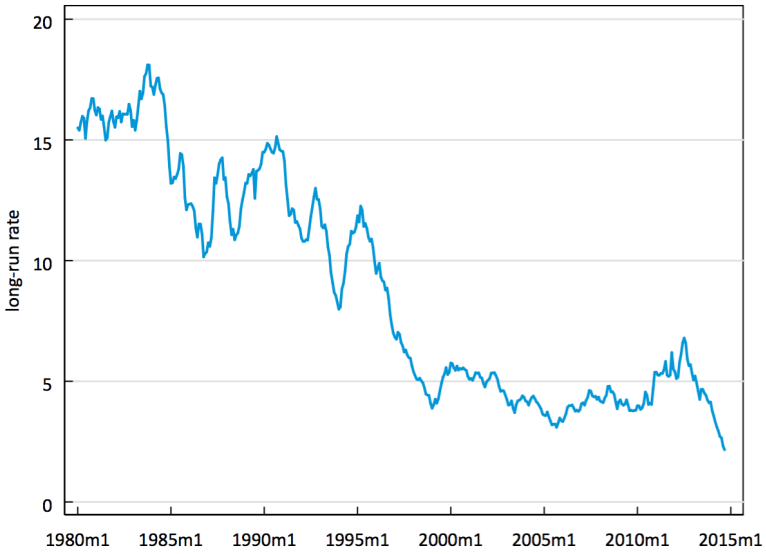


Differences $Y_t - Y_{t-1}$

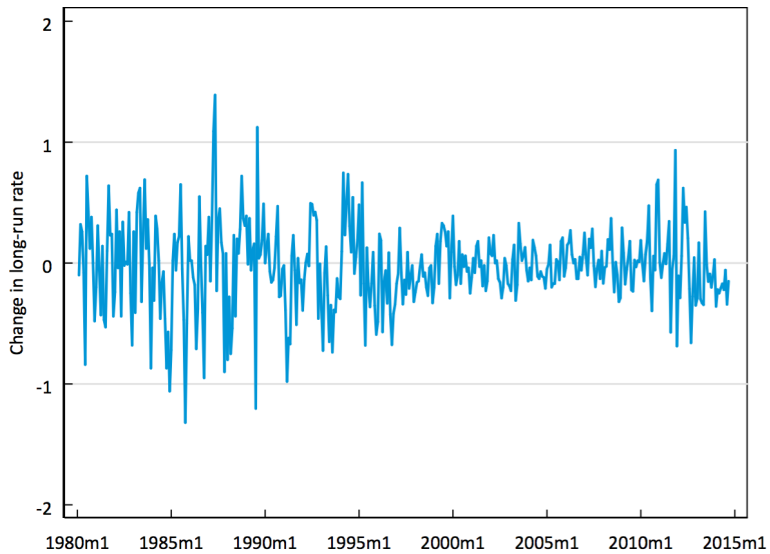


Testing Stationarity

Levels Y_t



Differences $Y_t - Y_{t-1}$



Intuitively, the best way to test for a potential stationarity break at time τ is estimate the model before τ and after τ

- If the model before and after are highly dissimilar, reject stationarity
- This is what the Chow Test formally does

Chow Test for Structural Breaks

Suppose you want to test if there was a break at specific time τ:

$$D_{\tau,t} = \begin{cases} 1 & \text{if } t \geq \tau \\ 0 & \text{if } t < \tau \end{cases}$$

AR(1): $Y_t = \beta_0 + \beta_1 Y_{t-1} + \delta_1 D_{\tau,t-1} + \gamma_1 \left(Y_{t-1} \times D_{\tau,t-1} \right) + u_t$

Chow test:

$$H_0 : \ \delta_1 = \gamma_1 = 0$$

If reject:

We have evidence that the relationship between Y_t and lag Y_{t-1} differs over time. Hence, we reject stationarity

QLR Test

We can use the Chow Test when testing for a break at a specific time τ

What if we want to test for any structural breaks across all time periods?

We calculate the **Quandt Likelihood Ratio Statistic**, which is the maximum Chow statistic across all τ in the central 70% of the time interval

$$\text{QLR} = \max_{\tau} \; F(\tau)$$

Calculating the Chow statistic for $\{\tau_0, \tau_0 + 1, \tau_0 + 2, \ldots, \tau_1 - 1, \tau_1\}$ to find the maximum manually can be very time-consuming

Fortunately, someone has written a `qlr` command in `STATA`

Good luck!