w241: Experiments and Causality

Covariates and Regression

David Reiley, David Broockman, D. Alex Hughes
UC Berkeley, School of Information
Updated: 2021-06-04

Returns to Schooling

Reading: *Mastering Metrics* pages 209–211.

- Section 6.1 gives another example of regression and OVB in observational data.
- This regression:
  - Includes a quadratic term.
  - The dependent variable is earnings.
  - The main covariate is experience.
  - It includes both a linear experience term and an experience-squared term:
  - This shows that earnings increase with experience but increase more slowly in later years.

Work Experience as a Covariate

Experience as a Covariate

Goal: Estimate the "returns to schooling"

- How much does an additional year of schooling cause a person to earn?
- Mincer includes *work experience* as a covariate:
  - People with less schooling but much more work experience might earn more than people with more schooling but no work experience.

Equation 6.2:

$$ln(Y_i) = \alpha + 0.70S_i + \epsilon_{i,short}$$
$$ln(Y_i) = \alpha + 0.107S_i + 0.81X_i - 0.0012X_i^2 + \epsilon_{i,long}$$

- The *short model* estimates an effect of 0.07 on years of schooling
- The *long model* estimates an effect of 0.107 on years of schooling?

Why are these estimates different?

Experience as a Covariate (cont'd)

Omitted Variables Bias

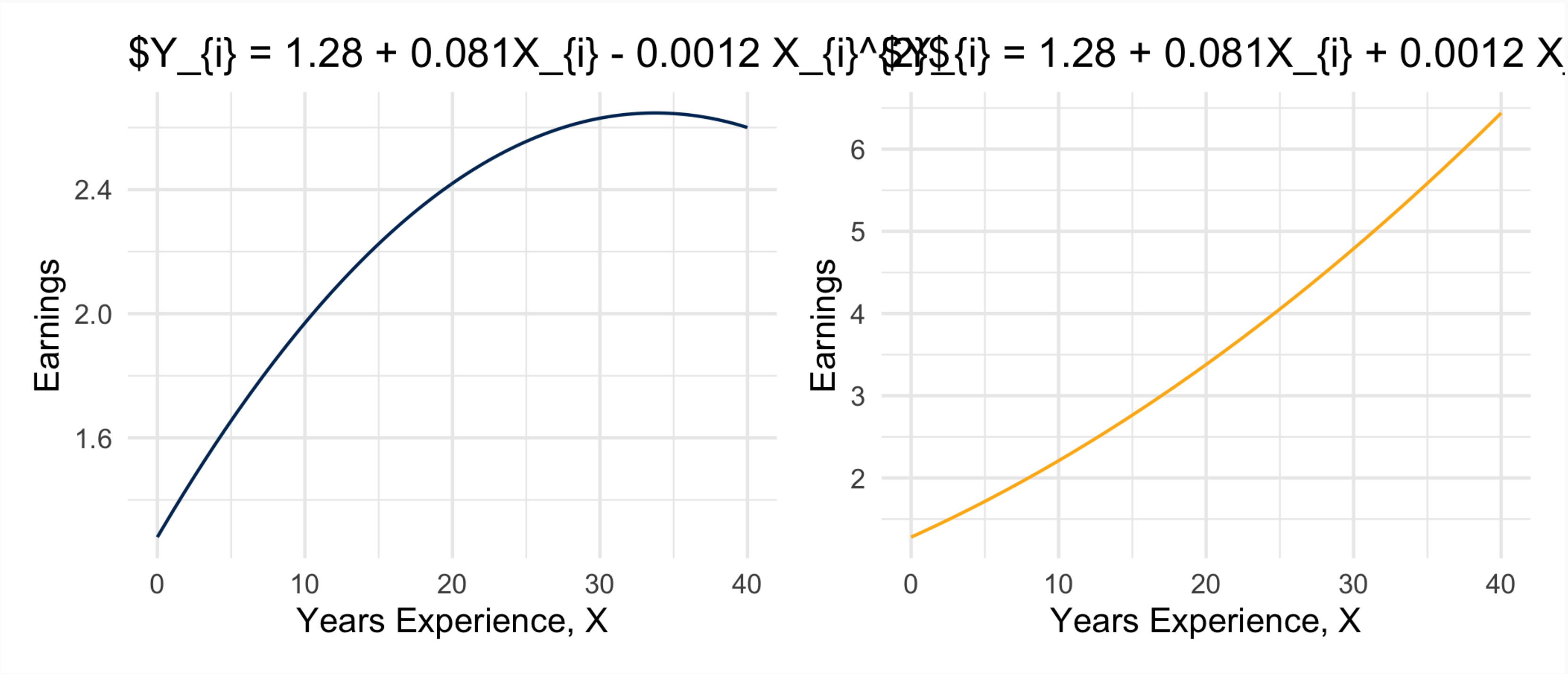
- Omitted variable bias, *OVB*, leads us to underestimate the returns to schooling. **Why?**

Obsesrvational Data

- The two regressors -- *schooling* and *experience* can be correlated with each other!
  1. Schooling and work experience are negatively correlated with one another
  2. Schooling and experience are positively correlated with earnings.
- If estimate the short model (we omit experience), its effect is measured as part of the schooling estimate
  - People with more schooling have less experience (negatively correlated)
  - When we increase schooling, earnings don't increase as much as they would if we were holding experience constant as a covariate.
  - Hence, with OVB we measure the coefficient of interest to be 7% instead of 11%.

Quadratic Specification: Equation 6.2

- The quadratic specification allows for flexibility in the fit: rather than a linear effect, it permits a changing effect at different levels of the covariate
- In this case, we estimate that the benefits of experience to accrue at a declining rate.



Reading

Read: *Mastering Metrics* 211 - 214

- Please read from the bottom of page 211 to the middle of page 214.

Omitted Variable Bias and Attenuation Bias

Is control for experience sufficient for **ceteris to be paribus**?

Is control for experience sufficient for **all else to be equal**?

Ability as an Omitted Variable

- Griliches expected Mincer's estimates to be overstated
  - Omitted ability from the regression!
  - Years of schooling are positively correlated with ability.
- When Griliches included IQ as a covariate, the estimated returns to schooling fell!
  - From 6.8% per year of schooling to 5.9% per year of schooling.
  - Without IQ, OVB caused an overestimate of returns to schooling.

All good?

- After controlling for IQ as a measure of ability, Have we now controlled for everything that might cause biased estimates?
- **No!**.
- There are more kinds of ability than IQ: emotional intelligence, curiosity, and many more.
- We still have omitted variables that are likely to be correlated with years of schooling.

Ability as an Omitted Variable (cont.)

- How do we know when we've got the *right* set of covariates so that we've got an unbiased estimate?
  - **We can't know!**
- We would need an experiment that randomly sends some students to more years of school than others.
- Then every possible omitted variable would be uncorrelated with years of schooling, eliminating OVB.

Attenuation Bias

Angrist and Pischke

- Imagine that we don't always correctly measure the treatment variable (years of schooling).
- With measurement error in the X variable, the resulting coefficient is biased toward zero.

Effects of Online Advertising

- Matched *Yahoo!* users to retail purchases using names and e-mail addresses.
- *Suppose the matching procedure allowed for nonexact matches.*
    - I might have some purchasers who I thought were in the control group, but who were really in the treatment group
    - **As a result** Some of the advertising effects would appear in the control group rather than in the treatment group,
    - The effects would look smaller than actual.

Reading

Optional Reading: *Mastering Metrics* pages 240-241

- This is the appendix to Chapter 6 and covers more about attenuation bias

Required Reading: *Mastering Metrics* pages 214-217

- Next, read *Mastering Metrics* pages 214–217 on bad controls.
- Bad controls are a type of covariate we do not want in our regressions when we analyze experiments

Bad Controls

When you analyze experimental results, do not include other outcome variables as covariates on the right-hand side of the regression.

Example: Random Assignment to College

Table 6.1 How Bad Control Creates Selection Bias

<table>
<tr><th rowspan="2">Type of worker</th><th colspan="2">Potential occupation</th><th colspan="2">Potential earnings</th><th colspan="2">Average earnings by occupation</th></tr>
<tr><th>Without college (1)</th><th>With college (2)</th><th>Without college (3)</th><th>With college (4)</th><th>Without college (5)</th><th>With college (6)</th></tr>
<tr><td>Always Blue (AB)</td><td>Blue</td><td>Blue</td><td>1,000</td><td>1,500</td><td rowspan="2">Blue 1,500</td><td>Blue 1,500</td></tr>
<tr><td>Blue White (BW)</td><td>Blue</td><td>White</td><td>2,000</td><td>2,500</td><td rowspan="2">White 3,000</td></tr>
<tr><td>Always White (AW)</td><td>White</td><td>White</td><td>3,000</td><td>3,500</td><td>White 3,000</td></tr>
</table>

- True average treatment effect (ATE) is \$500.

Example: Random Assignment to College

- Reminder, true ATE is $500

$$Y_i = \alpha + \beta E_i + \gamma W_i + \epsilon_i$$

- A regression on both a *college education* dummy, E_i, and a *white-collar occupation* dummy, W_i:
  - Yields a coefficient $\beta = 0$
  - Will mistakenly indicate that the return to college education is $0.
- Happens because:
  - Only the **most** talented non-college-educated workers will take white-collar jobs
  - Only the **least** talented college-educated workers will take blue-collar jobs.

More About Bad Controls

- We generally want to know the total effect of schooling on earnings.
  - Schooling helps you become a data scientist,
  - More educated data scientists earn more than less educated data scientists.
- Including the occupation covariate is therefore a bad idea: *It picks up only the latter kind of variation.*

Example: eBay Reputation

Does having a higher eBay reputation causes the seller to earn more revenue on eBay?

- Two eBay seller accounts:
  - One account with a low reputation
  - One account with a high reputation
- Measures:
  - Outcome, Y = auction price
  - Treatment, D = seller reputation
  - Covariate, X = number of bids

(Bad Controls) Estimating Equation

$$Y = \beta_0 + \beta_1 D_i + \beta_2 X_i + \epsilon_i$$

- Number of bids is a bad control.

General principle:

It is a bad idea to include posttreatment outcomes as covariates.

Big Picture in Estimating Causal Effects

Fundamentally Unanswerable Questions

Some research questions are poorly posed

> "What is the effect on earnings of being born in Africa instead of North America?"

- What experiment could possibly answer this question?

Questions to Ask

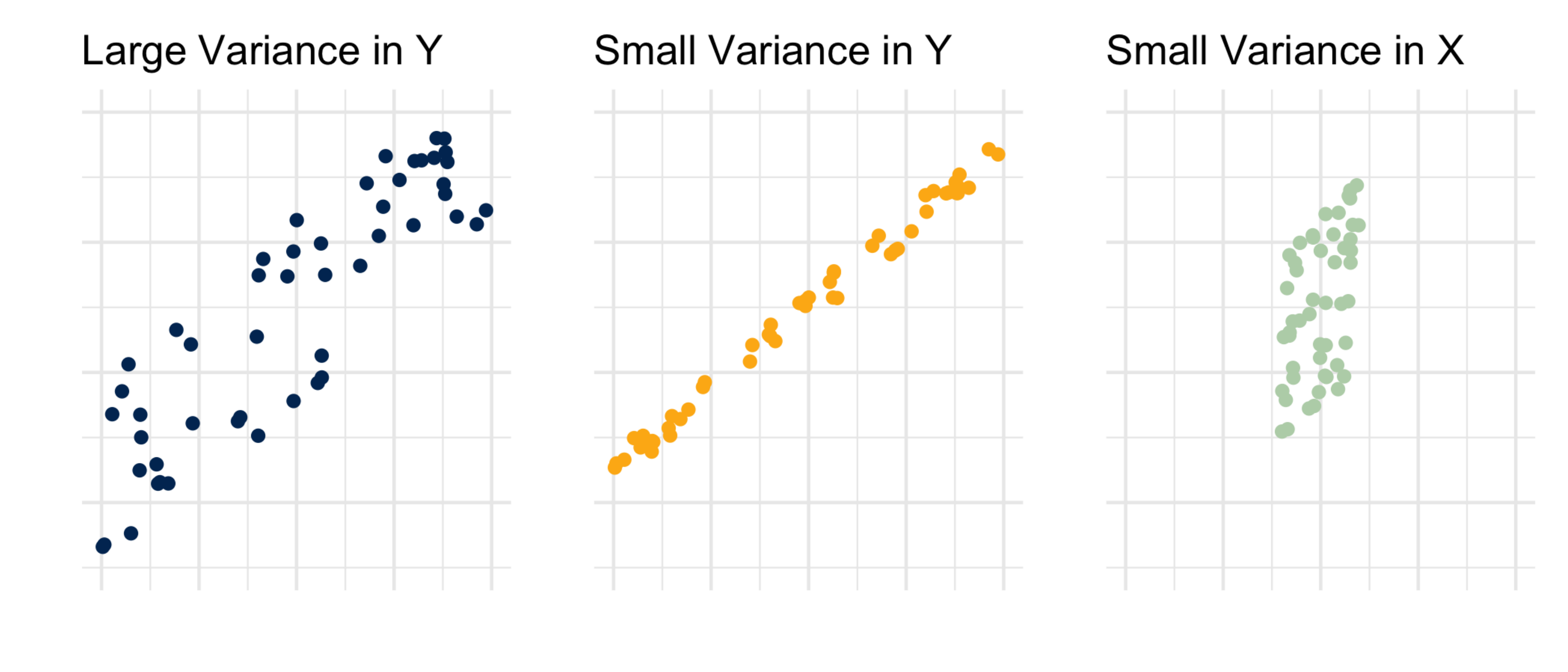
1. What is the causal relationship of interest?
2. What is the *ideal* experiment to measure this?
    - Even if you're doing observational research, *ask this question*!
    - If your question seemed, FUQ'd, how should you refine your question?
3. What is your *identification strategy*?
    - Where does variation come from?
    - Why is this variation independent of potential outcomes?
4. How are you computing your confidence intervals?

Reading

Reading: Read *Mastering Metrics*, Chapter 2

Appendix, pages 95-97

Robust Standard Errors and Confidence Intervals in Regression


Large Variance in Y
Small Variance in Y
Small Variance in X

Standard Errors and Confidence

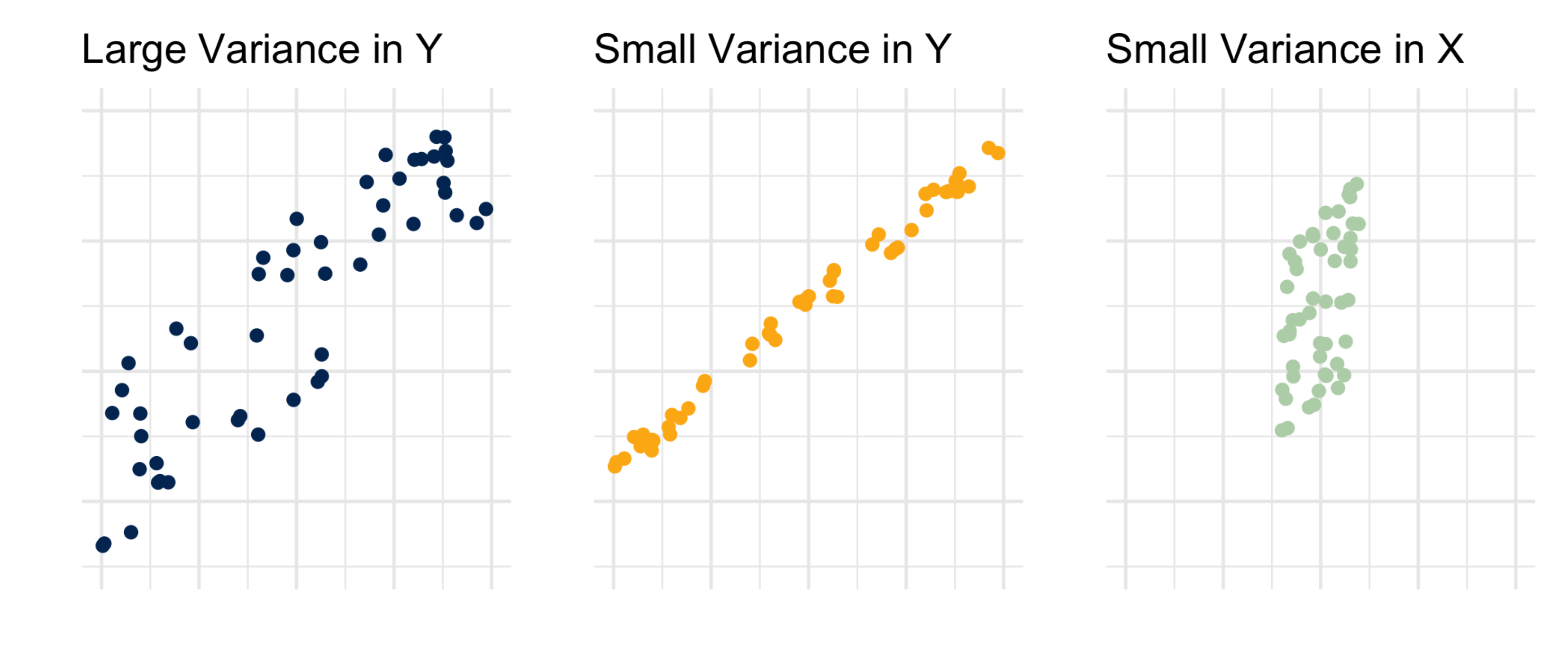
How reliably have we estimated our slope coefficient?

- Where does any *noise* come from?

Rules of Thumb about Standard Errors

- SEs are larger when variance in Y is larger
- SEs are smaller when variance in X is larger

```
large_variance_in_y | small_variance_in_y | small_variance_in_x
```



Standard Errors in Regression Output

- Treatment effect (with 95 percent confidence interval) $\approx$ slope coefficient ± 2 standard errors
- OLS standard errors assume each observation's idiosyncratic component, ϵ is iid (independent and identically distributed)
- Independence is sensible in a randomized experiment

Why should we expect all points to have the same variance?

Heteroskedasticity

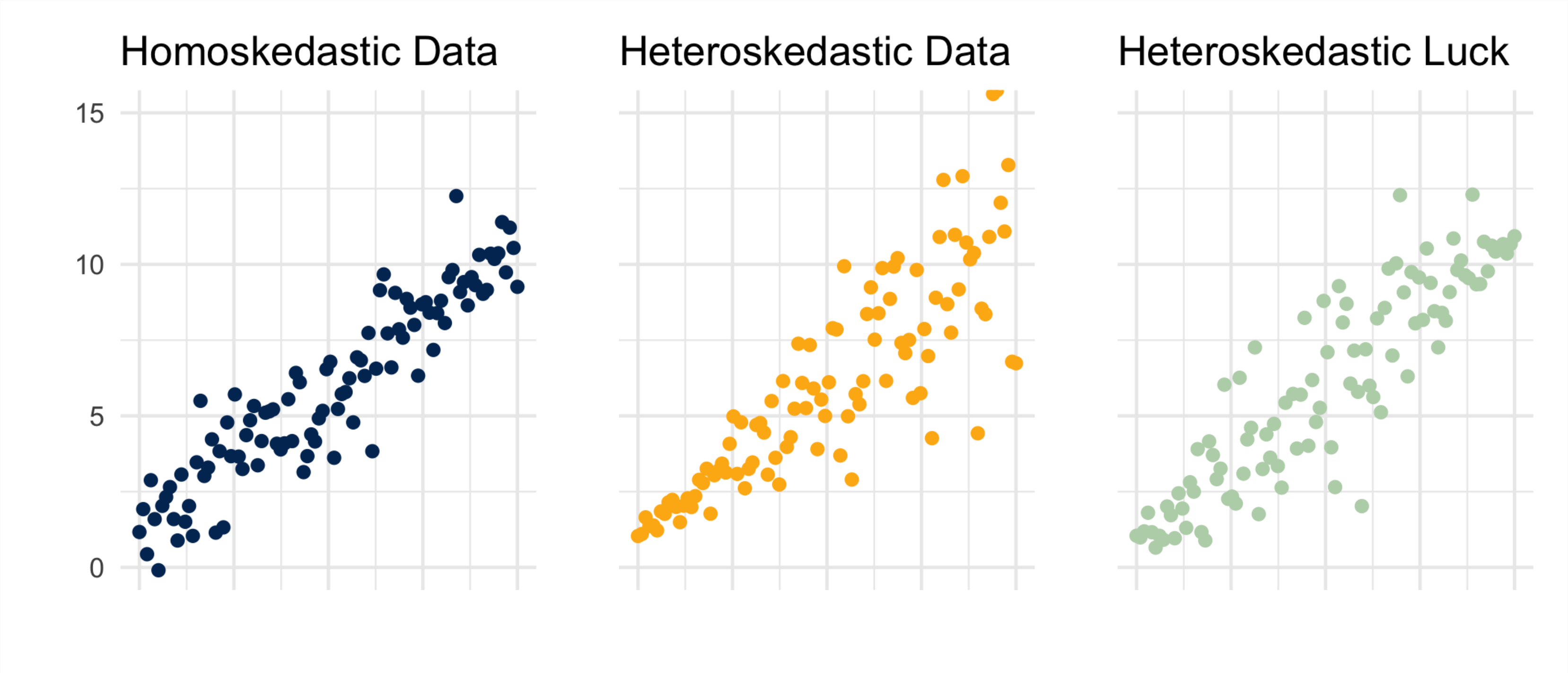
Heteroskedastic & Homoskedastic Errors

- **Hetero-skedastic**: Different observations have different error variances
- **Homo-skedastic**: Different observations have the same error variances
  - *We don't actually write it with the hyphen, but it makes it more clear to read*

OLS Defaults

- Homoskedasticity is the default assumption under which OLS standard errors are usually computed
- Vertical error variance causes uncertainty about line's true slope

Distributions of Data



- *Fanned Out* data means many lines could be fit, depending on the sample
- More accurate plot with accuracy *distant* from grand mean: endpoints anchor slope.
- Leverage: Data points nearer ends of regression line influence slope more.

Robust Standard Errors

- Estimate accurate confidence intervals, even when error variance varies with X
- Also known as *heteroskedasticity-robust* standard errors or *Eicher-White*, or *Huber-White* standard errors
- Do not require knowledge of:
    1. shape of heteroskedasticity; or,
    2. Which X-variables correlate with variance

Optional Reading

- Page 45 of *Mostly Harmless Econometrics* (the PhD version of *Mastering Metrics* describes technical details and matrix algebra).
- Shows that extreme values of X have more leverage on slope coefficient, β.

Accounting for Leverage

- When estimating variance of $\hat{\beta}$:
  - Take weighted sum of squared residuals (i.e., squared vertical deviations from regression line)
  - Divide by total variance in $X4
  - Weights in weighted sum correspond to leverage of each observation
  - Squared residuals, $\hat{\epsilon}^2$, weighted by squared horizontal deviations from the mean

Tennessee STAR Experiment

- Randomized at classroom level
- Each classroom's students had same teachers, similar backgrounds/experiences
- More similarity within each classroom than between classrooms
- Changing 20 similar students from control group to treatment group moves potential Y for all 20 at once
    - Result: more variance than if 20 randomly chosen people had been moved
- Clustered Standard Errorss account for lack of independence, asd RSEs account for heteroskedasticity
    - In Tennessee STAR experiment, CSEs are about three times larger than OLS standard errors
    - Further technical details, see MHE 8.2

Takeaways

Takewaways about Regression Uncertainty

- Best practice to always use *robust standard errors*
  - Do not assume homoskedastic errors
  - Pay only a small penalty if we're incorrect (i.e. the data is *actually* homoskedastic)
- If treatment assignmet is clustered, **must** use clustered standard errors to avoid unintentionally overstating precision of estimates

Multifactor Experiments

Multifactor Experiments

Reading: *Field Experiments* section 9.3.3

- Presents multi-factor experiments -- experiments that have more than a single treatment

Estimating effects in a multi-factor experiment

- Estimate regressions with interaction terms to estimate how much more *one* treatment matters when the *other* treatment is turned on

Example: The Visible Hand

Doleac and Stein (2013)

Example: The Visible Hand

- Conducted an experiment on Craigslist (this is now **very** hard to do successfully) to assess race- and class-based discrimination among consumers
  - Ran ads to sell iPods in different local markets
  - Measured average offer
  - Measured average number of offers

Treatment Dimensions

1. Race and Class
   - White hand
   - White hand with tattoo
   - Black hand
2. Ad quality
   - Grammar or spelling errors
   - No grammar or spelling errors
3. Asking price: $90, $110, or $130

Table C1

Correspondence Text

	High-quality advertisement	Low-quality advertisement
e-mail 1 (offer): ‘A’ text	Thank you for your interest in my iPod Nano. I’ve received a lot of responses, and would like to ell this quickly to the person who makes me the best offer. CASH ONLY, no trades. Is $[offer] your best offer? Thanks. [link to ad] [text of ad]	thank you for your interest in my ipod nano. i got a lot of responses, and would like to sell this quickly to the person who makes me the best offer. CASH ONLY, no trades. is $[offer] your best offer? thanks. [link to ad] [text of ad]
e-mail 1 (no offer): ‘A’ text	Thank you for your interest in my iPod Nano. I’ve received a lot of	thank you for your interest in my ipod nano.

Doleac and Stein (2013)

Results

- Both race and tattoo had much more treatment effect than ad quality.

Reading: Field Experiments Section 9.3.3

- Please read this section with the following question in mind

> "Do Black sellers hurt themselves more with bad grammar than White sellers do?"

- You can ignore the last two paragraphs containing subtle points.

Multifactor Designs

Four treatments summarized with a 2x2 table

- In this experiment, there were four different treatments
  1. *Colin* with Good Grammar
  2. *Colin* with Bad Grammar
  3. *Jose* with Good Grammar
  4. *Jose* with Bad Grammar

	Colin	**Jose**
Good Grammar	52%	37%
Bad Grammar	29%	34%

Craigslist Experiment Design

3 x 2 x 3

- 3 photo conditions
- 2 grammar conditions
- 3 price conditions

6 ad texts for 2 grammar conditions

- But only 2 grammar conditions were the treatments of interest; we aggregate other variants by grammar condition

Reading

Reading: *Field Experiments* Section 9.4 through page 304

- The first two pages present using regression to estimate the results of a multi-factor experiment
- In Equation 9.10, the expression between the two equals signs is compact notation for the following definition:
  - $Y = Y_i(0), \text{ if } d_i = 0$
  - $Y = Y_i(1), \text{ if } d_i = 1$

Regression Analysis of Multifactor Experiments

Regression Specification

Equation 9.11

Define the following symbols:

- NH_GG: Non-Hispanic, Good Grammar
- H_GG: Hispanic, Good Grammar
- NH_BG: Non-Hispanic, Bad Grammar
- H_BG: Hispanic, Bad Grammar

$$Y_i = \beta_1 NH\_GG + \beta_2 H\_GG + \beta_3 NH\_BG + \beta_4 H\_BG + u_i$$

Regression Specification (cont'd)

Equation 9.12

Define the following symbols

- J: Takes value 1 if letter sent from "Jose"; 0 if letter sent from "Colin"
- G: Takes value 1 if letter has "Good Grammar"; 0 if "Bad Grammar"

$$Y_i = \alpha + \beta(J_i) + \gamma(G_i) + \delta(J_i \times G_i) + u_i$$

- Equation 9.12 and 9.11 are equivalent
- *Related* estimated parameters -- see the equations that begin on page 306

Estimates from Equation 9.11

$$Y_i = \beta_1 NH\_GG + \beta_2 H\_GG + \beta_3 NH\_BG + \beta_4 H\_BG + u_i$$

- Equation 9.11 directly estimates averages within each treatment condition
- Coefficients $\beta_1, \ldots, \beta_4$ estimate numbers in table

	Colin	**Jose**
Good Grammar	52%	37%
Bad Grammar	29%	34%

Data Structure 9.11

Estimating Equation

$$Y_i = \beta_1 NH\_GG + \beta_2 H\_GG + \beta_3 NH\_BG + \beta_4 H\_BG + u_i$$

Data Structure

ID	Y	NH_GG	H_GG	NH_BG	H_BG
1	Yes	1	0	0	0
2	Yes	0	0	1	0
3	No	0	1	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
400	No	0	1	0	0

Estimating Coefficients for 9.12

$$Y_i = \alpha + \beta(J_i) + \gamma(G_i) + \delta(J_i \times G_i) + u_i$$

This specification measures *differences* between cells.

- α estimates the average response in the omitted category (Colin, Good Grammar).
- β estimates the effect of ethnicity *when there are no grammar errors*, `G=0`.
- γ estimates the effect of grammar, *when the ethnicity signal is Colin*, `J=0`.
- δ, the interaction coefficient, estimates how much more the grammar errors matter for Jose than for Colin.

Interaction Coefficient in Equation 9.12

$$Y_i = \alpha + \beta(J_i) + \gamma(G_i) + \delta(J_i \times G_i) + u_i$$

Suppose that `G=1`, the sender uses Good Grammar

- What happens when `J=0` $\longrightarrow$ `J=1`?
- Regression can make this analysis simpler
  - You can obtain results all at one time.
  - Results can be easier to interpret when coefficients are measuring differences instead of levels.

Estimated Coefficients

Estimtes are provided in Equation 9.16

$$\widehat{Y} = 0.52 - 0.15(J_i) - 0.23(G_i) + 0.20(J_iG_i)$$

Data Structure

ID	Y	J	G
1	Yes	0	0
2	Yes	1	0
3	No	1	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$
400	No	1	1

Interpreting Estimates in 9.16

Estimating Equation

$$\widehat{Y} = 0.52 - 0.15(J_i) - 0.23(G_i) + 0.20(J_i G_i)$$

Summary Table

	Colin	**Jose**
Good Grammar	52%	37%
Bad Grammar	29%	34%

Interpretation

- Regression coefficients found by subtracting numbers in the data table
- 0.52 is the fraction of letters sent that received a response *baseline condition*
- 0.15 the difference between `J=0` $\rightarrow$ `J=1`, when `G=0`.
- $0.52 - 0.37 = 0.15$

Interpreting Interaction Term

Estimating Equation

$$\widehat{Y} = 0.52 - 0.15(J_i) - 0.23(G_i) + 0.20(J_iG_i)$$

Summary Table

	Colin	**Jose**
Good Grammar	52%	37%
Bad Grammar	29%	34%

Interpreting Interation Term

$$\begin{aligned} 0.34 - 0.37 &= -0.03 \\ 0.29 - 0.52 &= -0.23 \\ -0.03 - (-0.23) &= \mathbf{0.20} \end{aligned}$$

Interpreting Interaction Term (cont'd)

Estimating Equation

$$\widehat{Y} = 0.52 - 0.15(J_i) - 0.23(G_i) + 0.20(J_iG_i)$$

Summary Table

	Colin	**Jose**
Good Grammar	52%	37%
Bad Grammar	29%	34%

Interpretation

- How much more does bad grammar, `G=1` matter with `J=1` vs. `J=0`?

Presenting Regression Results

Regression output automatically includes standard errors, for easy hypothesis testing.

Do grammar errors have less impact when letters are received from Jose vs. Colin?

- Perform a Wald-test on the regression coefficient: $\frac{\delta}{SE(\delta)}$
- The authors performed an F-test, which is unnecessary in this case with a single interaction term

Regression for more complex variable expression

- Regression gracefully handles non-binary categorical variables (where an F-test would be required)
    - One example is the amount of someone's schooling.

What to Remember

Bad Controls

- Do not include post-treatment covariates in the regression
- This can be especially tempting when you have a robust set of user-data

Standard Errors

- Robust standard errors are *always* a good idea
- They are of little cost if they are unnecessary, and are required with *heteroskedastic* errors
- Use clustered standard errors in regressions that have clustered treatment assignment; failing to do so will produce estimates that are incorrectly precise

Multifactor Experiments

- We can expand the complexity of treatment we assign by *crossing* treatment features
- Regression with dummy variables quickly, efficiently, and unbiasedly estimates the effects of these experiments.