Research Methodology for Public Policy

Lecture 9 — Regression Again - Multivariate, Assumptions, and Robustness

Dr. Colin Kuehl

2026-07-20

Northern Illinois University

SCHOOL OF PUBLIC POLICY
CHIANG MAI UNIVERSITY

Welcome Back

Warm-Up

- In your notes:

1. Write down your hypothesis for your final project?
2. How are you operationalizing your key variables?
3. What is your unit of analysis?
4. What are some potential confounding variables?

- On the sticky note:

1. Think through what you would like to learn from this class. What are you curious about? What will help your career? What have we missed? - Write down what you think we should cover in the remaining week that we haven't yet.

Today's Agenda

- Course Logistics
- Review: Bivariate regression: the model and the line of best fit
- Review: Interpreting coefficients and testing their significance
- Rejecting the Null
- Multivariate regression: why and how it differs from bivariate
- R^2: how well does the line fit?
- How to read a full real regression table
- Regression Assumptions
- Transformations
- Midterm Discussion

Course Logistics - Final Project

- Abstract and Initial Data due Tonight
- Feedback by Wednesday
- Presentations on Friday
  - 10 AM Start and then class lunch together?
  - Mock presentation tomorrow/Template posted
- Report due Friday - but not late until Saturday at Midnight

Course Logistics

- Midterms are graded, but will be returned tomorrow
  - Remember only 20% of course grade and lots of extra credit
  - Readings are an important part of this class/graduate school.
- Problem set #4 due Wednesday - Multivariate Regression

Course Logistics

- Schedule: Discussion
  - Today: Regression round #2
  - Tuesday: Experiments
  - Wednesday: Qualitative
  - Thursday: Qualitative #2/Ethics/Student Choice
- R Practicum
  - New skills vs work time?

Reviewing Regression's Building Blocks

Three Tools for Continuous Relationships

1. **Covariance**: do X and Y move together, and in which direction?
2. **Correlation**: a standardized version of covariance, always between -1 and 1
3. **Regression**: the workhorse — quantifies the relationship *and* lets us predict Y from X

- The quantitative focus for the rest of the course

Covariance and Correlation

Covariance

- **Covariance** tells us whether two variables are positively associated, negatively associated, or unrelated
- $\widehat{Cov}(X, Y) = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})$
- Problem: its scale depends on the units of X and Y, so it's hard to interpret on its own — a covariance of "350" tells you almost nothing by itself

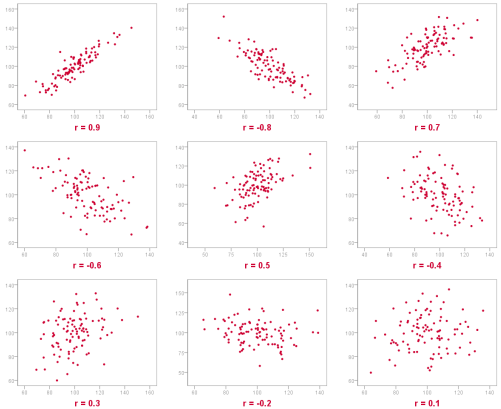
Correlation

- "Correlation is closely related to covariance. In essence, correlation standardizes covariance so it can be compared across variables."

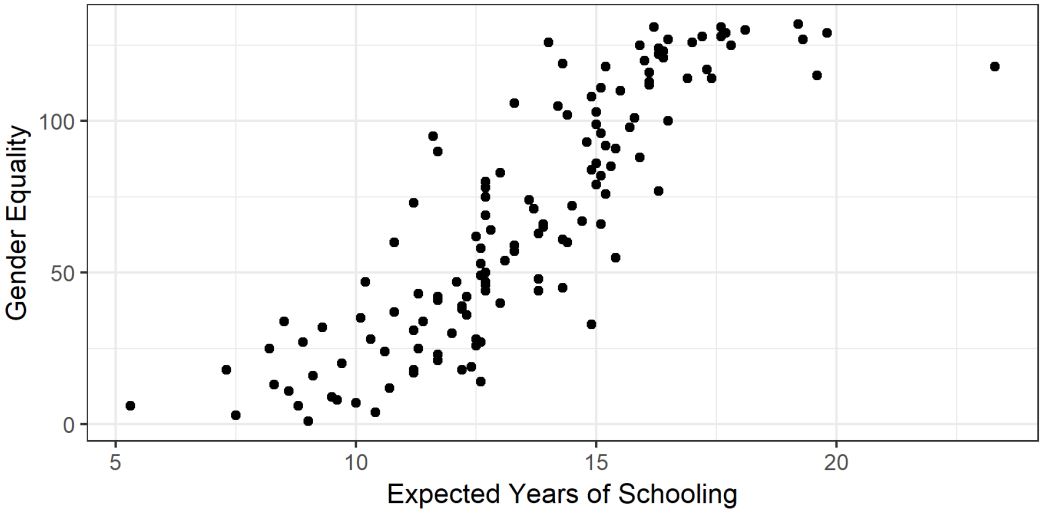
$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

- 1: perfect positive relationship; -1: perfect negative relationship; 0: no linear relationship

What Correlation Looks Like



Example


Gender Equality
100
50
0
5
10
15
20
Expected Years of Schooling

The Limits of Correlation

What Correlation Does and Doesn't Tell Us

- Tells us: direction (positive/negative) and a general sense of strength
- Does **not** tell us: the size of the effect in real units, whether the relationship is statistically significant, or anything about other variables
- Correlation only captures *linear* association — a strong non-linear relationship can show a correlation near zero

Bivariate Regression: The Idea

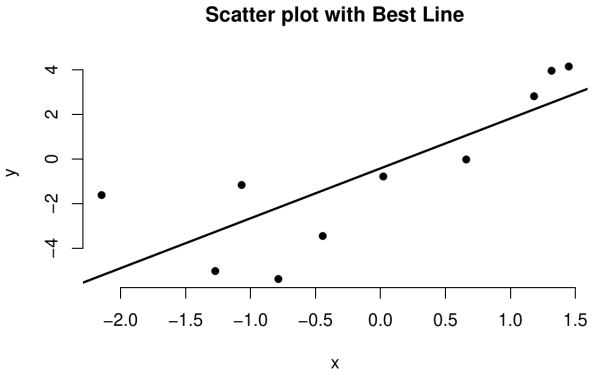
From Correlation to Regression

- Regression asks a sharper question than correlation: for a one-unit change in X, how much do we expect Y to change?
- "The intuition behind linear regression is simple: we want to find the line that best fits all of our data points."
- It gives us an equation we can use to *predict* Y, not just describe how strongly X and Y move together

Bivariate Regression

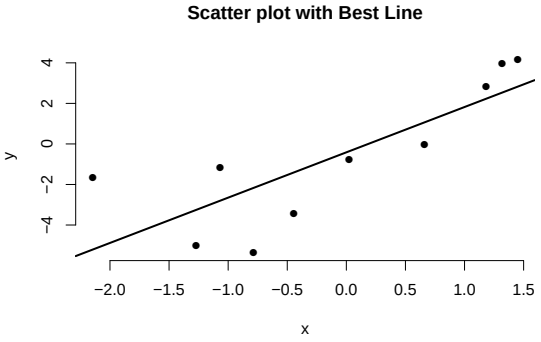
- One independent variable (X) predicts one dependent variable (Y): output is $y = f(x)$ — exactly the bivariate model from Lecture 2, $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
- The slope connects back to covariance: $\hat{\beta}_1 = \dfrac{\widehat{Cov}(X, Y)}{\widehat{Var}(X)}$ — regression and correlation are close cousins, not separate ideas

What Makes a Line "Best"?

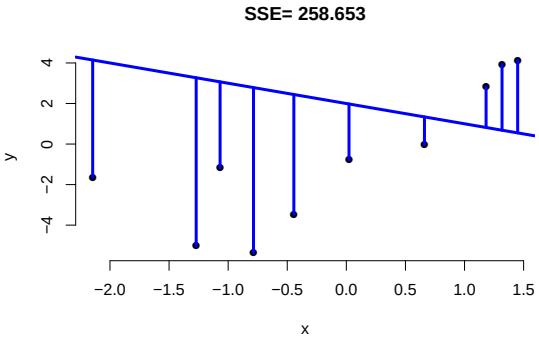


- We want $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ to predict Y_i as well as possible
- Among the infinitely many lines we could draw, only one minimizes total prediction error

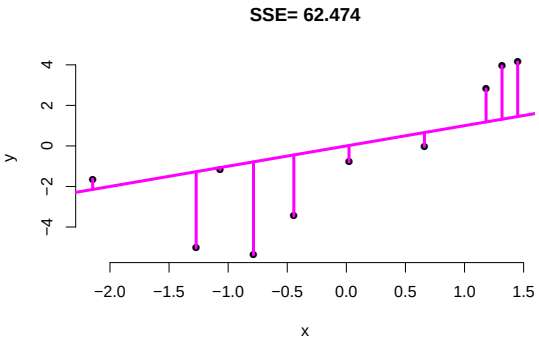
Finding the Best-Fitting Line



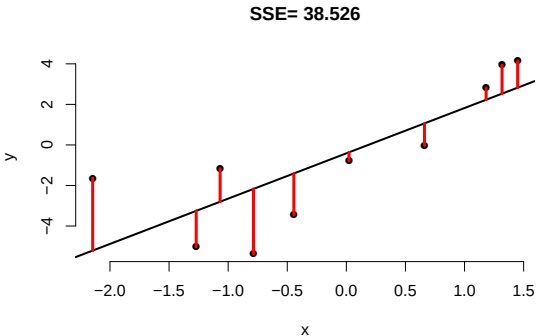
Finding the Best-Fitting Line



Finding the Best-Fitting Line



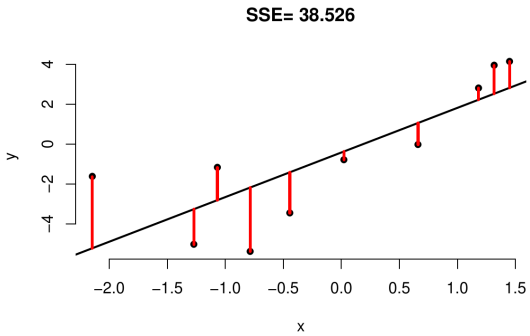
Finding the Best-Fitting Line



Residuals: How Far Off Are We?

- The **residual** is the gap between the actual value and the prediction: $\hat{\epsilon}_i = Y_i - \hat{Y}_i$
- Every point in a scatterplot has its own residual — some positive (model under-predicts), some negative (model over-predicts)

Why Squared Errors? Ordinary Least Squares

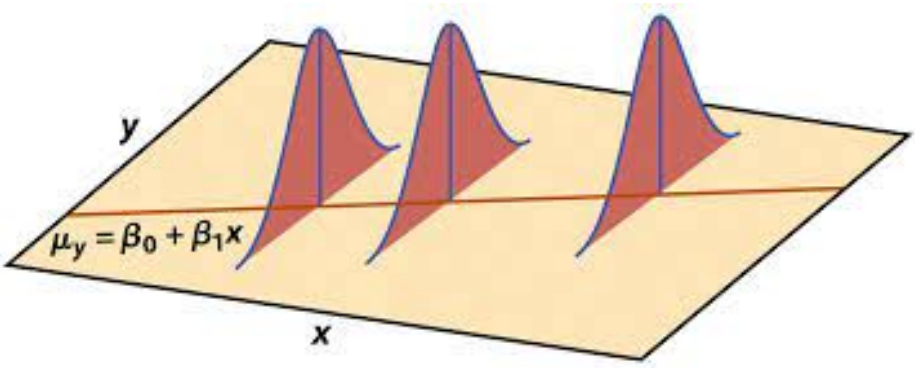


- Squaring makes every error positive (overshoots and undershoots both count) and penalizes large errors disproportionately more than small ones
- **Ordinary Least Squares (OLS)** chooses $\hat{\beta}_0, \hat{\beta}_1$ to minimize $SSE = \sum_i \left(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)\right)^2$ — among all possible lines, there is a unique

R code

```r
# a preview -- this is very simple in R
model <- lm(growth ~ years_school, data = country_data)
summary(model)
```

The Error Term Isn't Just "Noise"



- At every value of X, OLS assumes Y is scattered around the regression line in a predictable way (here, normally distributed)
- ϵ_i bundles together measurement error, an imperfectly specified model, and genuine stochastic variation — not just "everything we forgot"

A Caution

- This explains the data's relationship — not automatically a *causal* claim (only when exogeneous)
- Recall Lecture 2: is years of schooling unrelated to everything else in ϵ_i? Almost certainly not (institutions, wealth, geography...)
- We'll confront this directly with experiments (Lecture 9) and again with multivariate regression later today

A Regression Table

Anatomy of a Regression Table

```
> lm.out=lm(growth~yearsschool, data=dat)
> summary(lm.out)
Call:
lm(formula = growth ~ yearsschool, data = dat)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.95829    0.41846   2.290  0.02538 *
yearsschool  0.24703    0.08873   2.784  0.00708 **
```

What Each Column and Row Tells You

- $\hat{\beta}_0$ (intercept): predicted Y when every $X = 0$
- Each row below it: one predictor's coefficient, standard error (usually in parentheses), and significance stars
- More stars (or a smaller p-value) means stronger evidence the true coefficient isn't zero — not a bigger effect

Four Things to Interpret in a regression

1. **Significance** — is the p-value below your threshold (often 0.05)?
2. **Sign** — is the coefficient positive or negative?
3. **Size** — how big is the coefficient, in its original units?
4. **Substance** — what does this mean in terms a policymaker, not a statistician, would understand?

Significance

Is the Slope Really Different From Zero?

- Null hypothesis: $H_0 : \beta_1 = 0$ (no relationship in the population); alternative: $H_a : \beta_1 \neq 0$
- Same as when test any other parameter: how many standard errors is $\widehat{\beta}_1$ away from zero?
- The regression table reports a **standard error** for each coefficient; $t = \widehat{\beta}_1 / SE(\widehat{\beta}_1)$ — a large $|t|$ means the coefficient is many standard errors from zero, and the p-value converts that into "how surprising would this estimate be if the true effect were zero?"

Confidence Intervals on a Coefficient

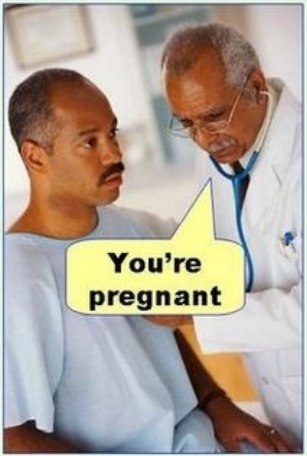
- Interpretation: we are 95% confident the true population slope falls in that range
- Because the *interval excludes zero, the coefficient is statistically significant at conventional levels* — the two tools (CI and p-value) always agree

Significance and error types

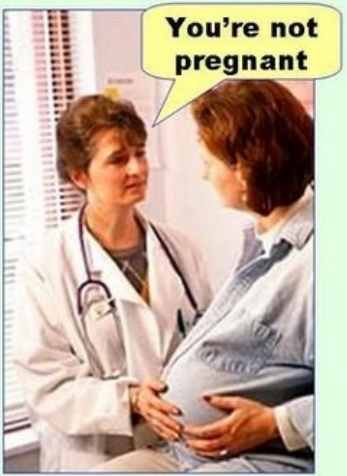
		REALITY	
		NULL HYPOTHESIS	
		TRUE	FALSE
STUDY FINDINGS	TRUE		Type II error (β) 'False negative'
	FALSE	Type I error (α) 'False positive'	

Significance and error types

Type I error
(false positive)



Type II error
(false negative)



Coefficients - Sign and Size

Reading the Equation

$$\widehat{\text{growth}}_i = \hat{\beta}_0 + \hat{\beta}_1 \, \text{years\_school}_i$$

- $\hat{\beta}_0$ (intercept): predicted Y when $X = 0$; $\hat{\beta}_1$
  - Caution on the intercept- We rarely interpret, unless X=0 is meaningful
- **(slope):** expected change in predicted Y per one-unit change in X

Four Things to Report

1. **Significance** — is the p-value below your threshold (often 0.05)?
2. **Sign** — is the coefficient positive or negative?
3. **Size** — how big is the coefficient, in its original units?
4. **Substance** — what does this mean in terms a policymaker, not a statistician, would understand?

Example

- $\widehat{\text{growth}}_i = 0.96 + 0.25\,\text{years\_school}_i$
- "A one-year increase in average schooling is associated with a 0.25 percentage-point increase in expected GDP growth, and this relationship is statistically significant at the 0.01 level"
- Translate every regression you produce across **four** registers: equation, table, graph, and plain words

Statistical vs. Substantive Significance

- **Statistically significant**: we're confident the coefficient isn't exactly zero
- **Substantively significant**: the effect is large enough to matter for policy, even if it is statistically significant
- A huge sample can make a tiny, meaningless effect statistically significant — always ask both questions, not just one

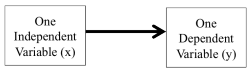
New Today: Multivariate Regression

Why One Predictor Is Rarely Enough

- Bivariate regression ignores every other variable that might affect Y — and those variables don't disappear, they hide in ϵ_i
- If an omitted variable affects *both* X and Y, the bivariate coefficient absorbs its influence too
- "This is a general rule: When we leave out variables that affect our main relationship, we tend to overestimate the regression coefficients of the variables in our regression. This is called omitted variable bias."
  - In other words, X will "soak up" the effect of Z.
  - Ice cream will "soak up" the effect of temperature

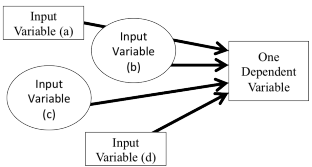
From One Predictor to Many

Single-Variable Regression:



Output: $y = f(x)$

Multiple Regression:



Output: $y = f(a, b, c, d)$

- Multiple regression adds inputs (a), (b), (c), (d)... all pointing at the same outcome
- "Multivariate OLS is used to estimate a model with multiple independent variables."

Why Multivariate?

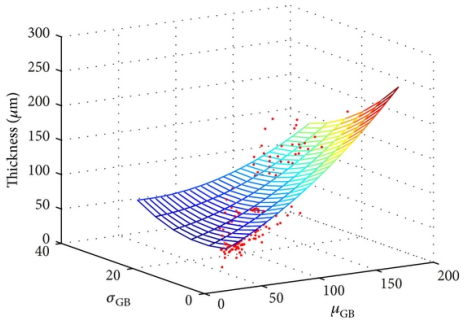
- to be more precise with our expectation for Y_i
- to "control for" or "remove the effect of" one variable (say Z_i) wheninterpreting the coefficient on the other
  - This is why we often differentiate independent variables between our primary variable and our control variables
- Remove confounders from the error term and put in model

The Multiple Regression Model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \epsilon_i$$

- Same logic as bivariate OLS: find the $\widehat{\beta}$'s that minimize the sum of squared residuals
- Geometrically, with two predictors we're now fitting a *plane* through the data, not a line — but the estimation principle doesn't change

Visualizing Three dimensions



The Multivariate Model

Building a Model

- With multivariate regression we move to thinking about the model
  - ie these are the set of things that matter for my outcome
- A model, like a map, is a simplified version of the real world with the key indicators included

What variables?

- Literature
- Theory
- Of course limited to data availability and subject to all caveats ondata quality and measurment validity
- Avoid the kitchen sink regression. Balance parsimony and power concerns with comprehensiveness

What Each Coefficient Now Means

REMEMBER THIS

1. Multivariate OLS is used to estimate a model with multiple independent variables.
2. Multivariate OLS fights endogeneity by pulling variables from the error term into the estimated equation.
3. As with bivariate OLS, the multivariate OLS estimation process selects $\hat{\beta}$'s in a way that minimizes the sum of squared residuals.

- Each $\widehat{\beta}_k$ is now a **partial effect**: the expected change in Y for a one-unit increase in X_k, **holding all other included variables constant**
- This "holding constant" language is new — and it's the whole point of going multivariate

An Example

- Education and inequality, controlling for GDP per capita:

$$\widehat{\text{Ineq}}_i = \hat{\beta}_0 + \hat{\beta}_1 \,\text{EDU}_i + \hat{\beta}_2 \,\text{GDP}_i$$

- EMPS Ch. 8's worked example: adding GDP per capita to the model drops the education coefficient from about 11.6 to 6.5
- That drop is omitted variable bias *being resolved* — part of education's apparent effect was really GDP's effect all along

"Robust" Coefficients

- A coefficient is called **robust** if it stays similar in size and significance across multiple model specifications
- A coefficient that survives several rounds of added controls is much more convincing than one that only shows up in a single bare-bones model
- This is why we report the results of multiple regression models in our table

A Full Regression Table

Panel A: RavenPack CSS

	(1)	(2)	(3)	(4)	(5)
alpha	0.6498***	0.6282***	0.4411***	0.5468***	0.4653***
	(3.88)	(4.09)	(3.37)	(3.16)	(3.32)
Mkt-Rf	-0.2239***	-0.1812***	-0.0944***	-0.1470***	-0.1156***
	(-4.30)	(-4.04)	(-3.10)	(-3.20)	(-2.98)
SMB		-0.0313	-0.0733	0.0029	-0.0916
		(-0.32)	(-0.84)	(0.03)	(-1.01)
HML		-0.6267***	-0.3936***	-0.6519***	-0.3078**
		(-6.18)	(-4.38)	(-3.98)	(-2.48)
RMW				0.1665	-0.0097
				(1.02)	(-0.07)
CMA				0.1599	-0.1825
				(0.61)	(-1.09)
MOM			0.3736***		0.3898***
			(7.96)		(8.86)
R-squared	0.1281	0.2881	0.4669	0.2865	0.4660
N	204	204	204	204	204

Model Fit: R^2

How Well Does the Line Fit?

- R^2 (the **coefficient of determination**) tells us what share of the variation in Y is "explained" by all the Xs
- Ranges from 0 (the line explains nothing) to 1 (the line fits perfectly)
- $R^2 = 0.73$: "73% of the variation" in the outcome is accounted for by the model

R^2

- R^2 can also be interpreted as the proportion of variance explained.
- A high R^2 is neither necessary nor sufficient for an analysis to be useful.
- A good model can have a low R^2 and a bad model can have a high R^2
- R^2 is our primary means of making comparisons between models

Adjusted R^2

- Adding additional variables necessarily increases the R^2
- Adjusted R^2takes the number of variables into account(using math beyond our scope)
- Interpreted the same way as normal R^2. Usually what is reported when presenting multivariate models.

R^2 Is Not the Whole Story

- A high R^2 does not mean the relationship is causal — it only means the line tracks the data closely
- A low R^2 does not mean the relationship is unimportant — a small, precisely estimated, policy-relevant effect can coexist with a lot of unexplained variation
- Always report R^2 *alongside* the coefficient, never as a substitute for it

Why Regression Is So Useful (and Where It Breaks)

Where Regression Shines, and Where It Breaks

Strengths

- **Generalizable** beyond the cases observed, if the sample represents the population well
- **Precise**: a number plus an uncertainty estimate, not just "yes" or "no"
- Easy to **control for other variables** in one model
- Supports **iteration**: add a variable, refit, learn something

Limitations

- **Measurement**: garbage in, garbage out
- **Average effects** can mask subgroup variation
- **Unrealistic causal claims** from correlational designs
- **Kitchen-sink regressions** risk data mining

Interpreting Dummy and Categorical Predictors

- Not every X is continuous — a **dummy variable** codes a category as 0/1 (e.g., democracy vs. not). Its coefficient is the average difference in Y between the two groups, holding other variables constant. More in Practicum

Dummy

- A dummy variable is equivalant to a difference in means between the two groups.

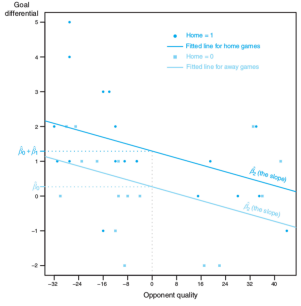


FIGURE 6.7: Fitted Values for Model with Dummy Variable and Control Variable: Manchester City Example

Questions

Key Terms & Looking Ahead

Key Terms from Today

- R^2 / coefficient of determination
- Multiple regression
- Omitted variable bias
- Partial effect / "holding constant"
- Robust coefficient
- Statistical vs. substantive significance

Next Time

Finishing quantitative and experiments

Questions?

Good luck on the midterm — see you in the Practicum this afternoon.

See You in the Practicum