Research Methodology for Public Policy

Lecture 7 — Thinking about Relationships and Regression

Dr. Colin Kuehl

2026-07-15

Northern Illinois University

SCHOOL OF PUBLIC POLICY
CHIANG MAI UNIVERSITY

Welcome Back

Today's Agenda

- Recap & questions from Lecture 6
- Warm-up: two-sample t-tests, p-values, and what "significant" means
- From "is there a difference?" to "how strong is the relationship?"
- Covariance and correlation — and the limits of correlation
- Bivariate regression: the model and the line of best fit
- Interpreting coefficients and testing their significance
- R^2: how well does the line fit?
- Multivariate regression: why and how it differs from bivariate
- How to read a real regression table

Where We've Been

- Lecture 6's confidence intervals let us ask: is a single estimate (a mean, a proportion) precise?
- A related question: is there a difference between two groups? (a t-test answers this — we'll build the intuition first thing today, and you'll see the mechanics in Practicum)
- Today's big question is different: for two *continuous* variables, how strong — and what shape — is their relationship?
- Lecture 2 introduced the core model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ as a way of formalizing causal claims — today we treat it as a statistical tool we can actually estimate
- *Note: Only material discussed in class will be covered on the midterm*

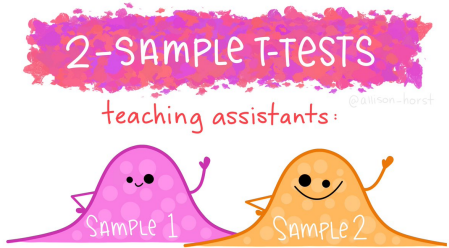
Three Tools for Continuous Relationships

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
- Today moves through all three, ending with how to read a real regression table

More Background: Significance and p-Values

Two Samples, Two Teaching Assistants



- Before we get to relationships between *continuous* variables, one more question about **differences between groups**
- Lecture 6 gave us the machinery: a sample mean, a standard error, and an interval of plausible values
- A **two-sample** t-test points that same machinery at two groups at once: do treatment and control villages have different average incomes? Do men and women report different levels of trust?

Start Here: What If the Means Are Really the Same?

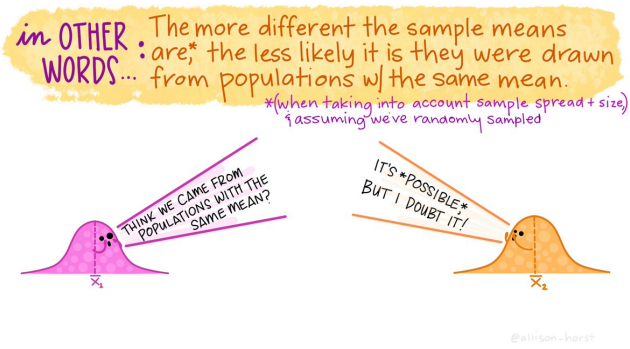
LET'S START HERE: if random samples are drawn from populations w/ the same mean... (i.e. the same population)

Then it is more likely that the 2 sample means will be close together...

...and it is less likely (but always possible!) that the sample means will be far apart.

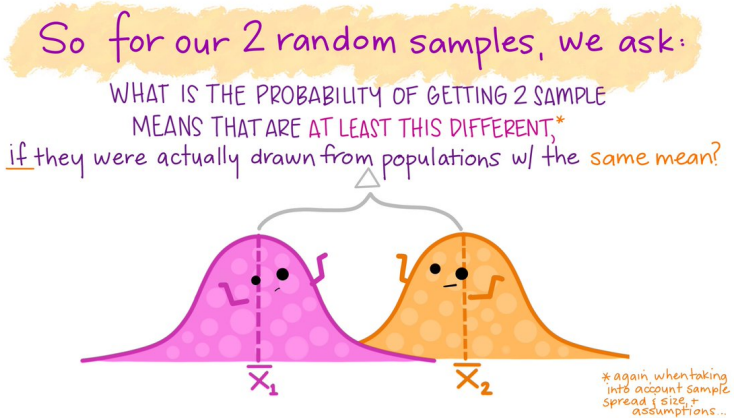
@allison_horst

In Other Words



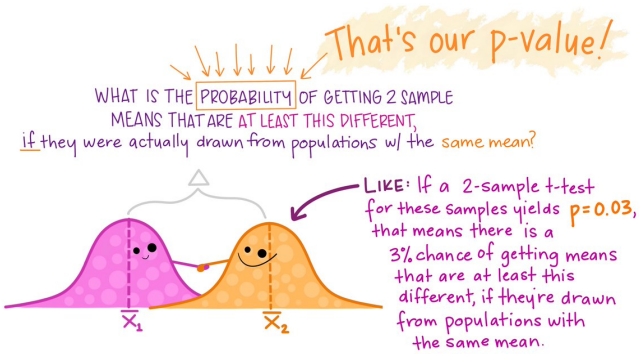
- Notice what does the work here: *sample variance* and *sample size* — exactly the ingredients of the standard error from yesterday

So, For Two Random Samples, We Ask:



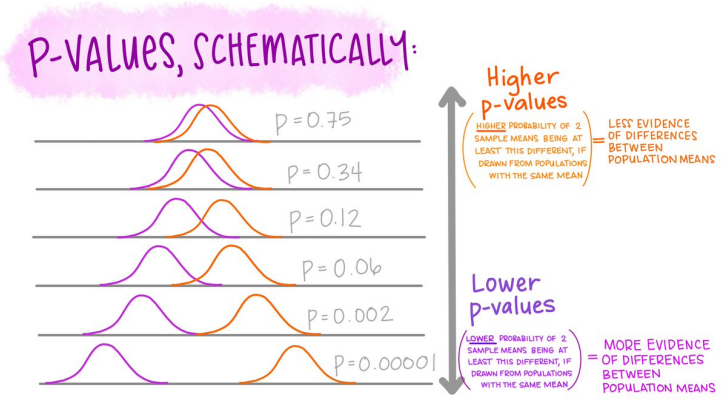
- If looking at comparing groups, we can run a T-test

So, For Two Random Samples, We Ask:



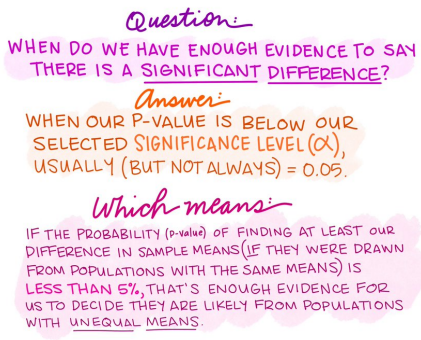
- The p-value is a probability **about the data, given an assumption** — not the probability that the assumption is true
- $p = 0.03$ "a 3% chance the means are equal"; or better said "if they *were* equal, we'd see a gap this big only 3% of the time"

p-Values, Schematically



- Same logic as a confidence interval, viewed from the other side: a small p-value corresponds to a 95% CI for the difference that **excludes zero**
- In this case we're putting a number to the probability of observing the difference

When Is a Difference "Significant"?



- $\alpha = 0.05$ is a *convention*, not a law of nature — $p = 0.049$ and $p = 0.051$ are essentially the same evidence
- Hold on to this logic: every p-value in today's regression tables answers the *same* question, just with $H_0 : \beta_1 = 0$ in place of "the two population means are equal"

Regression: Asking similar questions with continuous variables

Regression

- Regression allows us to ask similar question with continuous variables. ie is the change in Y associated with changes in X significant.
- We're still thinking probabilistically and want to define a point estimate and confidence
- But . . . we'll start with thinking about correlation

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

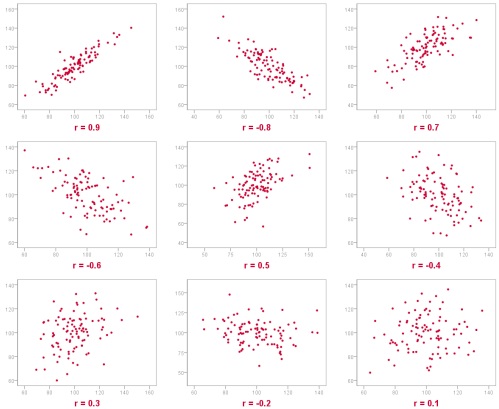
Correlation

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

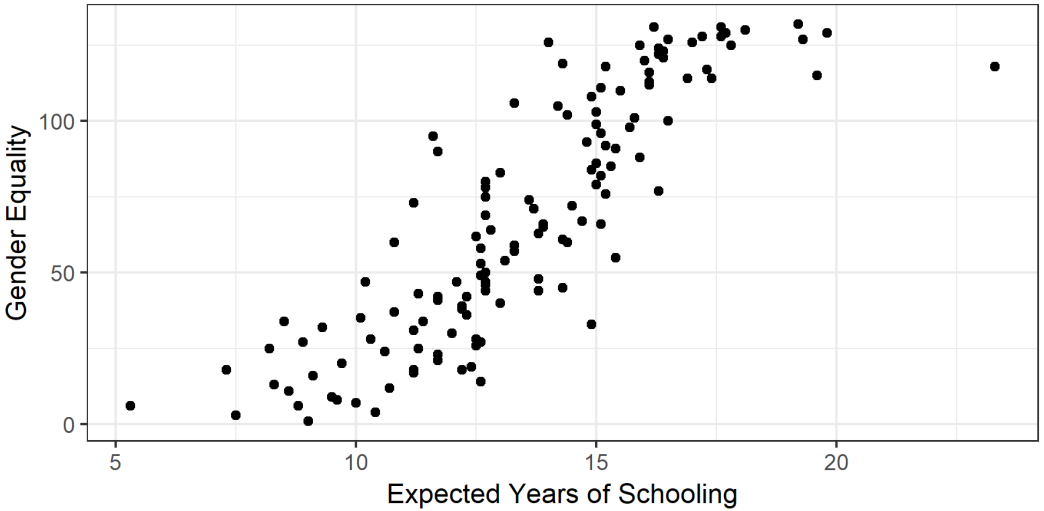
$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

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

What Correlation Looks Like



Example

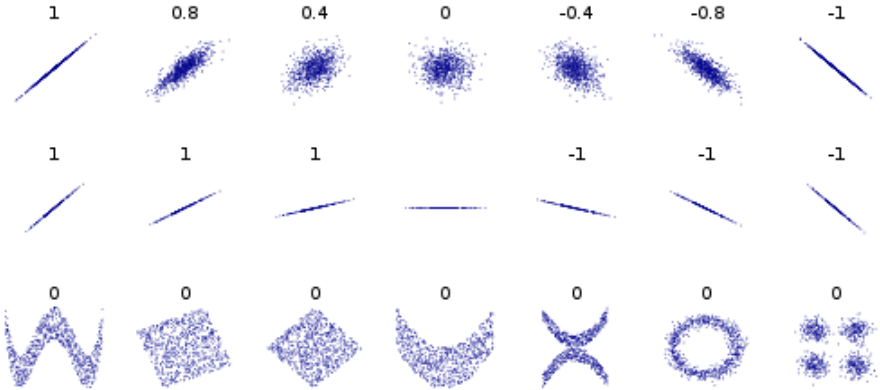

Gender Equality
100
50
0
5
10
15
20
Expected Years of Schooling

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

Visualizing Correlation



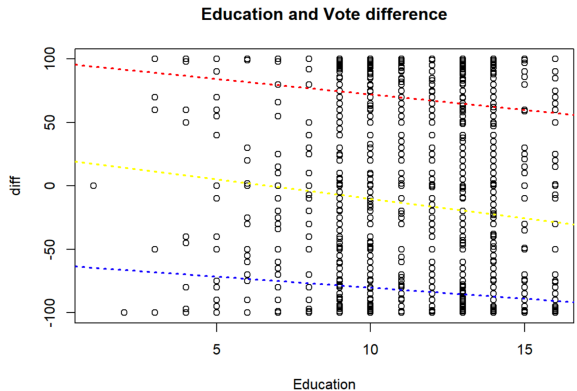
- All of these "0" panels share (about) the same correlation coefficient — and look nothing alike. Lesson: always plot your data
- A data-mining caution: with enough variables and pairs tested, some will correlate

Bivariate Regression: The Idea

From Correlation to Regression

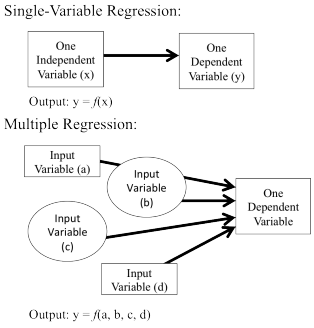
- Regression asks a sharper question than correlation: for a one-unit change in X, how much do we expect Y to change?
- "The intuition behind linear regression is simple: we want to find the line that best fits all of our data points."
- It gives us an equation we can use to *predict* Y, not just describe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

Our Focus Today



- Education level (X) and a measure of vote difference (Y) — three fitted lines for three groups
- Each colored line is a separate bivariate regression: same idea, different subsets of

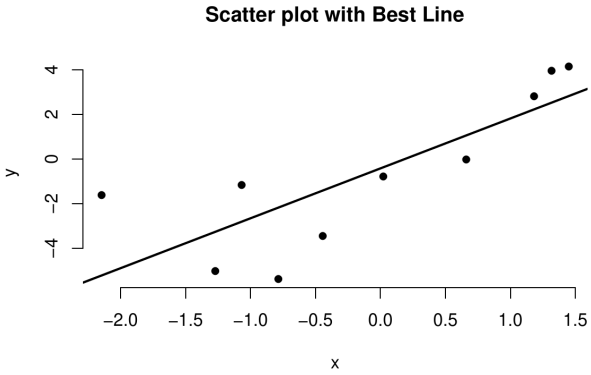
Single-Variable Regression



- 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
- Later we extend this to *multiple* independent variables at once

Finding the Best-Fitting Line

What Makes a Line "Best"?

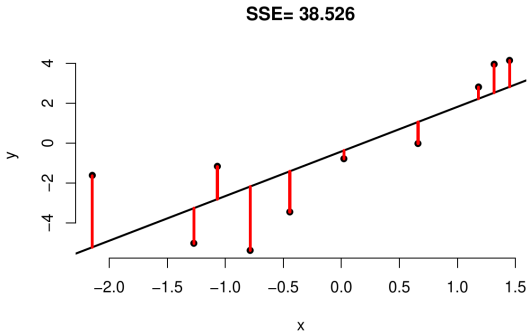


- We want $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$ to predict Y_i as well as possible
- Among the infinitely many lines we could draw, only one minimizes total prediction error

Residuals: How Far Off Are We?

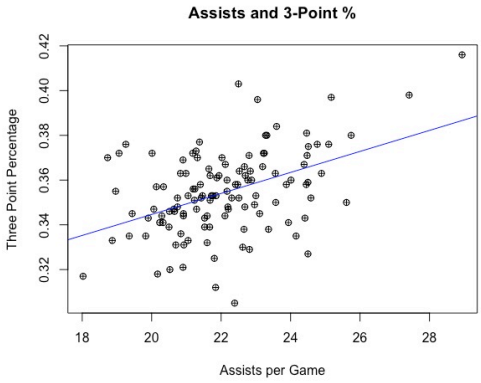
- The **residual** is the gap between the actual value and the prediction: $\hat{\epsilon}_i = Y_i - \hat{Y}_i$
- Serbia has 14.6 expected years of schooling and a gender-equality score of 106; the fitted line predicts about 80; the residual is roughly 25 — the model under-predicts Serbia's equality score by that much
- Every point in a scatterplot has its own residual — some positive (model under-predicts), some negative (model over-predicts)

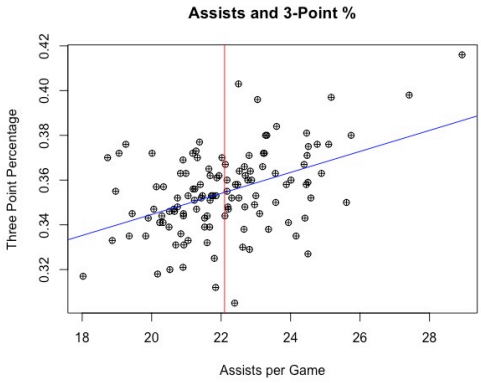
Why Squared Errors? Ordinary Least Squares



- 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

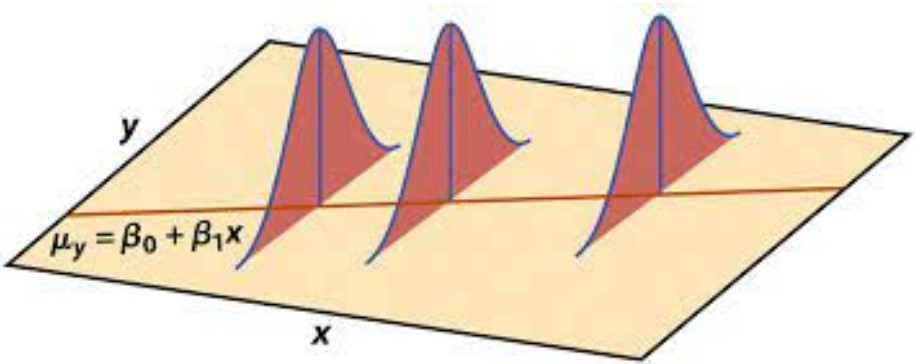
Another Worked Example: Basketball Assists and Shooting





- Left: the OLS line through assists-per-game and three-point shooting percentage
- Right: the same data against the simple mean of Y — regression always fits at least as well as just guessing the mean

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"

Interpreting Coefficients

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):** 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?

A Worked Interpretation

- $\widehat{\text{growth}}_i = 0.96 + 0.25\,\text{years\_school}_i$
- "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

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

Testing 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 $\hat{\beta}_1$ away from zero?
- The regression table reports a **standard error** for each coefficient; $t = \hat{\beta}_1 / SE(\hat{\beta}_1)$ — 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

How to Read 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?

Enough for Today?

Remember Our First Regression Table

Table 3
Performance and ancestral diversity.

	Dependent variable: goal difference							
	(1) OLS	(2) OLS	(3) OLS	(4) OLS	(5) IV	(6) IV	(7) IV	(8) IV
Variable of interest								
Bilateral diversity	0.075*	0.018	0.016	0.020	0.777**	1.543**	1.345**	1.786**
	(0.042)	(0.040)	(0.040)	(0.053)	(0.299)	(0.535)	(0.422)	(0.822)
Control variables								
Log of GDP/capita, home		-0.012	0.019	-0.039		0.107	0.079	-0.058
		(0.207)	(0.208)	(0.303)		(0.259)	(0.246)	(0.384)
Log of GDP/capita, away		-0.467**	-0.463**	-0.099		-0.727**	-0.630**	-0.200
		(0.216)	(0.217)	(0.301)		(0.291)	(0.267)	(0.430)
Log immig. stocks, 18y lag, home		0.067	0.050	0.032		0.156**	0.151**	0.196*
		(0.043)	(0.045)	(0.067)		(0.062)	(0.063)	(0.115)
Log immig. stocks, 18y lag, away		-0.096**	-0.100**	-0.064		-0.156**	-0.180**	-0.163*
		(0.045)	(0.047)	(0.065)		(0.060)	(0.060)	(0.094)
Population (mln), home			-0.004	-0.003			0.002	0.006
			(0.003)	(0.004)			(0.004)	(0.007)
Population (mln), away			-0.001	-0.003			-0.008*	-0.013*
			(0.003)	(0.004)			(0.004)	(0.007)
Observations	3877	3568	3568	2832	3877	3568	3568	2832
Kleibergen-Paap LM test					0.00	0.00	0.00	0.00
Kleibergen-Paap F test					51.57	22.32	33.24	9.76
Team FE	Yes	Yes	Yes	No	Yes	Yes	Yes	No
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Age controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Minute appearances		Yes	Yes	Yes		Yes	Yes	Yes
Geo-political controls			Yes				Yes	
Pair FE				Yes				Yes

Notes: Estimation sample: football national teams from the UEFA affiliation, performing in World Cup and EUROs in the years 1970-2018 (for columns 1 and 5) / years 1978-2018 (for all other columns). Dependent variable: Goal difference. *OLS* results appear on the left (columns 1 to 4). Column 5 to Column 8 display IV results. The first 3 columns of *OLS* and IV include individual team and year fixed effects, as well as a stage dummy Columns 4 and 8 replace individual-team with pair fixed effects. Standard errors are clustered at team pair level. For each IV specification, we present the p-value from the Kleibergen-Paap Lagrange Multiplier test for the instrument relevance, as well as the F-statistics from Kleibergen-Paap F-test for weak instruments. Stars correspond to the following p-values: * p < .10, ** p <

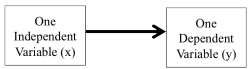
Multivariate Regression

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." *(EMPS Ch. 8)*

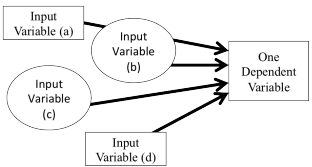
From One Predictor to Many

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." *(QRM Ch. 12, paraphrased)*

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

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

A Concrete 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

Why Fire Trucks Don't Cause Fire Deaths

- Classic illustration (QRM Ch. 12, paraphrased): more fire trucks sent to a fire is associated with more fire-related deaths — sending fewer trucks doesn't save lives
- Both trucks-sent and deaths are driven by a third variable, the size of the fire; adding "size of fire" to the model shrinks the truck coefficient toward zero — the same diagnostic as EMPS Ch. 8's education/GDP example
- Caution: if two predictors are highly correlated with *each other* (**multicollinearity**), OLS struggles to tell their separate effects apart — don't just throw in every variable you can find

"Robust" Coefficients

- A coefficient is called **robust** if it stays similar in size and significance across multiple model specifications
- EMPS Ch. 8's example: adding a migration-index variable produces "a 1-unit increase in the migration index results in a 0.27 increase... holding all other variables constant, but the finding is not statistically significant"
- 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

Dummy Predictors and a Look Ahead to Logistic Regression

- 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
- When Y itself is binary (won the election or didn't), OLS is the wrong tool — **logistic regression** models the *probability* that $Y = 1$, with coefficients often reported as odds ratios
- Same underlying logic — best-fitting model — adapted to a different outcome; not our focus today, but you will see it again

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 X
- Ranges from 0 (the line explains nothing) to 1 (the line fits perfectly)
- EMPS Ch. 8's example reports $R^2 = 0.73$: "73% of the variation" in the outcome is accounted for by the model

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

Key Terms & Looking Ahead

Key Terms from Today

- Two-sample t-test
- Null hypothesis / p-value
- Significance level (α)
- Covariance / correlation
- Bivariate regression
- OLS / residual
- Sum of squared errors
- Coefficient (intercept, slope)
- Standard error / t-statistic / p-value
- R^2 / coefficient of determination
- Multiple regression
- Omitted variable bias
- Partial effect / "holding constant"
- Robust coefficient
- Statistical vs. substantive significance

Before Next Session

Lecture 8: Midterm Exam

- Review Lectures 1–7 — bring at least three questions to class tomorrow
- Tonight's best use of time: redo one regression table from a paper you've read, out loud, in all four registers — equation, table, graph, plain words
- Get some sleep. You know this material better than you think you do

Discussion

Find a published policy paper with a simple regression result. Walk through its coefficient using all four registers — equation, table, graph, and plain words.

Questions?

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

See You in the Practicum