LECTURE 3.4

Qualitative Information

ECON30130 Econometrics I

Sam Deegan
University College Dublin

Autumn 2027

Where We Are

1. Foundations	2. Simple Regression	3. Multiple Regression	4. Broken Assumptions
1.1 Data and Variation	2.1 Least Squares	3.1 Estimation	4.1 Heteroskedasticity
1.2 Conditional Means	2.2 Units and Fit	3.2 Inference	4.2 Specification
1.3 Sampling Variation	2.3 Bias and Precision	3.3 Functional Form	
		3.4 Dummies	

The Last of Four Lectures on Multiple Regression
Lecture 3.3 changed the shape of the schooling term. Today a regressor can be a category rather than a number.

Last Week's Lecture

Lecture 3.3 logged earnings and asked which specification to keep:

1. **Logs and Elasticities.**
   A log reads the return as a proportion, reported in per cent.

2. **Choosing a Specification.**
   R^2 never falls as columns are added; adjusted R^2 can fall but tests nothing.

3. **Specification Search.**
   A p-value describes one test chosen in advance, not the best of twenty.

Graduates Earn 8.17 Euro an Hour More

Our 19 graduates average 19.1277 euro an hour, and our 101 non-graduates 10.9550.

> **Question**
> Graduates also have more schooling. How much of the 8.17 euro gap remains at equal schooling?

Lecture Plan

1. **A Single Dummy Variable.**
   A 0/1 column shifts the intercept and leaves the two lines parallel.
2. **Multiple Categories.**
   Five qualifications need four dummies, each a gap from the base you chose.
3. **Interactions.**
   A product term lets the slopes differ, and the schooling coefficient describes the base group.
4. **Testing a Difference in Slopes.**
   Equal slopes is a test on one coefficient, and 19 graduates cannot reject it.

SECTION 1

A Single Dummy Variable

A Category Enters the Model as a 0/1 Column

A category has no scale, so it enters as ones and zeros:

$$y_i = \beta_0 + \beta_1(x_i - x_0) + \delta_0 G_i + u_i \tag{1}$$

- x_0: the reference level of schooling, thirteen years in our population
- G_i: one if person i has sixteen or more years of schooling, zero otherwise
- δ_0: the vertical gap between graduates and non-graduates

Nineteen of our 120 people carry a one.

Setting the Dummy Shifts Only the Intercept

Setting G_i to zero or one changes the intercept, not the slope:

$$G_i = 0: \quad \beta_0 + \beta_1(x_i - x_0) \qquad\qquad G_i = 1: \quad (\beta_0 + \delta_0) + \beta_1(x_i - x_0) \tag{2}$$

> **Delta Is a Difference Between Two Conditional Means**
> $\delta_0 = E(y \mid G = 1, x) - E(y \mid G = 0, x)$, at the same schooling in both (Wooldridge, 2020, eq. 7.2 p.222). Least squares is unchanged by the dummy. Only the reading of its coefficient is new (Wooldridge, 2020, p. 224).

Two Parallel Lines Through One Sample

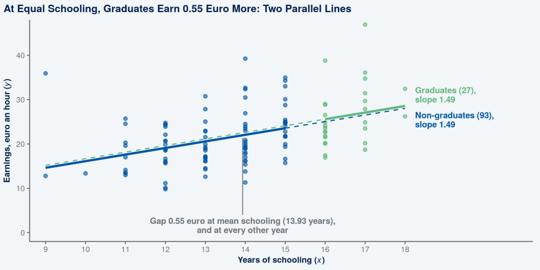

Figure 1: Our 120 by graduate status: two parallel fitted lines, 0.55 euro apart.

> **One Schooling Coefficient Forces Parallel Lines**
> Both slopes are 1.49, so the gap is 0.55 euro at every schooling level.

The Coefficient Is a Difference Between Two Averages

With schooling dropped, the intercept and the dummy reproduce the two group averages ($\bar{y}_0$ for non-graduates, $\bar{y}_1$ for graduates):

$$\hat{y}_i = \bar{y}_0 + \hat{\delta}_0 G_i \qquad \hat{\delta}_0 = \bar{y}_1 - \bar{y}_0 \tag{3}$$

Our 101 non-graduates average 10.9550 and our 19 graduates 19.1277.

> **A Constant and One Dummy Compare Two Means**
> The intercept is the non-graduate average. The coefficient is the gap between the two averages (Wooldridge, 2020, p. 224).

Worked example: our 120 people

Schooling Cuts the Graduate Premium from 8.17 To 2.86

Holding schooling fixed shrinks the graduate premium from 8.1726 to 2.8634:

- **Without Schooling:** $\hat{\delta}_0$ = 8.1726, with a standard error of 1.2806
- **With Schooling:** $\hat{\delta}_0$ = 2.8634, with a standard error of 1.6594

> **Graduates Have More Schooling by Definition**
> Most of the raw gap was schooling: every graduate has more schooling than every non-graduate. The drop is Lecture 3.1's partialling out, applied to a 0/1 column (Angrist & Pischke, 2015).

Dummy Coefficients When Earnings Are Logged

With log earnings, a dummy's coefficient is a proportional gap with two readings:

$$\text{shortcut: } 100\,\hat{\delta}_0 \qquad\qquad \text{exact: } 100\left(e^{\hat{\delta}_0} - 1\right) \tag{4}$$

Wooldridge's female coefficient of -0.297 means 29.7 per cent below men, or 25.7 exact (Wooldridge, 2020, p. 227).

> **The Shortcut Sits Between the Two Exact Figures**
> Measured from women, the exact figure is 34.6 per cent more for men. The shortcut's 29.7 sits between 25.7 and 34.6. So the shortcut can be reported without choosing a base group (Wooldridge, 2020, p. 228).

SECTION 2

Multiple Categories

Four Dummies for Five Qualifications

Highest qualification has five categories, so four dummies go in beside the intercept:

$$y_i = \beta_0 + \beta_1(x_i - x_0) + \theta_1 Q_{1i} + \theta_2 Q_{2i} + \theta_3 Q_{3i} + \theta_4 Q_{4i} + u_i \tag{5}$$

- Q_{ji}: one if person i holds qualification j, zero otherwise
- Q_{1i} to Q_{4i}: apprenticeship, certificate, degree and postgraduate
- θ_j: the gap between qualification j and the Leaving Certificate

Resetting the Base Changes All Four Coefficients

A new base changes all four coefficients and nothing about Ireland:

- **R's Default:** the base is the factor's first level, alphabetical unless set.
- **The Override:** `relevel()` sets the base, so spelling no longer decides the table.

> **Name the Omitted Category Under the Table**
> Without that name a reader cannot tell what each gap is measured from. $\theta_3 - \theta_2$ gives the degree-against-certificate gap. Our simulated people carry no qualification label, so nothing is fitted here.

Five Dummies Add up to the Intercept Column

Add a Leaving Certificate dummy (Q_{5i}), and the five dummies sum to one:

$$\sum_{j=1}^{5} Q_{ji} = 1 \qquad \text{for every person } i \tag{6}$$

The Intercept Is Already a Column of Ones
Each person holds exactly one qualification. So the five dummies together repeat the intercept column exactly.

Shifting Every Coefficient by κ Changes No Fitted Value

Raise the intercept by any constant (κ) and lower all five dummy coefficients by it:

$$(\beta_0 + \kappa) + \sum_{j=1}^{5} (\theta_j - \kappa)\, Q_{ji} = \beta_0 + \sum_{j=1}^{5} \theta_j Q_{ji} - \kappa \left(\sum_{j=1}^{5} Q_{ji} - 1 \right) \tag{7}$$

The Bracket Is Zero for Every Person

The five dummies sum to one, so the κ term vanishes. Every fitted value, every residual and the residual sum of squares are unchanged.

Use One Dummy Fewer Than There Are Categories

Infinitely many coefficient sets fit equally well, so no unique least-squares answer exists:

> **The Dummy Variable Trap**
> Wooldridge's assumption MLR.3, no perfect collinearity, fails exactly rather than nearly (Wooldridge, 2020, pp. MLR.3 p.80). With g categories and an intercept, $g-1$ dummies go in (Wooldridge, 2020, p. 229).

Some packages drop a dummy for you, and others only report the collinearity (Wooldridge, 2020, p. 228).

Three Qualification Coefficients and No Base Named

A table reports coefficients on `apprenticeship`, `degree` and `postgraduate`. Your co-author writes that the degree coefficient is the return to a degree.

> **Question**
> What is that number a comparison against? What else must the table print before you write about it?

The Degree Coefficient Is a Gap from the Omitted Category

The degree coefficient compares degree holders with whichever category was left out:

- **The Comparison:** degree against the omitted category, not a return to a degree.
- **The Missing Line:** the note under the table must name that category.

> **Degree Against Postgraduate Needs a Refit**
> Degree minus postgraduate is the difference of two printed coefficients. Its standard error needs their covariance, which no printout shows. Refit with degree as the base (Wooldridge, 2020, p. 229).

SECTION 3

Interactions

Interaction Terms Let the Slope Differ Across Groups

The product of schooling and the dummy lets the two groups' slopes differ:

$$y_i = \beta_0 + \beta_1(x_i - x_0) + \delta_0 G_i + \delta_1 \underbrace{(x_i - x_0)G_i}_{\text{the interaction term}} + u_i \tag{8}$$

- δ_1: the difference in slope between graduates and non-graduates

The Product Is Zero for Every Non-Graduate
For graduates the product equals centred schooling, which neither parent column repeats (Wooldridge, 2020, eq. 7.17 p.234).

Two Fitted Lines That Are Not Parallel

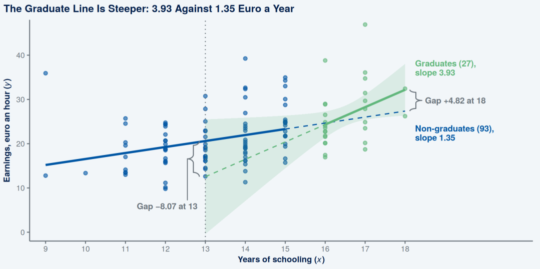

Figure 2: The same two groups, with slopes allowed to differ: 1.35 and 3.93 euro a year.

> **The Difference in Slope Is the Interaction Coefficient**
> The vertical gap between the lines is no longer one number. It grows by 2.58 euro with each extra year, $\hat{\delta}_1 = 2.58$.

Each Group's Fitted Line from One Model

Set $G_i = 0$, and only the non-graduates' line is left:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 (x_i - x_0) \tag{9}$$

Set $G_i = 1$: the intercept shifts by $\hat{\delta}_0$ and the slope by $\hat{\delta}_1$:

$$\hat{y}_i = (\hat{\beta}_0 + \hat{\delta}_0) + \underbrace{(\hat{\beta}_1 + \hat{\delta}_1)}_{\text{the graduates' slope}} (x_i - x_0) \tag{10}$$

Worked example: our 120 people

The Schooling Coefficient Now Describes the Base Group

Adding the product narrows what the schooling coefficient describes:

- **Without the Product:** 1.2042 euro a year, one return for all 120 people.
- **With the Product:** 1.2599 for the 101 non-graduates, 0.0531 for the 19 graduates.

> **The Main Term Describes Non-Graduates**
> A main term is the partial effect where the other variable is zero (Wooldridge, 2020, p. 193). Here that is $G_i = 0$. So 1.2599 is not an average of the two groups. R prints 1.2599 under the same row name as 1.2042.

The Dummy's Gap on Raw and Centred Schooling

On raw schooling, the dummy's coefficient (δ_0^{raw}) is the gap at zero years:

$$E(y \mid G = 1, x) - E(y \mid G = 0, x) = \delta_0^{\text{raw}} + \delta_1 \, x \tag{11}$$

Centred at thirteen, δ_0 is the gap at the Leaving Certificate:

$$E(y \mid G = 1, x) - E(y \mid G = 0, x) = \delta_0 + \delta_1 \, (x - x_0) \tag{12}$$

The Centred Coefficients at Thirteen Years

Nobody has zero years of schooling, so centring reports two coefficients at thirteen:

- **Intercept:** $\hat{\beta}_0$ = 11.2710, fitted earnings for a non-graduate at thirteen years.
- **Dummy:** $\hat{\delta}_0$ = 7.6357, the gap between the two lines at thirteen years.

> **No Graduate Has Thirteen Years**
> Graduates have sixteen years or more, by definition. So 7.6357 extends the graduate line three years below any graduate.

Centring Relocates Two Coefficients and Nothing Else

Replacing x_i by $x_i - x_0$ rewrites the intercept and the dummy's coefficient only:

Every Fitted Value Is Identical

Every fitted value, every residual and $\hat{\delta}_1$ are identical on both scales (Wooldridge, 2020, p. 193).

The centred fit prints the gap at thirteen with its own standard error.

SECTION 4

Testing a Difference in Slopes

The Equal-Slopes Null Restricts the Interaction Alone

The null of equal returns, $H_0 : \delta_1 = 0$, restricts the slopes alone:

> **Both Parent Terms Stay In**
> Drop G_i and the two lines must meet at thirteen years. Drop $(x_i - x_0)$ and the non-graduate slope must be zero.

Under H_0 the groups can still differ by one constant gap (Wooldridge, 2020, p. 234).

The t-Statistic for Equal Slopes

Whether the slopes differ depends on δ_1 alone, so the test uses $\hat{\delta}_1$:

$$t = \frac{\hat{\delta}_1}{\mathrm{se}(\hat{\delta}_1)} \tag{13}$$

> **Nineteen Graduates Cannot Separate the Slopes**
> Slopes of 1.2599 and 0.0531 look nothing alike. Equal slopes are still not rejected at five per cent. With 19 graduates, the standard error of 1.2566 exceeds the estimate's 1.2068.

Worked example: our 120 people

The Graduates’ Slope Needs Its Own Standard Error

The graduates’ slope adds two estimates, so its variance includes their covariance:

$$\mathrm{Var}(\hat{\beta}_1 + \hat{\delta}_1) = \mathrm{Var}(\hat{\beta}_1) + \mathrm{Var}(\hat{\delta}_1) + 2\,\mathrm{Cov}(\hat{\beta}_1, \hat{\delta}_1) \qquad (14)$$

The Covariance Is Not on the Printout

The printout gives standard errors for 1.2599 and -1.2068, not their covariance. Refit with graduates as the base to print 0.0531 with its standard error (Wooldridge, 2020, pp. 194, p.229).

The Return to Schooling for Women

One model carries log earnings, centred schooling, a women dummy and their interaction. You are asked for the return to schooling for women.

> **Question**
> Which of the three coefficients answers that? How would you get the standard error to put beside it?

The Return for Women Adds Two Coefficients

The women's return adds the schooling and interaction coefficients, in log points a year:

- **Schooling:** the return for men, the base group.
- **Women Dummy:** the gap in log earnings at thirteen years, not a return.
- **Interaction:** the women's return minus the men's.

> **The Sum Needs Its Own Standard Error**
> Adding the two standard errors, or their squares, leaves out twice the covariance. Refit with women as the base, and the sum prints with its standard error (Wooldridge, 2020, p. 229).

What Would Have to Be True?

The sixteen-year cut and a straight line within each group were our choices. Graduates are defined by schooling, so the groups share no year of schooling.

> **Question**
> What would have to be true for this number to mean what we just said it means? And what is the plausible story where it isn't?

Three Conditions Behind 1.2599 And 0.0531

Reading 1.2599 and 0.0531 as the groups' returns needs three things to hold:

- **Shape:** a straight line in schooling fits within each group.
- **The Cut:** sixteen years marks a real difference, not a round number.
- **Overlap:** graduates and non-graduates can be compared at the same schooling.

> **The Third Condition Fails by Construction**
> Graduates have sixteen years or more by definition. So 0.0531 rests on 19 people with 16 to 20 years. Our population gives everyone one slope, 1.30.

Summary: Dummies, Categories, Interactions and Tests

1. **A Single Dummy Variable.**
   The graduate premium is 8.1726 raw and 2.8634 with schooling held fixed.
2. **Multiple Categories.**
   Each coefficient is a gap from the omitted category, which the table must name.
3. **Interactions.**
   One model gives slopes of 1.2599 for non-graduates and 0.0531 for graduates.
4. **Testing a Difference in Slopes.**
   At $t = -0.96$, 19 graduates cannot reject equal slopes.

Before Lecture 4.1

1. **Play: the Specification App.**
   Switch the interaction on and off, and compare the two slopes.

2. **Apply: Problem Set 3.**
   One table of coefficients, written out in prose.

3. **Review: Wooldridge, Chapter 7.**
   Wooldridge (2020), sections 7-1 to 7-4, then 8-1 and 8-2 before Lecture 4.1.

Next Lecture

Lecture 4.1 asks what happens when the error’s spread differs across people:

1. **Heteroskedasticity.**
   The error variance changes with schooling, and the estimate stays centred on 1.30.
2. **Robust Standard Errors.**
   Squared residuals replace the unseen variances, with HC2 or HC3 in small samples.
3. **Serial Correlation.**
   Correlated errors leave the estimate centred too, and a missing variable is sought first.

Thank you

sam.deegan@ucdconnect.ie
sam-deegan.com

BACKUP

Appendix

Notation Used in This Lecture

Symbol	What it is	First met
i	which person, running over our 120	Lecture 1.1
x_i	years of schooling of person i	Lecture 1.1
y_i	hourly earnings of person i, in euro	Lecture 1.1
x	schooling, inside an expectation	Lecture 1.2
y	earnings, inside an expectation	Lecture 1.2
$E(y \mid G, x)$	expected earnings given the dummy and schooling	Lecture 1.2
$\mathrm{Var}(\cdot)$	the variance of an estimate across samples	Lecture 1.2
$\mathrm{Cov}(\cdot, \cdot)$	the covariance of two estimates across samples	Lecture 1.2
x_0	the reference level of schooling, thirteen years in our population	Lecture 2.1
β_0	base-group expected earnings at thirteen years	Lecture 2.1
β_1	the coefficient on centred schooling	Lecture 2.1
u_i	everything other than the regressors	Lecture 2.1

Notation Used in This Lecture

Symbol	What it is	First met
$\hat{y}_i$	the fitted value for person i	Lecture 2.1
$\hat{\beta}_0$	the fitted intercept, 11.2710	Lecture 2.1
$\hat{\beta}_1$	the fitted schooling slope, 1.2599	Lecture 2.1
$\mathrm{se}(\cdot)$	the standard error of an estimate	Lecture 2.3
H_0	the null hypothesis	Lecture 3.2
t	an estimate over its standard error	Lecture 3.2
$\log(\cdot)$	the natural logarithm	Lecture 3.3
e	the base of the natural logarithm	Lecture 3.3
Δ	the change in what follows	Lecture 3.3
$\%\Delta y$	the percentage change in earnings	Lecture 3.3
G_i	one for sixteen or more years of schooling	A Category Enters the Model as a 0/1 Column
δ_0	the gap between the two groups	A Category Enters the Model as a 0/1 Column

Notation Used in This Lecture

Symbol	What it is	First met
$\bar{y}_0$	the non-graduates' average earnings	The Coefficient Is a Difference Between Two Averages
$\bar{y}_1$	the graduates' average earnings	The Coefficient Is a Difference Between Two Averages
$\hat{\delta}_0$	the fitted graduate gap	The Coefficient Is a Difference Between Two Averages
Q_{ji}	one for qualification j	Four Dummies for Five Qualifications
Q_{1i}	apprenticeship	Four Dummies for Five Qualifications
Q_{2i}	certificate	Four Dummies for Five Qualifications
Q_{3i}	degree	Four Dummies for Five Qualifications
Q_{4i}	postgraduate	Four Dummies for Five Qualifications
j	which qualification, 1 to 5	Four Dummies for Five Qualifications
θ_j	the gap from the base	Four Dummies for Five Qualifications
Q_{5i}	the Leaving Certificate	Five Dummies Add Up to the Intercept Column

Notation Used in This Lecture

Symbol	What it is	First met
$\sum_{j=1}^{5}$	the sum over the dummies	Five Dummies Add Up to the Intercept Column
κ	any constant	Shifting Every Coefficient by κ Changes No Fitted Value
g	the number of categories	Use One Dummy Fewer Than There Are Categories
$(x_i - x_0)G_i$	the interaction term	Interaction Terms Let the Slope Differ Across Groups
δ_1	the slope difference	Interaction Terms Let the Slope Differ Across Groups
$\hat{\delta}_1$	the fitted slope difference	Two Fitted Lines That Are Not Parallel
δ_0^{raw}	the gap on raw schooling	The Dummy's Gap on Raw and Centred Schooling
β_0^{raw}	the raw-scale intercept	Derivation: Why the Centring Constant Cancels From the Slopes
$\hat{y}(x)$	the fitted value at x	Derivation: Where the Exact Percentage Change Comes From

Worked Example: Two Group Averages

With schooling dropped, the intercept and the dummy reproduce the two group averages:

$$\hat{y}_i = \bar{y}_0 + \hat{\delta}_0 \, G_i \qquad \hat{\delta}_0 = \bar{y}_1 - \bar{y}_0 \tag{15}$$

Our 101 non-graduates average 10.9550 and our 19 graduates 19.1277, and the fit on our 120 returns:

$$\hat{y}_i = \underbrace{10.9550}_{\bar{y}_0} + \underbrace{8.1726}_{\hat{\delta}_0} \, G_i \qquad \mathrm{se}(\hat{\delta}_0) = 1.2806 \tag{16}$$

Back to the slide

Worked Example: Each Group's Fitted Line

One model gives both lines, with $G_i = 1$ switching on the graduates' two shifts:

$$G_i = 0: \quad \hat{\beta}_0 + \hat{\beta}_1(x_i - 13) \qquad\qquad G_i = 1: \quad (\hat{\beta}_0 + \hat{\delta}_0) + (\hat{\beta}_1 + \hat{\delta}_1)(x_i - 13) \tag{17}$$

$$G_i = 0: \quad \hat{y}_i = 11.2710 + 1.2599\,(x_i - 13) \tag{18}$$

Set $G_i = 1$: the intercept rises by 7.6357 and the slope falls by 1.2068:

$$\hat{y}_i = \underbrace{(11.2710 + 7.6357)}_{18.9067} + \underbrace{(1.2599 - 1.2068)}_{0.0531}\,(x_i - 13) \tag{19}$$

Back to the slide

Worked Example: t-Statistic for Equal Slopes

Whether the slopes differ depends on δ_1 alone, so the test uses $\hat{\delta}_1$:

$$t = \frac{\hat{\delta}_1}{\mathrm{se}(\hat{\delta}_1)} \tag{20}$$

Our 120 people return $\hat{\delta}_1 = -1.2068$ with a standard error of 1.2566:

$$t = \frac{-1.2068}{1.2566} = -0.96 \tag{21}$$

Back to the slide

Derivation: Where the Exact Percentage Change Comes From

The body frame stated $\%\Delta y = 100(e^{\beta_1} - 1)$. That formula is the fitted model differenced across one year and exponentiated:

$$\log \hat{y}(x+1) - \log \hat{y}(x) = \beta_1 \quad \Longrightarrow \quad \frac{\hat{y}(x+1)}{\hat{y}(x)} = e^{\beta_1} \quad \Longrightarrow \quad \frac{\Delta \hat{y}}{\hat{y}} = e^{\beta_1} - 1 \tag{22}$$

The Shortcut Is the First Term of a Series

$e^{\beta_1} - 1 = \beta_1 + \beta_1^2/2 + \ldots$, so the shortcut drops everything past the first term. The leading error is $\beta_1^2/2$. That leading error is 0.00125 at $\beta_1 = 0.05$ and 0.125 at $\beta_1 = 0.50$. The gap column on the body frame grows at that same square rate (Wooldridge, 2020, p. 186).

Derivation: Why the Centring Constant Cancels from the Slopes

Replacing x_i by $x_i - x_0$ rewrites the intercept and the dummy coefficient. The two slopes are unchanged:

$$\beta_0^{\text{raw}} + \beta_1 x_i + \delta_0^{\text{raw}} G_i + \delta_1 x_i G_i = \underbrace{(\beta_0^{\text{raw}} + x_0\beta_1)}_{\beta_0} + \beta_1(x_i - x_0) + \underbrace{(\delta_0^{\text{raw}} + x_0\delta_1)}_{\delta_0} G_i + \delta_1(x_i - x_0)G_i \quad (23)$$

The Intercept and the Dummy Coefficient Are Relabelled

β_1 and δ_1 appear unchanged on both sides. No slope, fitted value or residual differs. Only the intercept and the dummy coefficient are relabelled. Centring reports those two coefficients at thirteen years of schooling instead of zero (Wooldridge, 2020, p. 193).

Notes and Tools

- **The App:** the specification app, with the four forms and the interaction switchable.
- **The Lecture in One Line:** `lm(log(y) ~ I(x - 13) * G)`, which carries both parent terms.
- **This Lecture's Lab:** `log()`, `factor()`, `relevel()`, `I(x - 13)`, and `y ~ x * G` for both parent terms and their product.

References

Angrist, J. D., & Pischke, J.-S. (2015). *Mastering 'metrics: The path from cause to effect*. Princeton University Press.

Wooldridge, J. M. (2020). *Introductory econometrics: A modern approach* (7th ed.). Cengage Learning.