Indicator Variables

Putting Qualitative Information into a Regression

Jake Anderson

March 21, 2026

Outline

Outline

Men Earn More Than Women on Average

From the Current Population Survey (CPS 2013):

	Mean hourly wage	*N*
Male	\$25.00	600
Female	\$20.78	600
Difference	\$4.22	

Question: How do we control for education when measuring this gap?

Gender is not a number on a continuous scale. It is a **category**. We need a way to put categorical information into a regression.

Outline

Indicator Variables: The Idea

Create a variable that equals 1 or 0 depending on group membership:

$$\text{female}_i = \begin{cases} 1 & \text{if person } i \text{ is female} \\ 0 & \text{if person } i \text{ is male} \end{cases}$$

Other names: **dummy variable**, **binary variable**, **dichotomous variable**.

Now put it into the wage regression:

$$\text{wage}_i = \beta_1 + \delta\,\text{female}_i + \beta_2\,\text{educ}_i + e_i$$

$\Longrightarrow$ This is a standard multiple regression. OLS applies exactly as before. The only new feature is that one regressor takes just two values.

What the Model Produces: Two Regression Functions

$$\text{wage}_i = \beta_1 + \delta\,\text{female}_i + \beta_2\,\text{educ}_i + e_i$$

For men (female $= 0$):

$$\text{E(wage)} = \beta_1 + \beta_2\,\text{educ}$$

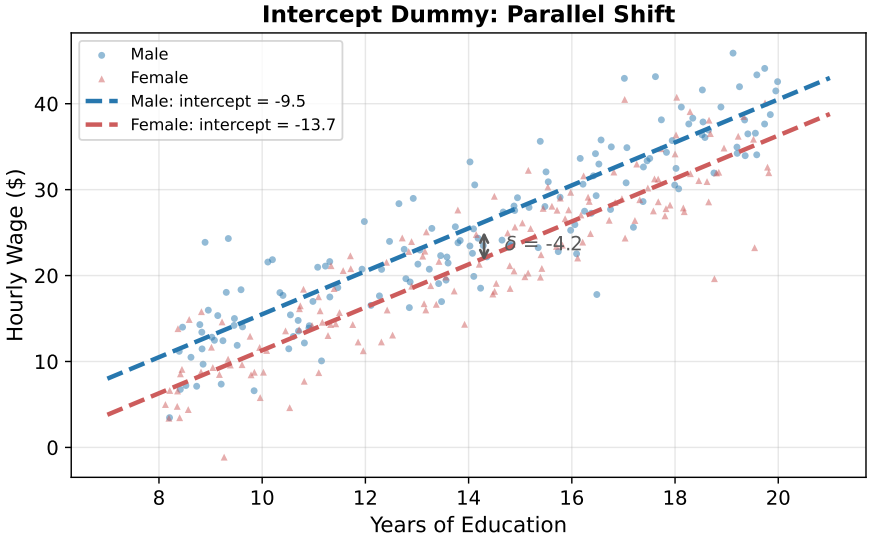
For women (female $= 1$):

$$\text{E(wage)} = (\beta_1 + \delta) + \beta_2\,\text{educ}$$

$\Longrightarrow$ Two **parallel lines** with the same slope (β_2) but different intercepts.

The indicator variable creates a **parallel shift** of δ dollars.

Parallel Lines: The Picture



Interpreting δ: The Average Group Difference

$$\text{wage}_i = \beta_1 + \delta\,\text{female}_i + \beta_2\,\text{educ}_i + e_i$$

δ is the difference in expected wages between women and men, **holding education constant**:

$$\delta = \text{E}(\text{wage} \mid \text{female} = 1, \text{educ}) - \text{E}(\text{wage} \mid \text{female} = 0, \text{educ})$$

- $\delta < 0$: women earn less than men at the same education level
- $\delta > 0$: women earn more
- $\delta = 0$: no gender wage gap after controlling for education

Testing the gap: $H_0 : \delta = 0$ vs. $H_1 : \delta \neq 0$ is a standard t-test.

The Reference Group

The group coded $D = 0$ is the **reference group** (or base group).

- With female_i: males are the reference group
- δ measures the difference *relative to males*

We could equivalently define $\text{male}_i = 1 - \text{female}_i$ and write:

$$\text{wage}_i = \beta_1^* + \delta^* \, \text{male}_i + \beta_2 \, \text{educ}_i + e_i$$

Now females are the reference group, and $\delta^* = -\delta$.

$\implies$ The choice of reference group changes the sign of the coefficient but not the model's predictions. Always ask: "relative to what?"

Outline

What If There Are More Than Two Groups?

Consider U.S. regions: Northeast, South, Midwest, West.

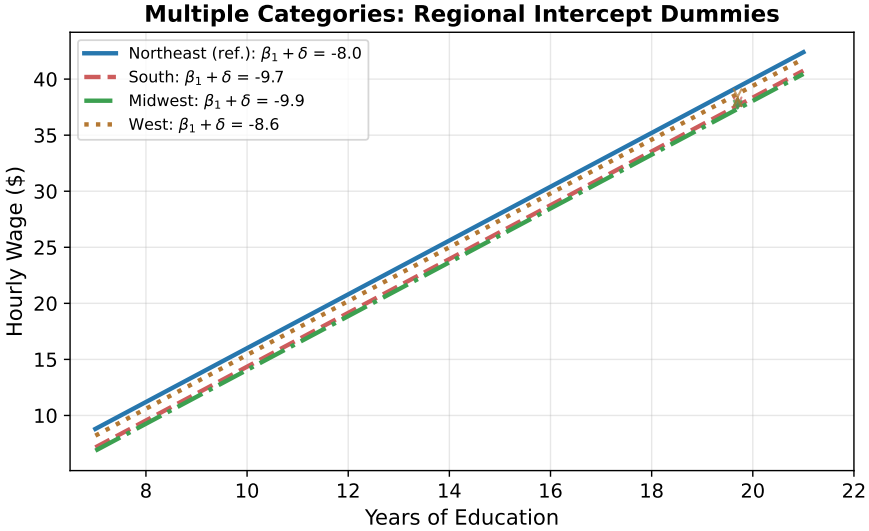
Wage equation with regional indicators:

$$\text{wage}_i = \beta_1 + \delta_1 \, \text{south}_i + \delta_2 \, \text{midwest}_i + \delta_3 \, \text{west}_i + \beta_2 \, \text{educ}_i + e_i$$

Each δ_j measures the wage difference relative to **Northeast** (the omitted category).

General rule: For g categories, include $g - 1$ indicator variables. One category must be omitted as the reference group.

Reference Group: The Picture



The Dummy Variable Trap

Why not include indicators for *all* categories?

Suppose we include dummies for all four regions *plus* the intercept:

$$\text{wage}_i = \beta_1 + \delta_1 \text{south}_i + \delta_2 \text{midwest}_i + \delta_3 \text{west}_i + \delta_4 \text{northeast}_i + \beta_2 \text{educ}_i + e_i$$

Problem: for every observation, exactly one region indicator equals 1 and the rest equal 0:

$$\text{south}_i + \text{midwest}_i + \text{west}_i + \text{northeast}_i = 1 = x_{1i}$$

$\implies$ The sum of the four indicator columns equals the intercept column. This is **exact collinearity**, and OLS cannot run.

The fix: always omit one category (or drop the intercept). This is the **dummy variable trap**.

Outline

Is the Return to Education the Same for Everyone?

The intercept-dummy model forces the **same slope** for both groups:

$$\text{wage}_i = \beta_1 + \delta\,\text{female}_i + \beta_2\,\text{educ}_i + e_i$$

But what if women get a different payoff per year of education than men?

We already covered this in Topic 16 (Interaction Terms). The solution is an **interaction between the indicator and the continuous variable**:

$$\text{wage}_i = \beta_1 + \delta\,\text{female}_i + \beta_2\,\text{educ}_i + \gamma\,(\text{female}_i \times \text{educ}_i) + e_i$$

The interaction term female $\times$ educ is called a **slope-indicator variable** (or slope dummy).

Two Completely Separate Regression Lines

$$\text{wage}_i = \beta_1 + \delta\,\text{female}_i + \beta_2\,\text{educ}_i + \gamma\,(\text{female}_i \times \text{educ}_i) + e_i$$

For men (female $= 0$):

$$\text{E(wage)} = \beta_1 + \beta_2\,\text{educ}$$

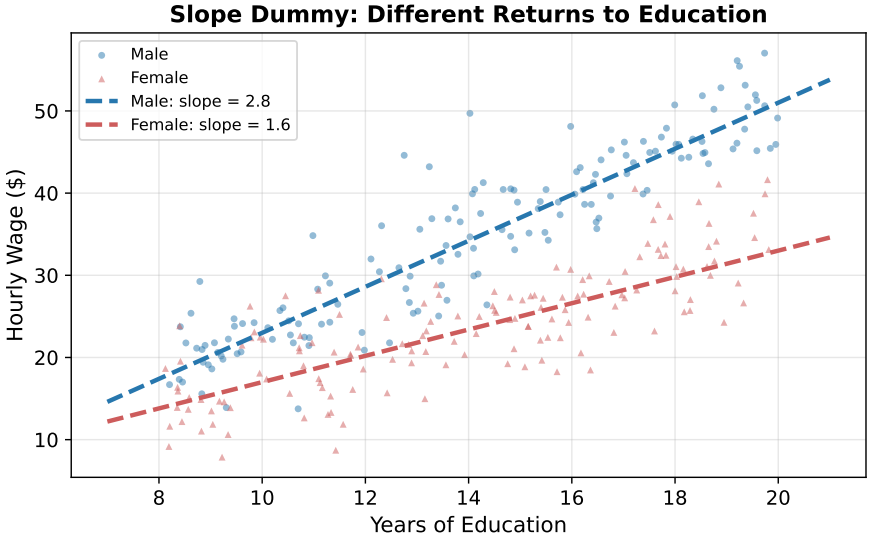
For women (female $= 1$):

$$\text{E(wage)} = \underbrace{(\beta_1 + \delta)}_{\text{intercept}} + \underbrace{(\beta_2 + \gamma)}_{\text{slope}}\,\text{educ}$$

$\Longrightarrow$ Each group now has its **own intercept and its own slope**.

γ measures how much the return to education differs for women relative to men.

Different Slopes: The Picture



Interpreting the Slope Dummy Coefficient

$$\text{wage}_i = \beta_1 + \delta\,\text{female}_i + \beta_2\,\text{educ}_i + \gamma\,(\text{female}_i \times \text{educ}_i) + e_i$$

- β_2: return to education **for men** (the reference group)
- $\beta_2 + \gamma$: return to education **for women**
- γ: the **difference** in returns (women minus men)
- δ: wage gap *at* educ $= 0$ (rarely meaningful alone)

The full wage gap at education level e:

$$\text{E}(\text{wage}_{\text{female}}) - \text{E}(\text{wage}_{\text{male}}) = \delta + \gamma \cdot e$$

$\Longrightarrow$ The gap is not a single number. You must specify *at what education level* to evaluate it.

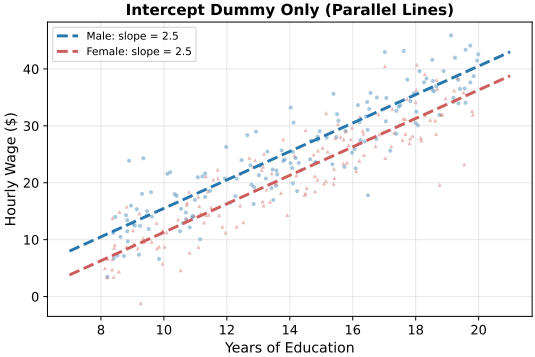
Outline

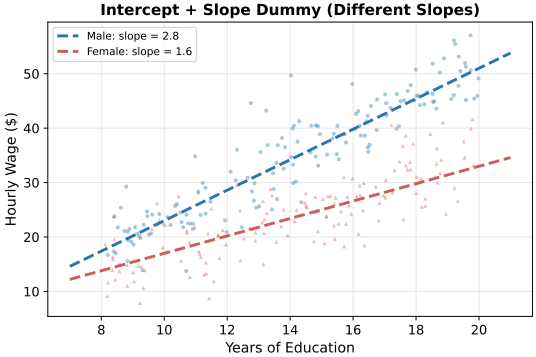
Building the Wage Model Step by Step

Model	Specification	Geometry
No gender	$\text{wage} = \beta_1 + \beta_2\,\text{educ} + e$	One line
Intercept dummy	$\text{wage} = \beta_1 + \delta\,\text{female} + \beta_2\,\text{educ} + e$	Two parallel lines
Intercept + slope dummy	$\text{wage} = \beta_1 + \delta\,\text{female} + \beta_2\,\text{educ} + \gamma(\text{female} \times \text{educ}) + e$	Two lines with different slopes

$\Longrightarrow$ Each step adds flexibility. But more flexibility uses more degrees of freedom. How do we decide which model to use?

Side-by-Side: Parallel Lines vs. Different Slopes





Outline

Testing Whether Groups Have the Same Regression

The fully interacted model allows both the intercept and slope to differ:

$$\text{wage}_i = \beta_1 + \delta\,\text{female}_i + \beta_2\,\text{educ}_i + \gamma(\text{female}_i \times \text{educ}_i) + e_i$$

Question: Do we actually need separate regressions, or is a single pooled regression adequate?

This is a joint F-test on all the indicator-related coefficients:

$$H_0 : \delta = 0 \text{ and } \gamma = 0$$

$$H_1 : \text{at least one is nonzero}$$

$\Longrightarrow$ This is the **Chow test**: an F-test for the equivalence of two regressions.

Chow Test: The Procedure

Step 1: Estimate the restricted model (pooled, no indicators):

$$\text{wage}_i = \beta_1 + \beta_2\,\text{educ}_i + e_i \qquad \longrightarrow \qquad SSE_R$$

Step 2: Estimate the unrestricted model (with all indicators):

$$\text{wage}_i = \beta_1 + \delta\,\text{female}_i + \beta_2\,\text{educ}_i + \gamma(\text{female}_i \times \text{educ}_i) + e_i \qquad \longrightarrow \qquad SSE_U$$

Step 3: Compute the F-statistic:

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} \sim F_{(J,\,N-K)}$$

where J = number of restrictions (here $J = 2$: $\delta = 0$ and $\gamma = 0$).

$\Longrightarrow$ This is exactly the F-test from Topic 18. The only new part is what we are testing.

Chow Test: General Case

With K parameters in the base model (including the intercept, as in the F-test convention from Topic 17), the fully interacted model adds a group indicator and $K-1$ interaction terms, for $J = K$ total restrictions:

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_K = 0$$

Equivalently, the unrestricted SSE_U equals the sum of SSEs from running **separate regressions** for each group:

$$SSE_U = SSE_{\text{male}} + SSE_{\text{female}}$$

Assumption required: equal error variances across groups. If variances differ, the standard F-test is not valid.

$\Longrightarrow$ The Chow test answers: "Should I run one regression on the pooled data, or separate regressions for each group?"

Example: Chow Test for Regional Differences (CPS Data)

From the textbook (Example 7.4): testing whether the wage equation differs for South vs. non-South workers.

- $SSE_R = 214{,}400.9$ (pooled model, no SOUTH interactions)
- $SSE_U = 213{,}774.0$ (sum of separate regressions)
- $J = 5$ restrictions, $N - K = 1190$

$$F = \frac{(214{,}400.9 - 213{,}774.0)/5}{213{,}774.0/1190} = 0.698 \qquad p = 0.625$$

Fail to reject H_0. The wage equation is not significantly different in the South.

$\Longrightarrow$ A single pooled regression is adequate. We do not need separate models.

Outline

Dummy Variables with $\ln(y)$

When the dependent variable is in logs, the indicator coefficient has a **percentage interpretation**:

$$\ln(\text{wage}_i) = \beta_1 + \beta_2\, \text{educ}_i + \delta\, \text{female}_i + e_i$$

Taking the difference between female $= 1$ and female $= 0$:

$$\ln(\text{wage}_\text{F}) - \ln(\text{wage}_\text{M}) = \delta$$

Rough approximation (for small $|\delta|$):

$$\text{Percentage difference} \approx 100\delta\%$$

For larger $|\delta|$, the approximation breaks down. The exact percentage difference is:

Exact formula:

$$\text{Percentage difference} = 100(e^{\delta} - 1)\%$$

Example: $\hat{\delta} = -0.178$. Rough: -17.8%. Exact: $100(e^{-0.178} - 1) = -16.3\%$.

$\implies$ For $|\delta| > 0.10$, use the exact formula.

Outline

Summary: What Indicator Variables Can Do

What you add	What changes	Test
D alone	Intercept shifts	t-test on δ
$D \times x$	Slope changes	t-test on γ
D and $D \times x$	Both change	Chow test (F-test)
$g - 1$ dummies	g group intercepts	Joint F-test

Takeaways:

- Indicator variables turn qualitative information into regressors
- An intercept dummy creates parallel regression lines; a slope dummy allows different slopes
- For g categories, use $g - 1$ indicators (omit one reference group)
- The Chow test is a joint F-test for whether separate regressions are needed
- In log-linear models, use $100(e^{\delta} - 1)\%$ for the exact percentage effect

Thank you!
jakeanderson@g.ucla.edu