

# Indicator Variables

Putting Qualitative Information into a Regression

Jake Anderson

March 21, 2026

# Outline

- 1 Motivation: The Gender Wage Gap
- 2 The Intercept Indicator (Dummy Variable)
- 3 Multiple Categories and the Dummy Variable Trap
- 4 Slope Dummies: Different Slopes for Different Groups
- 5 The Wage Gap: From Simple to Fully Flexible
- 6 The Chow Test: Do We Need Separate Regressions?
- 7 Indicators in Log-Linear Models
- 8 Summary

- 1 Motivation: The Gender Wage Gap
- 2 The Intercept Indicator (Dummy Variable)
- 3 Multiple Categories and the Dummy Variable Trap
- 4 Slope Dummies: Different Slopes for Different Groups
- 5 The Wage Gap: From Simple to Fully Flexible
- 6 The Chow Test: Do We Need Separate Regressions?
- 7 Indicators in Log-Linear Models
- 8 Summary

# Men Earn More Than Women on Average

From the Current Population Survey (CPS 2013):

	<b>Mean hourly wage</b>	<i>N</i>
Male	\$25.00	600
Female	\$20.78	600
Difference	\$4.22	

# Men Earn More Than Women on Average

From the Current Population Survey (CPS 2013):

	Mean hourly wage	<i>N</i>
Male	\$25.00	600
Female	\$20.78	600
Difference	\$4.22	

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

# Men Earn More Than Women on Average

From the Current Population Survey (CPS 2013):

	Mean hourly wage	<i>N</i>
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

- 1 Motivation: The Gender Wage Gap
- 2 The Intercept Indicator (Dummy Variable)**
- 3 Multiple Categories and the Dummy Variable Trap
- 4 Slope Dummies: Different Slopes for Different Groups
- 5 The Wage Gap: From Simple to Fully Flexible
- 6 The Chow Test: Do We Need Separate Regressions?
- 7 Indicators in Log-Linear Models
- 8 Summary

# Indicator Variables: The Idea

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

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

# 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**.

# 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:

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

# 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:

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

⇒ 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

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

# 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):**

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

# 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):**

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

**For women (female = 1):**

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

# 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):

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

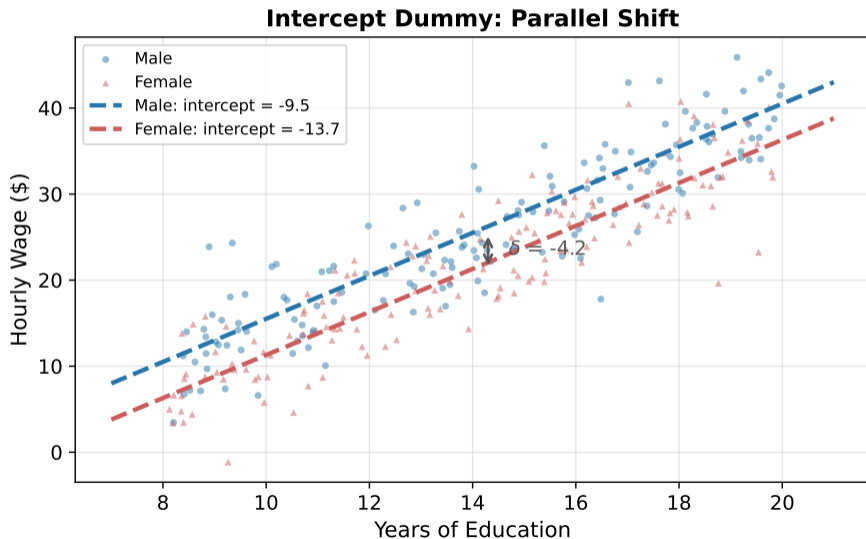
**For women** (female = 1):

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

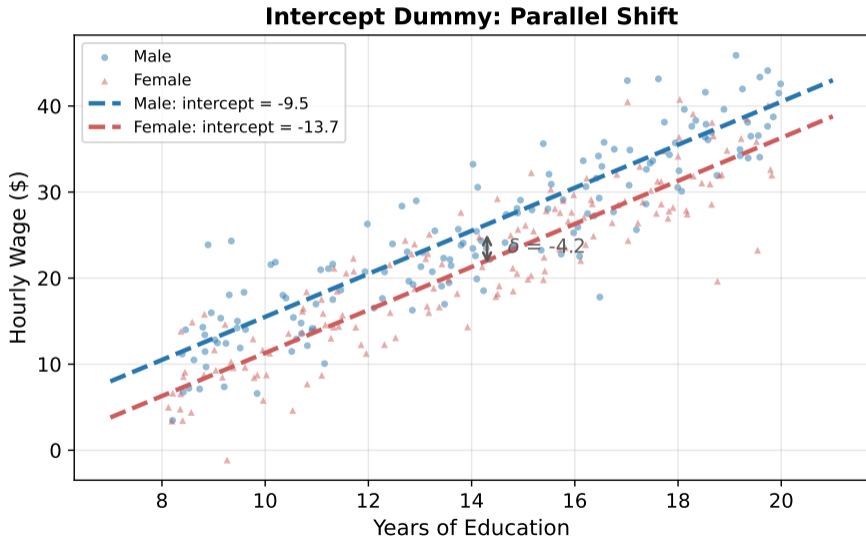
⇒ Two **parallel lines** with the same slope ( $\beta_2$ ) but different intercepts.

The indicator variable creates a **parallel shift** of  $\delta$  dollars.

# Parallel Lines: The Picture



# Parallel Lines: The Picture



## Interpreting $\delta$ : The Average Group Difference

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

# Interpreting $\delta$ : The Average Group Difference

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

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

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

## Interpreting $\delta$ : The Average Group Difference

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

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

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

- $\delta < 0$ : 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

## Interpreting $\delta$ : The Average Group Difference

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

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

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

- $\delta < 0$ : 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;: males are the reference group
- $\delta$  measures the difference *relative to males*

# The Reference Group

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

- With  $\text{female}_i$ ; males are the reference group
- $\delta$  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  $\delta^* = -\delta$ .

# The Reference Group

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

- With  $\text{female}_i$ : males are the reference group
- $\delta$  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  $\delta^* = -\delta$ .

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

# Outline

- 1 Motivation: The Gender Wage Gap
- 2 The Intercept Indicator (Dummy Variable)
- 3 Multiple Categories and the Dummy Variable Trap**
- 4 Slope Dummies: Different Slopes for Different Groups
- 5 The Wage Gap: From Simple to Fully Flexible
- 6 The Chow Test: Do We Need Separate Regressions?
- 7 Indicators in Log-Linear Models
- 8 Summary

# What If There Are More Than Two Groups?

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

Wage equation with regional indicators:

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

# 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  $\delta_j$  measures the wage difference relative to **Northeast** (the omitted category).

# What If There Are More Than Two Groups?

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

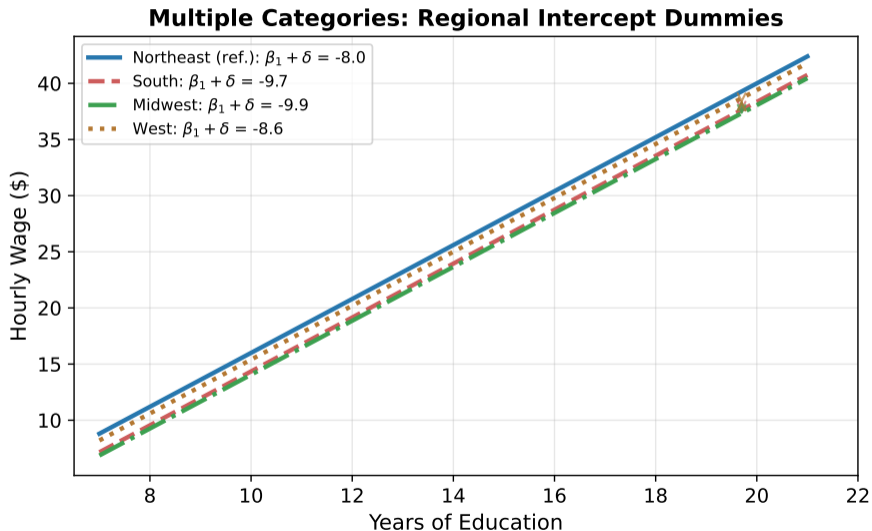
Wage equation with regional indicators:

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

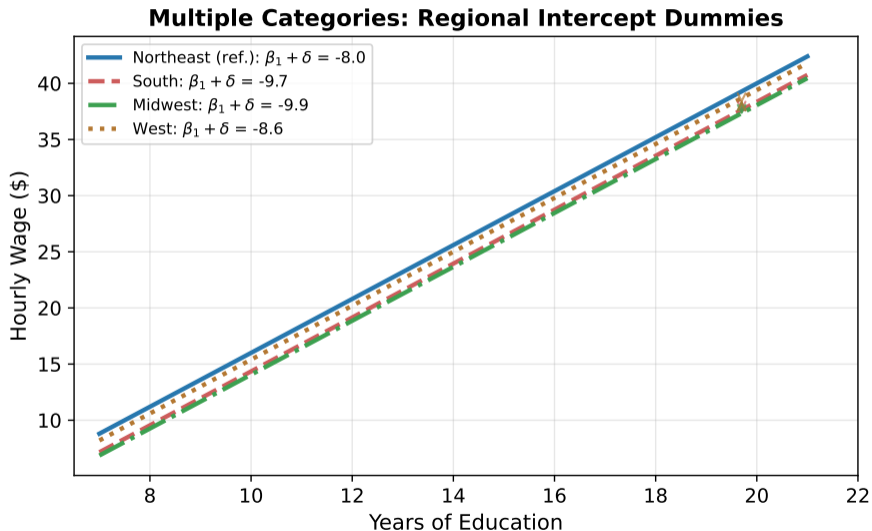
Each  $\delta_j$  measures the wage difference relative to **Northeast** (the omitted category).

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

# Reference Group: The Picture



# 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:

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

# 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:

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

# 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:

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

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

# 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:

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

⇒ 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

- 1 Motivation: The Gender Wage Gap
- 2 The Intercept Indicator (Dummy Variable)
- 3 Multiple Categories and the Dummy Variable Trap
- 4 Slope Dummies: Different Slopes for Different Groups**
- 5 The Wage Gap: From Simple to Fully Flexible
- 6 The Chow Test: Do We Need Separate Regressions?
- 7 Indicators in Log-Linear Models
- 8 Summary

# Is the Return to Education the Same for Everyone?

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

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

# 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?

# 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**:

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

# 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  $\text{female} \times \text{educ}$  is called a **slope-indicator variable** (or slope dummy).

## Two Completely Separate Regression Lines

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

## 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):

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

## 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):**

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

**For women (female = 1):**

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

## 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):

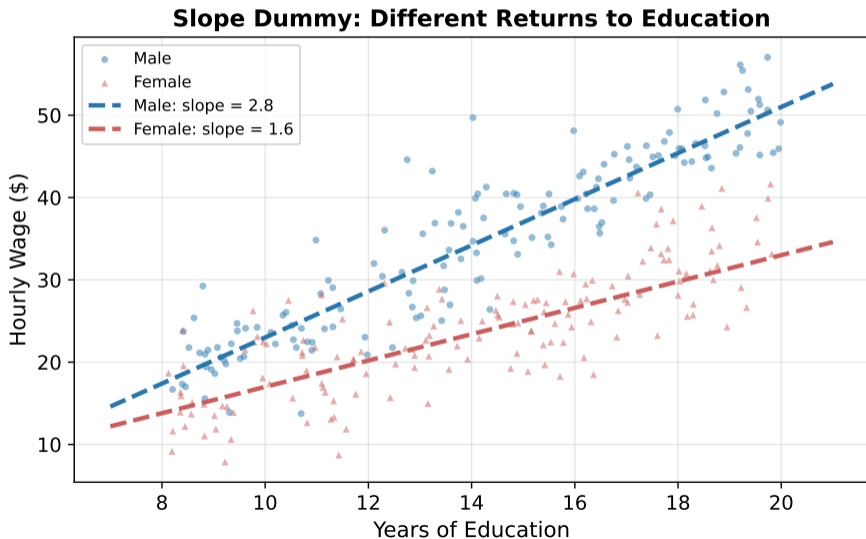
$$E(\text{wage}) = \beta_1 + \beta_2 \text{educ}$$

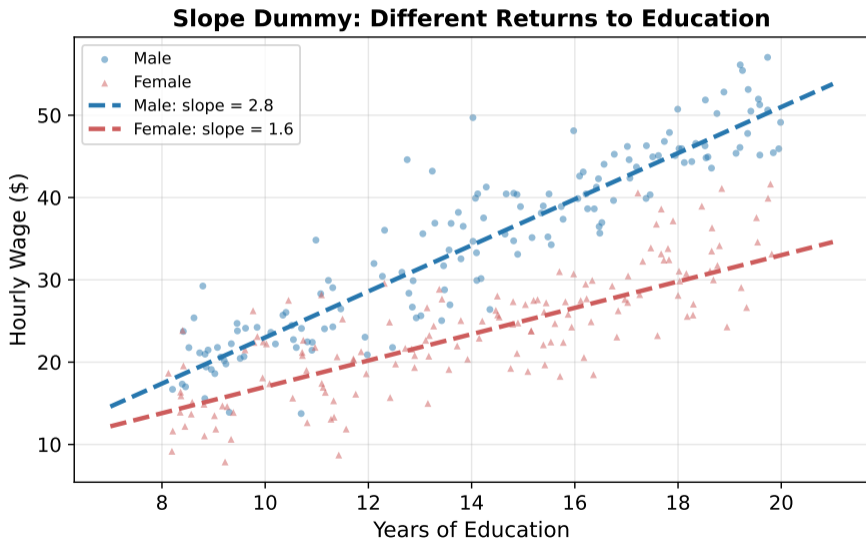
**For women** (female = 1):

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

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

$\gamma$  measures how much the return to education differs for women relative to men.





# Interpreting the Slope Dummy Coefficient

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

# Interpreting the Slope Dummy Coefficient

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

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

# Interpreting the Slope Dummy Coefficient

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

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

The full wage gap at education level  $e$ :

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

# Interpreting the Slope Dummy Coefficient

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

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

The full wage gap at education level  $e$ :

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

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

# Outline

- 1 Motivation: The Gender Wage Gap
- 2 The Intercept Indicator (Dummy Variable)
- 3 Multiple Categories and the Dummy Variable Trap
- 4 Slope Dummies: Different Slopes for Different Groups
- 5 The Wage Gap: From Simple to Fully Flexible**
- 6 The Chow Test: Do We Need Separate Regressions?
- 7 Indicators in Log-Linear Models
- 8 Summary

# Building the Wage Model Step by Step

Model	Specification	Geometry
No gender	$wage = \beta_1 + \beta_2 educ + e$	One line

# Building the Wage Model Step by Step

Model	Specification	Geometry
No gender	$wage = \beta_1 + \beta_2 educ + e$	One line
Intercept dummy	$wage = \beta_1 + \delta female + \beta_2 educ + e$	Two parallel lines

# Building the Wage Model Step by Step

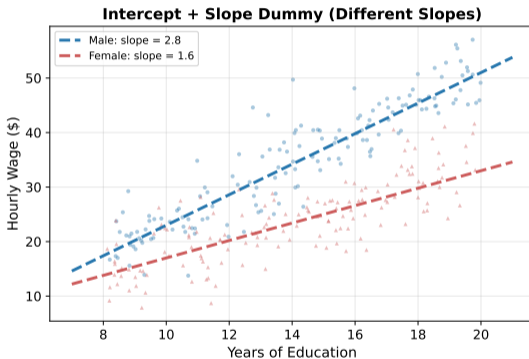
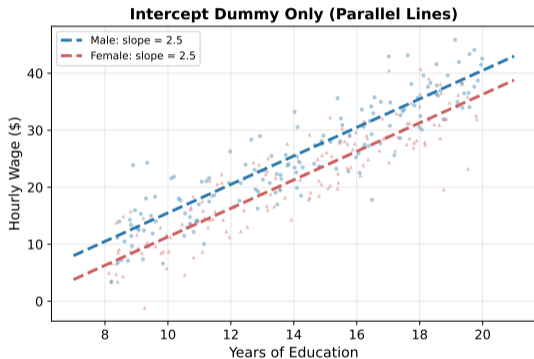
Model	Specification	Geometry
No gender	$wage = \beta_1 + \beta_2 educ + e$	One line
Intercept dummy	$wage = \beta_1 + \delta female + \beta_2 educ + e$	Two parallel lines
Intercept + slope dummy	$wage = \beta_1 + \delta female + \beta_2 educ + \gamma(female \times educ) + e$	Two lines with different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

# Building the Wage Model Step by Step

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

⇒ 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

- 1 Motivation: The Gender Wage Gap
- 2 The Intercept Indicator (Dummy Variable)
- 3 Multiple Categories and the Dummy Variable Trap
- 4 Slope Dummies: Different Slopes for Different Groups
- 5 The Wage Gap: From Simple to Fully Flexible
- 6 The Chow Test: Do We Need Separate Regressions?**
- 7 Indicators in Log-Linear Models
- 8 Summary

# Testing Whether Groups Have the Same Regression

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

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

# 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?

# 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:

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

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

# 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:

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

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

⇒ 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):

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

# Chow Test: The Procedure

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

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

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

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

# Chow Test: The Procedure

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

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

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

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

**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  $\gamma = 0$ ).

# Chow Test: The Procedure

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

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

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

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

**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  $\gamma = 0$ ).

$\implies$  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:

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

## 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:

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

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

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

## 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:

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

Equivalently, the unrestricted  $SSE_U$  equals the sum of  $SSE$ s 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.

## 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:

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

Equivalently, the unrestricted  $SSE_U$  equals the sum of  $SSE$ s 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.

⇒ 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,  $N - K = 1190$

## 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,  $N - K = 1190$

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

## 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,  $N - K = 1190$

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

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

## 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,  $N - K = 1190$

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

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

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

# Outline

- 1 Motivation: The Gender Wage Gap
- 2 The Intercept Indicator (Dummy Variable)
- 3 Multiple Categories and the Dummy Variable Trap
- 4 Slope Dummies: Different Slopes for Different Groups
- 5 The Wage Gap: From Simple to Fully Flexible
- 6 The Chow Test: Do We Need Separate Regressions?
- 7 Indicators in Log-Linear Models**
- 8 Summary

## Dummy Variables with $\ln(y)$

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

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

## 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  $\text{female} = 1$  and  $\text{female} = 0$ :

$$\ln(\text{wage}_F) - \ln(\text{wage}_M) = \delta$$

## 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  $\text{female} = 1$  and  $\text{female} = 0$ :

$$\ln(\text{wage}_F) - \ln(\text{wage}_M) = \delta$$

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

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

## 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  $\text{female} = 1$  and  $\text{female} = 0$ :

$$\ln(\text{wage}_F) - \ln(\text{wage}_M) = \delta$$

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

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

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

## 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:

$$\ln(\text{wage}_F) - \ln(\text{wage}_M) = \delta$$

**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:**

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

## 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:

$$\ln(\text{wage}_F) - \ln(\text{wage}_M) = \delta$$

**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

- 1 Motivation: The Gender Wage Gap
- 2 The Intercept Indicator (Dummy Variable)
- 3 Multiple Categories and the Dummy Variable Trap
- 4 Slope Dummies: Different Slopes for Different Groups
- 5 The Wage Gap: From Simple to Fully Flexible
- 6 The Chow Test: Do We Need Separate Regressions?
- 7 Indicators in Log-Linear Models
- 8 Summary**

## Summary: What Indicator Variables Can Do

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

## Summary: What Indicator Variables Can Do

What you add	What changes	Test
$D$ alone	Intercept shifts	$t$ -test on $\delta$
$D \times x$	Slope changes	$t$ -test on $\gamma$
$D$ 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