

# Regression Mechanics: Exponentials & Interactions

**RMDA II — Spring 2026**

Sebastian Tello-Trillo

University of Virginia

# Today's Roadmap

## Part I: Review & Non-Linearities

1. Reading output
2. Predicting  $Y$
3. Reference categories
4.  $X^2$  and marginal effects
5. Climate change application

## Part II: Interaction Terms

6. Motivation: bacteria & humidity
7. Binary  $\times$  binary
8. Binary  $\times$  continuous
9. Sub-sample equivalence

## Part III: Log Transformations

10. Why percentages matter
11. Log-linear, linear-log, log-log
12. The  $2 \times 2$  summary table
13. Combining everything

### Key Insight

Every tool today involves *non-linear* transformations of variables—but the model stays **linear in the parameters** ( $\beta$ 's), so OLS still works.

## What Does $\hat{\beta}$ Actually Tell Us?

---

# If I Hand You This Output, What Can You Do With It?

```
. reg hwage age wgtadj hgtadj male
```

Source	SS	df	MS	Number of obs	=	129,755
Model	3092320.87	4	773080.217	F(4, 129750)	=	9.15
Residual	1.0964e+10	129,750	84497.8969	Prob > F	=	0.0000
Total	1.0967e+10	129,754	84519.1242	R-squared	=	0.0003
				Adj R-squared	=	0.0003
				Root MSE	=	290.69

hwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.5461922	.1002411	5.45	0.000	.3497214	.7426629
wgtadj	-.024785	.0241487	-1.03	0.305	-.0721159	.022546
hgtadj	.2585054	.3378081	0.77	0.444	-.4035924	.9206032
male	3.273136	2.362267	1.39	0.166	-1.356865	7.903137
_cons	-17.30314	21.17435	-0.82	0.414	-58.80449	24.1982

Regression of hourly wages on age, weight, height, and a male dummy. Every number here is usable.

## Practice: Reading Regression Output

$$\text{hwage}_i = \hat{\alpha}_0 + \hat{\beta}_1 \text{Age} + \hat{\beta}_2 \text{Weight} + \hat{\beta}_3 \text{Height} + \hat{\beta}_4 \text{Male}$$

### Question

1. What is the **marginal effect** of being male (binary variable)?
2. What is the **marginal effect** of one more year of age (continuous)?
3. Does the marginal effect of age change at age 40 versus age 25?
4. **Predict** the hourly wage for a 40-year-old male, 150 lbs, 62 inches.

## Answers: Marginal Effects and Prediction

$$\text{hwage}_i = \hat{\alpha}_0 + \hat{\beta}_1 \text{Age} + \hat{\beta}_2 \text{Weight} + \hat{\beta}_3 \text{Height} + \hat{\beta}_4 \text{Male}$$

**Marginal effect of Male:**

$$\frac{\partial Y}{\partial \text{Male}} = \hat{\beta}_4$$

Since Male is binary,  $\hat{\beta}_4$  is the wage gap between males and females, holding other variables fixed.

**Marginal effect of Age:**

$$\frac{\partial Y}{\partial \text{Age}} = \hat{\beta}_1 = 0.546$$

One additional year  $\rightarrow$  \$0.55 higher wages. Same effect at age 25 as at age 40—the model forces it.

# Predicting $\hat{Y}$ : A Worked Example

```
. reg hwage age wgtadj hgtadj male
```

Source	SS	df	MS	Number of obs	=	129,755
Model	3092320.87	4	773080.217	F(4, 129750)	=	9.15
Residual	1.0964e+10	129,750	84497.8969	Prob > F	=	0.0000
Total	1.0967e+10	129,754	84519.1242	R-squared	=	0.0003
				Adj R-squared	=	0.0003
				Root MSE	=	290.69

hwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age	.5461922	.1002411	5.45	0.000	.3497214 .7426629
wgtadj	-.024785	.0241487	-1.03	0.305	-.0721159 .022546
hgtadj	.2585054	.3378081	0.77	0.444	-.4035924 .9206032
male	3.273136	2.362267	1.39	0.166	-1.356865 7.903137
_cons	-17.30314	21.17435	-0.82	0.414	-58.80449 24.1982

**Predicted wage** for a 40-year-old male, 150 lbs, 62 inches:

$$\hat{Y} = -17.303 + 0.546(40) + (-0.025)(150) + 0.259(62) + 3.273(1) = \$20.14$$

Plug in the values—the regression gives a prediction for any combination of  $X$ 's.

## Reference Categories

**Binary** = exactly two values

Male/female, treated/control

**Categorical** = multiple groups

Race, region, industry, religion

We create dummies for each group and **omit one**. Every coefficient is relative to that omitted group.

### Key Insight

The **reference category** is the group that Stata drops. If you include dummies for Northeast, South, and West, the Midwest becomes the baseline. All coefficients are interpreted relative to that omitted group.

# When Straight Lines Aren't Enough

---

## Does an Extra Inch of Height Always Have the Same Effect?

In a linear model, the answer is **yes**—by construction:

$$\text{wage} = \beta_0 + \beta_1 \text{height} + \dots \implies \frac{\partial \text{wage}}{\partial \text{height}} = \beta_1$$

### Question

But is that realistic? Being 5'1" vs. 5'2" might matter differently than being 6'1" vs. 6'2". We need a model where the slope *changes*.

## Three Non-Linear Specifications

$$(1) \text{ wage} = \beta_0 + \beta_1 \text{height} + \beta_2 \text{height}^2 + \beta_3 \text{weight} + \beta_4 \text{male}$$

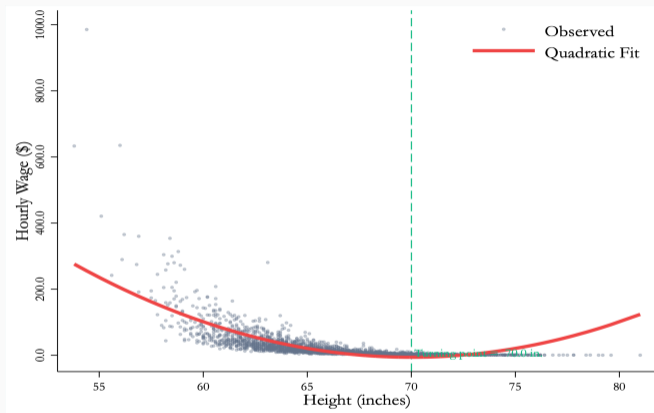
$$(2) \ln(\text{wage}) = \beta_0 + \beta_1 \text{education} + \beta_2 \text{age} + \beta_3 \text{female}$$

$$(3) \ln(\text{wage}) = \beta_0 + \beta_1 \text{male} + \beta_2 \ln(\text{educ}) + \beta_3 \text{male} \times \text{educ}$$

### Key Insight

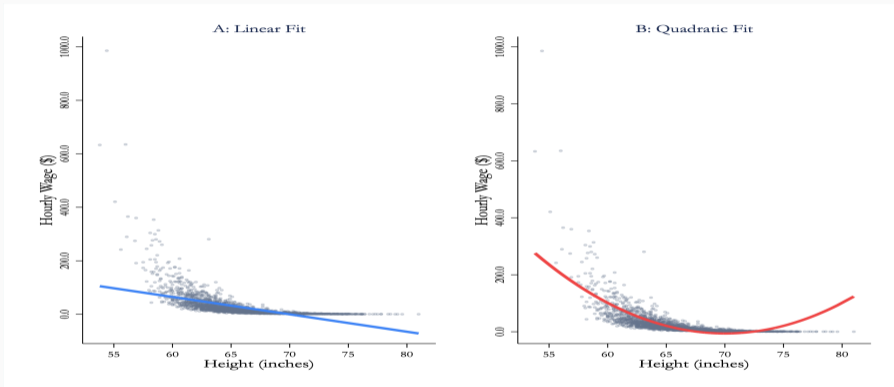
OLS requires linearity in **parameters** (the  $\beta$ 's), not in variables. Think of  $\text{height}^2$  as just another column in the data matrix. Everything still works.

# Height and Wages: The Data



Simulated data. A straight line misses the curvature—the quadratic fit captures the U-shape.

# Linear vs. Quadratic: See the Difference



The linear fit imposes a constant slope. The quadratic lets the relationship bend.

## $X^2$ : The Marginal Effect Changes

$$\text{wage} = \beta_0 + \beta_1 \text{height} + \beta_2 \text{height}^2 + \dots$$

Take the derivative:

$$\frac{\partial \text{wage}}{\partial \text{height}} = \beta_1 + 2\beta_2 \cdot \text{height}$$

**Worked example (A):**  $\text{wage} = -11 - 0.11 \text{ht} + 0.02 \text{ht}^2 + \dots$

At 60 inches:

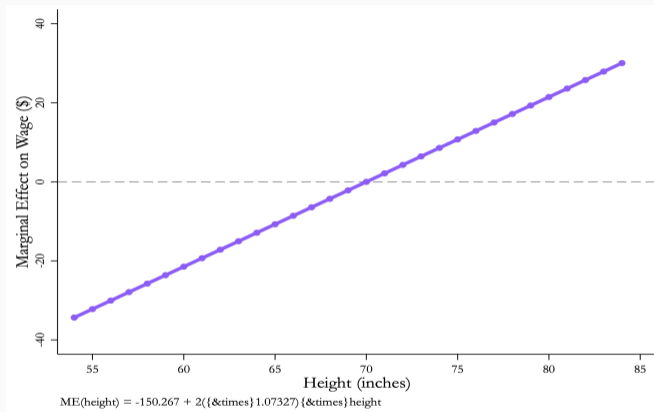
$$-0.11 + 0.04 \times 60 = \$2.29$$

At 72 inches:

$$-0.11 + 0.04 \times 72 = \$2.77$$

The marginal effect *depends on where you are*—that is the whole point of the quadratic.

# Marginal Effect of Height at Different Heights



The derivative plot shows where the effect is negative, zero (turning point), and positive.

## Worked Example (B): Height<sup>2</sup> and Wages (4-Step Process)

	Wages
Height	-2.1 (0.519)
Height <sup>2</sup>	0.04 (0.006)
<i>N</i>	100,254

### Warning

You cannot say “height increases wages” or “height decreases wages.” With a quadratic, the answer is: **it depends on the level of height**. Use the 4-step process next.

## Solution: Steps 1–2

**Step 1:** Write the marginal effect (derivative):

$$\frac{\partial \text{Wages}}{\partial \text{Height}} = -2.1 + 2(0.04) \cdot \text{Height} = -2.1 + 0.08 \cdot \text{Height}$$

**Step 2:** Find the turning point—set the derivative to zero:

### Warning

$$-2.1 + 0.08 \cdot \text{Height} = 0 \quad \implies \quad \text{Height} = \frac{2.1}{0.08} = 26.25 \text{ inches}$$

## Solution: Steps 3–4

**Step 3:** Minimum or maximum? Check the second derivative:

$$\frac{\partial^2 \text{Wages}}{\partial \text{Height}^2} = 0.08 > 0 \implies \text{Local minimum (U-shape)}$$

Positive second derivative  $\Rightarrow$  concave up  $\Rightarrow$  minimum. Negative  $\Rightarrow$  hill-shape (maximum).

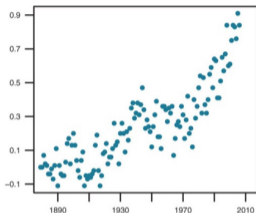
**Step 4:** Putting it together:

### Key Insight

Height **decreases** wages below 26.25 inches and **increases** wages above 26.25 inches. The relationship is U-shaped.

# Application: Climate Change and Polynomial Regression

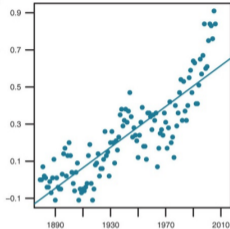
Temperature  
(deviation from  
average  
pre-industrial  
temperature,  
in Fahrenheit)



Year

(a)

Temperature  
(deviation from  
average  
pre-industrial  
temperature,  
in Fahrenheit)



Year

(b)

## Question

Temperature over time. Regressing on year and  $\text{year}^2$ , the derivative tells us the *rate of change*—and whether it is accelerating.

# Climate Change: The Regression

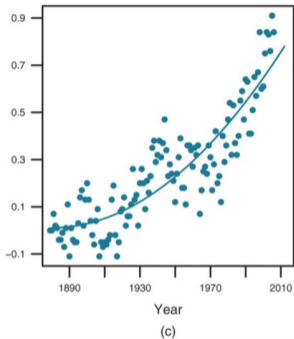


TABLE 7.1 Global Temperature, 1879–2012

	(a)	(b)
Year	0.006* (0.0003) [ $t = 18.74$ ]	-0.166* (0.031) [ $t = 5.31$ ]
Year <sup>2</sup>		0.000044* (0.000008) [ $t = 5.49$ ]
Constant	-10.46* (0.57) [ $t = 18.31$ ]	155.68* (30.27) [ $t = 5.14$ ]
$N$	128	128
$R^2$	0.73	0.78

Standard errors in parentheses.

\* indicates significance at  $p < 0.05$ , two-tailed.

## Policy Application

Has warming *accelerated*? The positive coefficient on Year<sup>2</sup> says yes—the marginal effect of time on temperature is increasing.

### Creating and estimating:

```
gen height_sq = height^2  
reg hwage height height_sq education female
```

### Computing marginal effects at specific values:

```
margins, dydx(height) at(height=(55(5)80))
```

The margins command computes  $\hat{\beta}_1 + 2\hat{\beta}_2 \times h$  at each value of  $h$  you specify. No manual calculation needed.

**Does the Effect Depend on Who  
You Are?**

---

# Does Education Pay Off the Same for Everyone?

Consider a regression of wages on education:

$$\text{wage} = \beta_0 + \beta_1 \text{education} + \beta_2 \text{female}$$

## Question

This model says the return to education is  $\beta_1$  for *everyone*—male or female, young or old. The Female dummy only shifts the intercept. Is that realistic?

Same logic: does the effect of bacteria on mold vary with humidity? Does the effect of exercise on health vary with age?

## Adding an Interaction Term

$$\text{Mold} = \beta_0 + \beta_1 \text{Bacteria} + \beta_2 \text{Humidity} + \beta_3 (\text{Humidity} \times \text{Bacteria})$$

Now the marginal effect of bacteria depends on humidity:

$$\frac{\partial \text{Mold}}{\partial \text{Bacteria}} = \beta_1 + \beta_3 \times \text{Humidity}$$

**What does  $\beta_1$  represent?**

The effect of bacteria when Humidity = 0.

**What does  $\beta_3$  represent?**

How much the bacteria effect *changes* per unit of humidity.

## Binary × Binary: Gender and Race

$$\text{Wage} = \beta_0 + \beta_1 \text{Female} + \beta_2 \text{Black}$$

$\beta_1$ : wage gap for females vs. males

$\beta_2$ : wage gap for Black vs. white

$\beta_0 + \beta_1$ : predicted wage for white females

$\beta_0 + \beta_1 + \beta_2$ : predicted wage for Black females

Now add the interaction:

$$\text{Wage} = \beta_0 + \beta_1 \text{Female} + \beta_2 \text{Black} + \beta_3 (\text{Female} \times \text{Black})$$

Now  $\beta_1$  is the female effect *for white individuals only*, and  $\beta_2$  is the Black effect *for males only*.

## Interpreting $\beta_3$ : The Interaction Coefficient

$$\text{Wage} = \beta_0 + \beta_1 \text{Female} + \beta_2 \text{Black} + \beta_3 (\text{Female} \times \text{Black})$$

- $\beta_3$ : the *additional* effect of being both female and Black, beyond what each contributes alone
- ME of race:  $\frac{\partial Y}{\partial \text{Black}} = \beta_2 + \beta_3 \cdot \text{Female}$
- ME of gender:  $\frac{\partial Y}{\partial \text{Female}} = \beta_1 + \beta_3 \cdot \text{Black}$

### Key Insight

$\beta_3$  tells us the race effect *varies by gender*, or equivalently, the gender effect *varies by race*. It's symmetric.

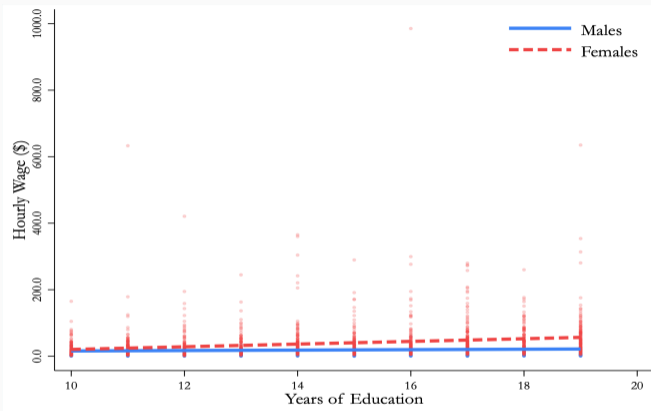
## The Interaction Matrix: A Systematic Approach

$$\text{Income} = \beta_0 + \beta_1 \text{Female} + \beta_2 \text{White} + \beta_3 (\text{White} \times \text{Female})$$

Predicted Income	Female	Male	Female – Male
White	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_0 + \beta_2$	$\beta_1 + \beta_3$
Non-White	$\beta_0 + \beta_1$	$\beta_0$	$\beta_1$
White – Non-White	$\beta_2 + \beta_3$	$\beta_2$	$\beta_3$

Bottom-right corner:  $\beta_3$  is the *difference-in-differences*—how the gender gap differs across racial groups.

# Education and Wages by Gender: Different Slopes



Without the interaction, both groups share one slope. With it, each group gets its own return to education.

## Sub-Sample Regression Equivalence

These three approaches produce the same fitted values:

(a)  $\text{Wage} = \alpha_0 + \alpha_1 \text{Educ}$  (females only)

(b)  $\text{Wage} = \gamma_0 + \gamma_1 \text{Educ}$  (males only)

(c)  $\text{Wage} = \beta_0 + \beta_1 \text{Fem} + \beta_2 \text{Educ} + \beta_3 (\text{Fem} \times \text{Educ})$

### Key Insight

The fully interacted model (c) is equivalent to running separate regressions by group—but lets you **test** whether slopes differ ( $H_0: \beta_3 = 0$ ) in a single step.

### Creating and estimating:

```
gen fem_educ = female * education
reg hwage education female fem_educ age weight
```

### Equivalent sub-sample approach:

```
reg hwage education if female == 0
reg hwage education if female == 1
```

### Testing whether the interaction matters:

```
reg hwage education female fem_educ
test fem_educ = 0
```

# Thinking in Percentages

---

## A \$1,000 Raise Means Different Things

### Question

A \$1,000 raise for someone earning \$20,000/year is a 5% increase—life-changing.

A \$1,000 raise for someone earning \$500,000/year is a 0.2% increase—barely noticed.

When levels vary enormously, we want to think in **percentages**, not dollars.

# What Is the Natural Log?

## Definition:

- $\ln(x) = \log_e(x)$ , where  $e \approx 2.718$
- Inverse of the exponential:  $e^{\ln(x)} = x$
- Key property:  $\ln(a) - \ln(b) \approx \frac{a-b}{b}$   
(for small changes, the log difference  $\approx$  percent change)

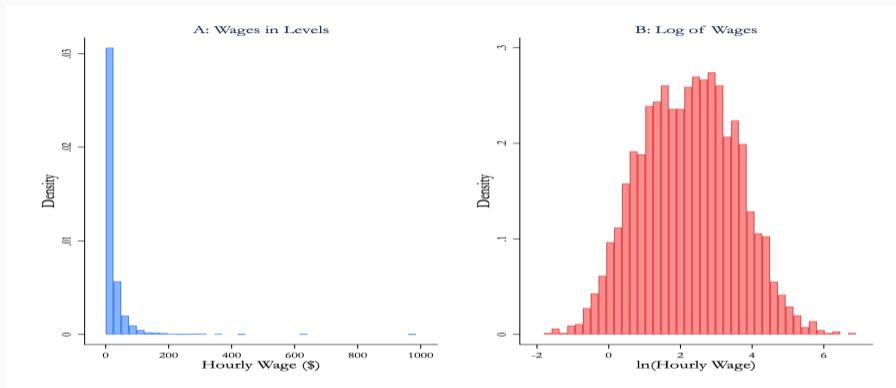
## Why use it?

Log-transformed coefficients have natural **percentage** interpretations.

## Key Insight

Euler's number  $e \approx 2.71828$  is  $\pi$ 's lesser-known cousin—just as fundamental. The natural log is the inverse of  $e^x$ .

# Why Log? Compare the Distributions



Wages in levels are right-skewed (panel A). The log transformation produces a much more symmetric distribution (panel B), which is better-suited for OLS.

# Log-Linear: ln on Y Only

$$\ln(\text{test scores}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots$$

**Interpretation:** A one-unit increase in  $X_1 \approx \beta_1 \times 100\%$  change in  $Y$ .

Exact:  $(e^{\beta_1} - 1) \times 100\%$

lgwrite	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
female	.114718	.0195341	5.87	0.000	.076194	.153242
read	.0066305	.0012689	5.23	0.000	.0041281	.0091329
math	.0076792	.0013873	5.54	0.000	.0049432	.0104152
intercept	3.135243	.0598109	52.42	0.000	3.017287	3.253198

$\log(\text{write}) = \beta_0 + \beta_1 * \text{female} + \beta_2 * \text{read} + \beta_3 * \text{math}$

## Question

Using this output, what is the effect of being female on writing scores? What about one more point of math?

- Writing scores are  $\approx 12\%$  higher for females than males
- A one-point increase in reading  $\rightarrow \approx 0.66\%$  increase in writing
- A one-point increase in math  $\rightarrow \approx 0.76\%$  increase

### What about a 10-point increase in math?

#### Approximation:

$$0.0077 \times 10 = 0.077 \rightarrow 7.7\%$$

#### Exact:

$$(e^{0.0077 \times 10} - 1) \times 100 = 7.98\%$$

The approximation works well for small  $\beta$ 's. For large changes, use the exact formula.

# Linear-Log: ln on X Only

write	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
female	5.388777	.9307948	5.79	0.000	3.555118	7.224436
lgnath	20.94097	3.430907	6.10	0.000	14.17473	27.7072
lgread	16.85218	3.063376	5.50	0.000	10.81076	22.89359
intercept	-99.16397	10.80406	-9.18	0.000	-120.4711	-77.85685

Written in equation, we have

$$\text{write} = \beta_0 + \beta_1 \cdot \text{female} + \beta_2 \cdot \text{lgnath} + \beta_3 \cdot \text{lgread}$$

**Interpretation:** A 1% increase in  $X \rightarrow$  a  $\beta/100$  unit change in  $Y$ .

**Approximation:**

A 1% increase in reading  $\rightarrow$  0.16 pt  
increase in writing

**Exact formula:**

$$\beta \cdot \ln\left(1 + \frac{\%}{100}\right)$$

10% increase:  $16.85 \times \ln(1.10) = 1.61$

# Log-Log: ln on Both X and Y

lgwrite	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
female	.1142399	.0194712	5.87	0.000	.07584 .1526399
lgnath	.4085269	.0720791	5.67	0.000	.2663866 .5506872
read	.0066086	.0012561	5.26	0.000	.0041113 .0090859
intercept	1.928101	.2469391	7.81	0.000	1.441102 2.415099

Written as an equation, we can describe the model:

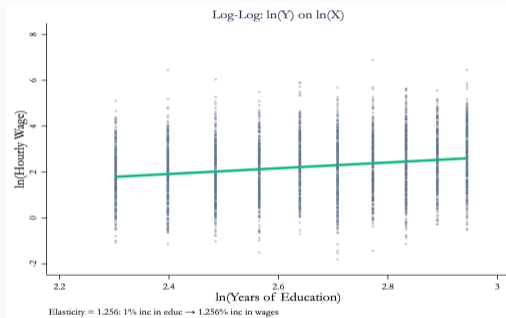
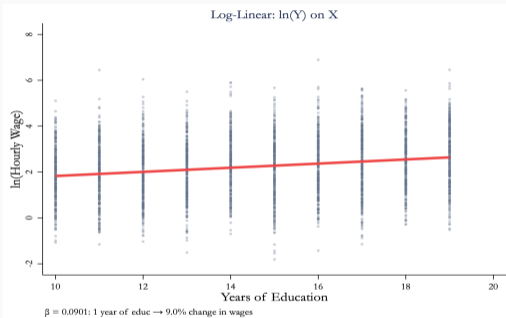
$$\log(\text{write}) = \beta_0 + \beta_1 \text{female} + \beta_2 \log(\text{math}) + \beta_3 \text{read}$$

**Interpretation:** A 1% increase in  $X \rightarrow$  a  $\beta\%$  change in  $Y$ . This is an **elasticity**.

- A 1% increase in math  $\rightarrow \approx 0.41\%$  increase in writing
- A 10% increase in math  $\rightarrow \approx 4.1\%$  increase in writing

Exact:  $((1 + \%/100)^\beta - 1) \times 100$ . For 10%:  $(1.10^{0.409} - 1) \times 100 = 3.97\%$ .

# Visualizing Log Transformations: Our Data



Left: log-linear ( $\ln$  of wage on years of education). Right: log-log (both logged). The annotations show the interpretation directly.

## The 2×2 Summary Table (Memorize This)

	<b>Y (level)</b>	<b>ln(Y)</b>
<b>X (level)</b>	$\beta = \text{unit change in } Y \text{ per unit change in } X$	$\beta \times 100 = \% \text{ change in } Y \text{ per unit change in } X$
<b>ln(X)</b>	$\beta/100 = \text{unit change in } Y \text{ per } 1\% \text{ change in } X$	$\beta = \% \text{ change in } Y \text{ per } 1\% \text{ change in } X \text{ (elasticity)}$

This is the single most important reference for log interpretation. Mnemonic: log on  $Y \rightarrow$  multiply  $\beta$  by 100; log on  $X \rightarrow$  divide  $\beta$  by 100; both logged  $\rightarrow \beta$  is the elasticity directly.

## The “100% Change” Intuition

### With $X$ (levels):

A unit increase is the same absolute amount everywhere.

### With $\ln(X)$ :

A percentage increase is the same *proportional* amount everywhere.

Going from 20 to 40 is the same log difference as going from 40 to 80—both are 100% increases.

```
. display ln(0.1*40) - ln(0.1*20)  
.69314718
```

```
. display ln(0.1*80) - ln(0.1*40)  
.69314718
```

# Exercise: Combining Logs and Quadratics

```
. reg lnwage female black onw age agesq, robust
```

Linear regression		Number of obs	=	156,255
		F(5, 156249)	=	6760.97
		Prob > F	=	0.0000
		R-squared	=	0.1713
		Root MSE	=	.7339

lnwage	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]
female	-.3029419	.0037289	-81.24	0.000	-.3102503 -.2956334
black	-.1814776	.0059094	-30.71	0.000	-.1930599 -.1698953
onw	.013633	.0070694	1.93	0.054	-.0002229 .0274889
age	.1239595	.0010714	115.70	0.000	.1218596 .1260595
agesq	-.0013023	.0000131	-99.78	0.000	-.0013278 -.0012767
_cons	3.99952	.0207002	193.21	0.000	3.958948 4.040092

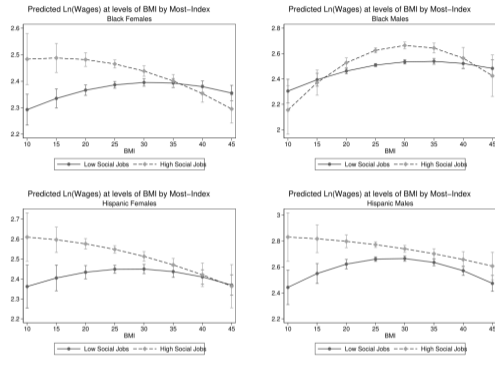
## Question

What is the marginal effect of age on weekly earnings for a 40-year-old? Show every step.

Hint:  $Y$  is in logs and the model includes both age and  $\text{age}^2$ .

# Application: BMI and Wages by Job Type

Figure 4: Relationship between BMI and wages across race and gender groups



The BMI–wage relationship differs across job types. An interaction between BMI and job category captures this heterogeneity.

## Summary

---

## What We Learned Today

1.  **$X^2$  terms** make the marginal effect depend on the level of  $X$ —use derivatives
2. **Interaction terms** let the effect of one variable vary by another
3. **Binary  $\times$  binary** interactions: fill in the  $2 \times 2$  matrix
4. **Fully interacted models** = sub-sample regressions, but allow formal testing
5. **Log-linear**:  $\beta \times 100 = \% \text{ change in } Y \text{ per unit change in } X$
6. **Linear-log**:  $\beta/100 = \text{unit change in } Y \text{ per } 1\% \text{ change in } X$
7. **Log-log**:  $\beta = \text{elasticity } (1\% \rightarrow \beta\%)$

### Policy Application

**Next class:** We apply these tools to real-world policy questions and introduce additional regression diagnostics.