Week 10

Questions of Description

Paul Laskowski and Alex Hughes

January 12, 2023

UC Berkeley, School of Information

Questions of Description

Questions of Description

What is the shape of the relationship between a country's economic output and Internet access?

Questions of Description

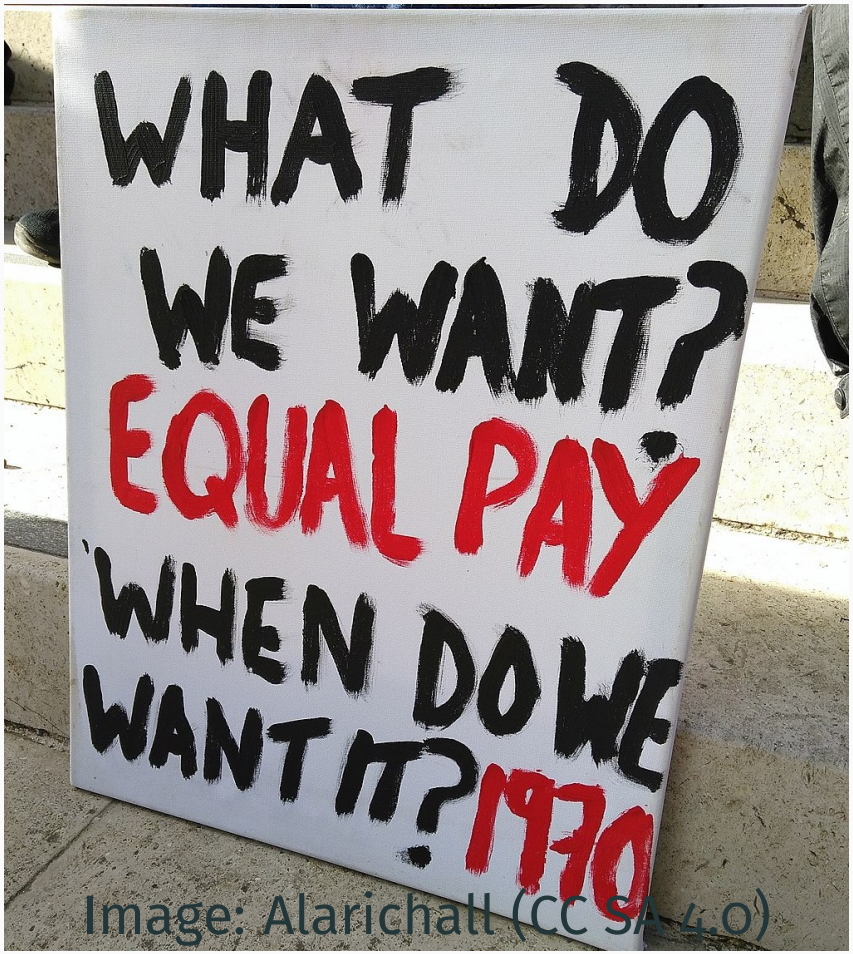


Image: Alarichall (CC SA 4.0)

How big is the pay gap in the United States?

Questions of Description

How does the pay gap depend on the age of the worker?

Description

Plan for the Week

Preamble

- Three modes of model building

Three sections about descriptive modeling

1. Capturing nonlinear relationships
2. Measurement with controls
3. Modeling conditional effects

Plan for the Week (cont.)

At the end of this week, you will be able to:

- Understand how three major modes of model building lead to very different models
- Balance design goals when creating a model for description
- Plan a set of model specifications for a regression table

What Is Linear Model Building?

What Is Linear?

Model	Linear?
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$	
$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$	
$\ln Y = \beta_0 + \beta_1 \ln X + \epsilon$	
$Y = \beta_0 X^{\beta_1} + \epsilon$	

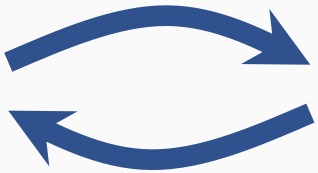
What Is Linear Model Building?

How do you select from all the possible linear models?

- Which variables to include, which to exclude
- Whether and how to transform each variable
- Whether to create new variables
- Whether to multiply variables together

The Model-Building Process

Propose Model



Evaluate Model

- Modeling goals
- Observations from data
- Background knowledge
- Background theory

- Diagnostic plots
- Measures of fit
- Tests of model assumptions

Modes of Model Building

An Important Question to Keep in Mind:

What are your goals?

Three Modes of Model Building

1. Predictive modeling: What value can we expect for data we haven't seen?
2. Descriptive modeling: How can we make sense of the patterns in data?
3. Explanatory modeling: How can we measure effects inside a causal theory?

Modes of Model Building

Predictive Modeling

Predictive Modeling

Prediction: making guesses for unknown values

- For new people, new pictures of cats, or other units
- For future time periods

A key focus of machine learning and time series analysis

Key Goals of Predictive Modeling

1. Accurately predict values
2. Be interpretable by humans (usually less important)

Implications for linear regression

- Hundreds of variables, or more
- Variable selection by algorithm
    - False discovery: Out of many variables, some will look important by chance.
- Coefficients usually don't have meaning

Modes of Model Building

Descriptive Modeling

Descriptive Modeling

Description: summarizing or representing data in a compact, human-understandable way

- Gain understanding by interpreting the model's internal structure
- Popular in statistics, but "not commonly used for theory building and testing in other disciplines"

—Shmueli, Galit. *To Explain or to Predict?*

Key Goals of Descriptive Modeling

1. Measure complex concepts
2. Capture features of the world
3. Simplify phenomena to make them understandable
4. Highlight associations
5. Generate hypotheses

Implications for linear regression

- Parsimonious models
- Use of logarithms and other transforms with clear interpretation
- Long build process relying on EDA

Modes of Model Building

Explanatory Modeling

Explanatory Modeling

Explanation: using data to test or estimate parameters in a causal theory

- Explanation lets us reason about actions
- Common mode in economics, political science, epidemiology, psychology, environmental science...
- Causal assumptions are required to make causal claims

Key Goals of Explanatory Modeling

1. Measure a causal effect.
2. Evaluate a theory.
3. Predict the consequences of potential actions.

Implications for linear regression

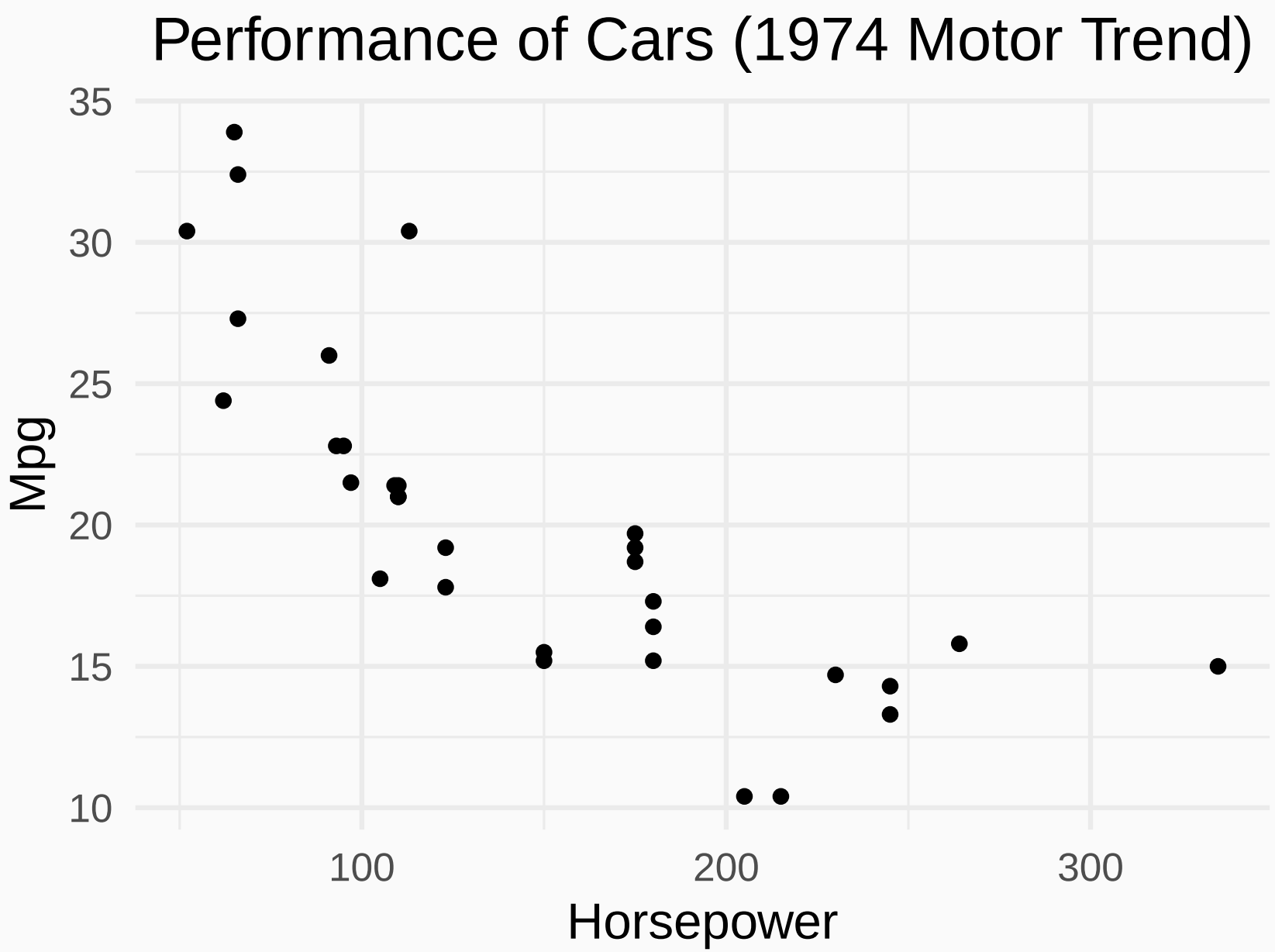
- Variable selection is guided by causal theory
- Models oriented to perform a specific measurement
- Key challenge is the operationalization gap between theoretical constructs and variables

Three Modes of Model Building

1. Predictive modeling: What value can we expect for data we haven't seen?
2. Descriptive modeling: How can we make sense of the patterns in data?
3. Explanatory modeling: How can we measure effects inside a causal theory?

Introduction to Descriptive Modeling

How Would You Describe These Data?



Introduction to Descriptive Modeling

Description: summarizing or representing data in a compact, human-understandable way

Linear regression is a tool that can help us answer:

- What patterns exist in a dataset?
- What is the shape of a specific relationship?
- What is the size of a feature or effect?

What Does It Take to Build a Descriptive Model?

Skill, art, and a lot of iteration

- No causal theory to constrain model choice
- Often many ways to operationalize concepts
- Need to balance many competing goals
- Takes time to build intuition in many dimensions

Logarithms

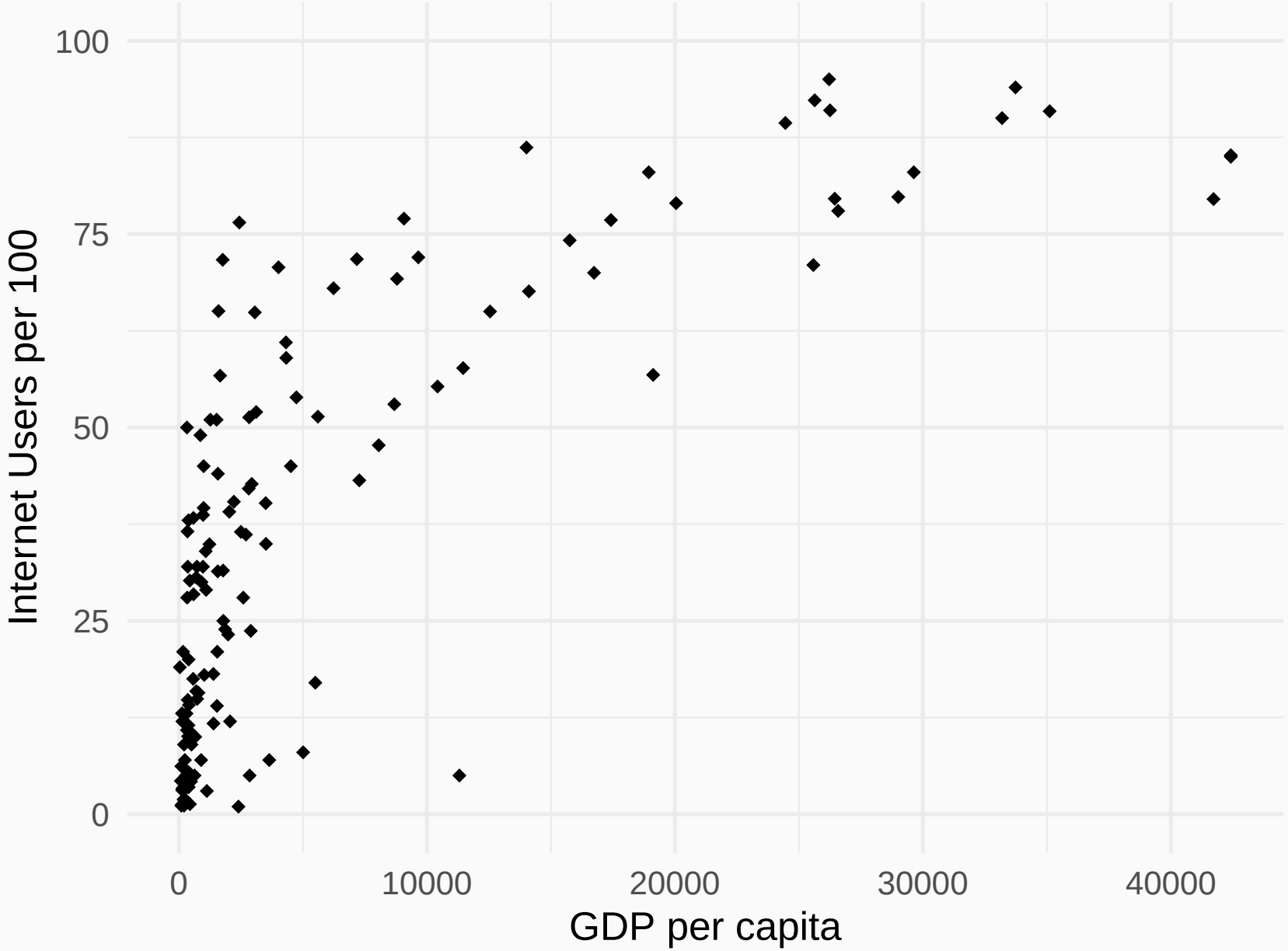
A Question of Development

How does the economic productivity of a country relate to Internet access?

Data taken from the World Bank

- GDP per capita
- Internet users per 100

Is OLS Regression Appropriate?



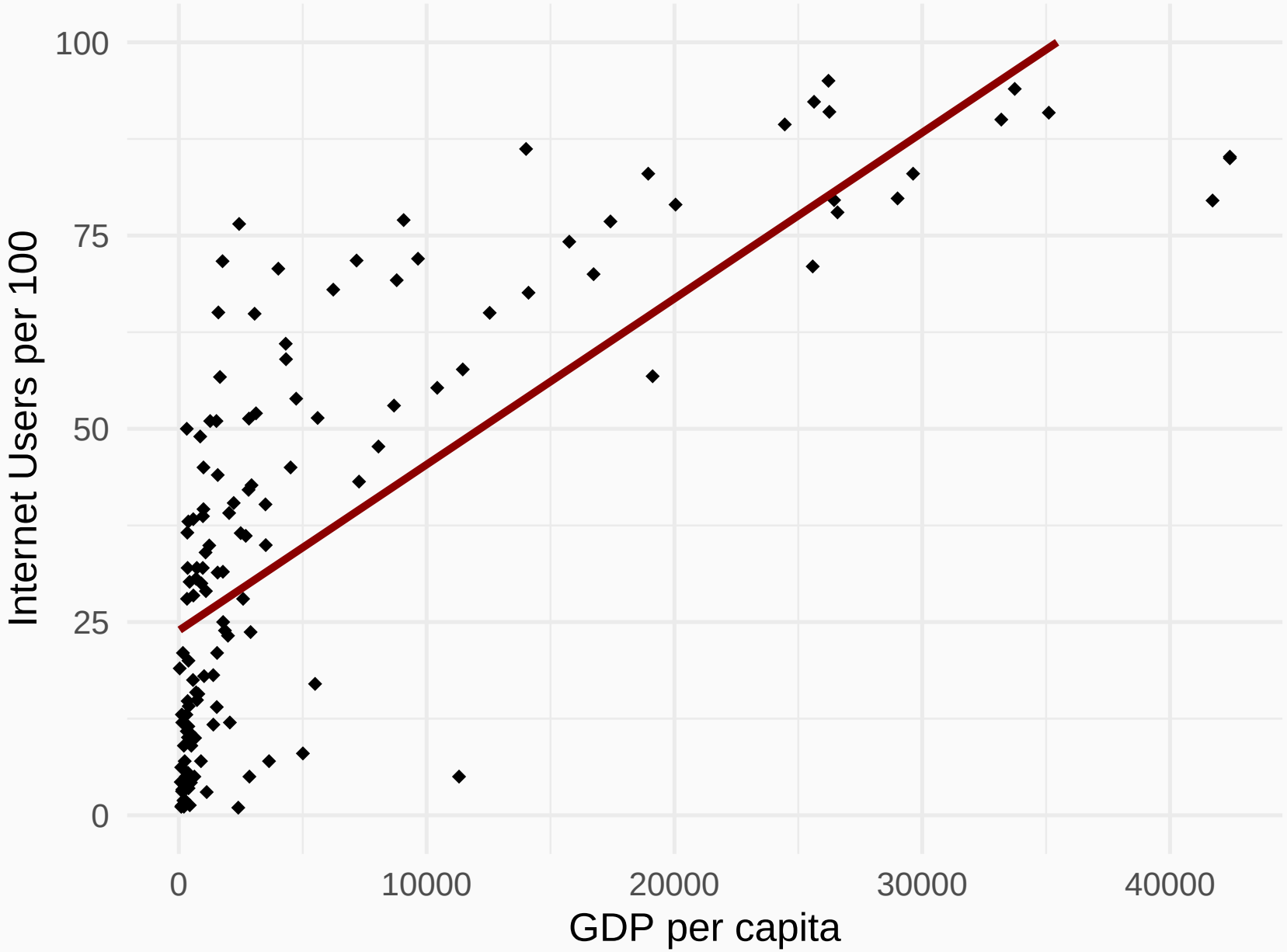
Interpreting the Level-Level Model

$$\widehat{users} = 2.4 + .0021\ GDP$$

Interpretation 1: A country with an extra $1 of GDP per capita is predicted to have another .0021 Internet users per 100.

Interpretation 2: A country with an extra $1,000 of GDP per capita is predicted to have another 2.1 Internet users per 100.

Does OLS Regression Meet Our Goals?



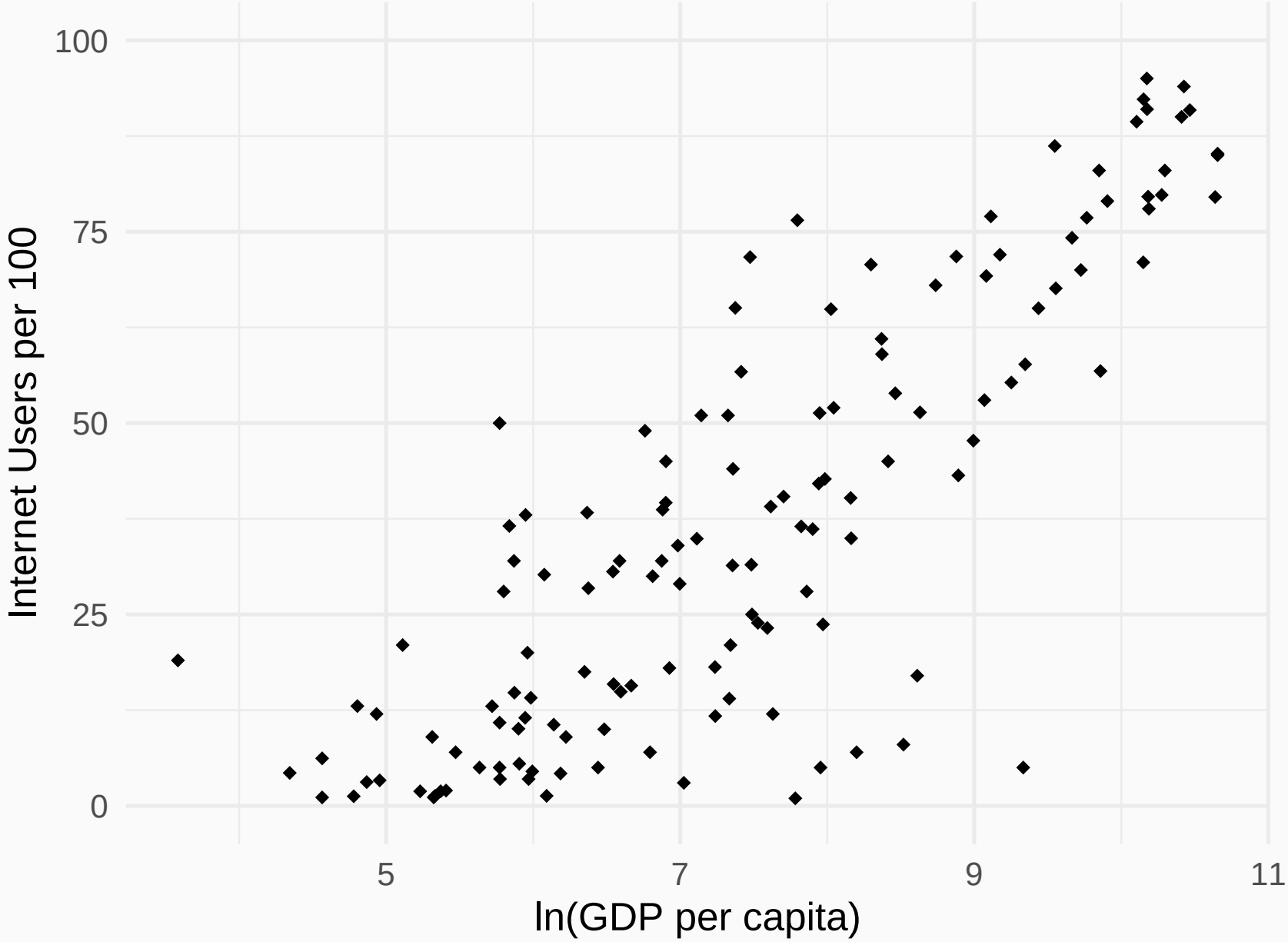
Variable Transformation

Variable transformation

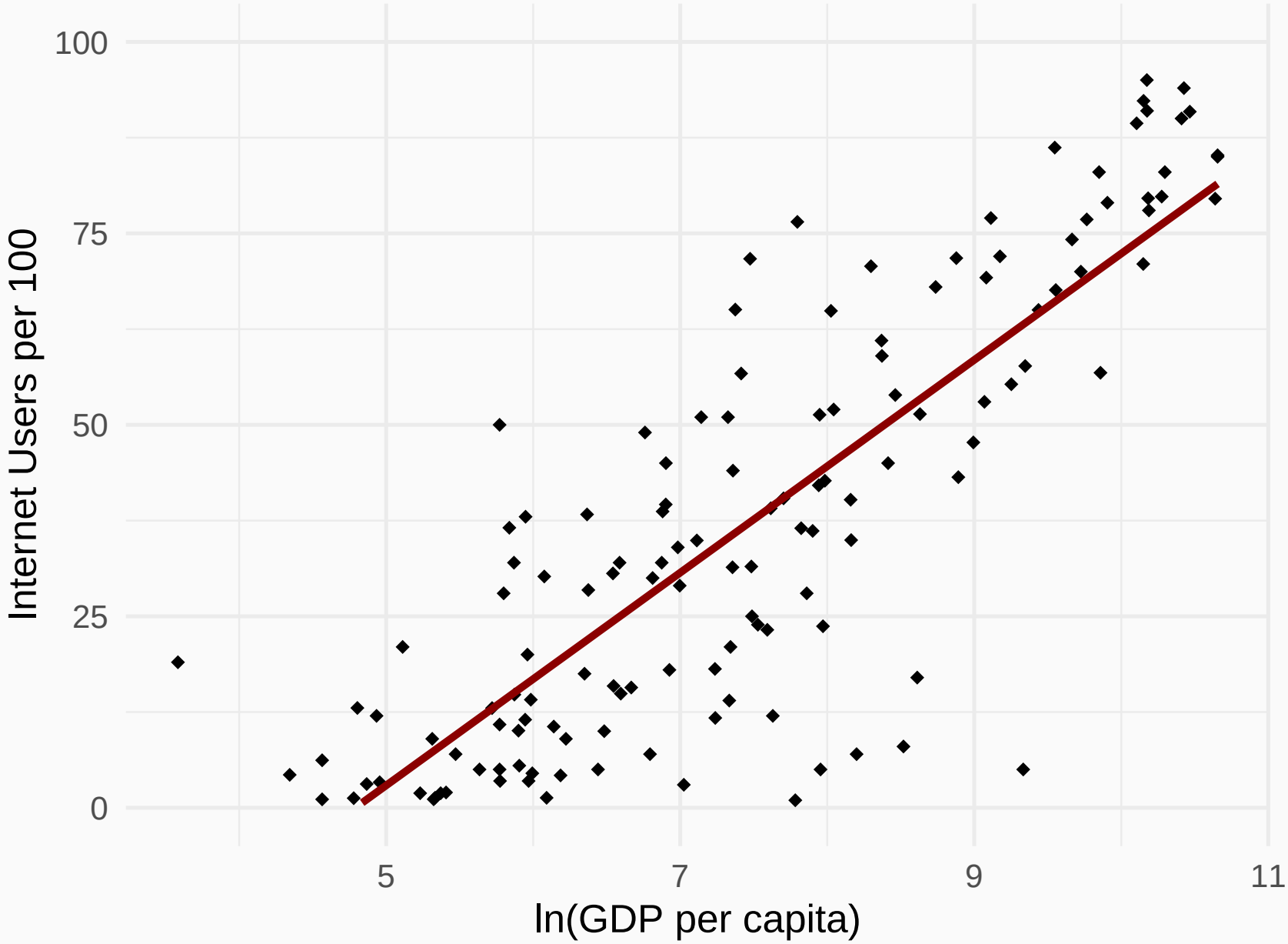
Replace X with $f(X)$ for some $f : \mathbb{R} \to \mathbb{R}$

- $\ln(X)$
- $\log_{10}(X)$
- Indicator functions
- X^a
- Polynomials(X)

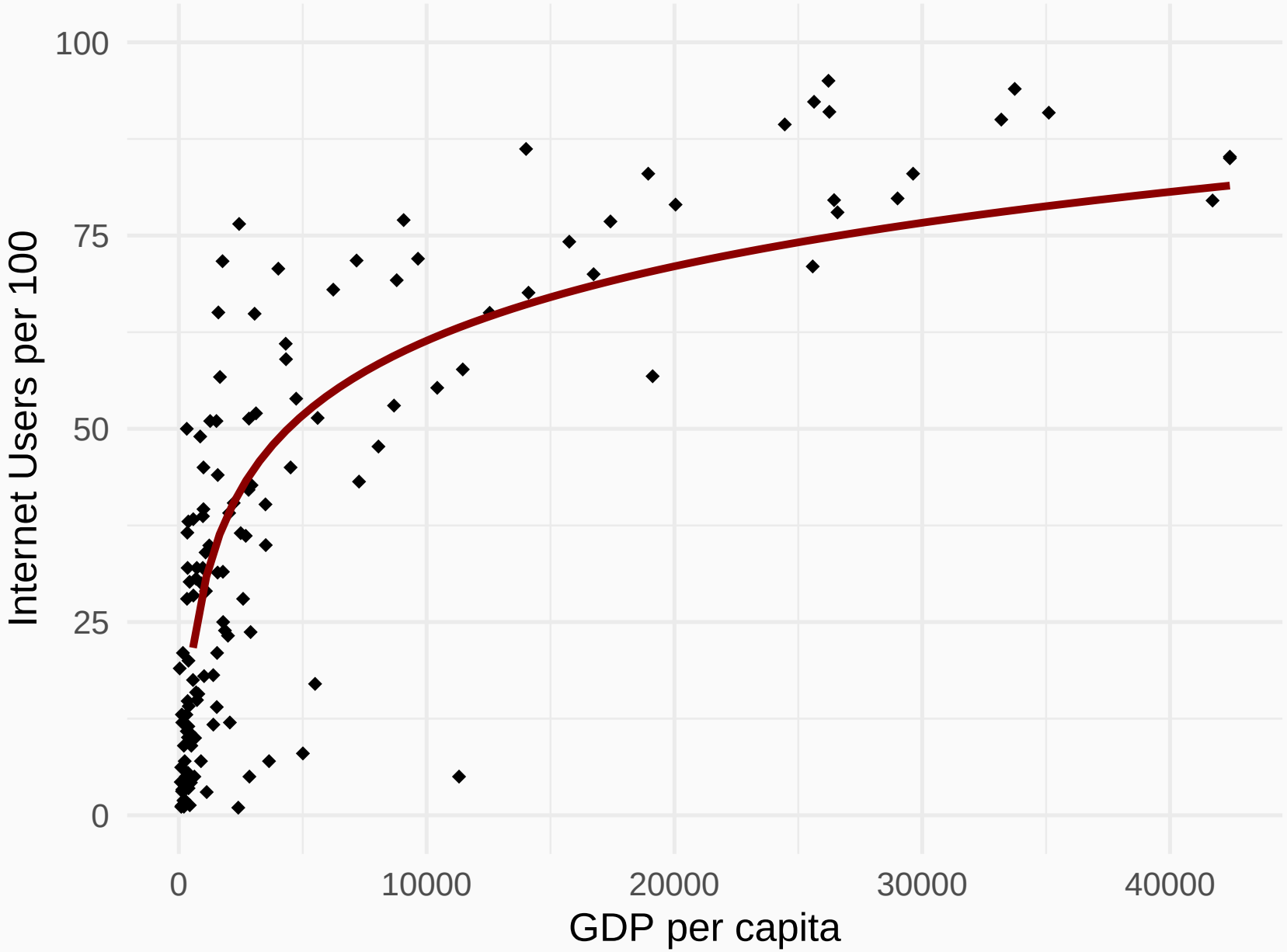
Taking the Log of GDP



Taking the Log of GDP



Model: $\widehat{users} = \beta_0 + \beta_1 \ln GDP$ $R^2 = .68$



Interpreting and Applying Logarithms

Interpreting Logarithms

Base 10 log

$$\widehat{users} = -66.5 + 32.0 \log_{10} GDP$$

Interpretation: adding a 0 to GDP associated with 32 more Internet users per 100

Base e log

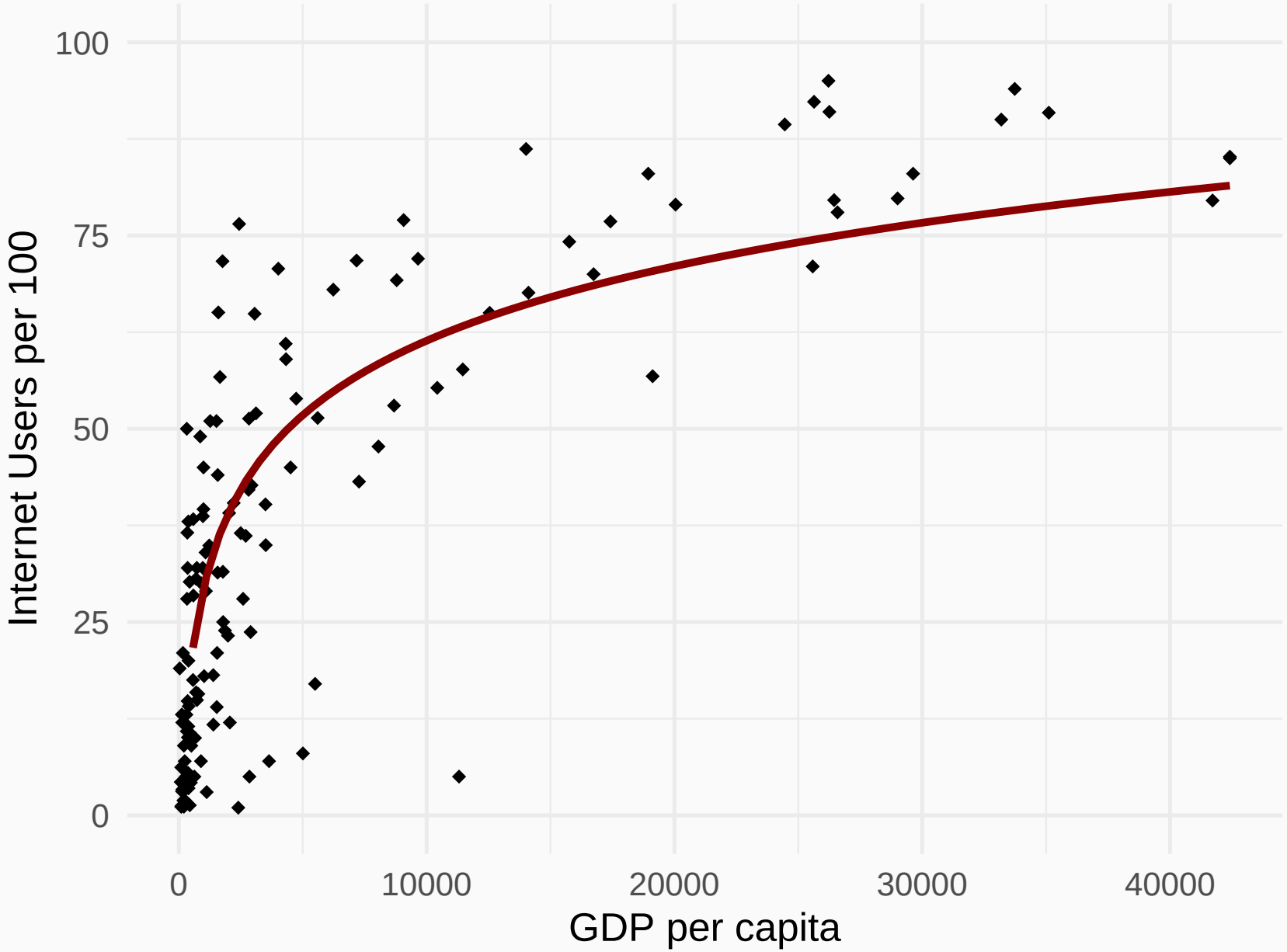
$$\widehat{users} = -66.5 + 13.9 \ln GDP$$

Interpretation: multiplying GDP by a small $1 + \alpha$ associated with 13.9α more Internet users per 100

Interpreting Logarithms (cont.)

$$\frac{\partial \widehat{users}}{\partial GDP} = \frac{\partial}{\partial GDP}[-66.5 + 13.9 \ln GDP]$$

Model: $\widehat{users} = \beta_0 + \beta_1 \ln GDP$ $R^2 = .68$



Other Uses of Logarithms

Log-linear form: $\ln Y = \beta_0 + \beta_1 X$

- Interpretation: $\frac{\Delta Y}{Y} \approx \beta_1 \Delta X$
- Example: add 0.1 to $X \rightarrow$ add $0.1\beta_1 \cdot Y$ to Y

Log-log form: $\ln Y = \beta_0 + \beta_1 \ln X$

- Constant elasticity model
- Interpretation: $\frac{\Delta Y}{Y} \approx \beta_1 \frac{\Delta X}{X}$
- Example: increase X by 1% $\rightarrow$ increase Y by β_1%

When to Apply Logs

What makes a variable a good candidate for a log transform?

- Always positive
    - Don’t add a constant to make variable positive
- Spans multiple orders of magnitude
- Clustered near zero with high outliers
- Percent changes are meaningful

Power Transformations

Power transformation: Replace X with X^α for some $\alpha \in \mathbb{R}$.

- $\sqrt{X}$ behaves similarly to $\ln X$.
- X^2, X^3,... spreads values far from zero further apart.

Interpretation is more difficult than with logs.

When to Apply Transformations

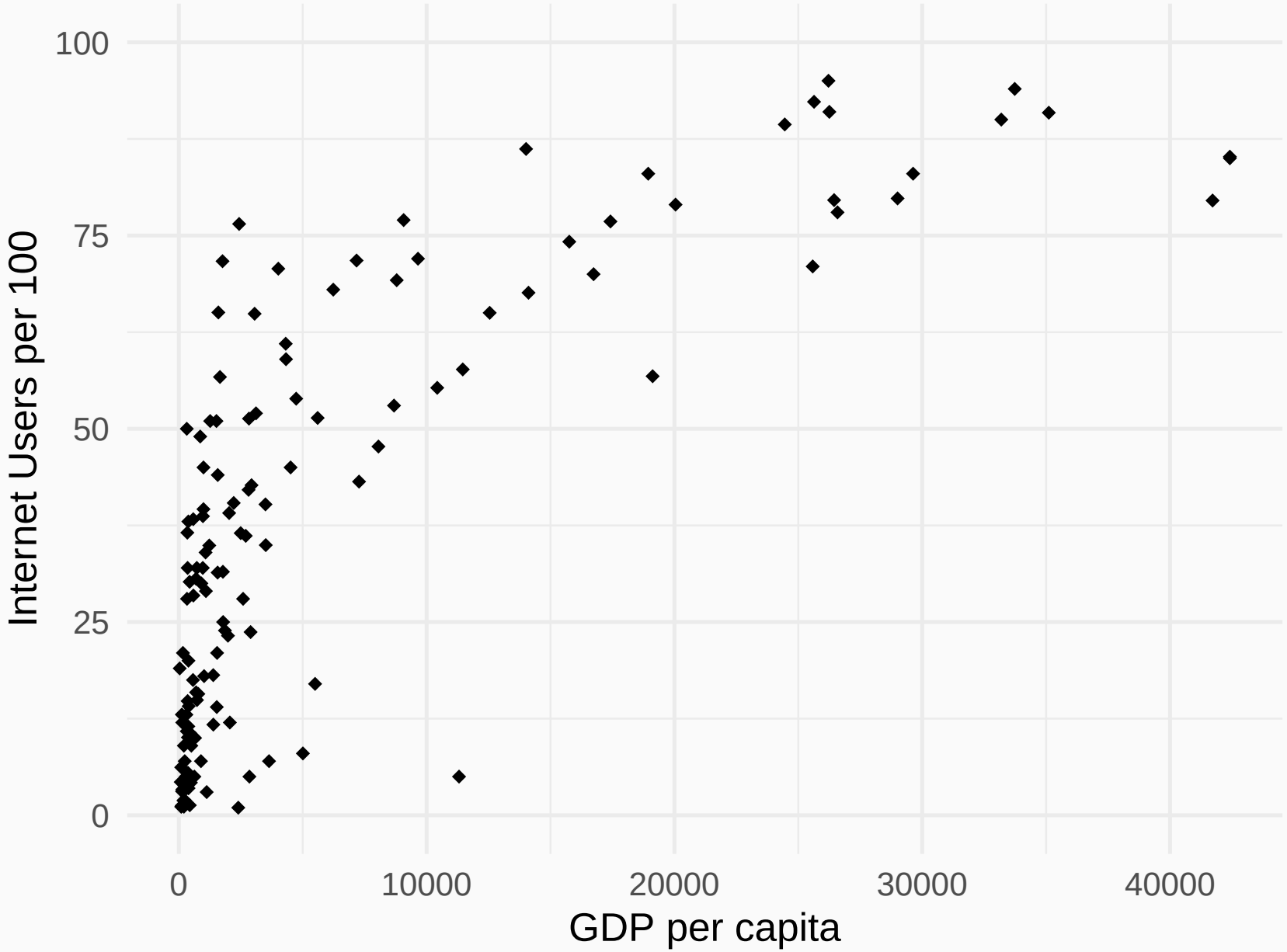
Question: Should you always transform your variables to make them normal?

- Normal variables are **never** a requirement for OLS regression.
  - Even in the classical linear model, *errors* are normal, not variables.

$\Longrightarrow$ Don't focus too much on normality. Capturing relationships is more important.

Polynomials

How Can We Improve Predictive Accuracy?



Polynomial Regression

Motivation: Polynomials can approximate any continuous function (e.g. Weierstrass theorem)

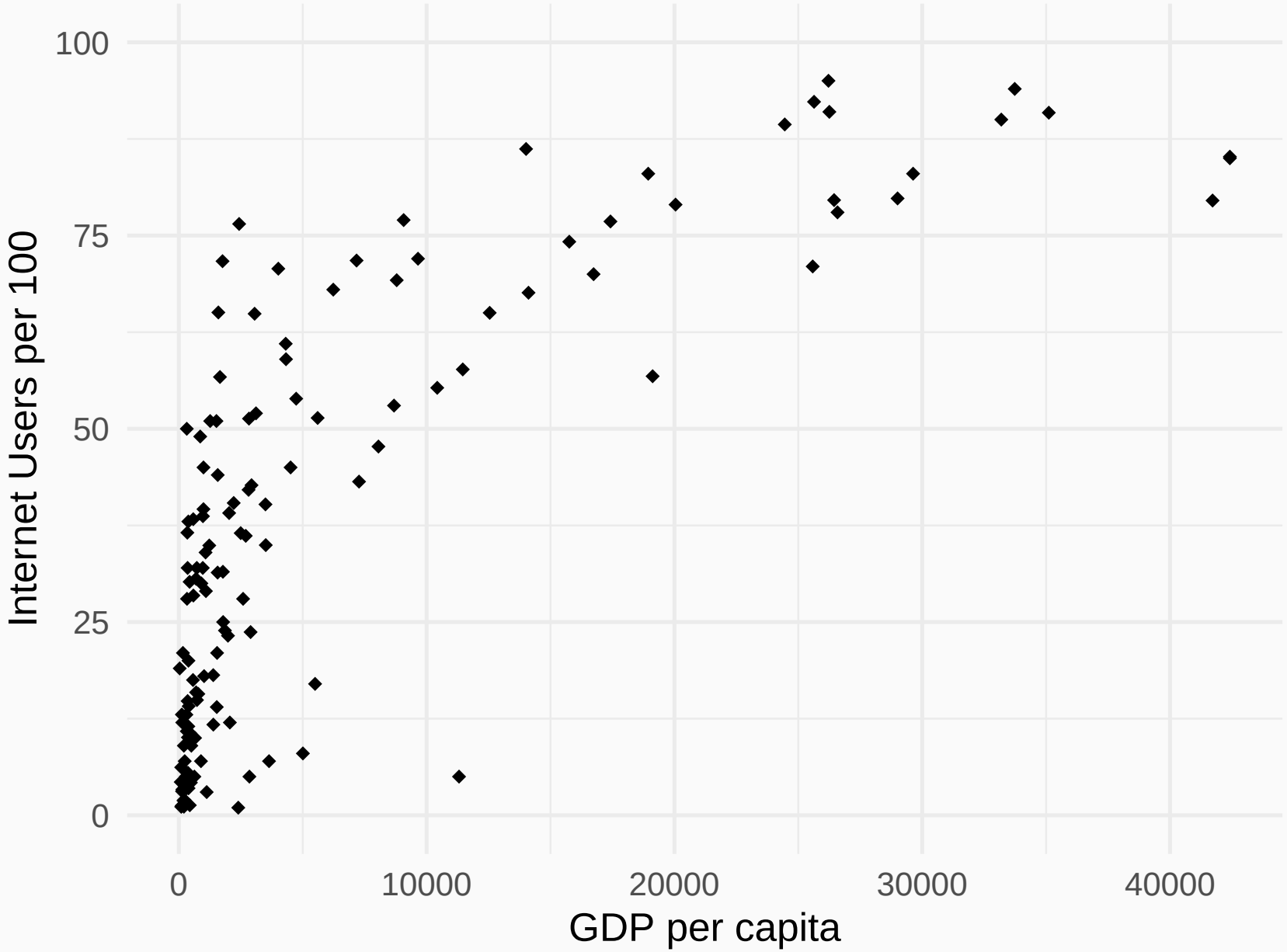
Quadratic: $\widehat{Users} = \beta_0 + \beta_1 GDP + \beta_2 GDP^2$

- OLS finds the parabola that minimizes MSE.

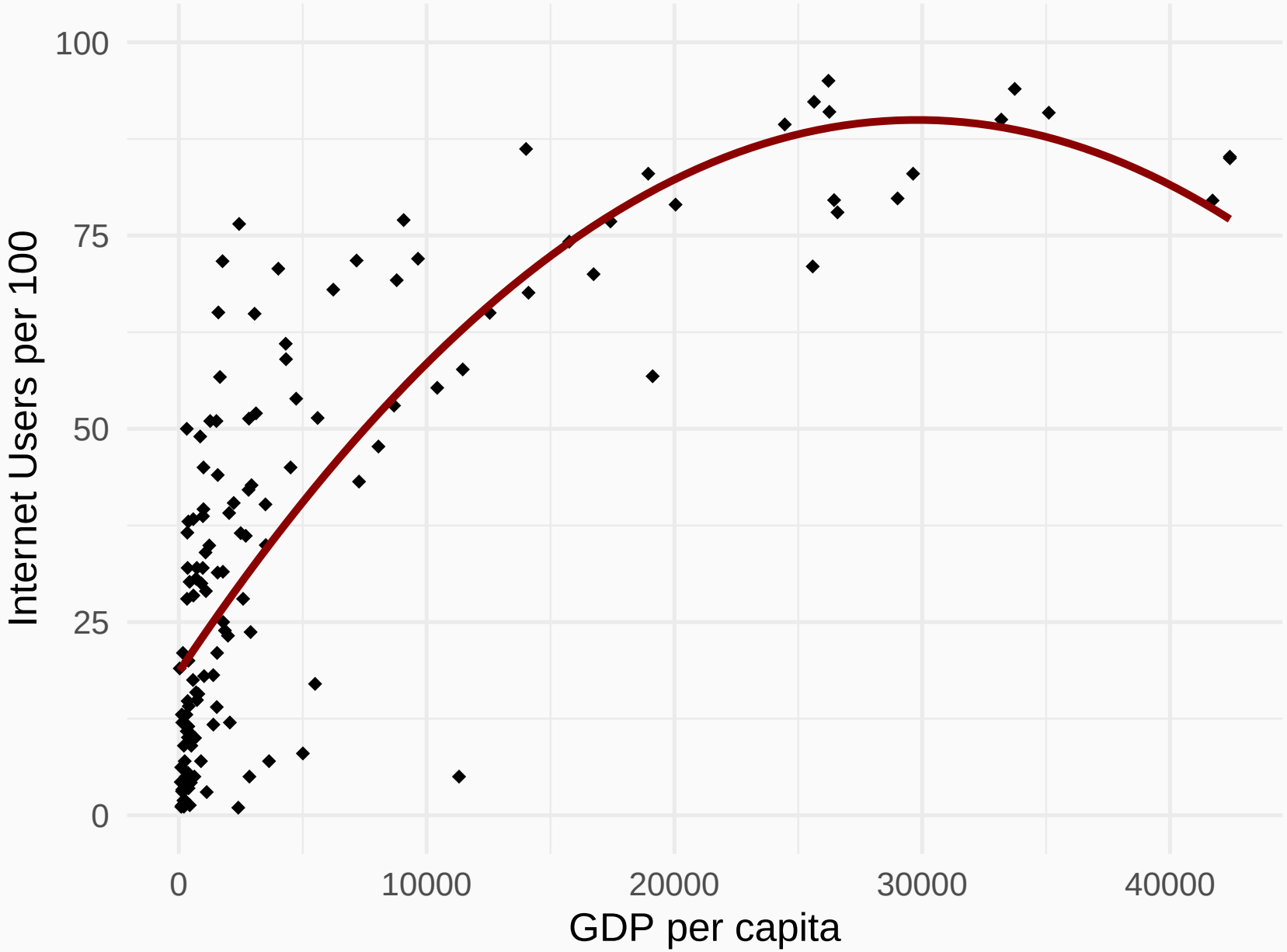
Cubic: $\widehat{Users} = \beta_0 + \beta_1 GDP + \beta_2 GDP^2 + \beta_3 GDP^3$

- OLS finds the cubic function that minimizes MSE.

A Parabolic Pattern?



Model: $\widehat{users} = \beta_0 + \beta_1 GDP + \beta_2 GDP^2$ $R^2 = .66$



Interpreting the Quadratic Specification

$$\widehat{Users} = 18.5 + .0048\ GDP - 8.0 \cdot 10^{-8}\ GDP^2$$

$$\frac{\widehat{\partial Users}}{\partial GDP} = .0048 - 1.6 \cdot 10^{-7}\ GDP$$

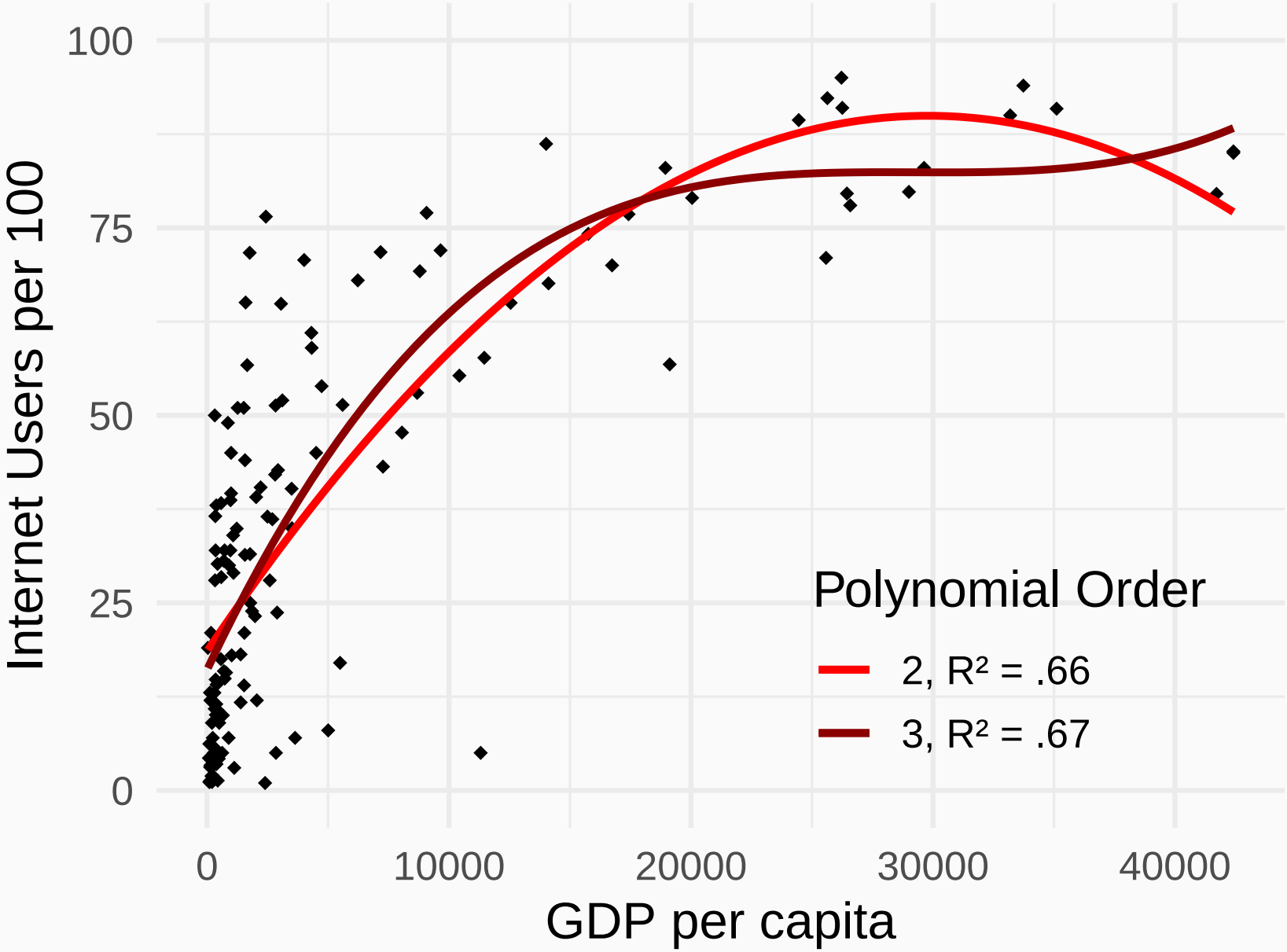
Example 1: If $GDP = 10,000$, then $\frac{\widehat{\partial Users}}{\partial GDP} = .0032$.

(Extra \$1,000 in GDP associated with 3.2 extra users)

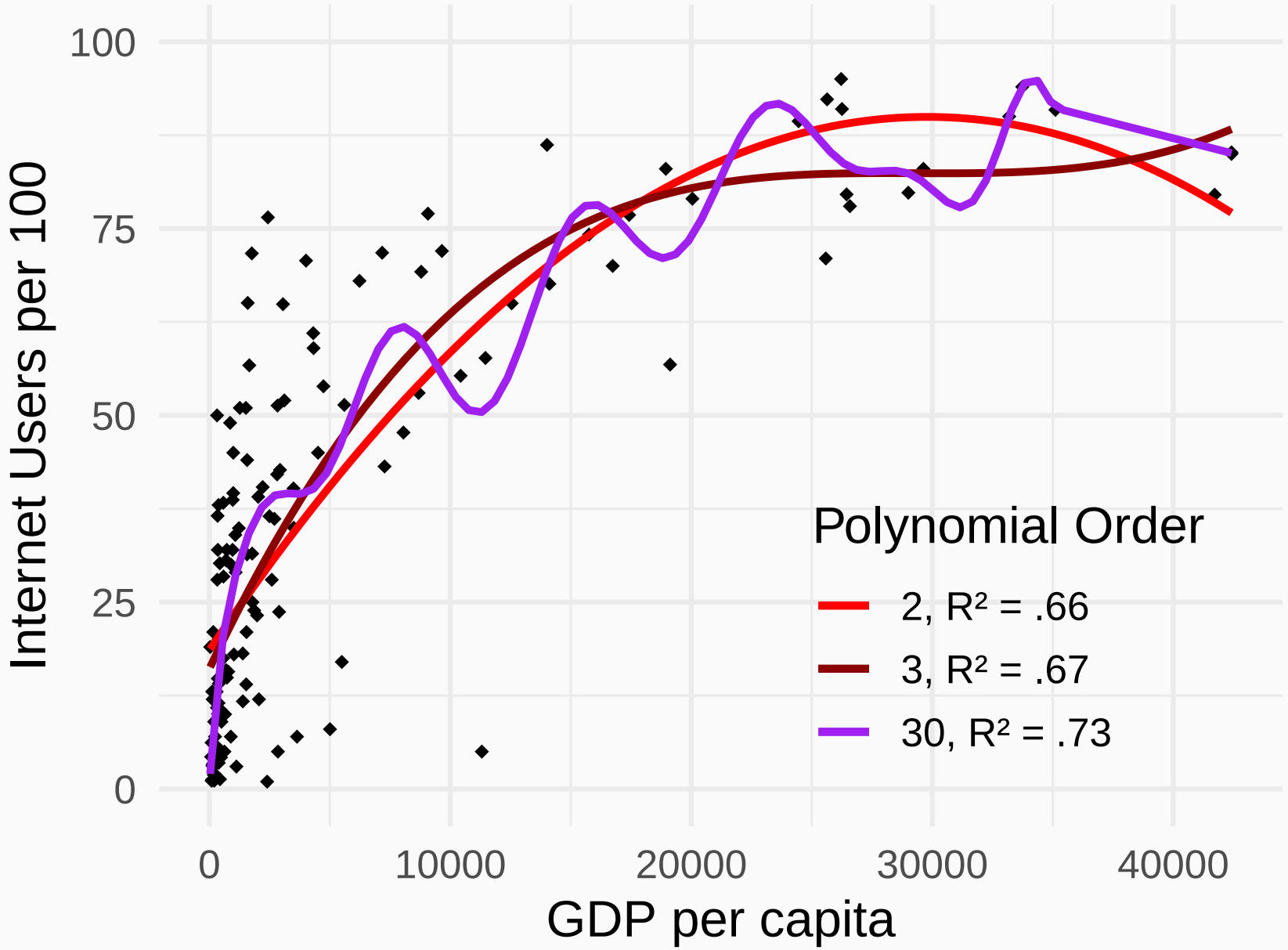
Example 2: If $GDP = 20,000$, then $\frac{\widehat{\partial Users}}{\partial GDP} = .0016$.

(Extra \$1,000 in GDP associated with 1.6 extra users)

Higher-Order Polynomials



Higher-Order Polynomials



Measurement With Controls

Measurement With Controls

Note: This is a placeholder slide for an introduction that we will provide to the section. We're just placing it here for organization.

Interpreting Indicator Variables

Example: Measuring the Wage Gap

78 cents on the dollar: The facts about the gender wage gap

The gender pay gap is even worse if you're a woman with a college degree

US women made economic strides in 2018, but pay gap persists

Example: Measuring the Wage Gap (cont.)

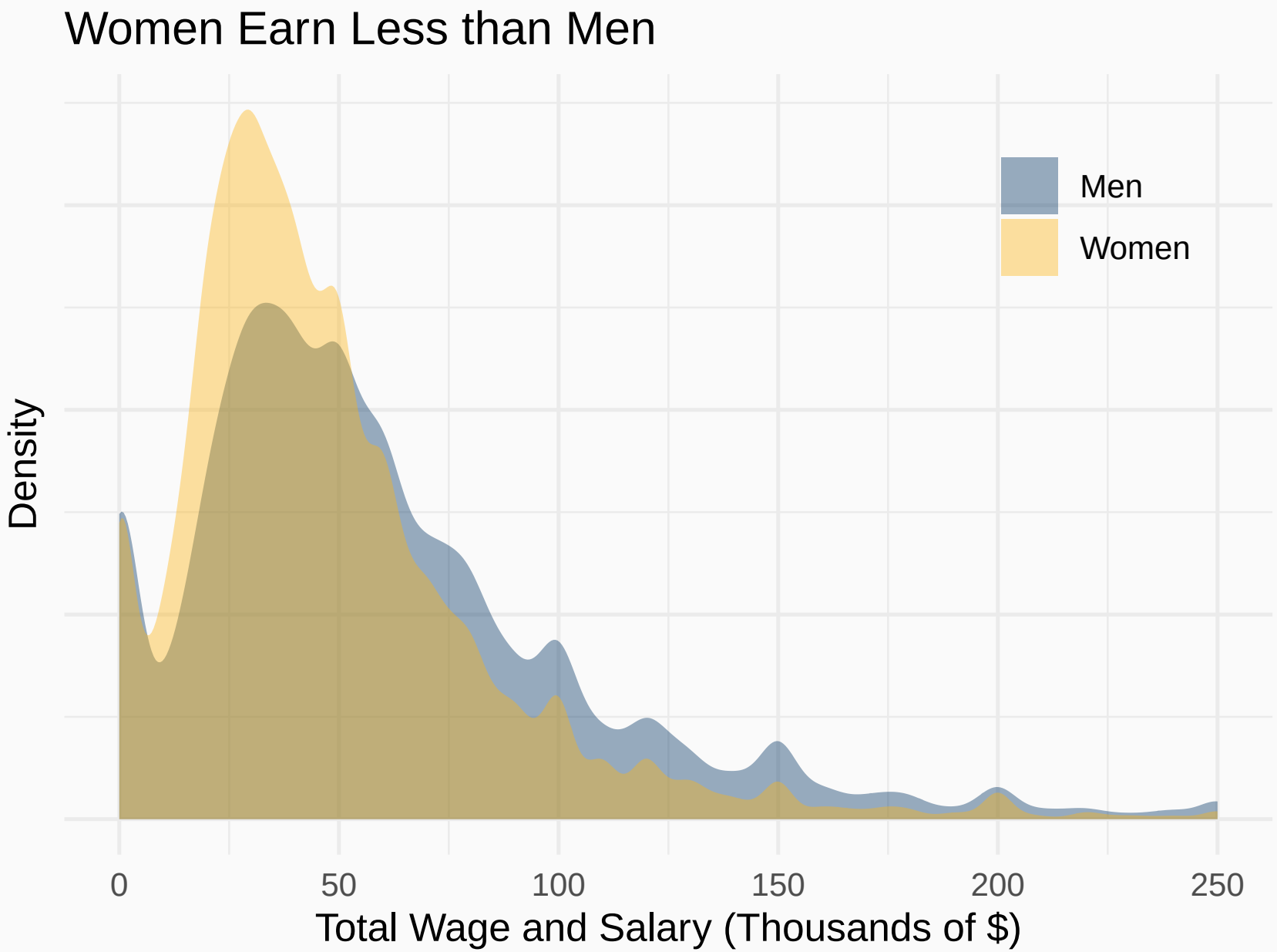
Data: Current Population Survey, 2019 Annual Social and Economic Supplement (ASEC)

Key variables:

- Status: full-/part-time work status
  - 1: not in labor force
  - 2: full-time hours (35 or more), usually full-time

    ⋮
- Pay: total wage and salary earnings
- Sex
  - 1: male
  - 2: female

Overall Comparison



Direct Application of t-Test

Mean for men: $68,700

Mean for women: $52,100

Mean difference: $16,600

(76 cents to the dollar)

$t = 27.938, df = 61,733, p < 2.2e - 16$

The Same Analysis in a Regression Framework, Part I

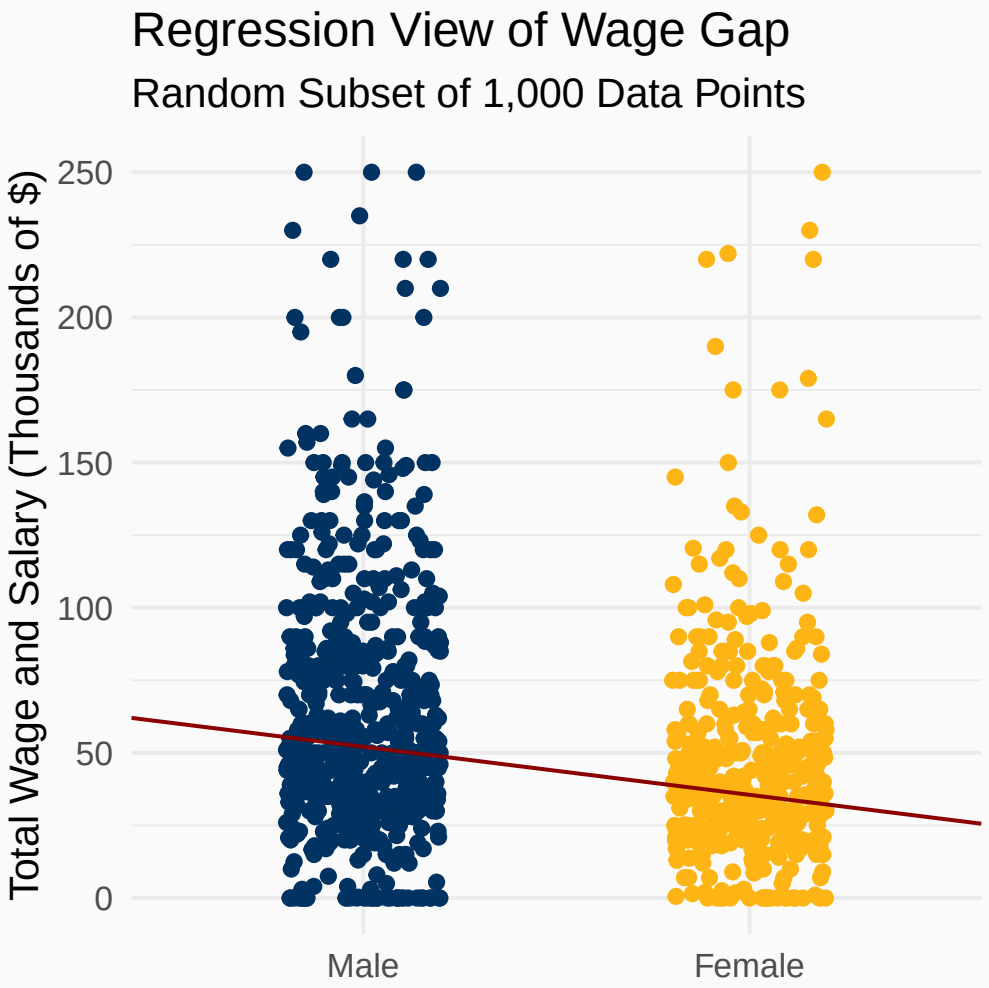
$$\widehat{Pay} = \beta_0 + \beta_1 Female$$

For men, $\widehat{Pay} = \beta_0 + \beta_1(0) = \beta_0$.

For women, $\widehat{Pay} = \beta_0 + \beta_1(1) = \beta_0 + \beta_1$.

$\implies$ β_1 represents the difference between groups.

The Same Analysis in a Regression Framework, Part II



$$\Delta Y = \frac{\text{slope}}{\Delta X}$$

The Same Analysis in a Regression Framework, Part III

	Dependent Variable: Pay
Female	$-16{,}561^{***}$
	(618)
Intercept	$68{,}667^{***}$
	(408)
Observations	62,110
Note:	*p$<$0.05; **p$<$0.01; ***p$<$0.001

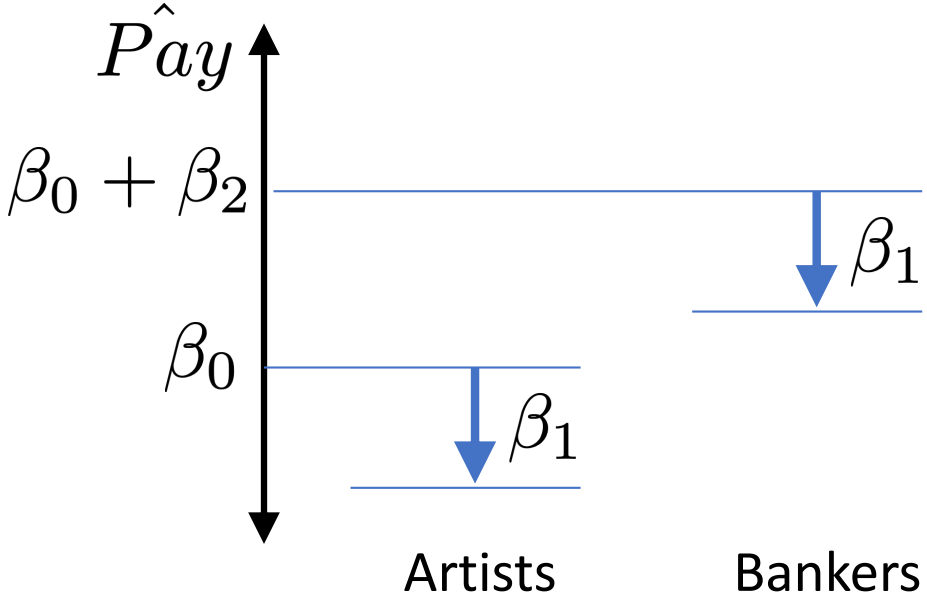
Indicator Variables as Controls

Controlling for Occupation, Part I

How can we compare men and women in the same occupation?

Controlling for Occupation, Part II

$$\widehat{Pay} = \beta_0 + \beta_1 Female + \beta_2 Banker + \beta_3 Engineer + \ldots$$



Controlling for Occupation, Part III

	Dependent Variable: Pay	
	(1)	(2)
Female	−16,561***	−15,035***
	(618)	(679)
Intercept		183,292***
		(3,440)
occup20		84,255***
		(3,714)
occup40		76,881***
		(13,132)
occup50		108,112***
		(3,753)
⋮		⋮

Controlling for Occupation, Part IV

	Dependent Variable: Pay	
	(1)	(2)
Female	−16,561***	−15,035***
	(618)	(679)
Intercept	68,667***	7,110**
	(408)	(2,321)
Occupation FE	No	Yes
Observations	62,110	62,110
Note:	*p<0.05; **p<0.01; ***p<0.001	

Metric Variables as Controls

Age as a Covariate

Ways to add age to a specification

Linear effect: $\widehat{Pay} = \beta_0 + \beta_1 Female + \beta_2 Age$

Polynomial effect: $\widehat{Pay} = \beta_0 + \beta_1 Female + \beta_2 Age + \beta_3 Age^2$

Understanding the Linear Age Effect

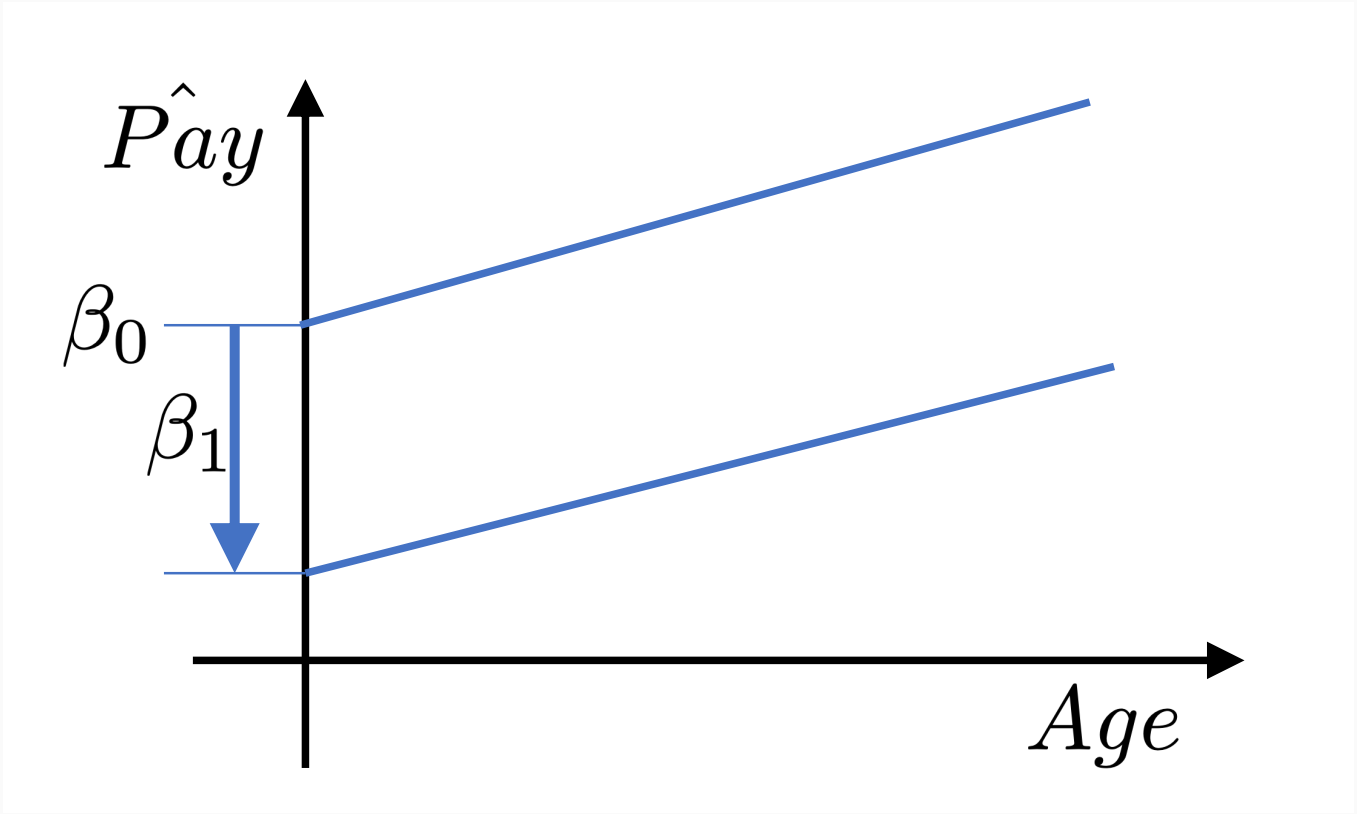
For men: $\widehat{Pay} = \beta_0 + \beta_1(0) + \beta_2 Age = \beta_0 + \beta_2 Age$

For women: $\widehat{Pay} = \beta_0 + \beta_1(1) + \beta_2 Age = \beta_0 + \beta_1 + \beta_2 Age$

Understanding the Linear Age Effect

For men: $\widehat{Pay} = \beta_0 + \beta_1(0) + \beta_2 Age = \beta_0 + \beta_2 Age$

For women: $\widehat{Pay} = \beta_0 + \beta_1(1) + \beta_2 Age = \beta_0 + \beta_1 + \beta_2 Age$

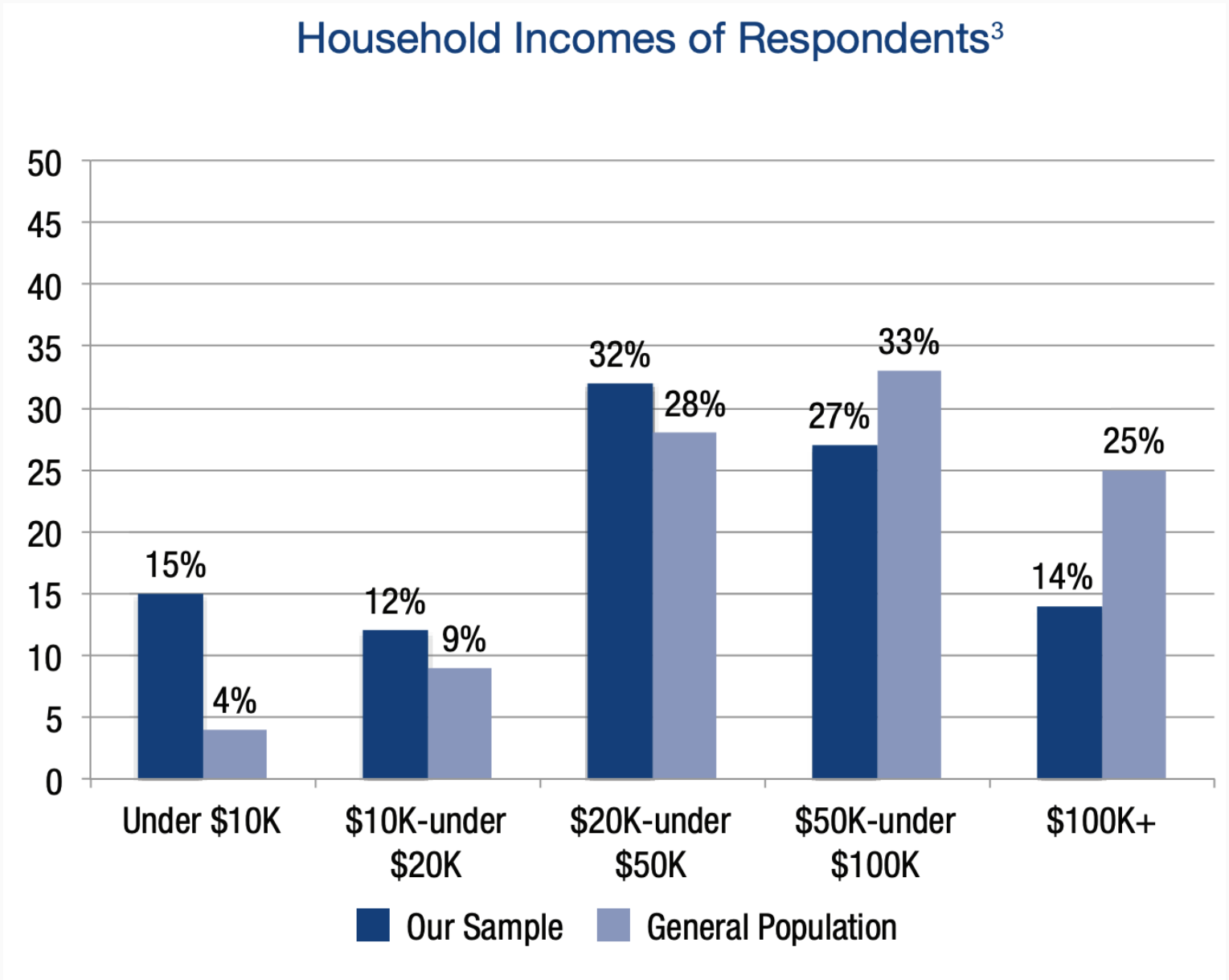


Controlling for Age

	Dependent Variable: Pay		
	(1)	(2)	(3)
Female	−16,561***	−15,035***	−15,063***
	(618)	(679)	(677)
Age			469***
			(22)
Intercept	68,667***	7,110**	−10,140***
	(408)	(2,321)	(2,447)
Occupation FE	No	Yes	Yes
Observations	62,110	62,110	62,110
Note:			*p<0.05; **p<0.01; ***p<0.001

Measuring With More Categories

National Transgender Discrimination Survey



Example of a More Inclusive Question

How would you describe yourself?

- Female
- Male
- Transgender female
- Transgender male
- Non-binary/nonconforming
- Prefer not to answer

Analyzing More Gender Categories

$$\widehat{Pay} = \beta_0 + \beta_1 Female + \beta_2 Transgender\_Male + \beta_3 Transgender\_Female + \beta_4 Nonbinary$$

$\implies$ Report each β to describe the pay gap for a different gender identity.

Reporting Overall Effect

Law of total variance:

$$V[Pay] = V\Big[E[Pay|Gender]\Big] + E\Big[V[Pay|Gender]\Big]$$

$$V[Pay] = \text{Gender-Explained Variance + Other Variance}$$

$$\eta^2 = \frac{\text{Gender-Explained Variance}}{\text{Total Variance}}$$

$$\text{Gender-Explained Standard Deviation} = \sqrt{V\Big[E[Pay|Gender]\Big]}$$

F-test: Null is that all genders have equal pay.

Planning Multiple Specifications

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Log working hours	-0.175***	-0.152***	-0.151***	-0.151***	-0.137***	-0.128***	-0.113***	-0.139***
	(0.007)	(0.013)	(0.013)	(0.013)	(0.013)	(0.012)	(0.013)	(0.011)
Trend			0.001***	0.001***	0.001***	0.001***	0.001***	0.001***
			(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Day of the week dummies (reference: Monday)								
Tuesday					0.004	0.007*	-0.023	-0.452***
					(0.004)	(0.003)	(0.024)	(0.045)
Wednesday					-0.000	0.001	0.029	-0.427***
					(0.004)	(0.004)	(0.025)	(0.052)
Thursday					-0.001	0.001	0.066***	-0.609***
					(0.004)	(0.003)	(0.025)	(0.063)
Friday					0.004	0.006*	0.011	-0.424***
					(0.004)	(0.004)	(0.026)	(0.050)
Saturday					0.109***	0.110***	-0.471***	-0.890***
					(0.008)	(0.007)	(0.069)	(0.074)
Sunday					0.809***	0.417***	0.806***	0.608***
					(0.058)	(0.063)	(0.039)	(0.040)
Age								-0.001***
								(0.000)
Tenure								0.008***
								(0.000)
Male								-0.003
								(0.004)
Individual fixed effects	No	Yes	Yes	Yes	Yes	Yes	Yes	No
Team fixed effects	No	No	No	Yes	Yes	Yes	Yes	Yes
Hour-of-the-day dummies	No	No	No	No	No	Yes	Yes	Yes
Day fixed effects	No	No	No	No	No	No	Yes	Yes
R-squared	0.085	0.094	0.152	0.160	0.198	0.285	0.385	0.403
N	33,123	33,123	33,123	33,123	33,123	33,123	33,123	31,525
Individuals		332	332	332	332	332	332	332

Collewet & Sauermann. (2017). *Working Hours and Productivity*

Why Report Multiple Specifications?

Modeling decisions are tough.

- Is it worthwhile to control for education, even if standard errors increase?
- Is it worth switching from a log to a square root to get a much better fit?
- Should we include occupation, even if that absorbs some of the concept we want to measure?

Would your results be different with different choices?

$\Longrightarrow$ Try different paths to see if your results are **robust**.

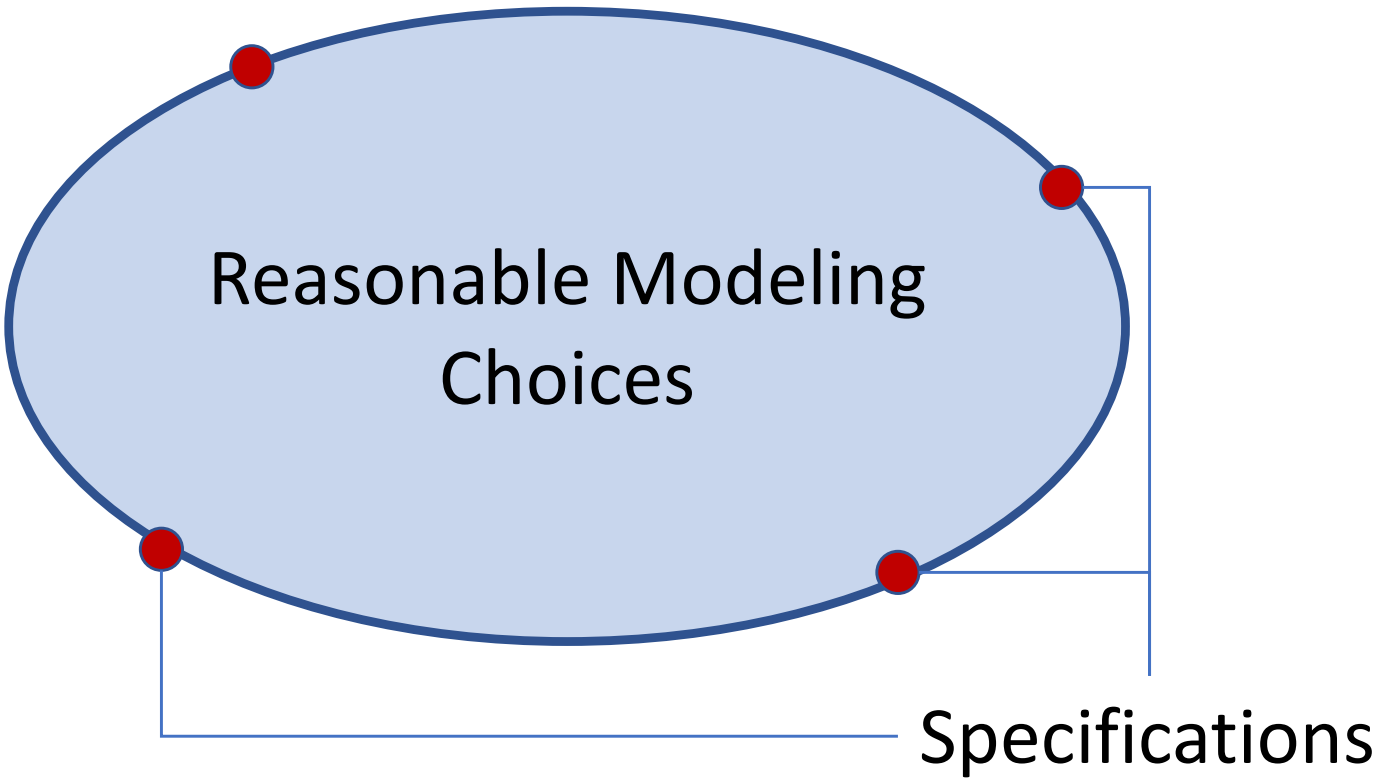
Why Report Multiple Specifications? (cont.)

Error rate inflation: a tendency for effects to appear large (and p-values small), especially when a researcher uses the same data for exploration and testing

p-Hacking: a deliberate attempt to generate significant results by altering the model specification and other researcher degrees of freedom

$\Longrightarrow$ Report multiple specifications to guard against error rate inflation.

How Should You Think About Specifications?



An Example Specification Table

	Dependent Variable: Pay		
	(1)	(2)	(3)
Female	−16,561***	−15,035***	−15,063***
	(618)	(679)	(677)
Age			469***
			(22)
Intercept	68,667***	7,110**	−10,140***
	(408)	(2,321)	(2,447)
Occupation FE	No	Yes	Yes
Observations	62,110	62,110	62,110
Note:			*p<0.05; **p<0.01; ***p<0.001

Modeling Conditional Effects

Conditional Effects

Conditional Effects

Idea: The relationships we measure may be different for different people or units.

- A vaccine may reduce infection rates more in adults than in children.
- Network access may be more associated with loan availability in poor countries.
- The pay gap may be different for people of different ages.

Conditional Effects of Age, Part I

Interaction term: a variable in a regression formed by multiplying two other variables together

$$\widehat{Pay} = \beta_0 + \beta_1 Female + \beta_2 Age + \beta_3 Female \cdot Age$$

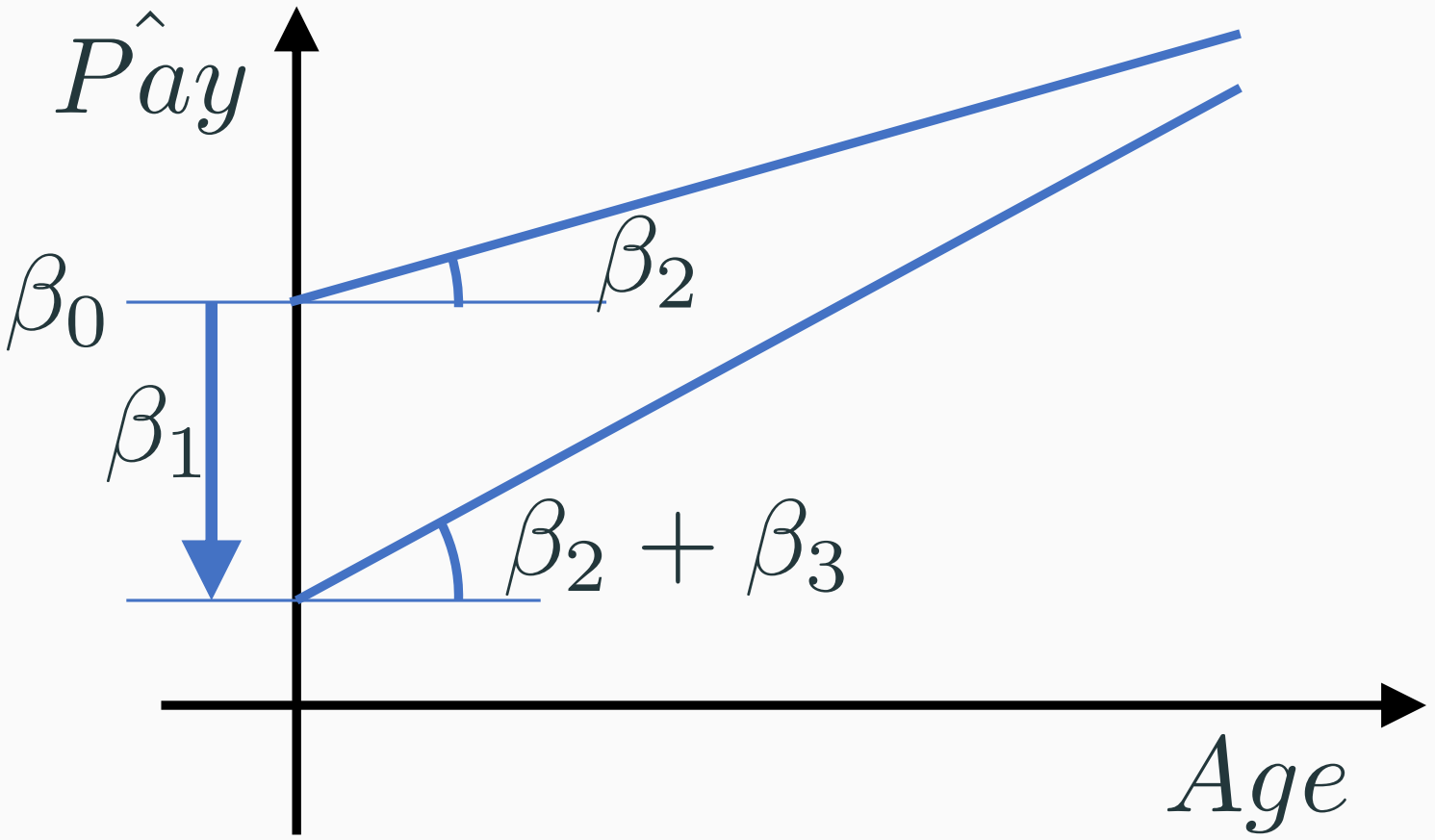
For men:

$$\widehat{Pay} =$$

For women:

$$\widehat{Pay} =$$

Conditional Effects of Age, Part II



Conditional Effects of Age, Part III

	Dependent Variable: Pay
Female	−6,230**
	(1,967)
Age	559***
	(29)
Female:Age	−207***
	(43)
Intercept	−13,971***
	(2,575)
Occupation FE	No
Observations	62,110
Note:	*p<0.05; **p<0.01; ***p<0.001

Interaction Terms

Interactions for Indicator Variables

Use old content: 12.9 Interaction Terms for Indicator Variables, Part 1

If there's extra time, I would redo this using the wage gap example.

Guidelines for Polynomial and Interaction Terms

Use old content: 12.12 Guidelines for Polynomial and Interaction Terms

Could redo some of the discussion of goals if there is time.