

Classical Assumptions

EC 320: Introduction to Econometrics

Kyle Raze

Fall 2019

Prologue

Housekeeping

Lab Attendance

Problem Set 2 grades

Problem Set 3

- Due this Wednesday by 17:00.
- For computational problems, submit an HTML document generated by R Markdown.

Agenda

Last Week

How does OLS estimate a regression line?

- Minimize RSS.

What are the direct consequences of minimizing RSS?

- Residuals sum to zero.
- Residuals and the explanatory variable X are uncorrelated.
- Mean values of X and Y are on the fitted regression line.

Whatever do we mean by *goodness of fit*?

- What information does R^2 convey?

Agenda

Today

Under what conditions is OLS *desirable*?

- **Desired properties:** Unbiasedness, efficiency, and ability to conduct hypothesis tests.
- **Cost:** Six **classical assumptions** about the population relationship and the sample.

Returns to Schooling

Policy Question: How much should the state subsidize higher education?

- Could higher education subsidies increase future tax revenue?
- Could targeted subsidies reduce income inequality and racial wealth gaps?
- Are there positive externalities associated with higher education?

Empirical Question: What is the monetary return to an additional year of education?

- Focuses on the private benefits of education. Not the only important question!
- Useful for learning about the econometric assumptions that allow causal interpretation.

Returns to Schooling

Step 1: Write down the population model.

$$\log(\text{Earnings}_i) = \beta_1 + \beta_2 \text{Education}_i + u_i$$

Step 2: Find data.

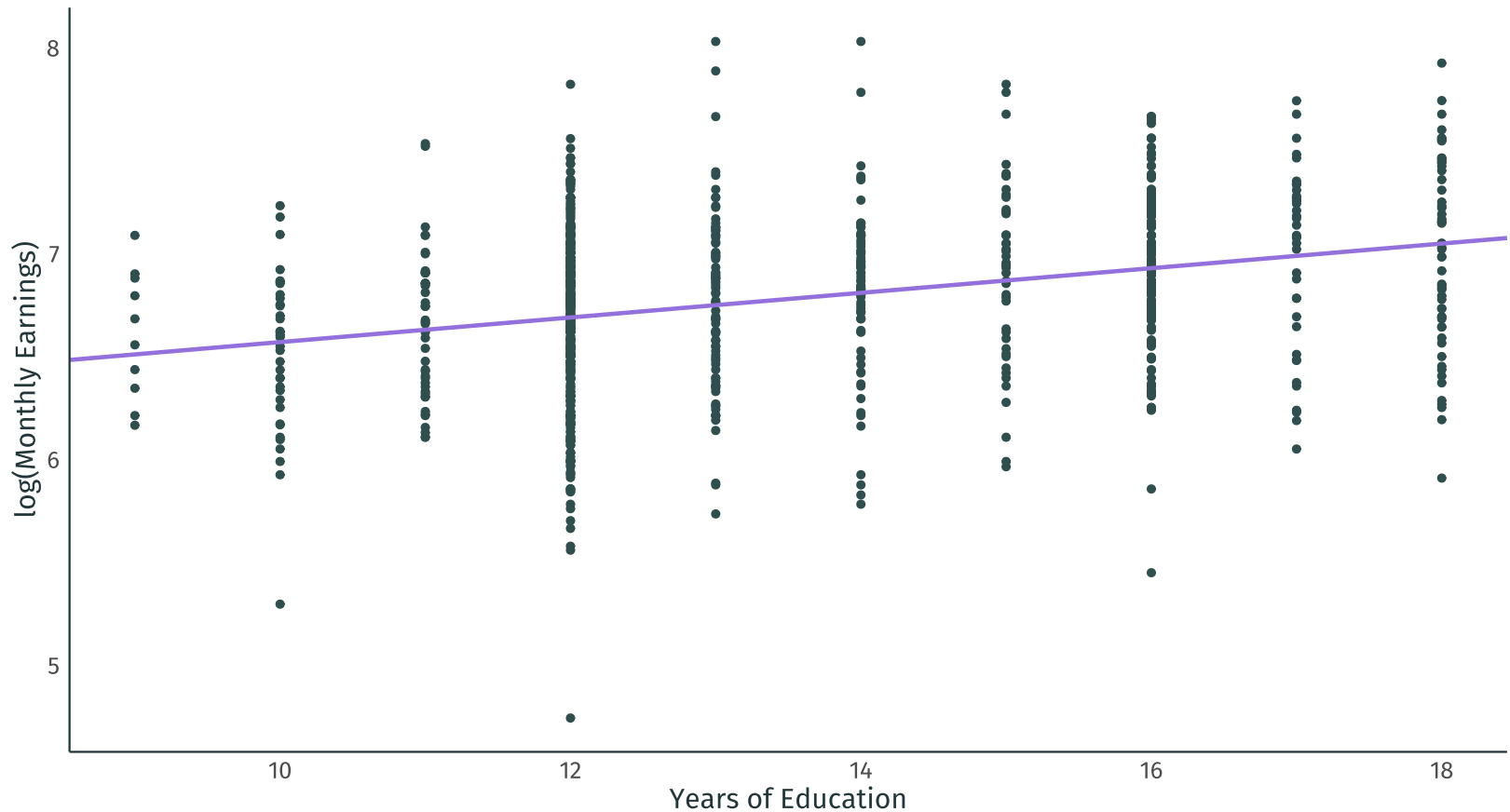
- *Source:* Blackburn and Neumark (1992).

Step 3: Run a regression using OLS.

$$\log(\text{Earnings}_i) = \hat{\beta}_1 + \hat{\beta}_2 \text{Education}_i$$

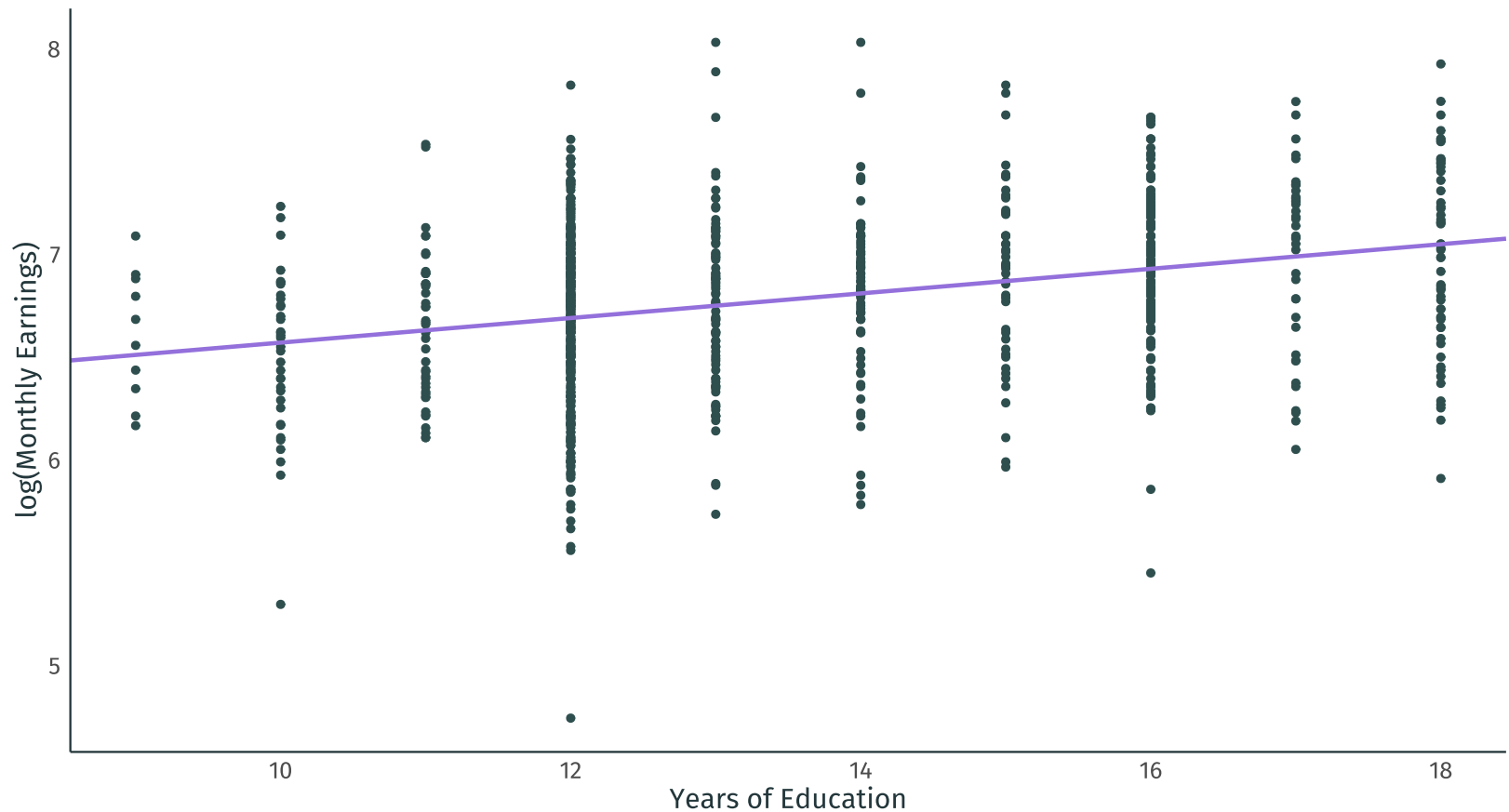
Returns to Schooling

$$\log(\text{Earnings}_i) = 5.97 + 0.06 \times \text{Education}_i$$



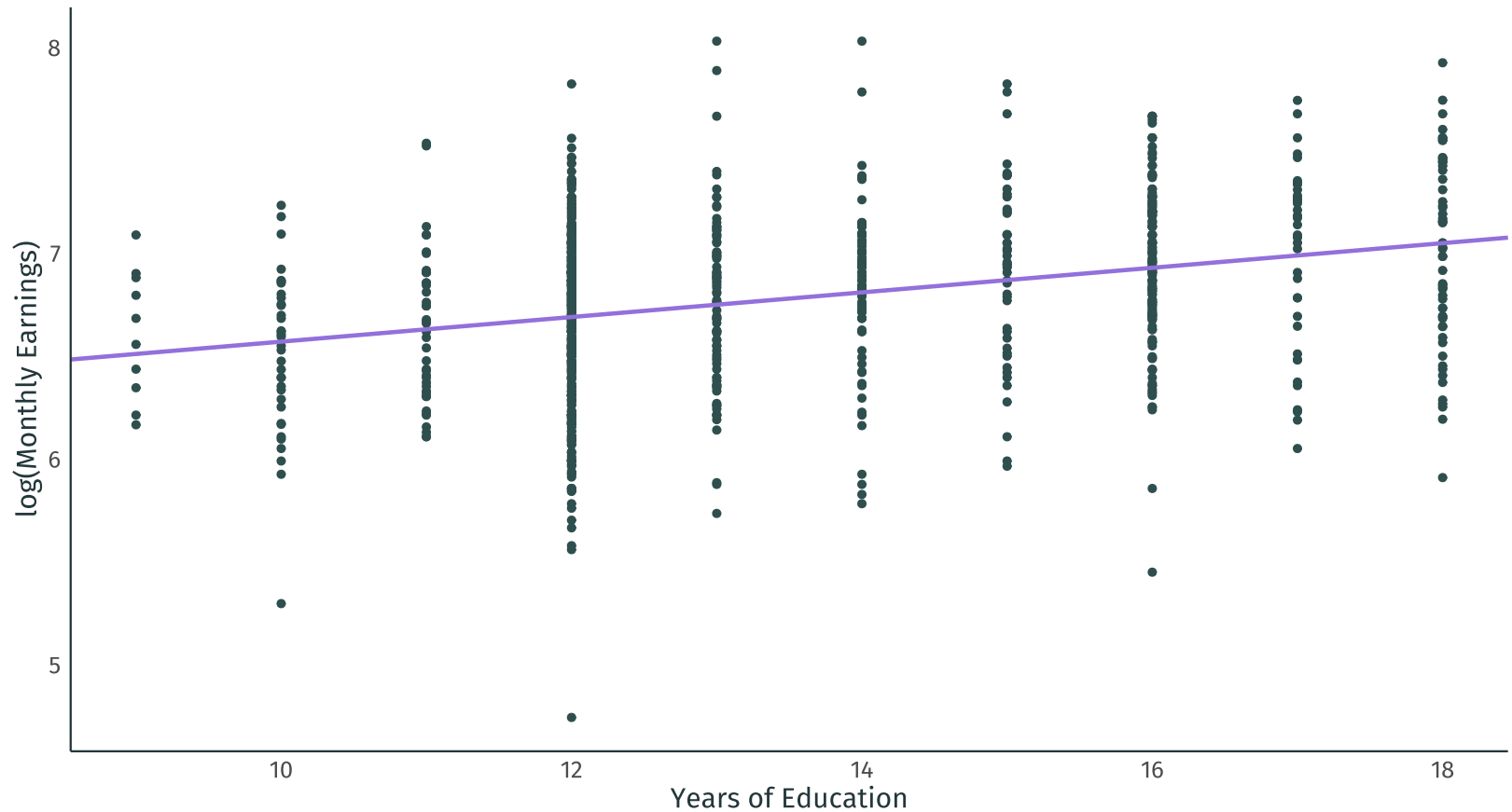
Returns to Schooling

Additional year of school associated with a **6%** increase in earnings.



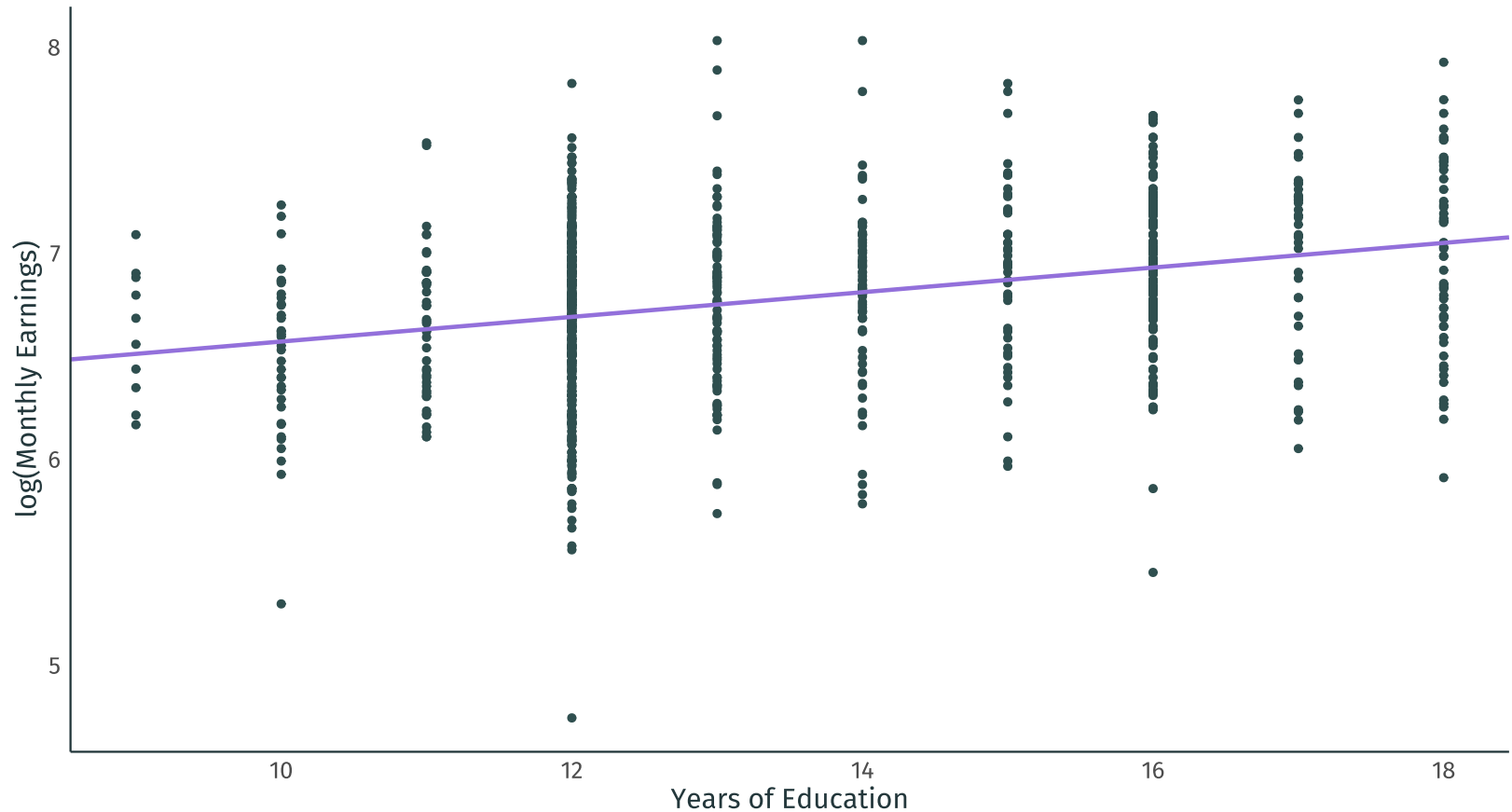
Returns to Schooling

$$R^2 = 0.097.$$



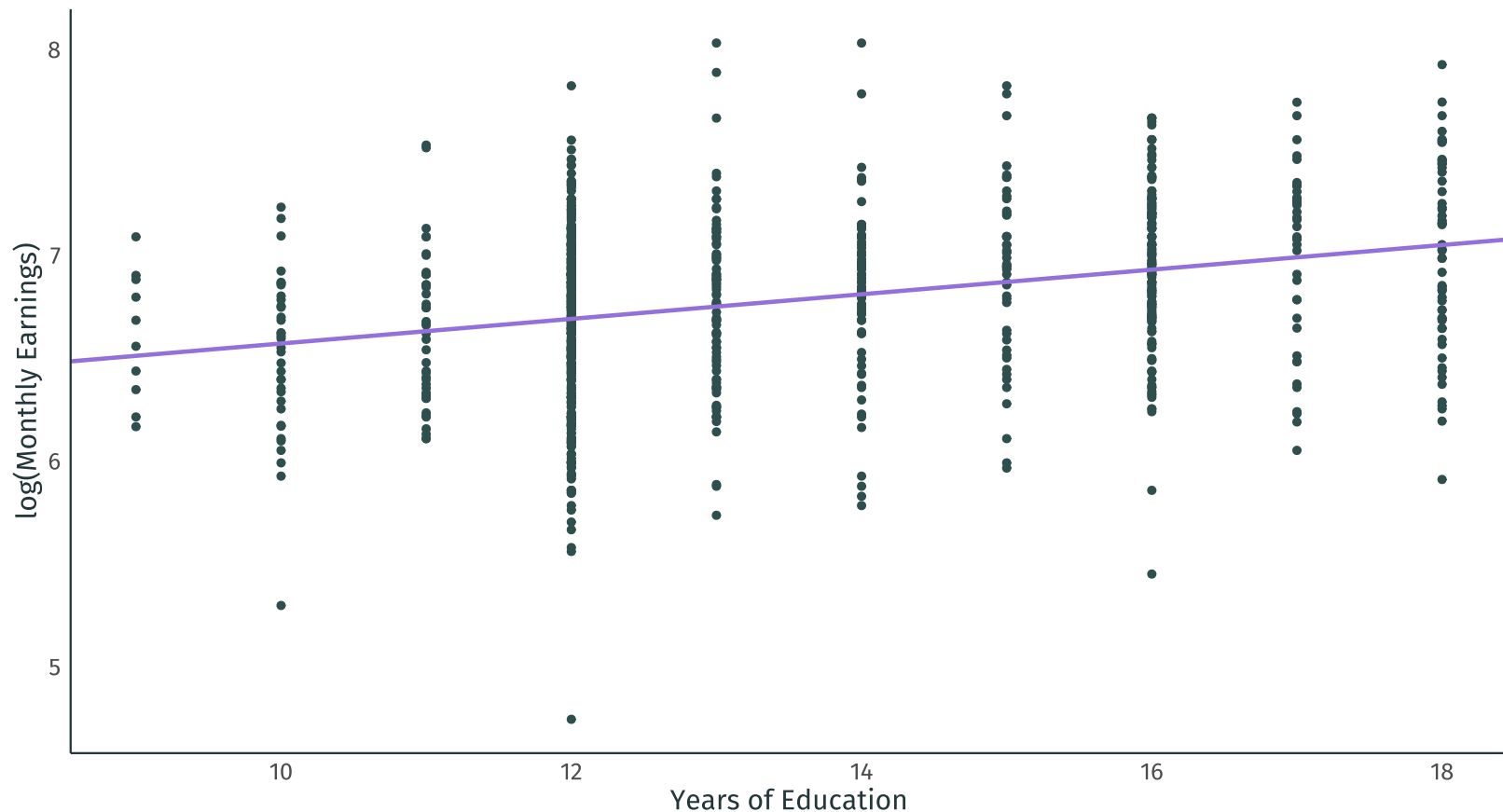
Returns to Schooling

Education explains **9.7%** of the variation in wages.



Returns to Schooling

What must we **assume** to interpret $\hat{\beta}_2 = 0.06$ as the return to schooling?



Residuals vs. Errors

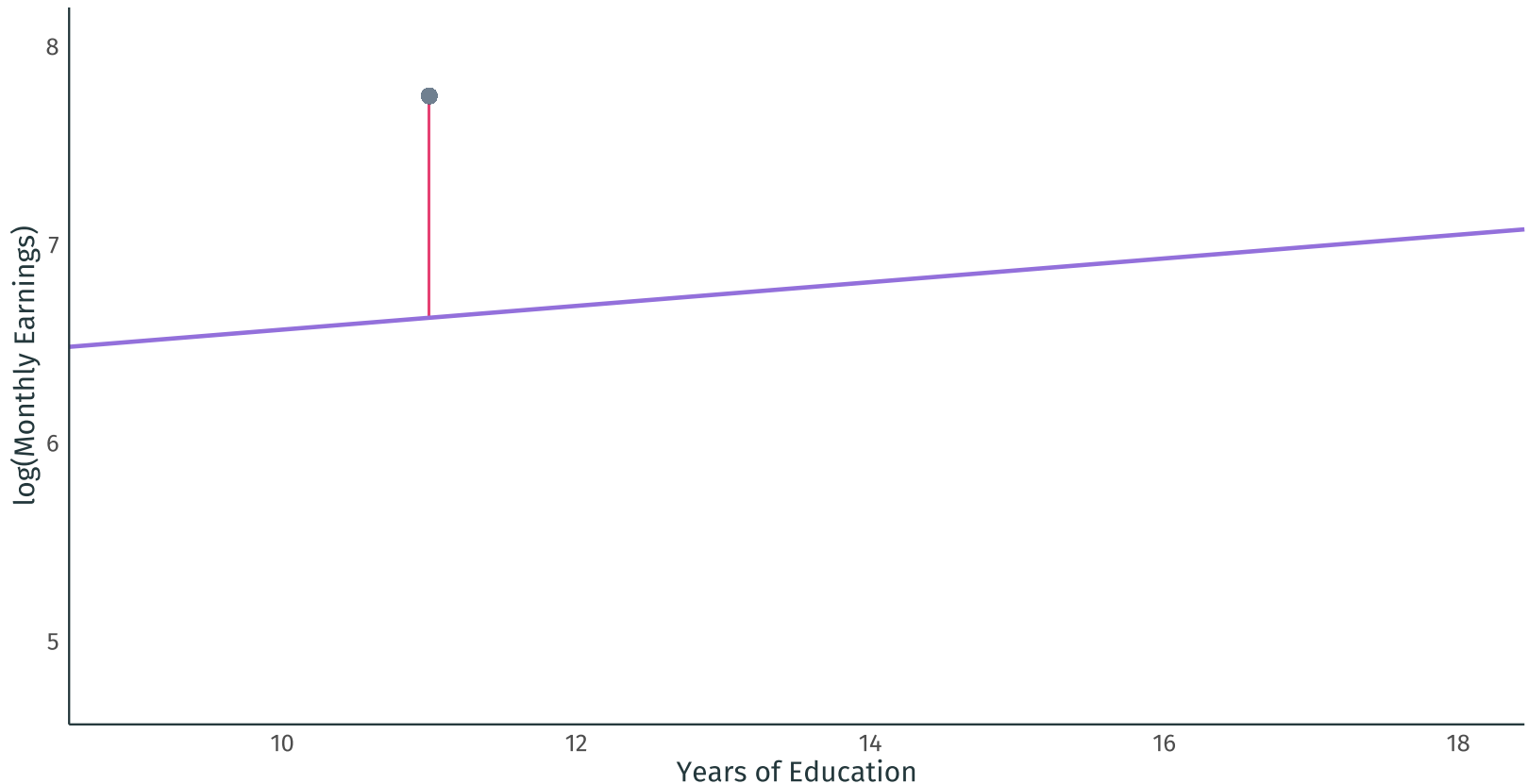
The most important assumptions concern the error term u_i .

Important: An error u_i and a residual $\hat{u}_i$ are related, but different.

- **Error:** Difference between the wage of a worker with 16 years of education and the **expected wage** with 16 years of education.
- **Residual:** Difference between the wage of a worker with 16 years of education and the **average wage** of workers with 16 years of education.
- **Population vs. sample.**

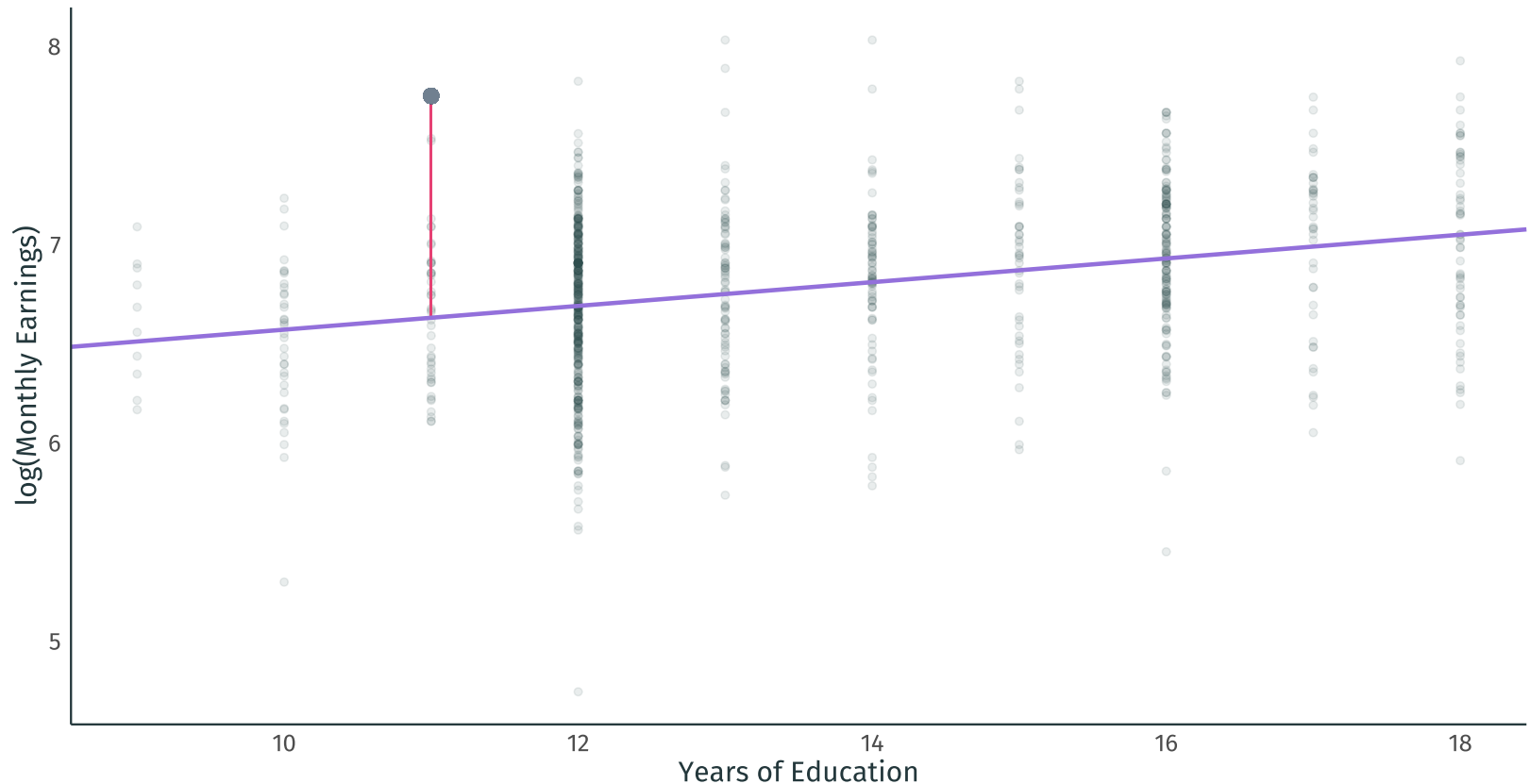
Residuals vs. Errors

A **residual** tells us how a **worker's** wages compare to the average wages of workers in the **sample** with the same level of education.



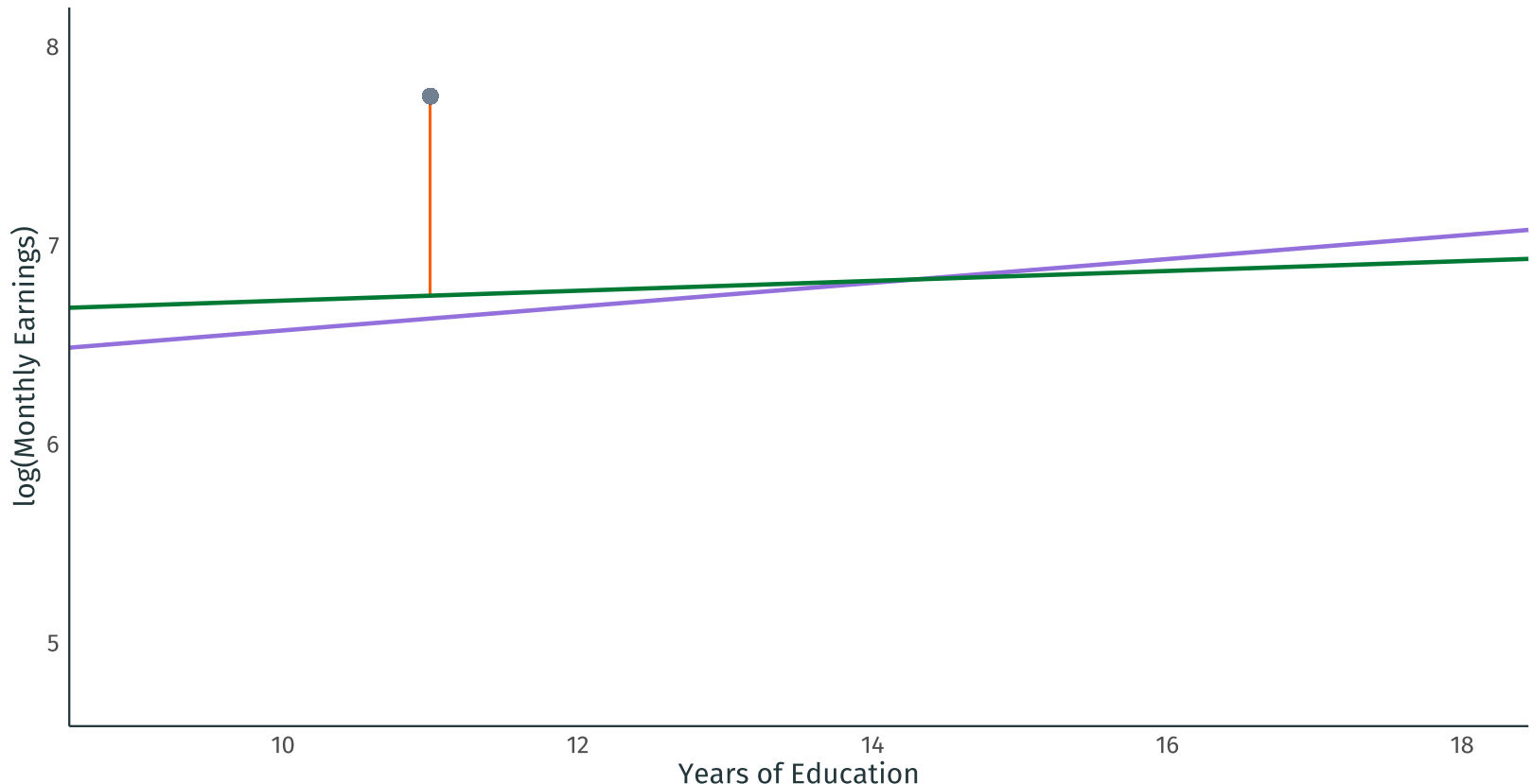
Residuals vs. Errors

A **residual** tells us how a **worker's** wages compare to the average wages of workers in the **sample** with the same level of education.



Residuals vs. Errors

An **error** tells us how a **worker's** wages compare to the expected wages of workers in the **population** with the same level of education.



Classical Assumptions

Classical Assumptions of OLS

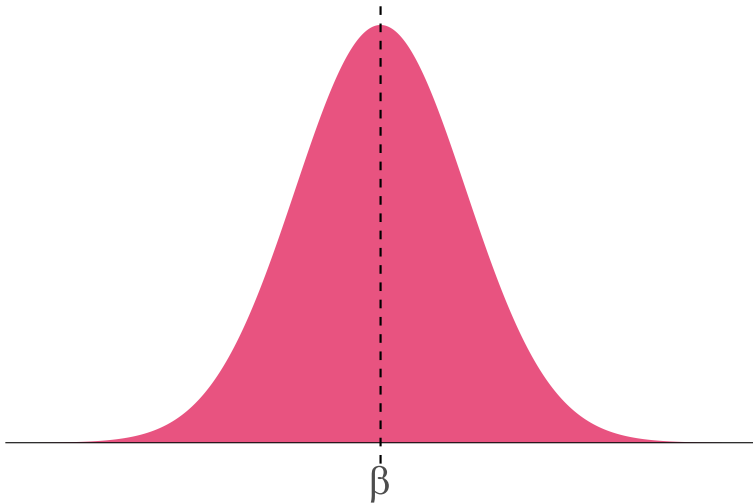
1. **Linearity:** The population relationship is **linear in parameters** with an additive error term.
2. **Sample Variation:** There is variation in X .
3. **Random Sampling:** We have a random sample from the population of interest.
4. **Exogeneity:** The X variable is **exogenous** (i.e., $E(u|X) = 0$).
5. **Homoskedasticity:** The error term has the same variance for each value of the independent variable (i.e., $\text{Var}(u|X) = \sigma^2$).
6. **Normality:** The population error term is normally distributed with mean zero and variance σ^2 (i.e., $u \sim N(0, \sigma^2)$)

When Can We Trust OLS?

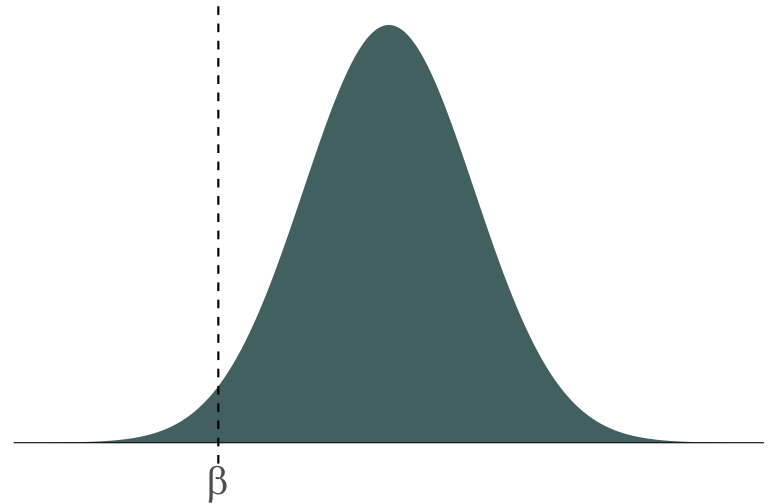
Bias

An estimator is **biased** if its expected value is different from the true population parameter.

Unbiased estimator: $E[\hat{\beta}] = \beta$



Biased estimator: $E[\hat{\beta}] \neq \beta$



When is OLS Unbiased?

Assumptions

1. **Linearity:** The population relationship is **linear in parameters** with an additive error term.
2. **Sample Variation:** There is variation in X .
3. **Random Sampling:** We have a random sample from the population of interest.
4. **Exogeneity:** The X variable is **exogenous** (i.e., $E(u|X) = 0$).

Result

OLS is unbiased.

Linearity

Assumption

The population relationship is **linear in parameters** with an additive error term.

Examples

- $\text{Wage}_i = \beta_1 + \beta_2 \text{Experience}_i + u_i$
- $\log(\text{Happiness}_i) = \beta_1 + \beta_2 \log(\text{Money}_i) + u_i$
- $\sqrt{\text{Convictions}_i} = \beta_1 + \beta_2 (\text{Early Childhood Lead Exposure})_i + u_i$
- $\log(\text{Earnings}_i) = \beta_1 + \beta_2 \text{Education}_i + u_i$

Linearity

Assumption

The population relationship is **linear in parameters** with an additive error term.

Violations

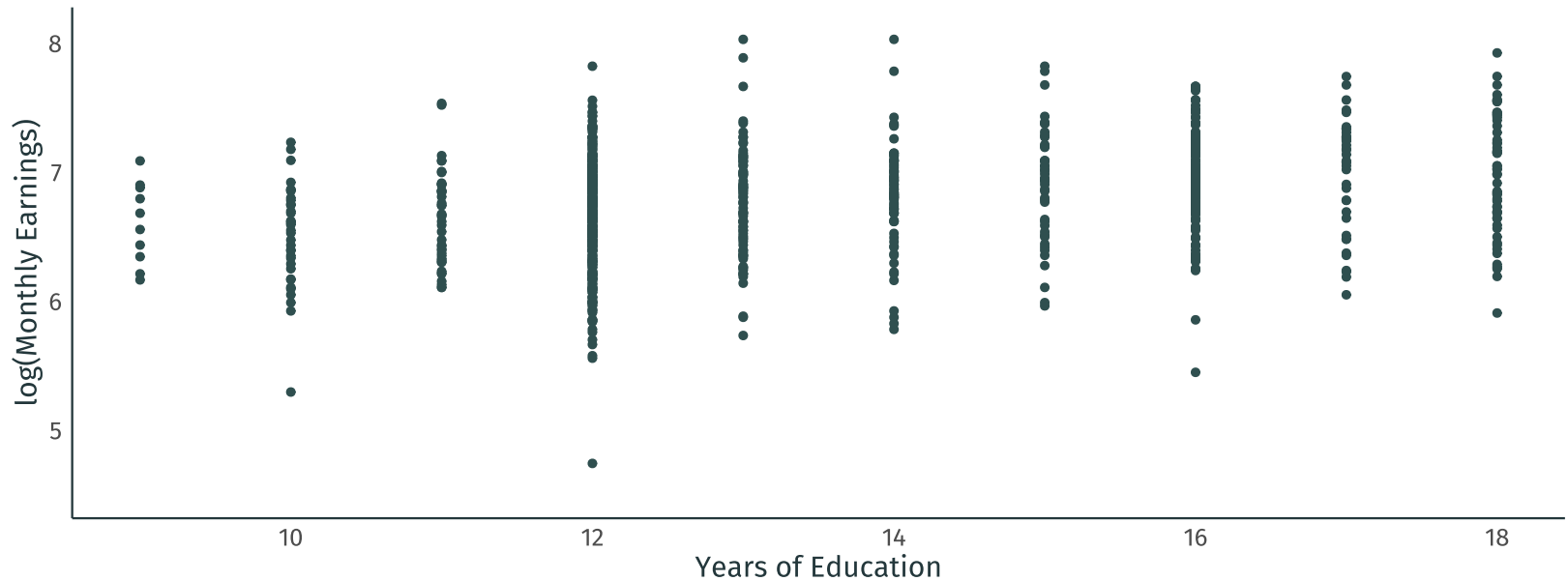
- $\text{Wage}_i = (\beta_1 + \beta_2 \text{Experience}_i) u_i$
- $\text{Consumption}_i = \frac{1}{\beta_1 + \beta_2 \text{Income}_i} + u_i$
- $\text{Population}_i = \frac{\beta_1}{1 + e^{\beta_2 + \beta_3 \text{Food}_i}} + u_i$
- $\text{Batting Average}_i = \beta_1 (\text{Wheaties Consumption}_i)^{\beta_2} + u_i$

Sample Variation

Assumption

There is variation in X .

Example

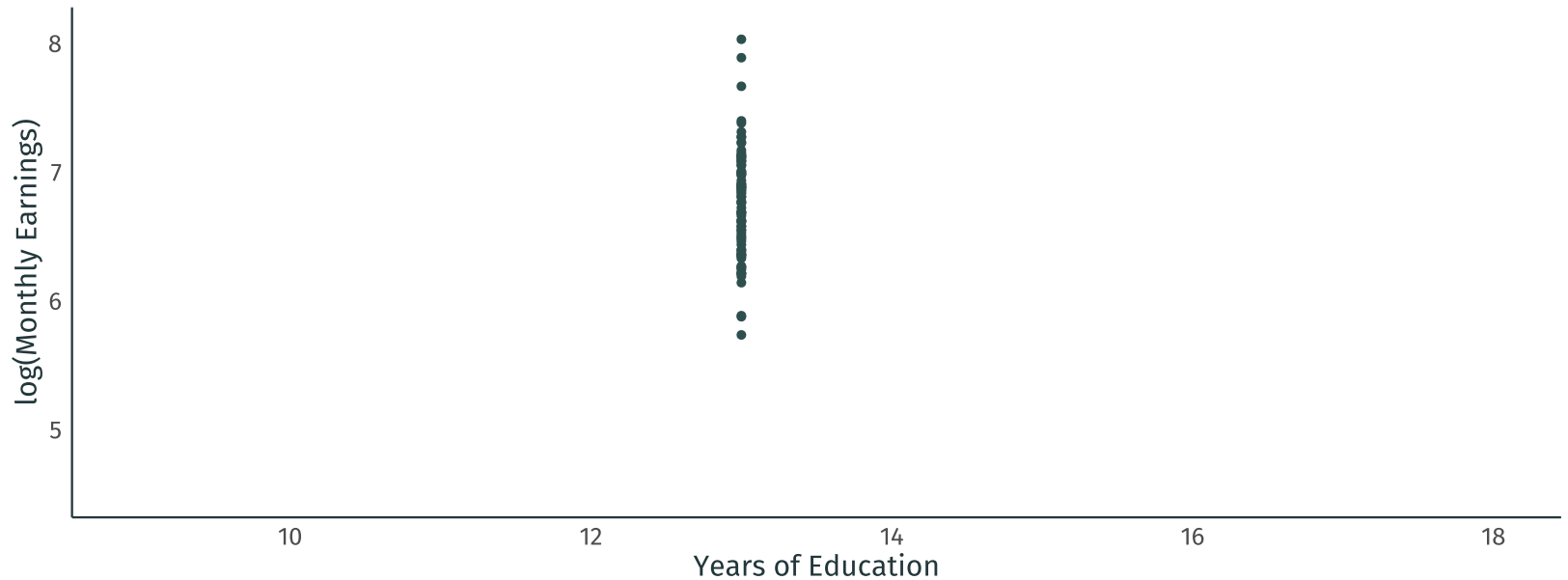


Sample Variation

Assumption

There is variation in X .

Violation



Random Sampling

Assumption

We have a random sample from the population of interest.

Examples

Random sampling generates many cross-sectional datasets (especially surveys).

- Government surveys (e.g., Current Population Survey, American Community Survey).
- Scientific surveys (e.g., General Social Survey, American National Election Study).
- High-quality political polls (e.g., YouGov, Quinnipiac University, Gallup).

Random Sampling

Assumption

We have a random sample from the population of interest.

Violations

- Data collected from non-probability sampling (*e.g.* snowball sampling).
- Most (all?) time-series data.
- Self-selected samples.

Exogeneity

Assumption

The X variable is **exogenous**: $E(u | X) = 0$.

- For *any* value of X , the mean of the error term is zero.

The most important assumption!

Really two assumptions bundled into one:

1. On average, the error term is zero: $E(u) = 0$.
2. The mean of the error term is the same for each value of X :
 $E(u | X) = E(u)$.

Exogeneity

Assumption

The X variable is **exogenous**: $E(u|X) = 0$.

- The assignment of X is effectively random.
- **Implication:** **no selection bias** and **no omitted-variable bias**.

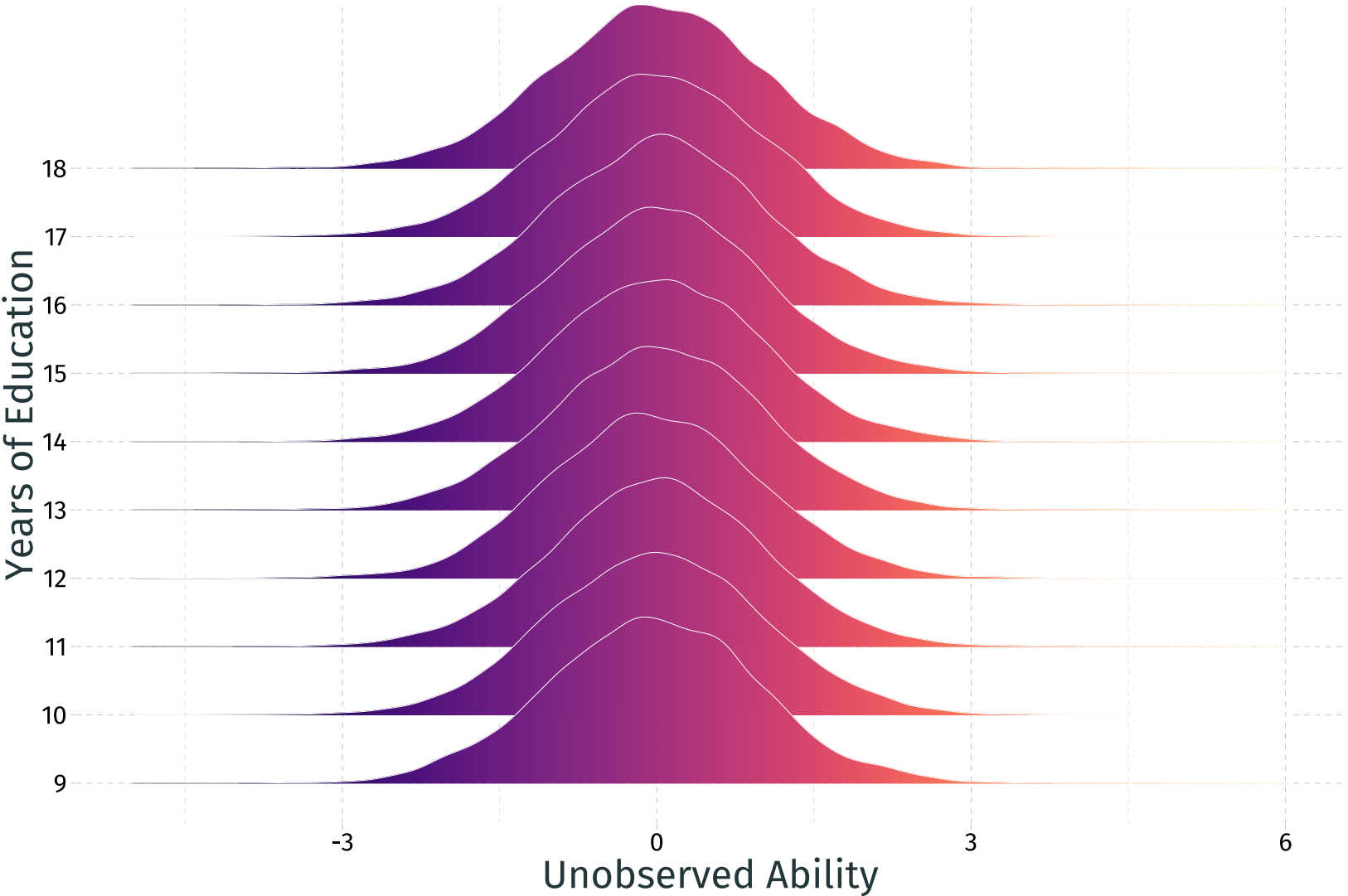
Examples

In the labor market, an important component of u is unobserved ability.

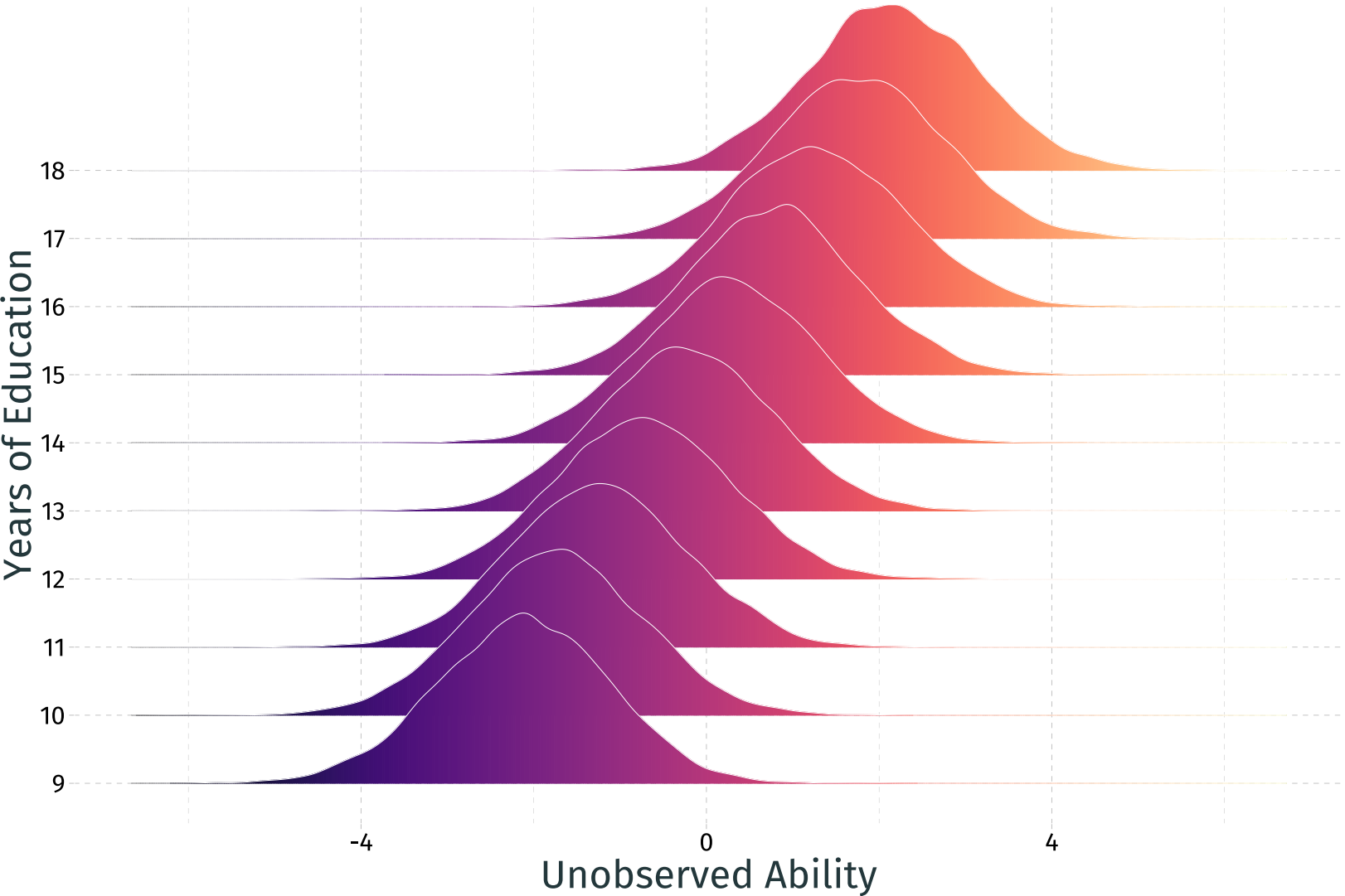
- $E(u | \text{Education} = 12) = 0$ and $E(u | \text{Education} = 20) = 0$.
- $E(u | \text{Experience} = 0) = 0$ and $E(u | \text{Experience} = 40) = 0$.
- **Do you believe this?**

Graphically...

Valid exogeneity, *i.e.*, $E(u \mid X) = 0$



Invalid exogeneity, *i.e.*, $E(u | X) \neq 0$



Variance Matters, Too

Why Variance Matters

Unbiasedness tells us that OLS gets it right, *on average*.

- But we can't tell whether our sample is "typical."

Variance tells us how far OLS can deviate from the population mean.

- How tight is OLS centered on its expected value?

The smaller the variance, the closer OLS gets to the true population parameters *on any sample*.

- Given two unbiased estimators, we want the one with smaller variance.

OLS Variance

To calculate the variance of OLS, we need:

1. The same four assumptions we made for unbiasedness.
2. **Homoskedasticity.**

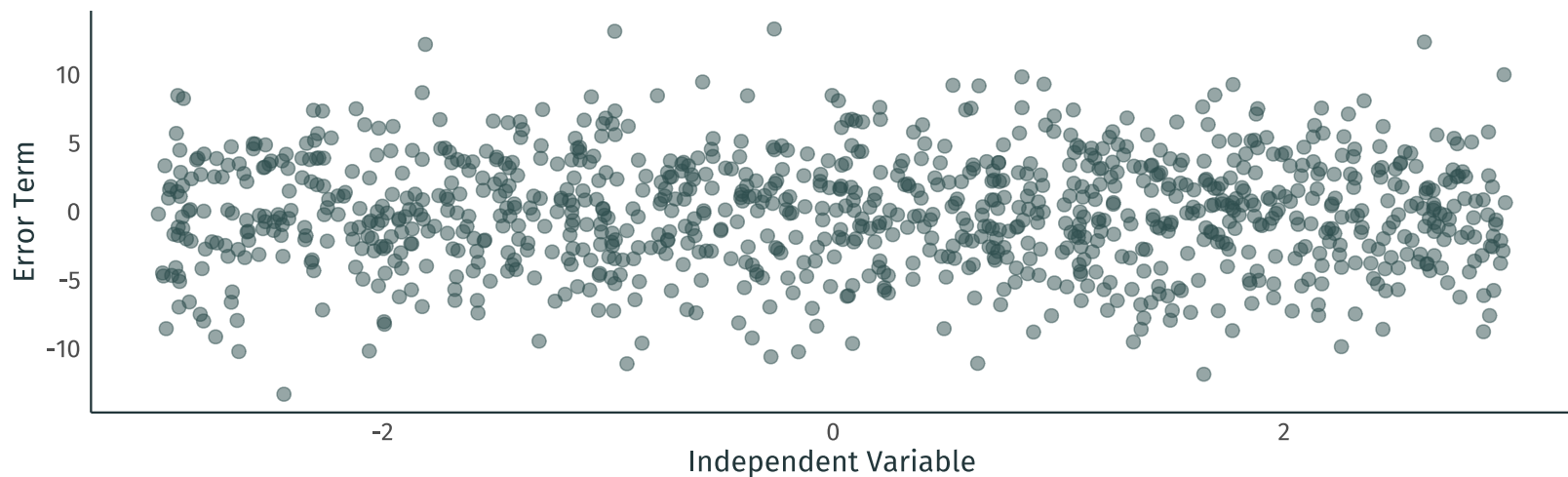
Homoskedasticity

Assumption

The error term has the same variance for each value of the independent variable:

$$\text{Var}(u | X) = \sigma^2.$$

Example



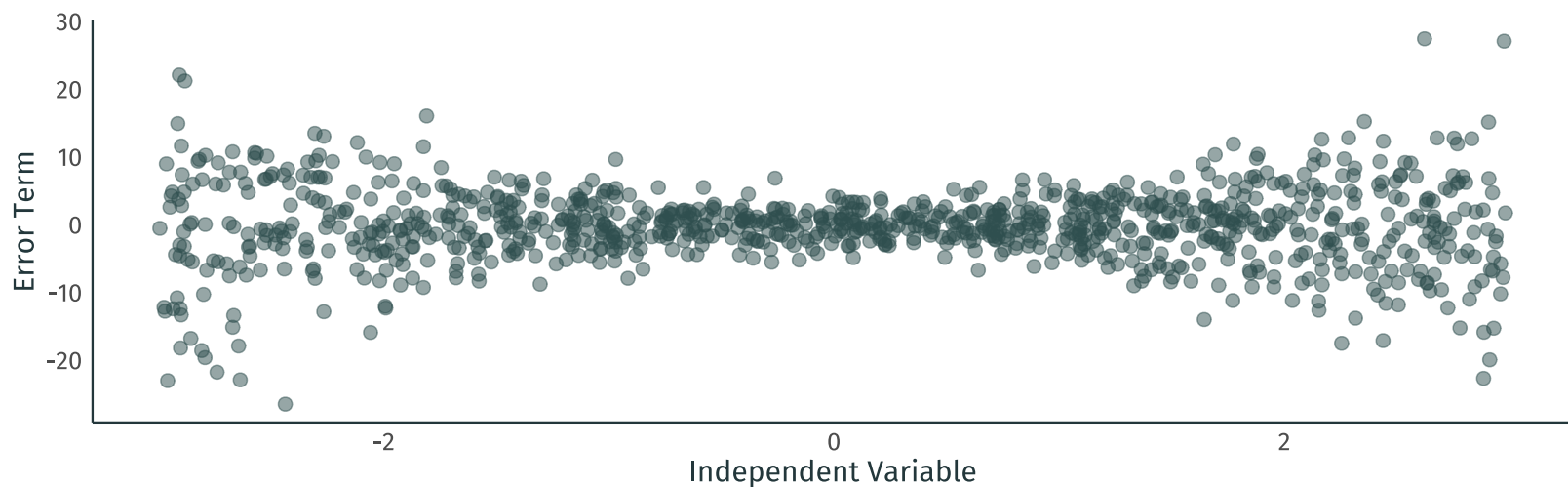
Homoskedasticity

Assumption

The error term has the same variance for each value of the independent variable:

$$\text{Var}(u | X) = \sigma^2.$$

Violation: Heteroskedasticity



OLS Variance

Variance of the slope estimator:

$$\text{Var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

- As the error variance increases, the variance of the slope estimator increases.
- As the variation in X increases, the variance of the slope estimator decreases.
- Larger sample sizes exhibit more variation in $X \Rightarrow \text{Var}(\hat{\beta}_2)$ falls as n rises.

Gauss-Markov

Gauss-Markov Theorem

OLS is the **Best Linear Unbiased Estimator (BLUE)** when:

1. **Linearity:** The population relationship is **linear in parameters** with an additive error term.
2. **Sample Variation:** There is variation in X .
3. **Random Sampling:** We have a random sample from the population of interest.
4. **Exogeneity:** The X variable is **exogenous** (*i.e.*, $E(u | X) = 0$).
5. **Homoskedasticity:** The error term has the same variance for each value of the independent variable (*i.e.*, $\text{Var}(u | X) = \sigma^2$).

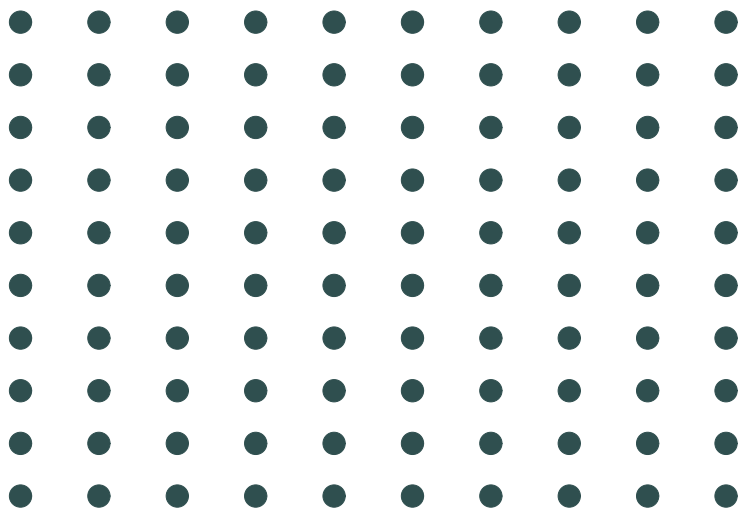
Gauss-Markov Theorem

OLS is the **Best Linear Unbiased Estimator (BLUE)**

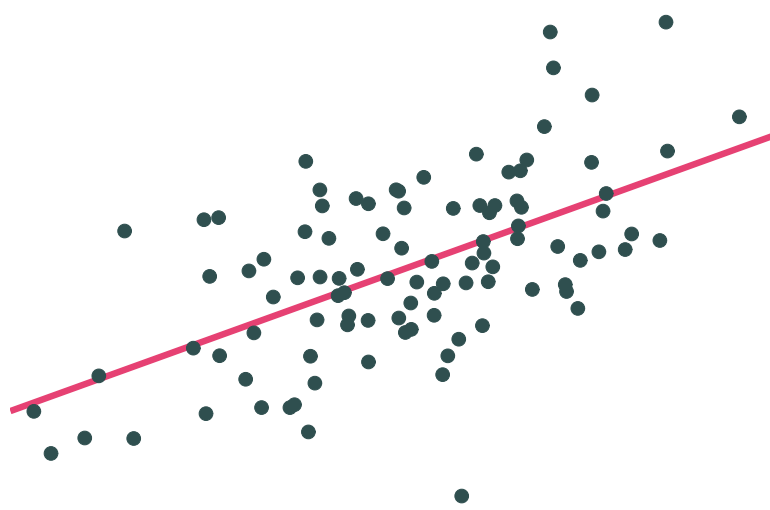
Population vs. Sample, Revisited

Population vs. Sample

Question: Why do we care about *population vs. sample*?



Population



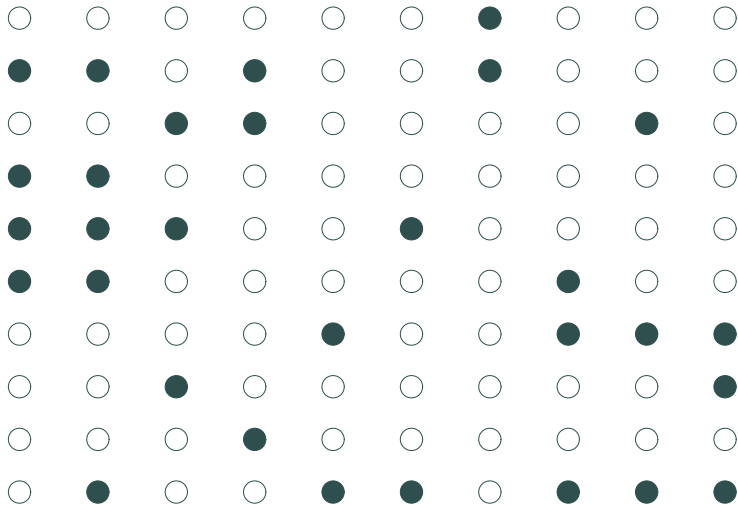
Population relationship

$$y_i = 2.53 + 0.57x_i + u_i$$

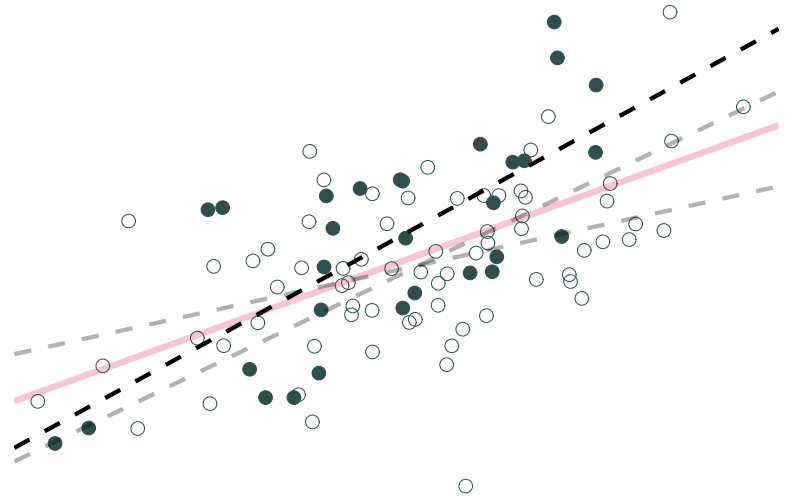
$$y_i = \beta_1 + \beta_2 x_i + u_i$$

Population vs. Sample

Question: Why do we care about *population vs. sample*?



Sample 3: 30 random individuals



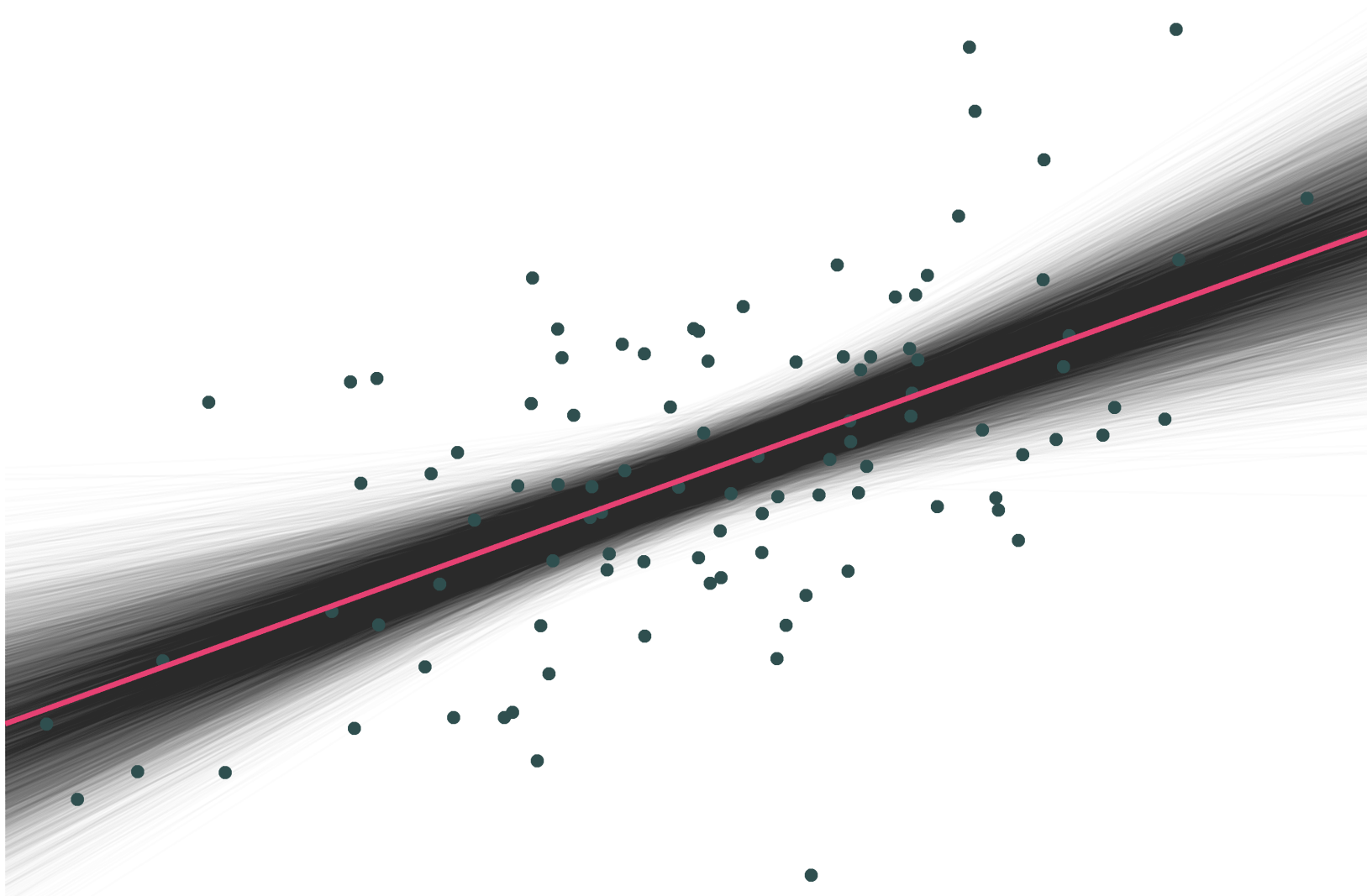
Population relationship

$$y_i = 2.53 + 0.57x_i + u_i$$

Sample relationship

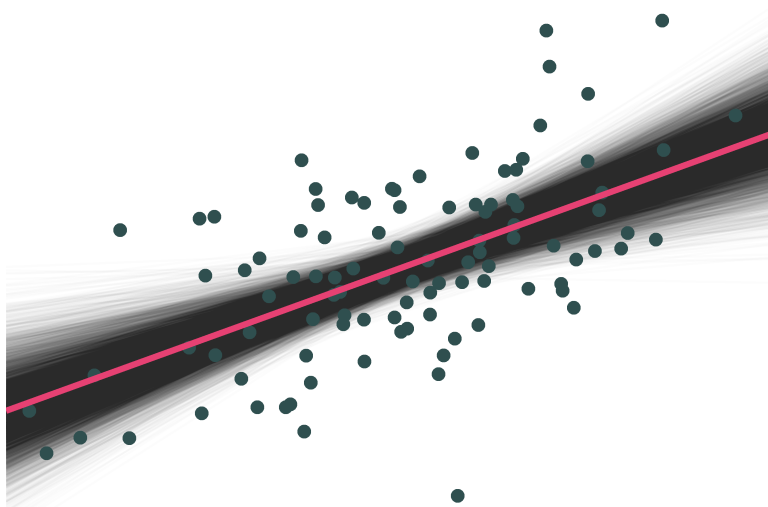
$$\hat{y}_i = 1.44 + 0.86x_i$$

Repeat **10,000 times** (Monte Carlo simulation).



Population vs. Sample

Question: Why do we care about *population vs. sample*?



- On **average**, the regression lines match the population line nicely.
- However, **individual lines** (samples) can miss the mark.
- Differences between individual samples and the population create **uncertainty**.

Population vs. Sample

Question: Why do we care about *population vs. sample*?

Answer: Uncertainty matters.

$\hat{\beta}_1$ and $\hat{\beta}_2$ are random variables that depend on the random sample.

We can't tell if we have a "good" sample (similar to the population) or a "bad sample" (very different than the population).

Next time, we will leverage all six classical assumptions, including **normality**, to conduct hypothesis tests.