Classical Assumptions

EC 320: Introduction to Econometrics

Winter 2022

Prologue

Housekeeping

- Problem Set 02 due today by 11:59 pm on Canvas
- The solution to the problem set will be released on Wednesday.
- Midterm grade appeal until tomorrow. Solution posted tomorrow. No appeals will be addressed after the solution is being posted.

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? "The proportion of the variance explained by the regression line"

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.

$$\log(\text{Earnings}_i) = \beta_0 + \beta_1 \text{Education}_i + u_i$$
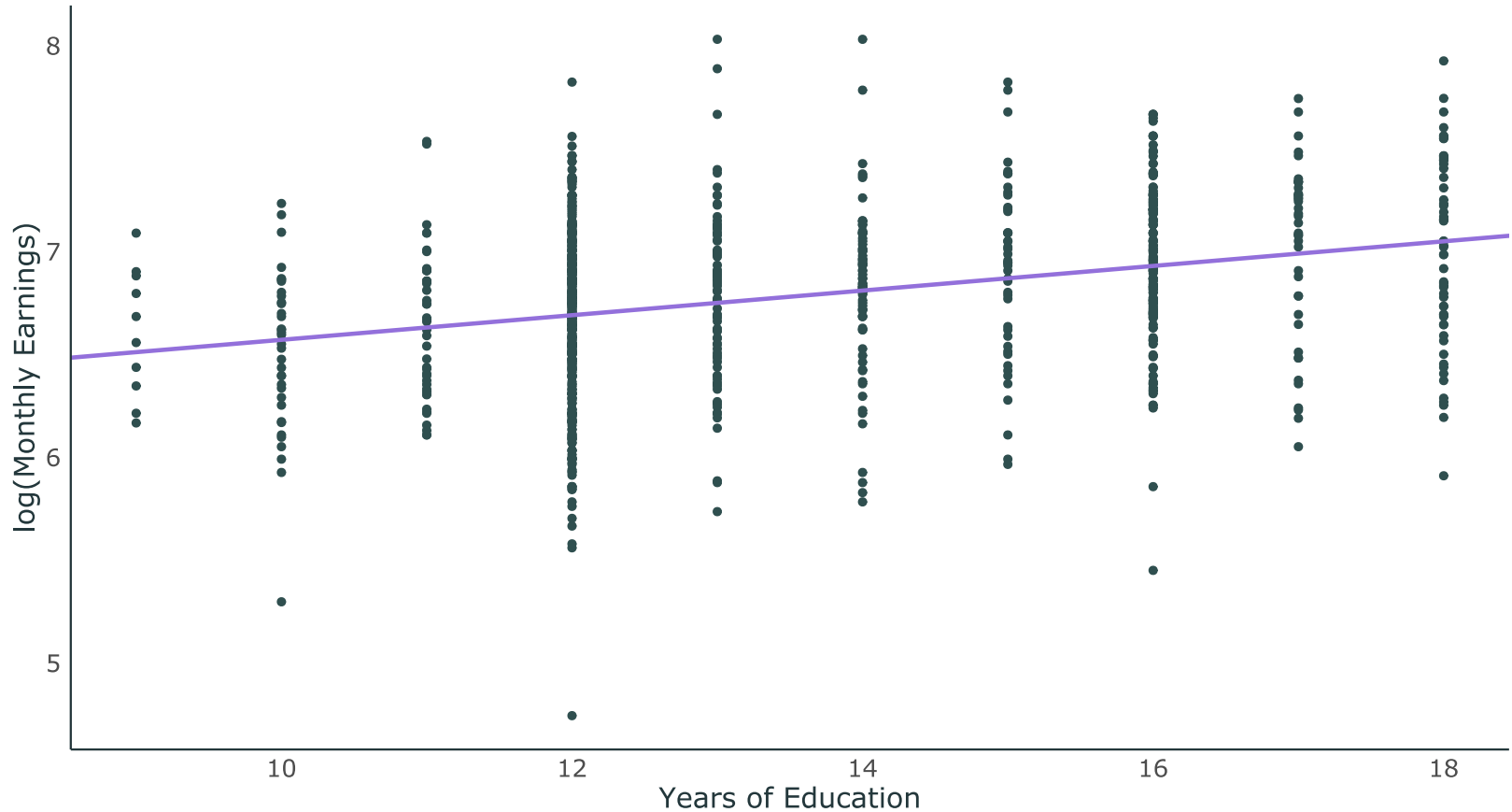
Step 2: Find data.

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

Step 3: Run a regression using OLS.

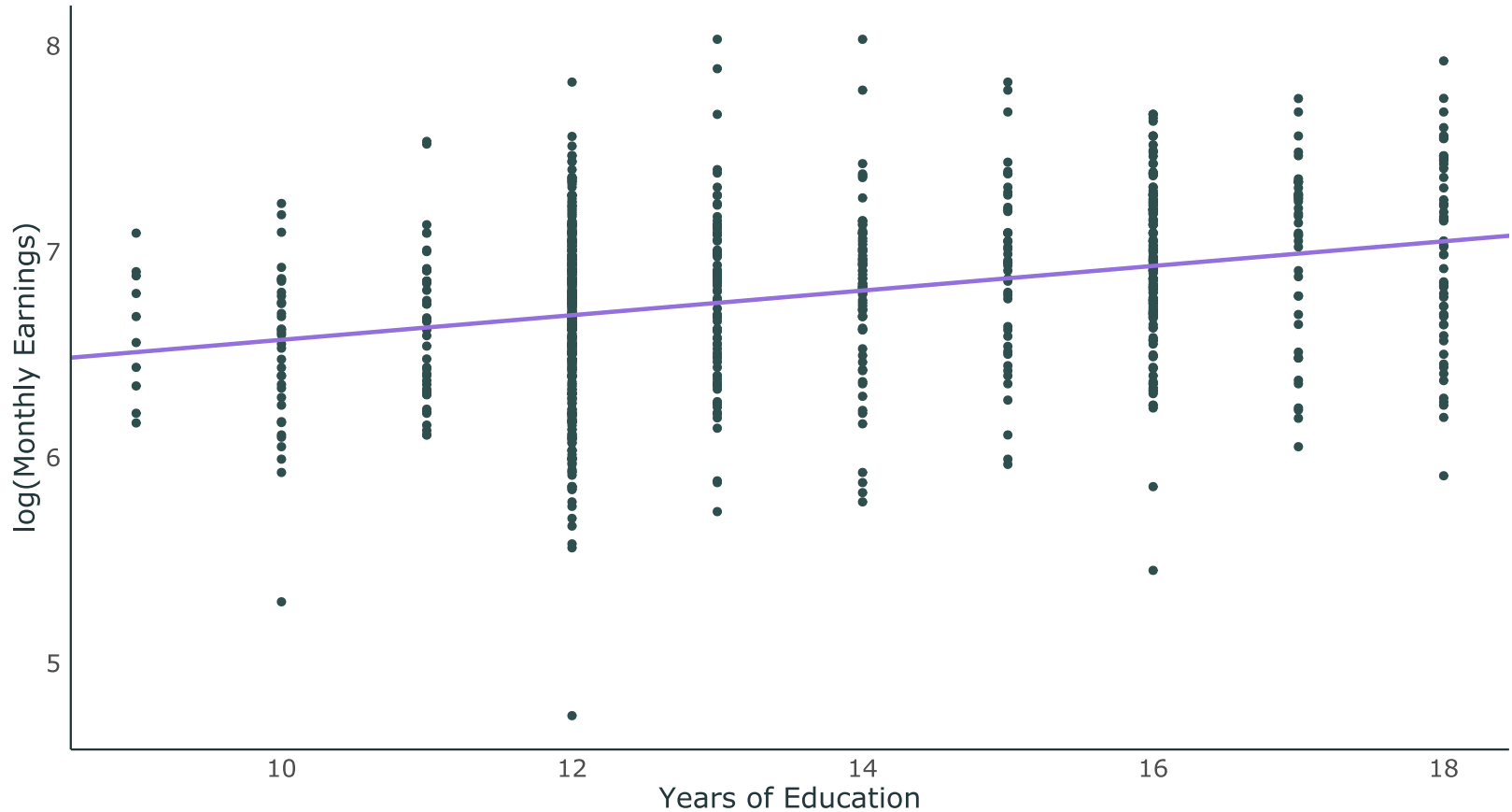$$\log(\hat{\text{Earnings}}_i) = \hat{\beta}_0 + \hat{\beta}_1 \text{Education}_i$$

Returns to Schooling

$$\log(\hat{\text{Earnings}}_i) = \mathbf{5.97} + \mathbf{0.06} \times \text{Education}_i.$$
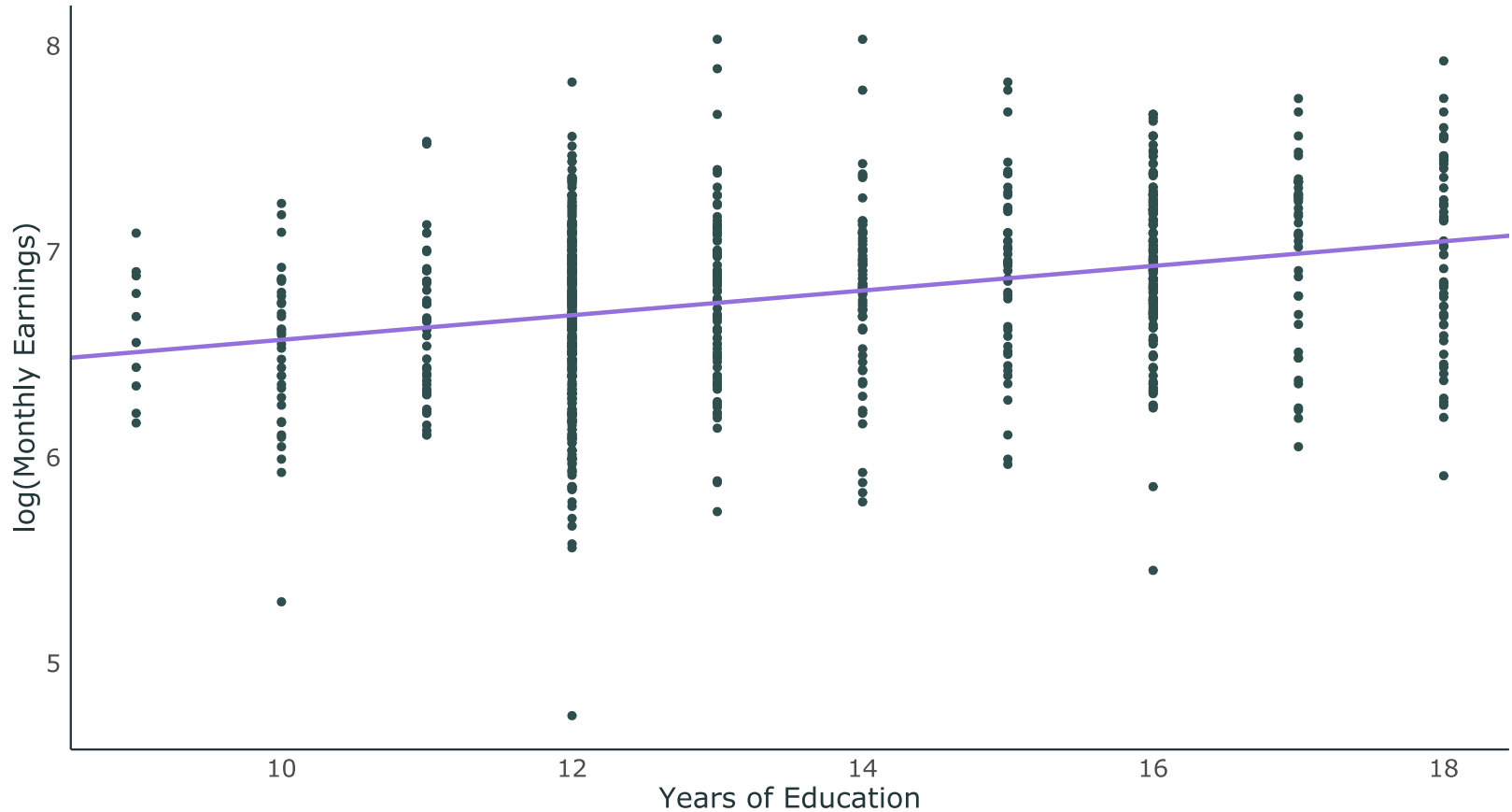
Returns to Schooling

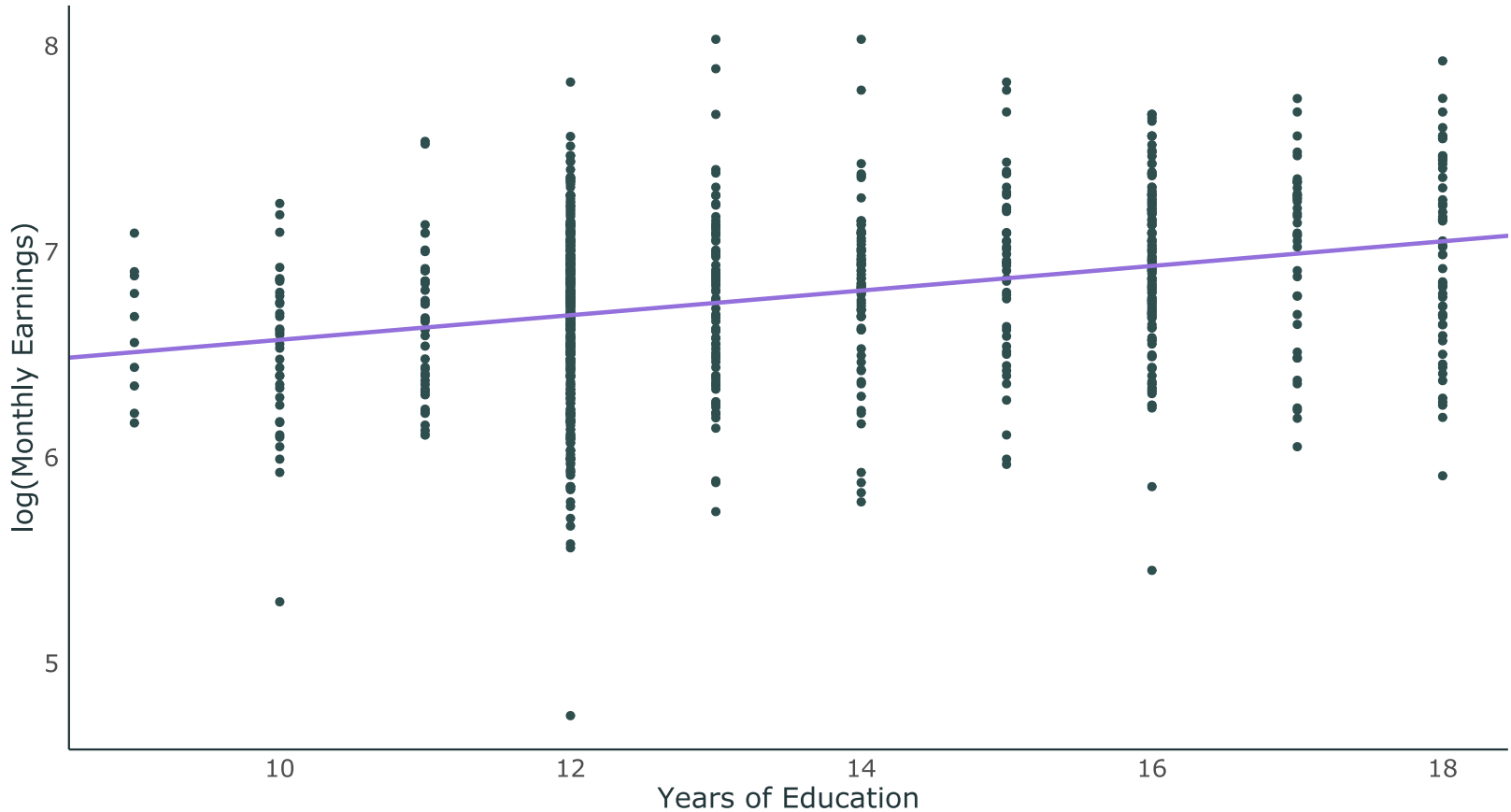Additional year of school associated with a **6%** increase in earnings.
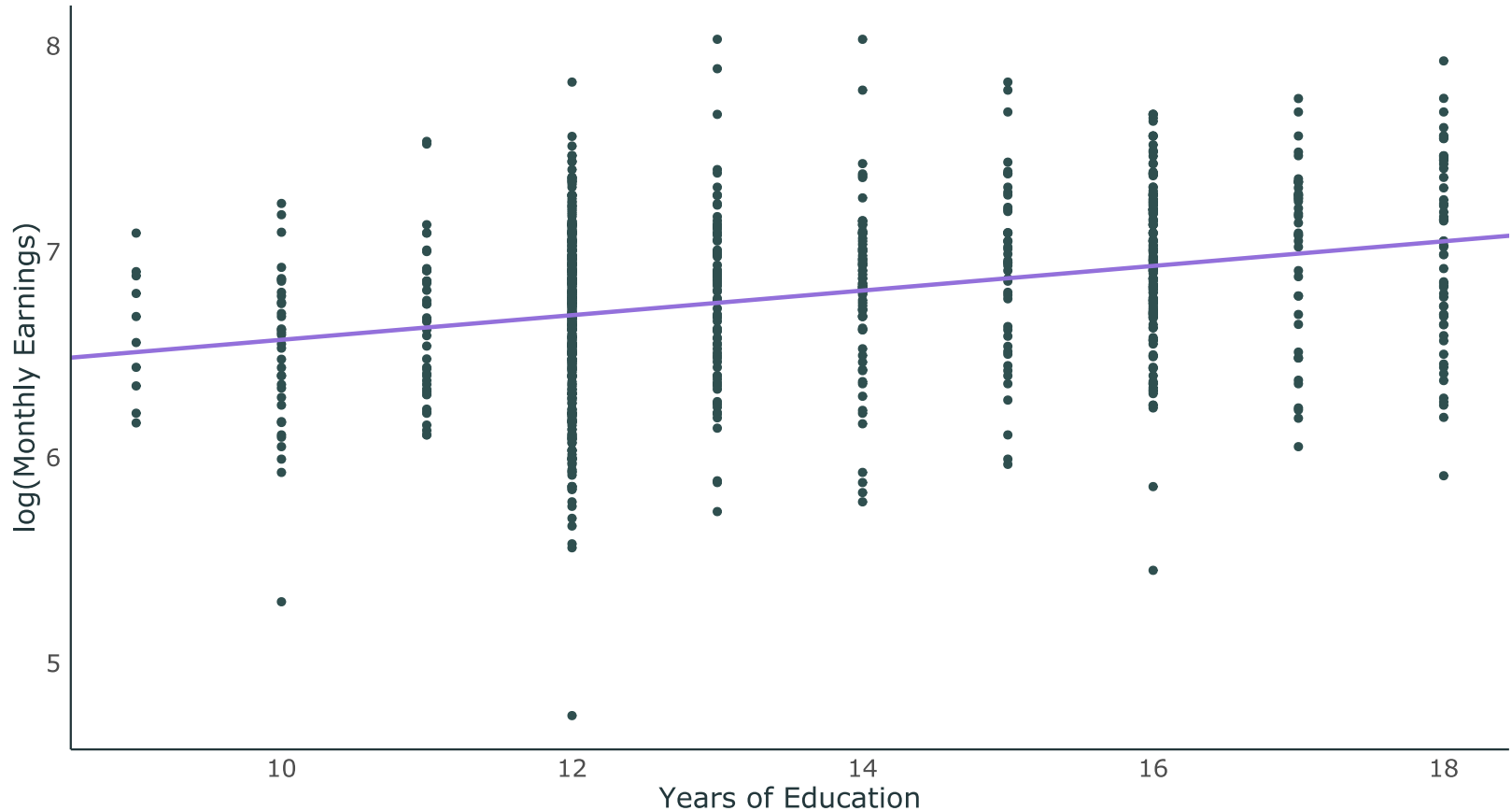
Returns to Schooling

$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}_1 = $ **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.

Population

- $Y_i = \beta_0 + \beta_1 X_i + u_i$
- **Error:** Difference between the wage of a worker with 16 years of education and the **expected wage** with 16 years of education.

Sample

- $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i$
- **Residual:** Difference between the wage of a worker with 16 years of education and the **average wage** of workers with 16 years of education.
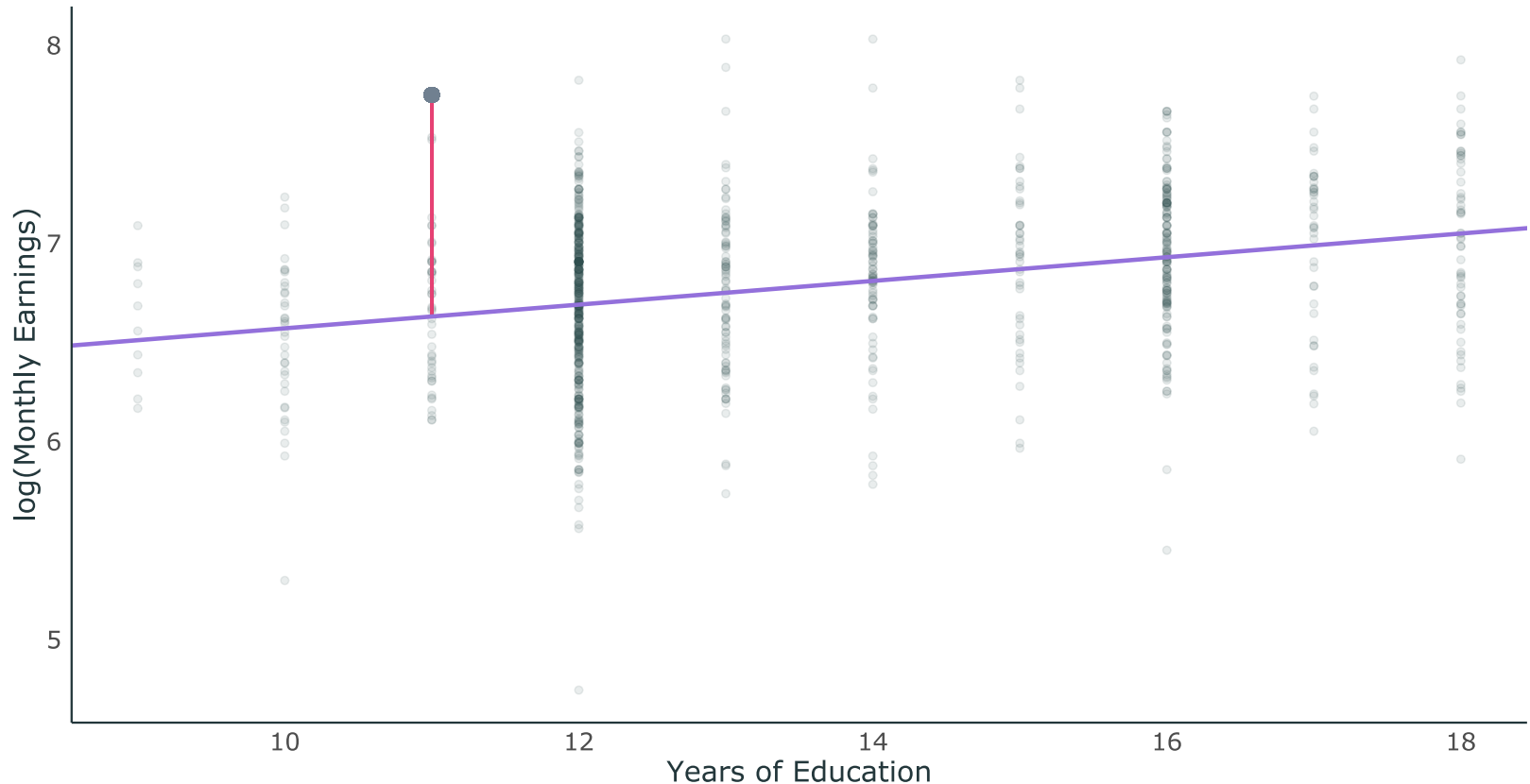
Residuals *vs.* Errors

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.
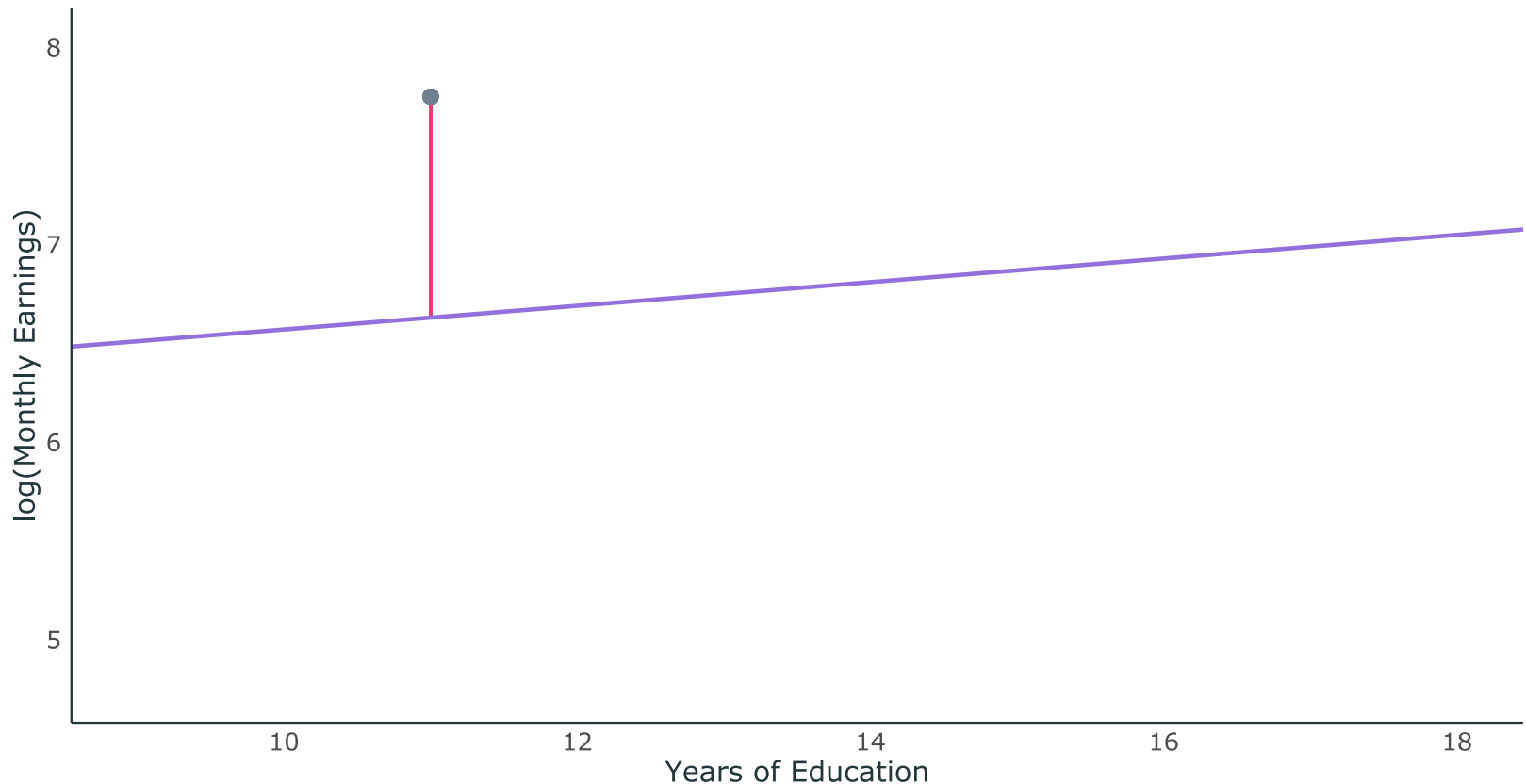
Residuals *vs.* Errors

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.
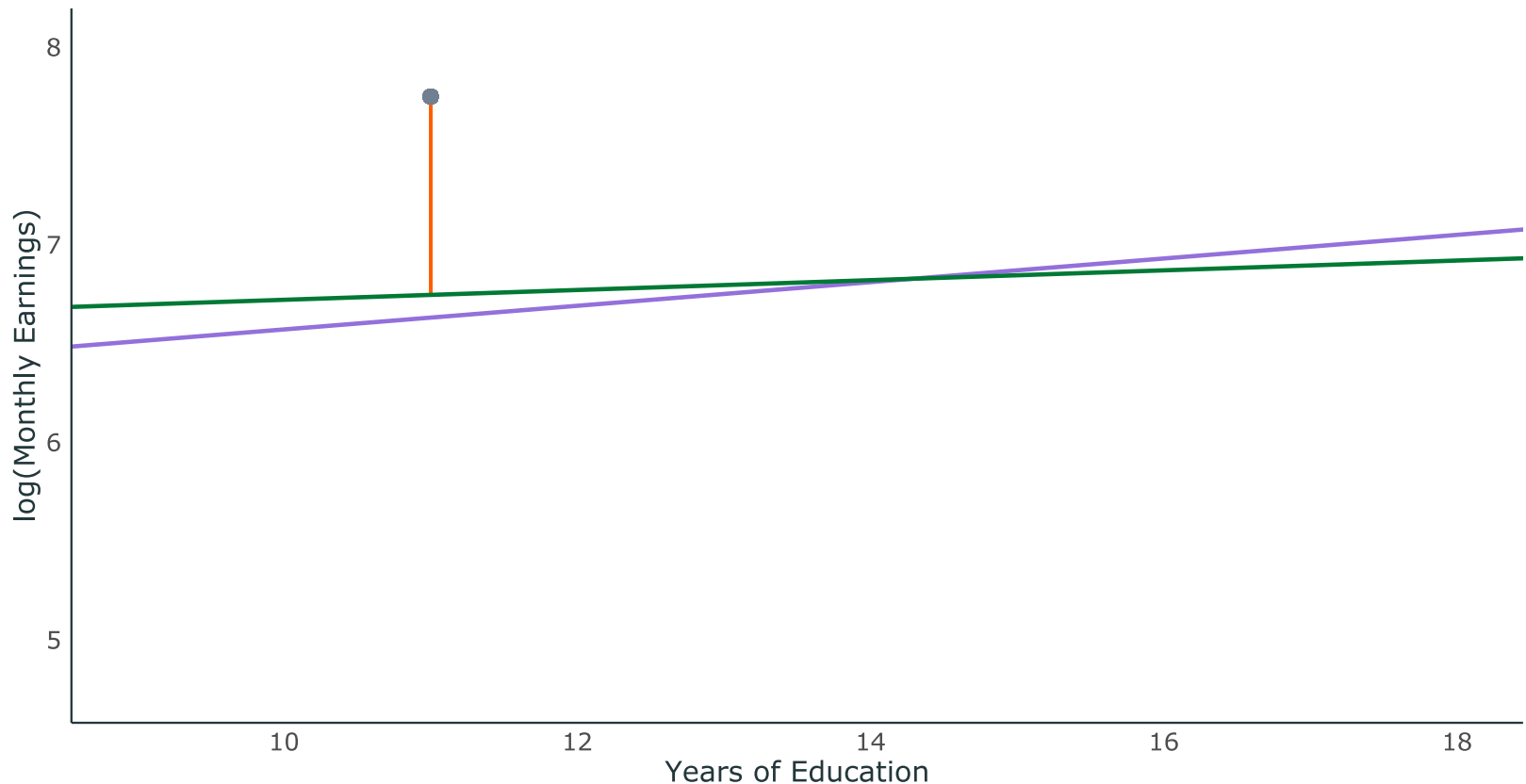
Residuals *vs.* Errors

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

- **A1. Linearity:** The population relationship is **linear in parameters** with an additive error term.
- **A2. Sample Variation:** There is variation in X.
- **A3. Exogeneity:** The X variable is **exogenous** (*i.e.*, $\mathbb{E}(u|X) = 0$)[†]
- **A4. Homoskedasticity:** The error term has the same variance for each value of the independent variable (*i.e.*, $\mathrm{Var}(u|X) = \sigma^2$).
- **A5. Non-autocorrelation:** The values of error terms are independent from one another (*i.e.*, $E[u_i u_j] = 0, \forall i \text{ s.t. } i \neq j$).[†]
- **A6. Normality:** The population error term is normally distributed with mean zero and variance σ^2 (*i.e.*, $u \sim N(0, \sigma^2)$)

[†] **Random Sampling:** Notice up until now, our underlying data type used for analysis was cross-sectional data. Under random sampling, this yields u_i and u_j independent for any two observations i and j, which is what non-autocorrelation assumption implies. However here I explicitly state **non-autocorrelation** to 1) generalize this case to account for a case we use time series data, 2) to be consistent with the notations from the textbook. Also it could be shown that the errors for different observations are independent conditional on the explanatory variables in the cross-sectional sample under random sampling.
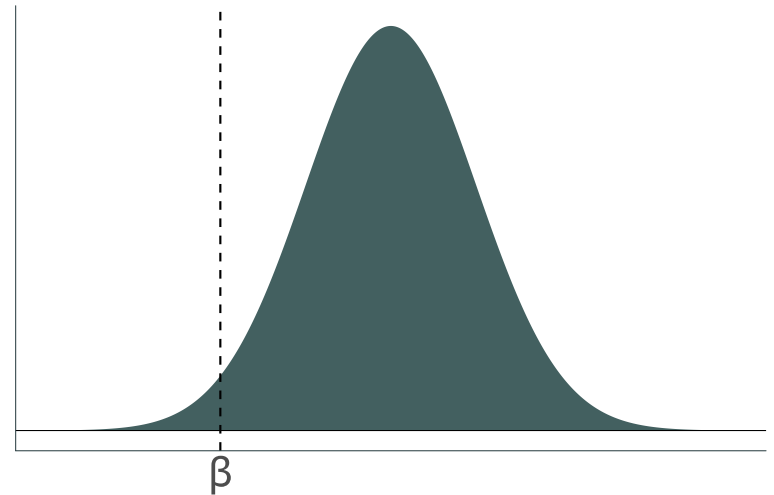
When Can We Trust OLS?

Bias

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

Unbiased estimator: $\mathbb{E}\left[\hat{\beta}\right] = \beta$
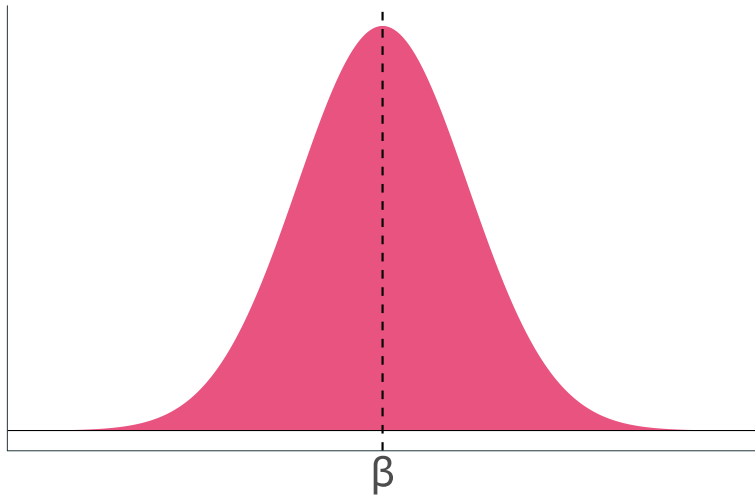
Biased estimator: $\mathbb{E}\left[\hat{\beta}\right] \neq \beta$

When is OLS Unbiased?

Assumptions

A1. Linearity: The population relationship is **linear in parameters** with an additive error term.

A2. Sample Variation: There is variation in X.

A3. Exogeneity: The X variable is **exogenous** (*i.e.*, $\mathbb{E}(u|X) = 0$). [†]

Result

OLS is unbiased.

[†] Should A3 be true, random sampling should be performed in cross sectional data, meaning that A3 implies random sampling in a sense.

Linearity

Assumption

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

Examples

- $\text{Wage}_i = \beta_0 + \beta_1 \text{Experience}_i + u_i$

- $\log(\text{Happiness}_i) = \beta_0 + \beta_1 \log(\text{Money}_i) + u_i$

- $\sqrt{\text{Convictions}_i} = \beta_0 + \beta_1 (\text{Early Childhood Lead Exposure})_i + u_i$

- $\log(\text{Earnings}_i) = \beta_0 + \beta_1 \text{Education}_i + u_i$

Linearity

Assumption

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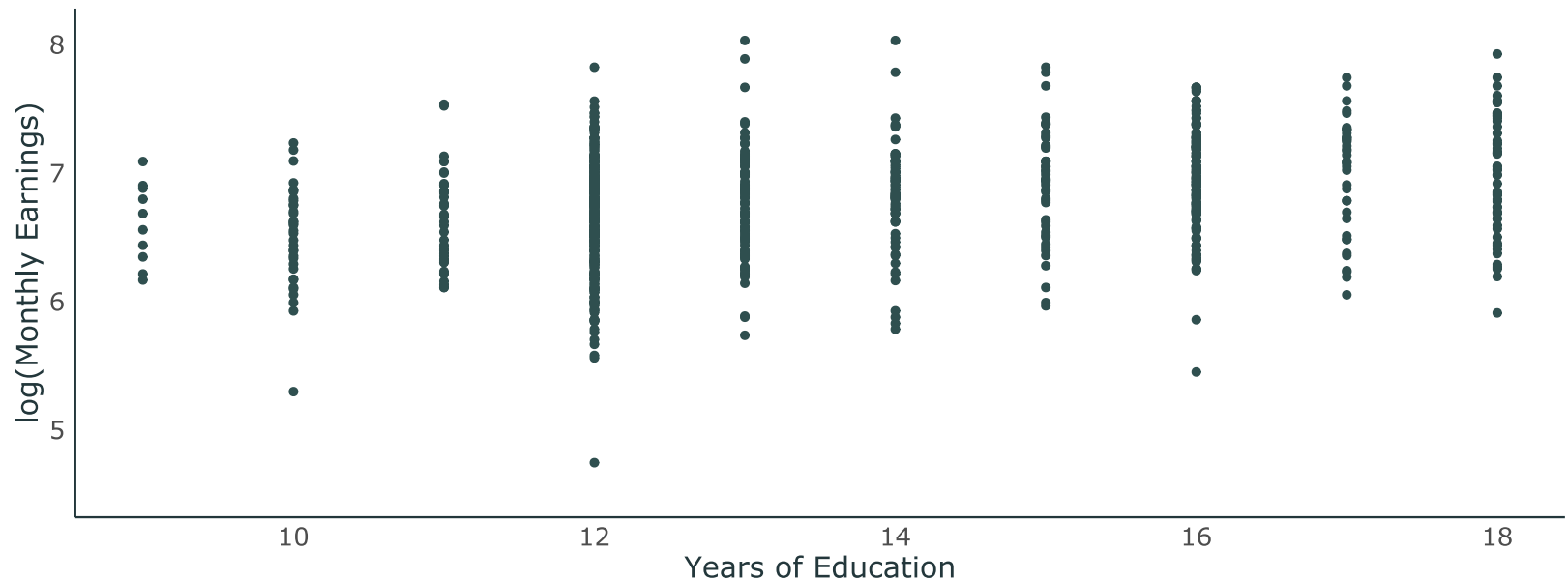
Violations

- $\text{Wage}_i = (\beta_0 + \beta_1 \text{Experience}_i) u_i$

- $\text{Consumption}_i = \dfrac{1}{\beta_0 + \beta_1 \text{Income}_i} + u_i$

- $\text{Population}_i = \dfrac{\beta_0}{1 + e^{\beta_1 + \beta_3 \text{Food}_i}} + u_i$

- $\text{Batting Average}_i = \beta_0 (\text{Wheaties Consumption})_i^{\beta_1} + u_i$

Sample Variation

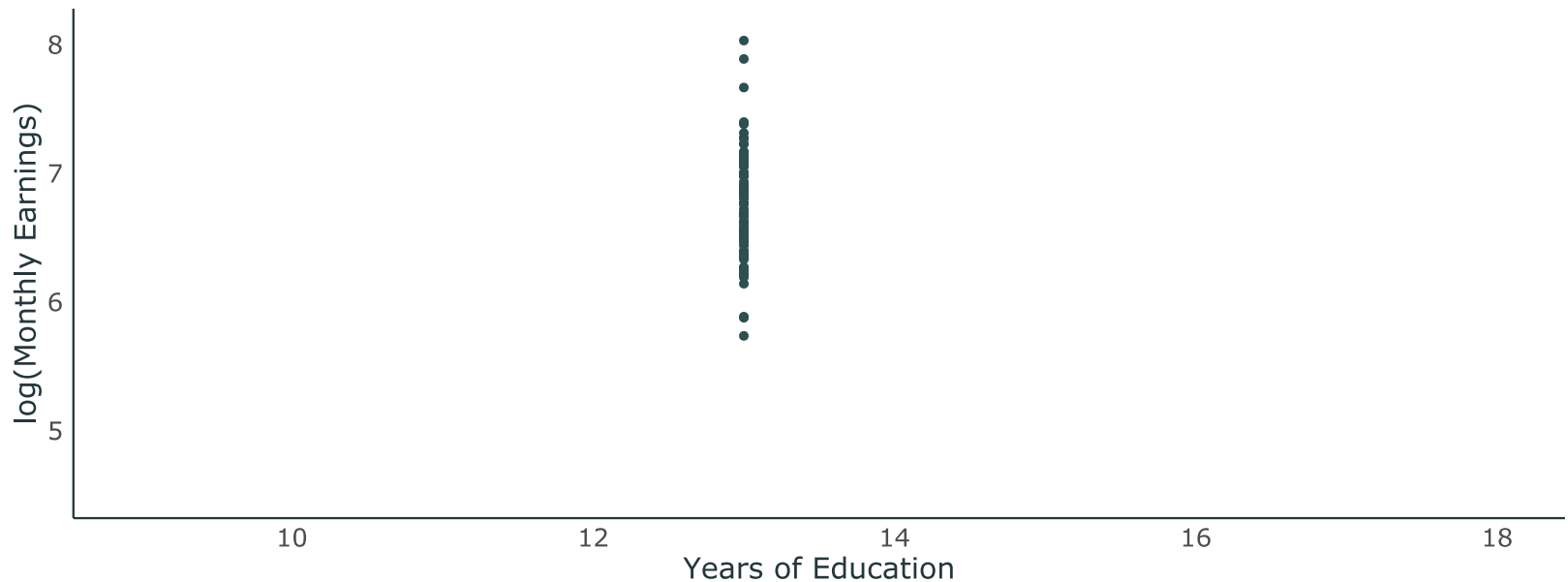## Assumption

There is variation in X.

Example

Sample Variation

Assumption

There is variation in X.

Violation

Exogeneity

Assumption

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

- 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: $\mathbb{E}(u) = 0$.

2. The mean of the error term is the same for each value of X:
   $\mathbb{E}(u|X) = \mathbb{E}(u)$.

Exogeneity

Assumption

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

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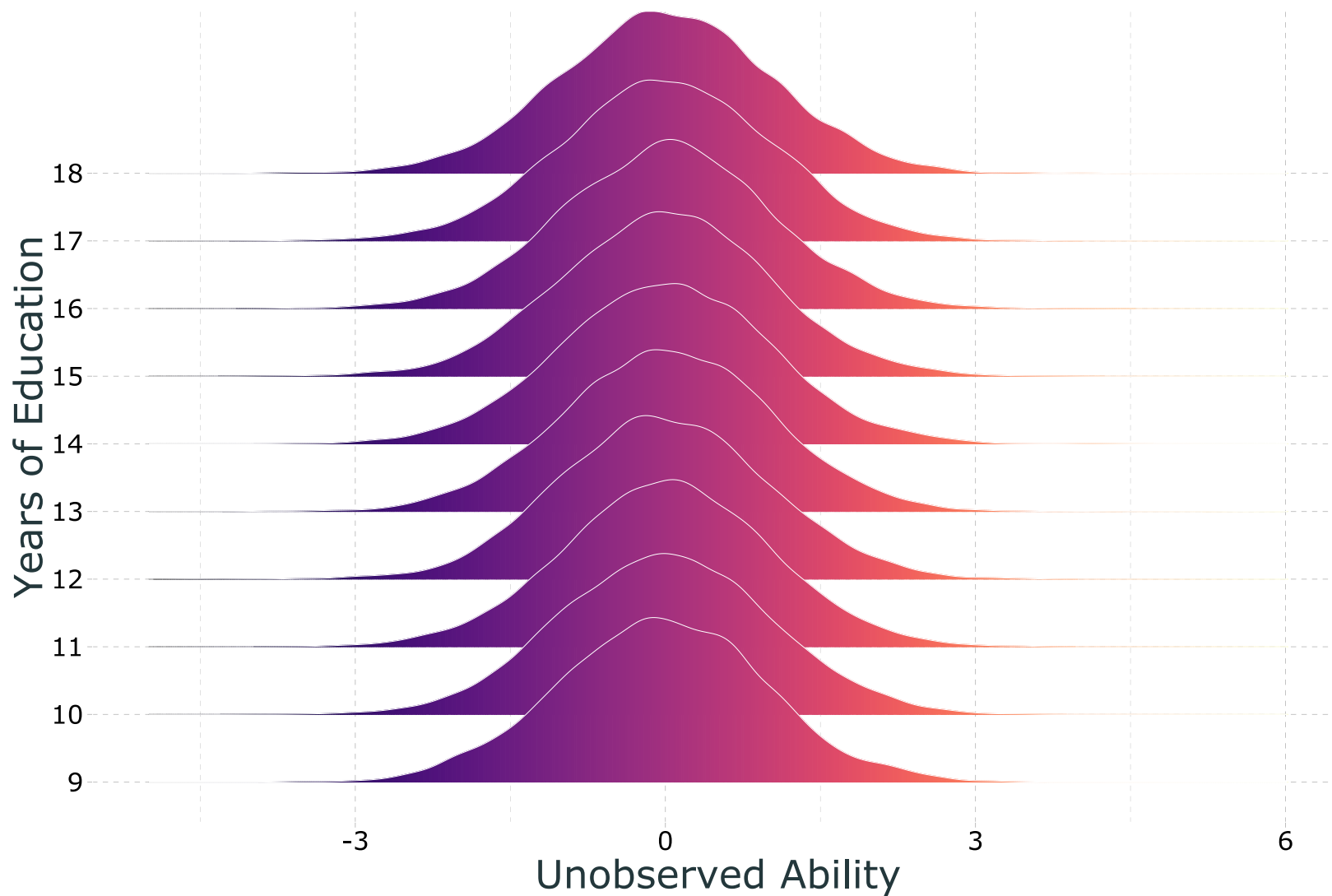- **Implication: no selection bias** and **no omitted-variable bias**.

Examples

In the labor market, an important component of u is unobserved ability.
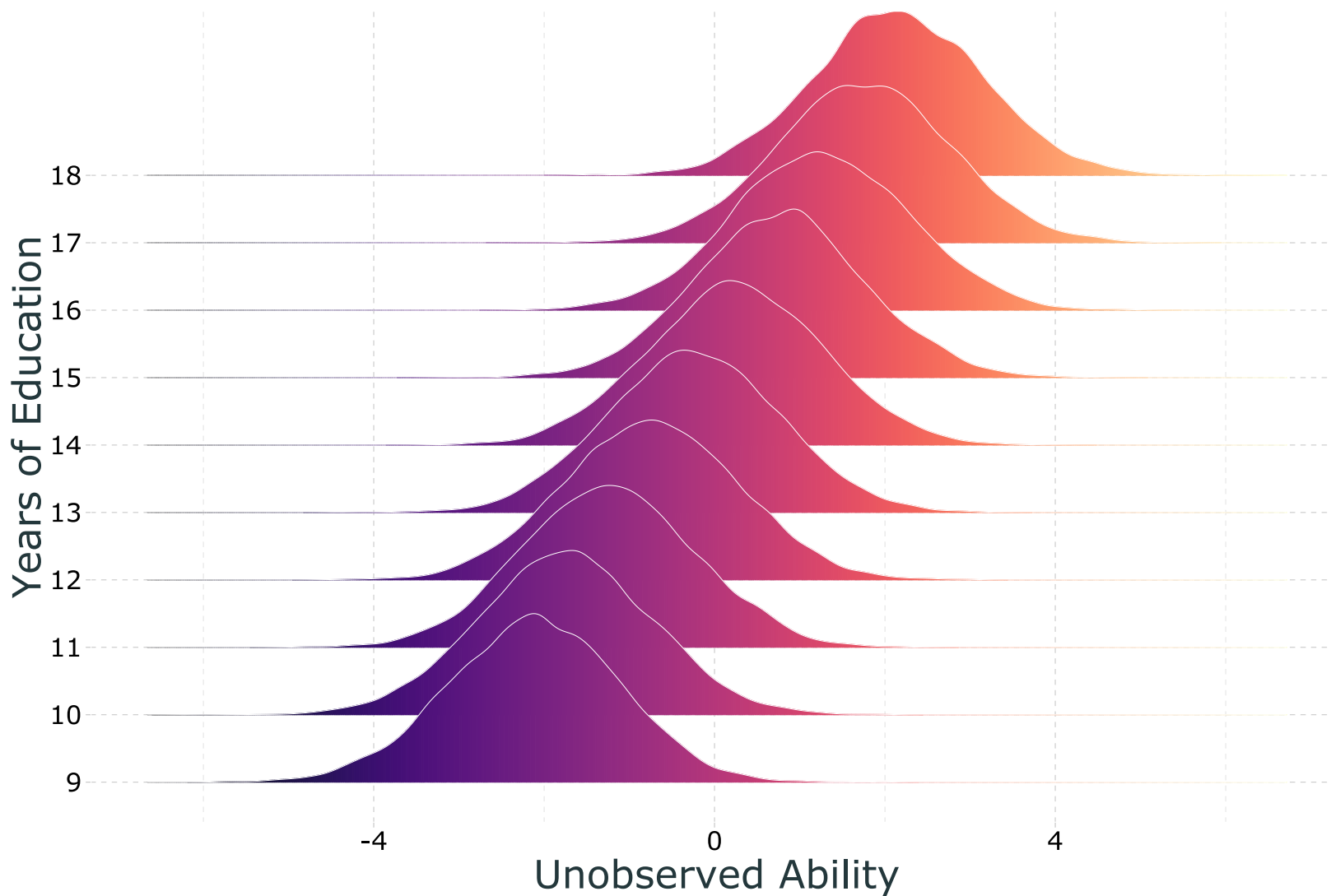
- $\mathbb{E}(u|\text{Education} = 12) = 0$ and $\mathbb{E}(u|\text{Education} = 20) = 0$.
- $\mathbb{E}(u|\text{Experience} = 0) = 0$ and $\mathbb{E}(u|\text{Experience} = 40) = 0$.
- **Do you believe this?**

Graphically...

Invalid exogeneity, *i.e.*, $\mathbb{E}(u \mid X) \neq 0$

Years of Education

Unobserved Ability

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 parameter.

- 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.

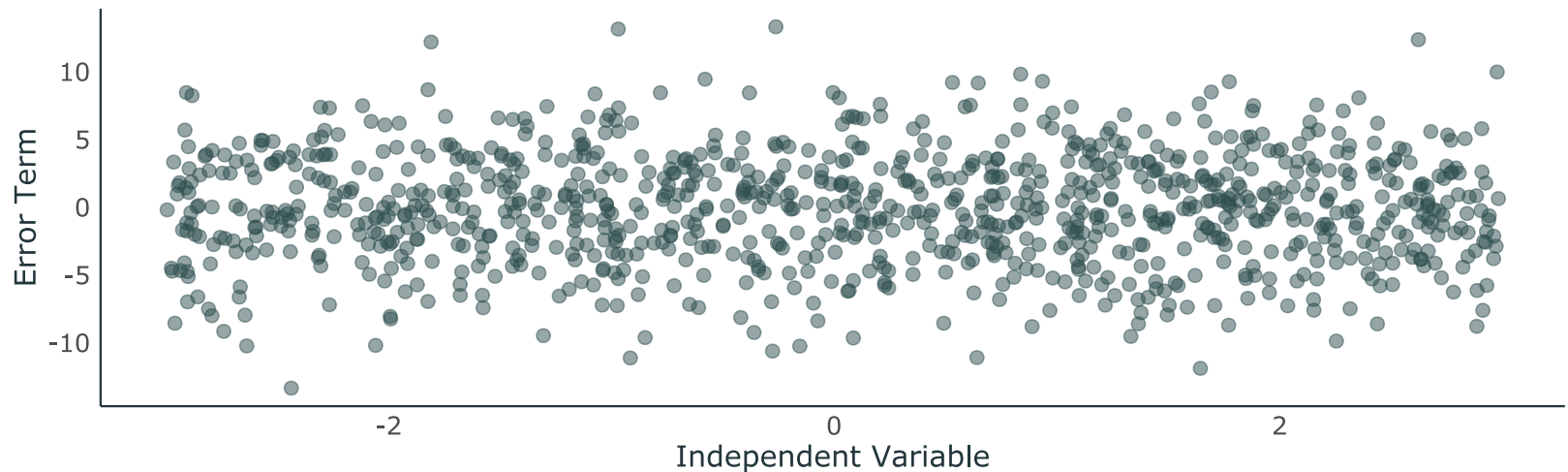2. **Homoskedasticity**

3. **Non-autocorrelation**

Homoskedasticity

Assumption

The error term has the same variance for each value of the independent variable:
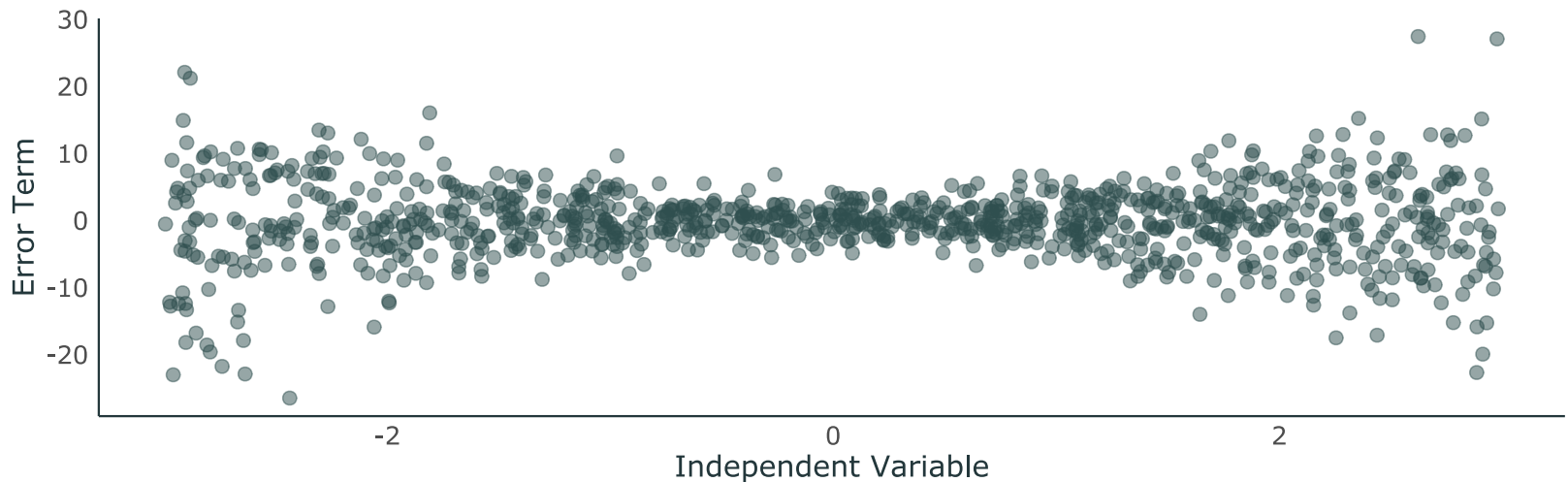
$$\mathrm{Var}(u|X) = \sigma^2.$$

Example

Homoskedasticity

Assumption

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

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

Violation: Heteroskedasticity

Non-Autocorrelation

Assumption

The population covariance between u_i and u_j is zero, meaning that any individual's error term is drawn independently of other error terms.

$$\text{Cov}(u_i, u_j) = E[(u_i - \mu_u)(u_j - \mu_u)]$$
$$= E[u_i u_j] = E[u_i]E[u_j] = 0, \text{where } i \neq j$$

- This implies no systematic association between error term values for any pair of individuals. The values of the error term should be independent of one another. If this assumption is not satisfied, OLS gives inefficient estimates.
- **Example:** The magnitude and the sign of disturbance term in one observation should not lead to a tendency of determining magnitude and the sign of the disturbance term of the other.
- **Violation:** Errors that are correlated with time (autocorrelation)

OLS Variance

Variance of the slope estimator:

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

- 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 $X \implies \text{Var}(\hat{\beta}_1)$ falls as n rises.

Gauss-Markov

Gauss-Markov Theorem

OLS is the **Best Linear Unbiased Estimator** under assumptions A1-A5:

- **A1. Linearity:** The population relationship is **linear in parameters** with an additive error term.

- **A2. Sample Variation:** There is variation in X.

- **A3. Exogeneity:** The X variable is **exogenous** (*i.e.*, $\mathbb{E}(u|X) = 0$).

- **A4. Homoskedasticity:** The error term has the same variance for each value of the independent variable (*i.e.*, $\mathrm{Var}(u|X) = \sigma^2$).

- **A5. Non-Autocorrelation:** Any pair of error terms are drawn independently of each other (*i.e.*, $\mathrm{E}(u_i u_j) = 0 \ \forall \ i \ \text{s.t.} \ i \neq j$)
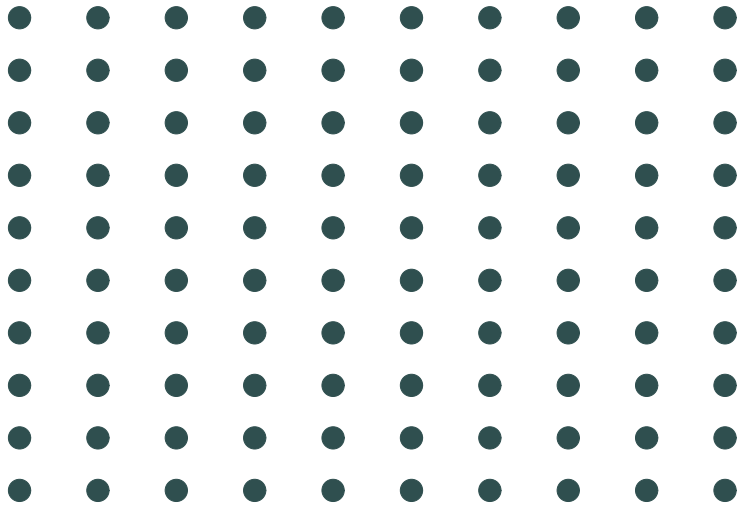
Gauss-Markov Theorem
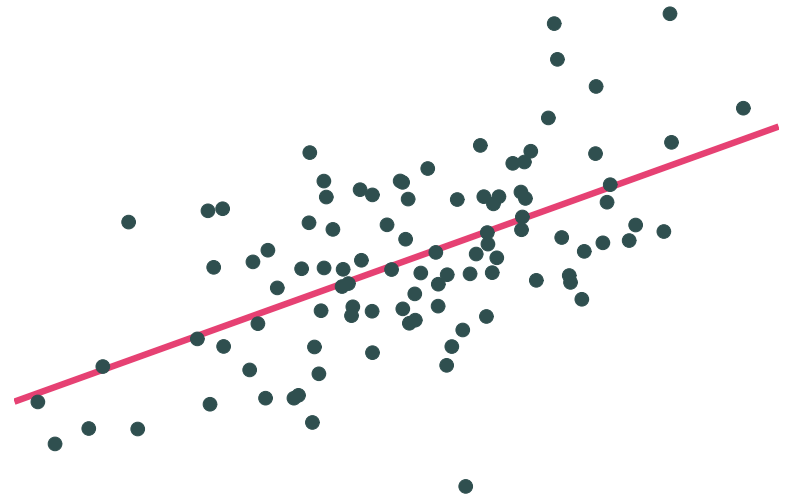
OLS is the **Best Linear Unbiased Estimator (BLUE)**

Population *vs.* Sample, Revisited

Population *vs.* Sample

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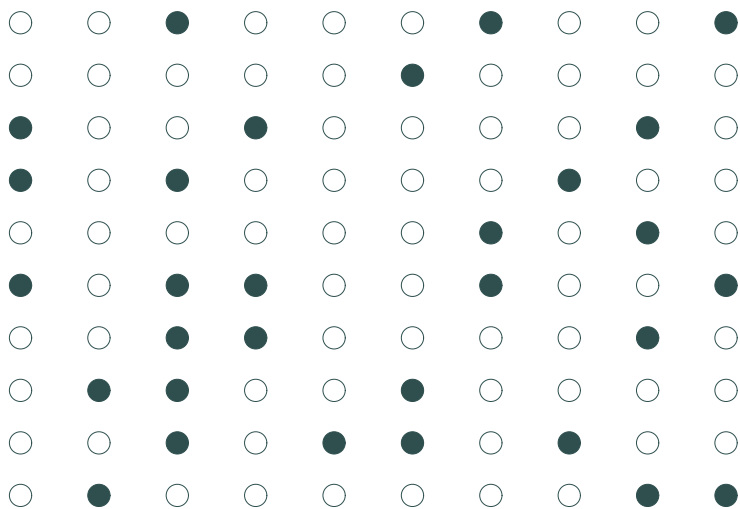

Population

Population relationship

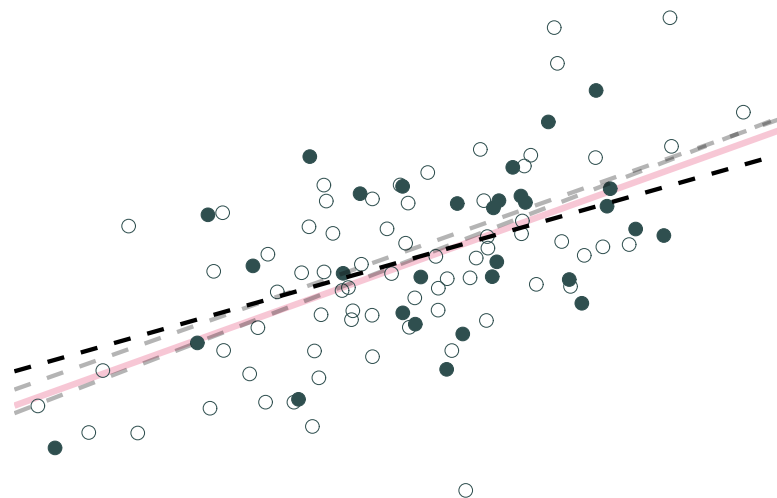$$y_i = 2.53 + 0.57x_i + u_i$$

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

Population *vs.* Sample

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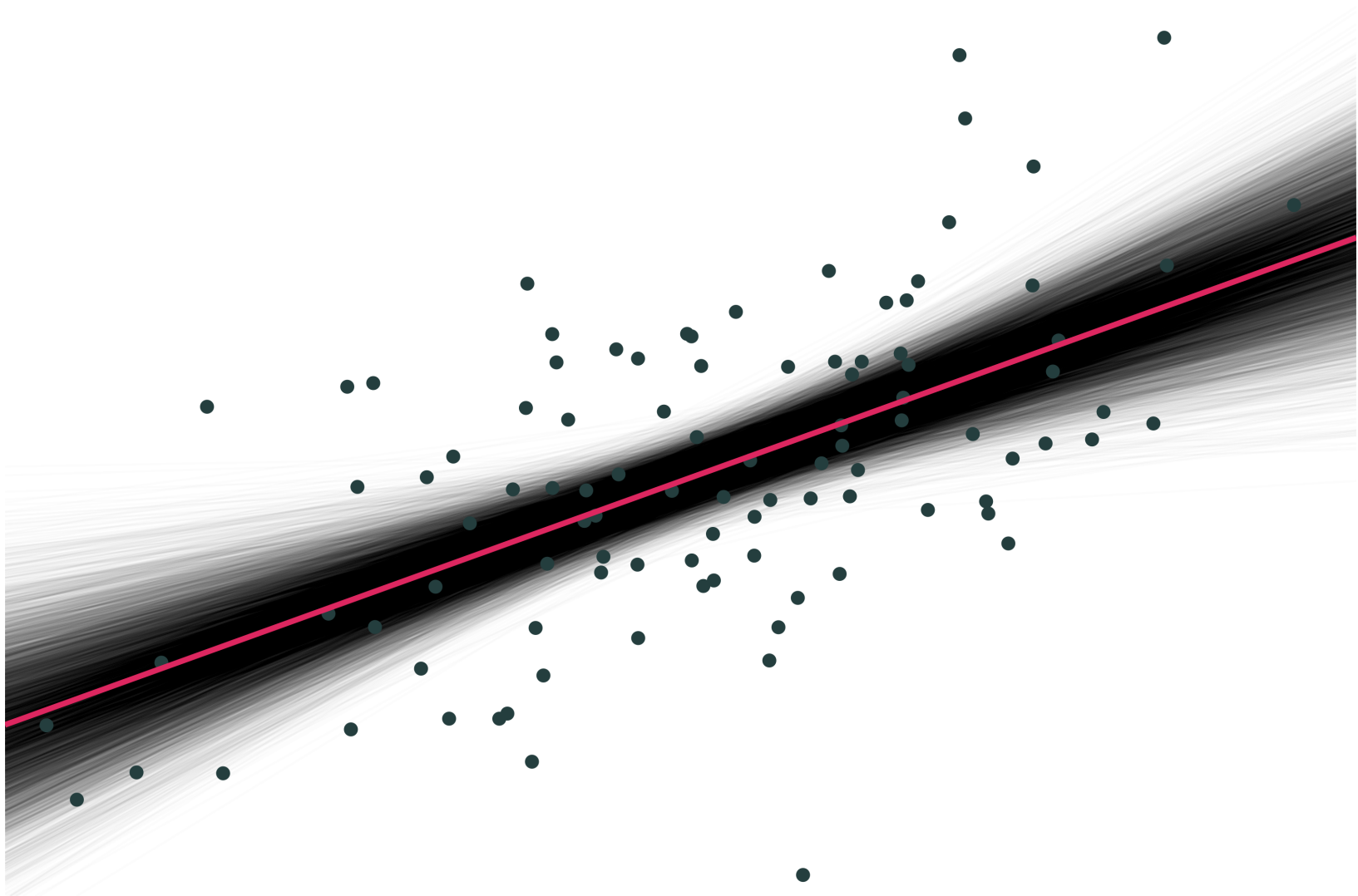

Sample 3: 30 random individuals

Population relationship

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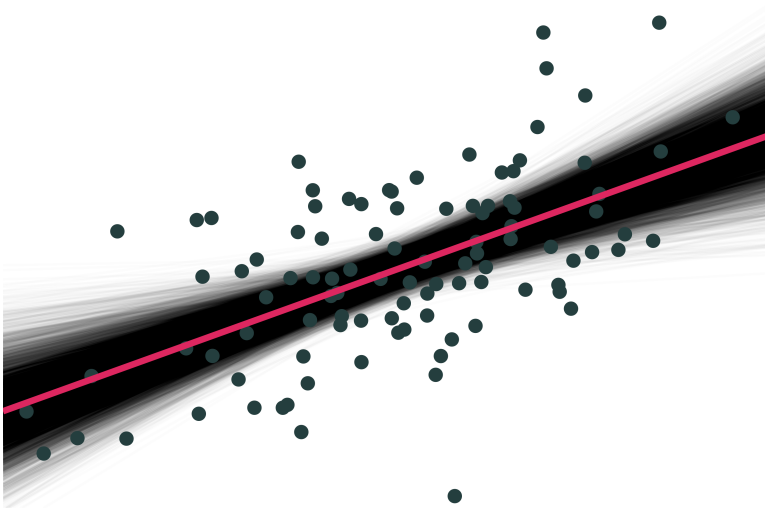
Sample relationship

$$\hat{y}_i = 3.21 + 0.45x_i$$

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.

$\hat{\beta}_0$ and $\hat{\beta}_1$ 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.