Heteroskedasticity, Part II

EC 421, Set 05

Edward Rubin

R showcase

Posit Cloud

- Cloud-based RStudio IDE.
- Free tier available. Also student plans.
- No installation of R/RStudio required.
- Easily share projects.
- Runs in your browser. Code on your phone!

Prologue

Schedule

Last Time

Heteroskedasticity: Issues and tests

Today

- **First assignment** due today
- Living with heteroskedasticity

Upcoming

- **Second assignment** released soon.

Goals

- Develop **intuition** for econometrics.
- Learn how to ***apply*** econometrics—strengths, weaknessed, *etc.*
- Learn **R**.

Goals

- Develop **intuition** for econometrics.
- Learn how to ***apply*** econometrics—strengths, weaknessed, *etc.*
- Learn **R**.

R does the calculations and has already memorized the formulas.

I want you to know what the formulas mean, when/why we use them, and when they fail/work.

Goals

- Develop **intuition** for econometrics.
- Learn how to ***apply*** econometrics—strengths, weaknessed, *etc.*
- Learn **R**.

R does the calculations and has already memorized the formulas.

I want you to know what the formulas mean, when/why we use them, and when they fail/work.

This course has the potential to be one of the most useful/valuable/applicable/marketable classes that you take at UO.

Heteroskedasticity

Review

Review

Three review questions

Question 1: What is the difference between u_i and e_i?

Question 2: We spend *a lot* of time discussing u_i^2. Why?

Question 3: We also spend *a lot* of time discussing e_i^2. Why?

Review

Question 1: What is the difference between u_i and e_i?

Answer 1:

Review

Question 1: What is the difference between u_i and e_i?

Answer 1:

u_i gives the **population disturbance** for the i^{th} observation.

Review

Question 1: What is the difference between u_i and e_i?

Answer 1:

u_i gives the **population disturbance** for the i^{th} observation. u_i measures how far the i^{th} observation is from the **population** line, *i.e.*,

$$u_i = y_i - \underbrace{(\beta_0 + \beta_1 x_i)}_{\text{Population line}}$$

Heteroskedasticity

Review

Question 1: What is the difference between u_i and e_i?

Answer 1:

u_i gives the **population disturbance** for the i^{th} observation. u_i measures how far the i^{th} observation is from the **population** line, *i.e.*,

$$u_i = y_i - \underbrace{(\beta_0 + \beta_1 x_i)}_{\text{Population line}}$$

e_i gives the **regression residual (error)** for the i^{th} observation.

Review

Question 1: What is the difference between u_i and e_i?

Answer 1:

u_i gives the **population disturbance** for the i^{th} observation. u_i measures how far the i^{th} observation is from the **population** line, *i.e.*,

$$u_i = y_i - \underbrace{(\beta_0 + \beta_1 x_i)}_{\text{Population line}}$$

e_i gives the **regression residual (error)** for the i^{th} observation. e_i measures how far the i^{th} observation is from the **sample regression** line, *i.e.*,

$$e_i = y_i - \underbrace{\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)}_{\text{Sample reg. line}=\hat{y}} = y_i - \hat{y}_i$$

Heteroskedasticity

Review

Question 2: We spend *a lot* of time discussing $\boldsymbol{u_i^2}$. Why?

Answer 2:

One of major assumptions is that our disturbances (the $\boldsymbol{u_i}$'s) are homoskedastic (they have constant variance), *i.e.*, $\mathbf{Var}(\boldsymbol{u_i}|\boldsymbol{x_i}) = \boldsymbol{\sigma^2}$.

We also assume that the mean of these disturbances is zero, $\boldsymbol{E}[u_i|x_i] = 0$.

By definition, $\mathbf{Var}(u_i|x_i) = \boldsymbol{E}\left[u_i^2 - \underbrace{\boldsymbol{E}[u_i|x_i]}_{=0}^2 \middle| x_i\right] = \boldsymbol{E}\left[u_i^2 \middle| x_i\right]$

Thus, if we want to learn about the variance of $\boldsymbol{u_i}$, we can focus on $\boldsymbol{u_i^2}$.

Review

Question 3: We also spend *a lot* of time discussing e_i^2. Why?

Answer 3:

We cannot observe u_i (or u_i^2).

But u_i^2 tells us about the variance of u_i.

We use e_i^2 to learn about u_i^2 and, consequently, σ_i^2.

Review: Current assumptions

Review: Current assumptions

1. Our sample (the x_k's and y_i) was **randomly drawn** from the population.

Review: Current assumptions

1. Our sample (the x_k's and y_i) was **randomly drawn** from the population.

2. y is a **linear function** of the β_k's and u_i.

Review: Current assumptions

1. Our sample (the x_k's and y_i) was **randomly drawn** from the population.

2. y is a **linear function** of the β_k's and u_i.

3. There is no perfect **multicollinearity** in our sample.

Heteroskedasticity

Review: Current assumptions

1. Our sample (the x_k's and y_i) was **randomly drawn** from the population.

2. y is a **linear function** of the β_k's and u_i.

3. There is no perfect **multicollinearity** in our sample.

4. The explanatory variables are **exogenous**: $\boldsymbol{E}[u|X] = 0$ ($\implies \boldsymbol{E}[u] = 0$).

Review: Current assumptions

1. Our sample (the x_k's and y_i) was **randomly drawn** from the population.

2. y is a **linear function** of the β_k's and u_i.

3. There is no perfect **multicollinearity** in our sample.

4. The explanatory variables are **exogenous**: $\boldsymbol{E}[u|X] = 0 \; (\implies \boldsymbol{E}[u] = 0)$.

5. The disurbances have **constant variance** σ^2 and **zero covariance**, *i.e.*,
   - $\boldsymbol{E}\left[u_i^2|X_i\right] = \text{Var}(u_i|X_i) = \sigma^2 \implies \text{Var}(u_i) = \sigma^2$
   - $\text{Cov}\left(u_i,\, u_j|X_i,\, X_j\right) = \boldsymbol{E}\left[u_i u_j|X_i,\, X_j\right] = 0$ for $i \neq j$

Review: Current assumptions

1. Our sample (the x_k's and y_i) was **randomly drawn** from the population.

2. y is a **linear function** of the β_k's and u_i.

3. There is no perfect **multicollinearity** in our sample.

4. The explanatory variables are **exogenous**: $\boldsymbol{E}[u|X] = 0 \; (\implies \boldsymbol{E}[u] = 0)$.

5. The disurbances have **constant variance** σ^2 and **zero covariance**, *i.e.*,
   - $\boldsymbol{E}\left[u_i^2|X_i\right] = \text{Var}(u_i|X_i) = \sigma^2 \implies \text{Var}(u_i) = \sigma^2$
   - $\text{Cov}\left(u_i,\, u_j|X_i,\, X_j\right) = \boldsymbol{E}\left[u_i u_j|X_i,\, X_j\right] = 0$ for $i \neq j$

6. The disturbances come from a **Normal** distribution, *i.e.*, $u_i \overset{\text{iid}}{\sim} \text{N}\left(0, \sigma^2\right)$.

Heteroskedasticity

Review

Today we're focusing on assumption #5:

> 5. The disurbances have **constant variance** σ^2 and **zero covariance**, *i.e.*,
>
> - $\boldsymbol{E}\left[u_i^2|X_i\right] = \text{Var}(u_i|X) = \sigma^2 \implies \text{Var}(u_i) = \sigma^2$
> - $\text{Cov}\left(u_i,\, u_j|X_i,\, X_j\right) = \boldsymbol{E}\left[u_i u_j|X_i,\, X_j\right] = 0$ for $i \neq j$

Heteroskedasticity

Review

Today we're focusing on assumption #5:

> 5. The disurbances have **constant variance** σ^2 and **zero covariance**, *i.e.*,
>
> - $\boldsymbol{E}\left[u_i^2 \middle| X_i\right] = \mathrm{Var}(u_i|X) = \sigma^2 \implies \mathrm{Var}(u_i) = \sigma^2$
> - $\mathrm{Cov}\left(u_i,\, u_j \middle| X_i,\, X_j\right) = \boldsymbol{E}\left[u_i u_j \middle| X_i,\, X_j\right] = 0$ for $i \neq j$

Specifically, we will focus on the assumption of **constant variance** (also known as *homoskedasticity*).

Heteroskedasticity

Review

Today we're focusing on assumption #5:

> 5. The disurbances have **constant variance** σ^2 and **zero covariance**, *i.e.*,
>
> - $\boldsymbol{E}\left[u_i^2 \middle| X_i\right] = \mathrm{Var}(u_i|X) = \sigma^2 \implies \mathrm{Var}(u_i) = \sigma^2$
> - $\mathrm{Cov}\left(u_i,\, u_j \middle| X_i,\, X_j\right) = \boldsymbol{E}\left[u_i u_j \middle| X_i,\, X_j\right] = 0$ for $i \neq j$

Specifically, we will focus on the assumption of **constant variance** (also known as *homoskedasticity*).

Violation of this assumption: Our disturbances have different variances.

Heteroskedasticity

Review

Today we're focusing on assumption #5:

> 5. The disurbances have **constant variance** σ^2 and **zero covariance**, *i.e.*,
>
> - $\boldsymbol{E}\left[u_i^2 \middle| X_i\right] = \text{Var}(u_i|X) = \sigma^2 \implies \text{Var}(u_i) = \sigma^2$
> - $\text{Cov}\left(u_i,\, u_j \middle| X_i,\, X_j\right) = \boldsymbol{E}\left[u_i u_j \middle| X_i,\, X_j\right] = 0$ for $i \neq j$

Specifically, we will focus on the assumption of **constant variance** (also known as *homoskedasticity*).

Violation of this assumption: Our disturbances have different variances.

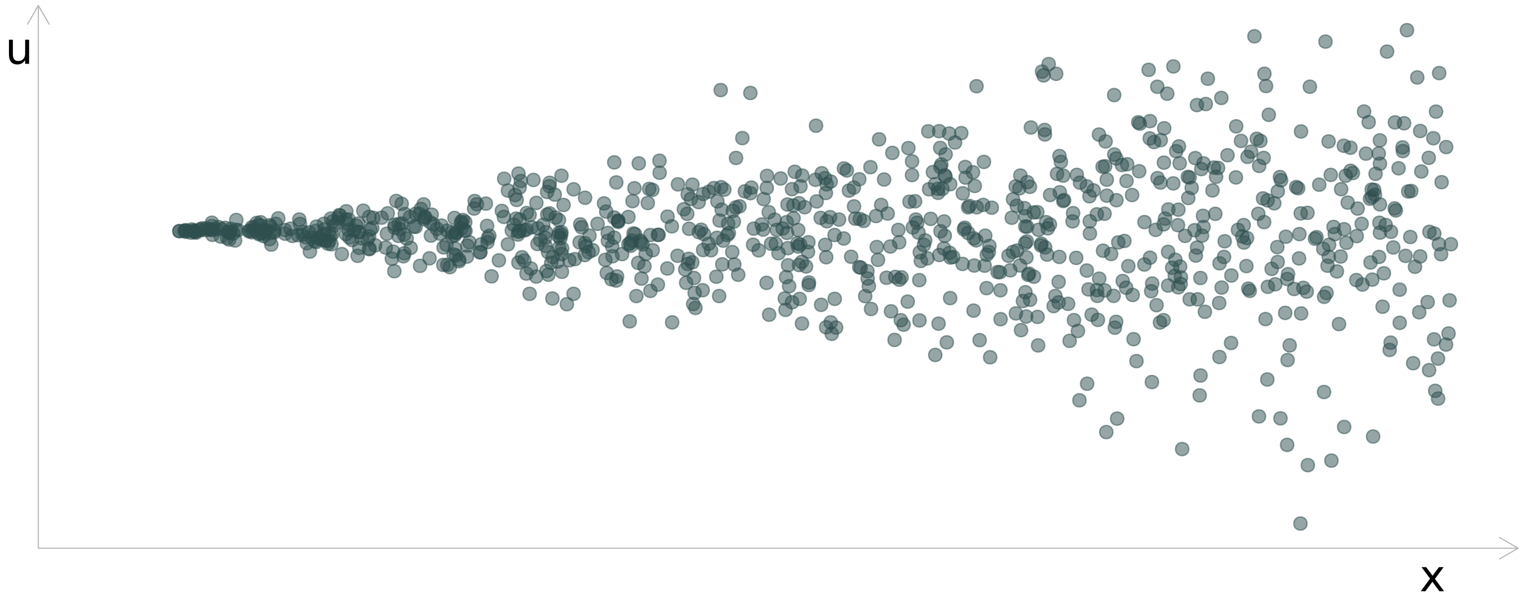
Heteroskedasticity: $\text{Var}(u_i) = \sigma_i^2$ and $\sigma_i^2 \neq \sigma_j^2$ for some $i \neq j$.

Heteroskedasticity

Review

Classic example of heteroskedasticity: The funnel

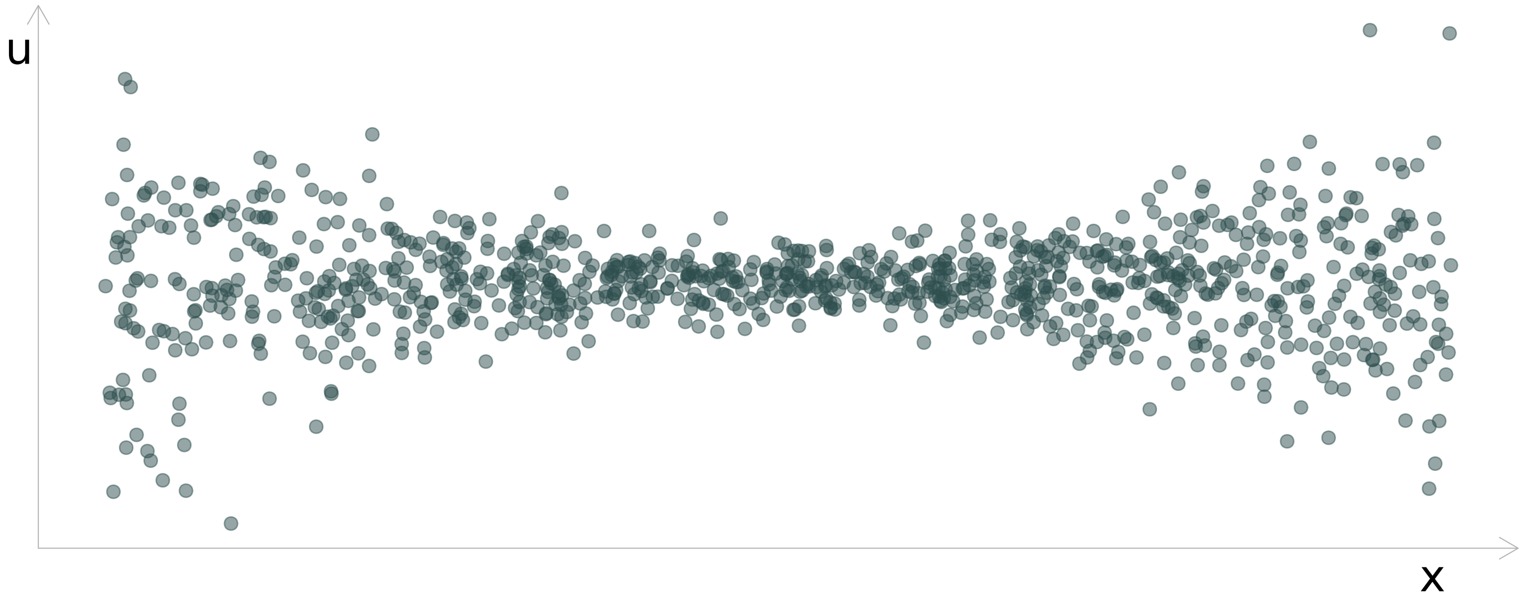
Variance of u increases with x



Review

Another example of heteroskedasticity: (double funnel?)

Variance of u increasing at the extremes of x

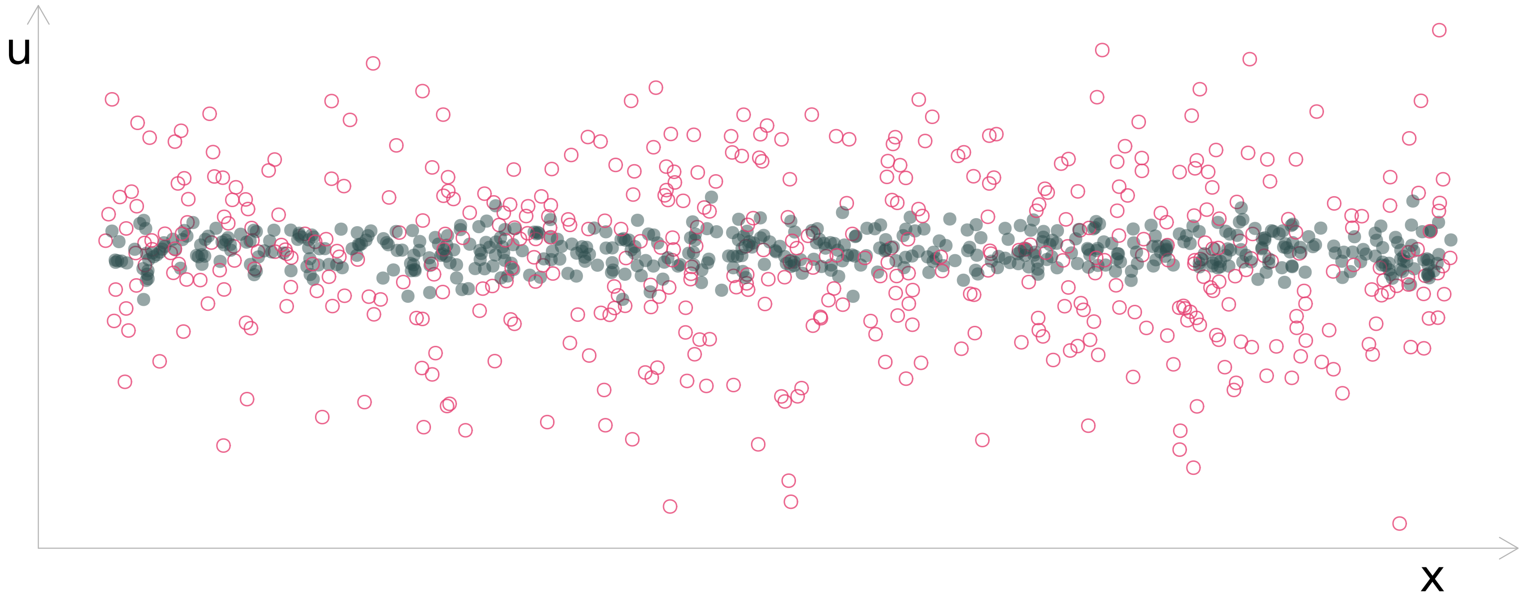


Heteroskedasticity

Review

Another example of heteroskedasticity:

Differing variances of u by group



Heteroskedasticity

Review

Heteroskedasticity is present when the variance of u changes with any combination of our explanatory variables x_1 through x_k.

Testing for heteroskedasticity

We have some tests that may help us detect heteroskedasticity.

- Goldfeld-Quandt
- White
- (There are others, *e.g.*, Breusch-Pagan)

Testing for heteroskedasticity

We have some tests that may help us detect heteroskedasticity.

- Goldfeld-Quandt
- White
- (There are others, *e.g.*, Breusch-Pagan)

What do we do if we detect it?

Living with heteroskedasticity

Living with heteroskedasticity

In the presence of heteroskedasticity, OLS is

- still **unbiased**
- **no longer the most efficient** unbiased linear estimator

On average, we get the right answer but with more noise (less precision).

Also: Our standard errors are biased.

Living with heteroskedasticity

In the presence of heteroskedasticity, OLS is

- still **unbiased**
- **no longer the most efficient** unbiased linear estimator

On average, we get the right answer but with more noise (less precision).

Also: Our standard errors are biased.

Options:

1. Check regression **specification**.
2. Find a new, more efficient **unbiased estimator** for β_j's.
3. Live with OLS's inefficiency; find a **new variance estimator**.
    - Standard errors
    - Confidence intervals
    - Hypothesis tests

Living with heteroskedasticity

Misspecification

As we've discussed, the specification† of your regression model matters a lot for the unbiasedness and efficiency of your estimator.

Response #1: Ensure your specification doesn't cause heteroskedasticity.

† *Specification:* Functional form and included variables.

Misspecification

Example: Let the population relationship be

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + u_i$$

with $\boldsymbol{E}[u_i|x_i] = 0$ and $\mathrm{Var}(u_i|x_i) = \sigma^2$.

However, we omit x^2 and estimate

$$y_i = \gamma_0 + \gamma_1 x_i + w_i$$

Then

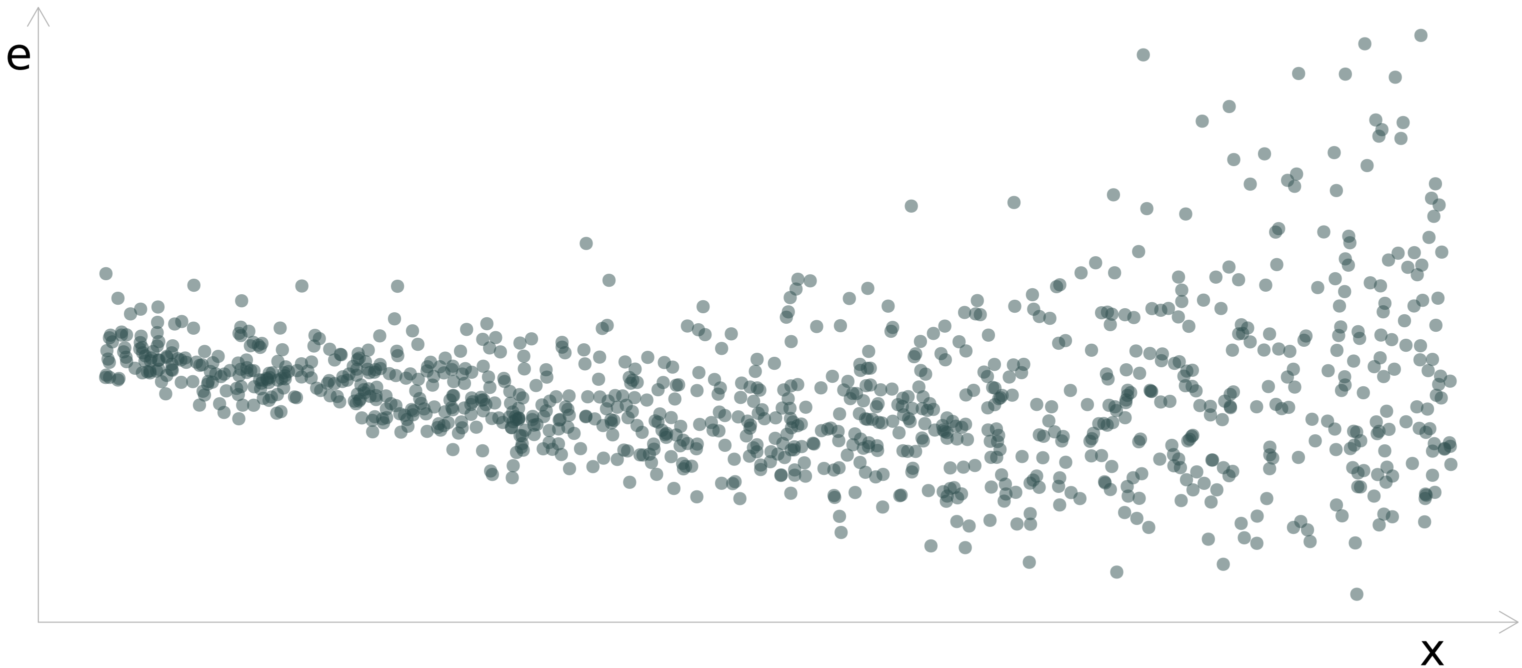
$$w_i = u_i + \beta_2 x_i^2 \implies \mathrm{Var}(w_i) = f(x_i)$$

I.e., the variance of w_i changes systematically with x_i (heteroskedasticity).

Living with heteroskedasticity

Misspecification

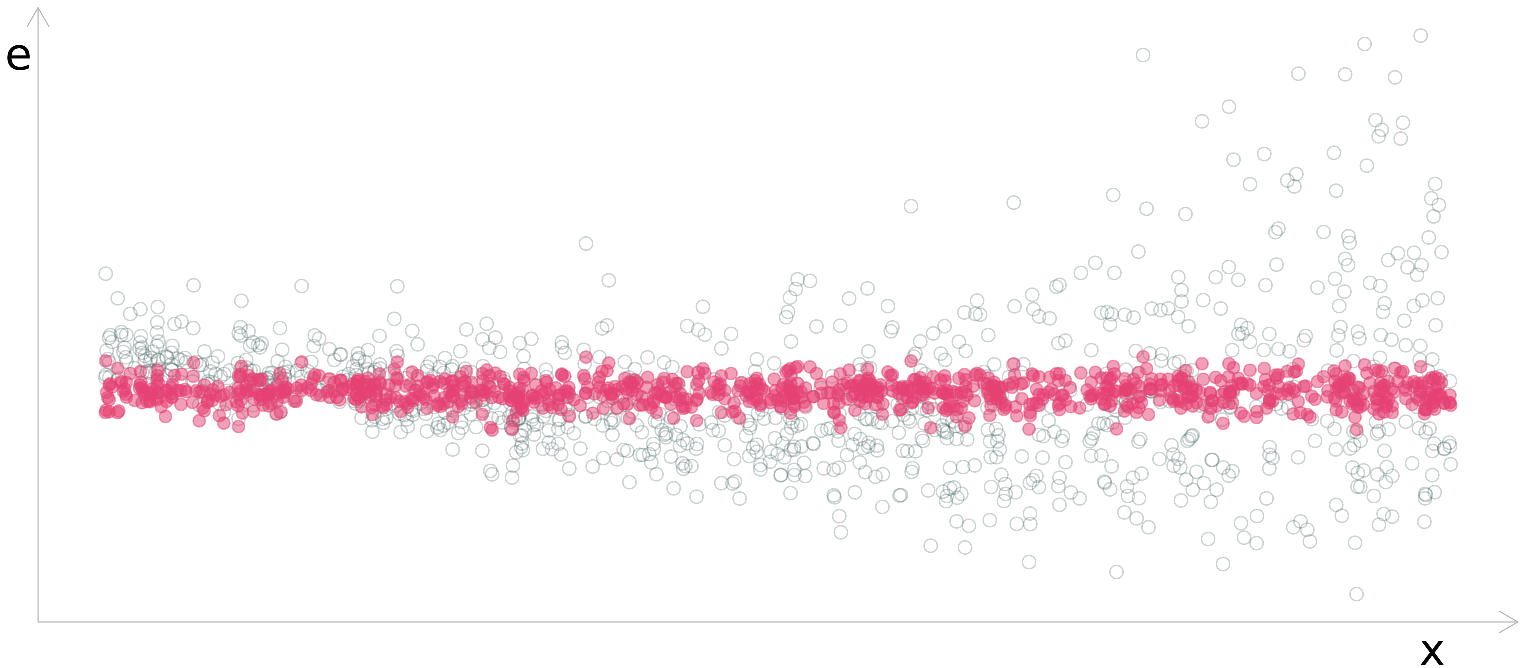
Truth: $\log(y_i) = \beta_0 + \beta_1 x_i + u_i$ Misspecification: $y_i = \beta_0 + \beta_1 x_i + v_i$



Living with heteroskedasticity

Misspecification

Truth: $\log(y_i) = \beta_0 + \beta_1 x_i + u_i$ Misspecification: $y_i = \beta_0 + \beta_1 x_i + v_i$



Living with heteroskedasticity

Misspecification

More generally:

Misspecification problem: Incorrect specification of the regression model can cause heteroskedasticity (among other problems).

Misspecification

More generally:

Misspecification problem: Incorrect specification of the regression model can cause heteroskedasticity (among other problems).

Solution: 💡 Get it right (*e.g.*, don't omit x^2).

Misspecification

More generally:

Misspecification problem: Incorrect specification of the regression model can cause heteroskedasticity (among other problems).

Solution: 💡 Get it right (*e.g.*, don't omit x^2).

New problems:

- We often don't know the *right* specification.
- We'd like a more formal process for addressing heteroskedasticity.

Misspecification

More generally:

Misspecification problem: Incorrect specification of the regression model can cause heteroskedasticity (among other problems).

Solution: 💡 Get it right (*e.g.*, don't omit x^2).

New problems:

- We often don't know the *right* specification.
- We'd like a more formal process for addressing heteroskedasticity.

Conclusion: Specification often will not "solve" heteroskedasticity. However, correctly specifying your model is still really important.

Weighted least squares

Weighted least squares (WLS) presents another approach.

Response #2: Increase efficiency by weighting our observations.

Living with heteroskedasticity

Weighted least squares

Weighted least squares (WLS) presents another approach.

Response #2: Increase efficiency by weighting our observations.

Let the true population relationship be

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{1}$$

with $u_i \sim N\left(0,\, \sigma_i^2\right)$.

Weighted least squares

Weighted least squares (WLS) presents another approach.

Response #2: Increase efficiency by weighting our observations.

Let the true population relationship be

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{1}$$

with $u_i \sim N\big(0,\, \sigma_i^2\big)$.

Now transform (1) by dividing each observation's data by σ_i, *i.e.*,

$$\frac{y_i}{\sigma_i} = \beta_0 \frac{1}{\sigma_i} + \beta_1 \frac{x_i}{\sigma_i} + \frac{u_i}{\sigma_i} \tag{2}$$

Living with heteroskedasticity

Weighted least squares

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{1}$$

$$\frac{y_i}{\sigma_i} = \beta_0 \frac{1}{\sigma_i} + \beta_1 \frac{x_i}{\sigma_i} + \frac{u_i}{\sigma_i} \tag{2}$$

Whereas (1) is heteroskedastic,

Weighted least squares

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{1}$$

$$\frac{y_i}{\sigma_i} = \beta_0 \frac{1}{\sigma_i} + \beta_1 \frac{x_i}{\sigma_i} + \frac{u_i}{\sigma_i} \tag{2}$$

Whereas (1) is heteroskedastic, (2) **is homoskedastic**.

$\therefore$ OLS is efficient and unbiased for estimating the β_k in (2)!

Weighted least squares

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{1}$$

$$\frac{y_i}{\sigma_i} = \beta_0 \frac{1}{\sigma_i} + \beta_1 \frac{x_i}{\sigma_i} + \frac{u_i}{\sigma_i} \tag{2}$$

Whereas (1) is heteroskedastic, (2) **is homoskedastic**.

$\therefore$ OLS is efficient and unbiased for estimating the β_k in (2)!

Why is (2) homoskedastic?

Weighted least squares

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{1}$$

$$\frac{y_i}{\sigma_i} = \beta_0 \frac{1}{\sigma_i} + \beta_1 \frac{x_i}{\sigma_i} + \frac{u_i}{\sigma_i} \tag{2}$$

Whereas (1) is heteroskedastic, (2) **is homoskedastic**.

$\therefore$ OLS is efficient and unbiased for estimating the β_k in (2)!

Why is (2) homoskedastic?

$$\textbf{Var}\left(\frac{u_i}{\sigma_i}\middle| x_i\right) =$$

Living with heteroskedasticity

Weighted least squares

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{1}$$

$$\frac{y_i}{\sigma_i} = \beta_0 \frac{1}{\sigma_i} + \beta_1 \frac{x_i}{\sigma_i} + \frac{u_i}{\sigma_i} \tag{2}$$

Whereas (1) is heteroskedastic, (2) **is homoskedastic**.

$\therefore$ OLS is efficient and unbiased for estimating the β_k in (2)!

Why is (2) homoskedastic?

$$\mathrm{Var}\left(\frac{u_i}{\sigma_i}\middle| x_i\right) = \frac{1}{\sigma_i^2}\mathrm{Var}(u_i|x_i) =$$

Weighted least squares

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{1}$$

$$\frac{y_i}{\sigma_i} = \beta_0 \frac{1}{\sigma_i} + \beta_1 \frac{x_i}{\sigma_i} + \frac{u_i}{\sigma_i} \tag{2}$$

Whereas (1) is heteroskedastic, (2) **is homoskedastic**.

$\therefore$ OLS is efficient and unbiased for estimating the β_k in (2)!

Why is (2) homoskedastic?

$$\text{Var}\left(\frac{u_i}{\sigma_i}\middle| x_i\right) = \frac{1}{\sigma_i^2}\text{Var}(u_i|x_i) = \frac{1}{\sigma_i^2}\sigma_i^2 =$$

Weighted least squares

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{1}$$

$$\frac{y_i}{\sigma_i} = \beta_0 \frac{1}{\sigma_i} + \beta_1 \frac{x_i}{\sigma_i} + \frac{u_i}{\sigma_i} \tag{2}$$

Whereas (1) is heteroskedastic, (2) **is homoskedastic**.

$\therefore$ OLS is efficient and unbiased for estimating the β_k in (2)!

Why is (2) homoskedastic?

$$\mathrm{Var}\left(\frac{u_i}{\sigma_i} \middle| x_i \right) = \frac{1}{\sigma_i^2} \mathrm{Var}(u_i | x_i) = \frac{1}{\sigma_i^2} \sigma_i^2 = 1$$

Weighted least squares

WLS is great, but we need to know σ_i^2, which is generally unlikely.

We can *slightly* relax this requirement—instead requiring

1. $\mathbf{Var}(u_i|x_i) = \sigma_i^2 = \sigma^2 h(x_i)$
2. We know $h(x)$.

Weighted least squares

WLS is great, but we need to know σ_i^2, which is generally unlikely.

We can *slightly* relax this requirement—instead requiring

1. $\mathbf{Var}(u_i|x_i) = \sigma_i^2 = \sigma^2 h(x_i)$

2. We know $h(x)$.

As before, we transform our heteroskedastic model into a homoskedastic model. This time we divide each observation's data† by $\sqrt{h(x_i)}$.

† Divide *all* of the data by $\sqrt{h(x_i)}$, including the intercept.

Living with heteroskedasticity

Weighted least squares

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{1}$$

$$\frac{y_i}{\sqrt{h(x_i)}} = \beta_0 \frac{1}{\sqrt{h(x_i)}} + \beta_1 \frac{x_i}{\sqrt{h(x_i)}} + \frac{u_i}{\sqrt{h(x_i)}} \tag{2}$$

with $\mathbf{Var}(u_i|x_i) = \sigma^2 h(x_i)$.

Weighted least squares

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{1}$$

$$\frac{y_i}{\sqrt{h(x_i)}} = \beta_0 \frac{1}{\sqrt{h(x_i)}} + \beta_1 \frac{x_i}{\sqrt{h(x_i)}} + \frac{u_i}{\sqrt{h(x_i)}} \tag{2}$$

with $\mathbf{Var}(u_i|x_i) = \sigma^2 h(x_i)$.

Now let's check that (2) is indeed homoskedastic.

$$\mathbf{Var}\left(\frac{u_i}{\sqrt{h(x_i)}}\middle| x_i\right) =$$

Weighted least squares

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{1}$$

$$\frac{y_i}{\sqrt{h(x_i)}} = \beta_0 \frac{1}{\sqrt{h(x_i)}} + \beta_1 \frac{x_i}{\sqrt{h(x_i)}} + \frac{u_i}{\sqrt{h(x_i)}} \tag{2}$$

with $\mathbf{Var}(u_i|x_i) = \sigma^2 h(x_i)$.

Now let's check that (2) is indeed homoskedastic.

$$\mathbf{Var}\left(\frac{u_i}{\sqrt{h(x_i)}} \middle| x_i \right) = \frac{1}{h(x_i)} \mathbf{Var}(u_i|x_i) =$$

Living with heteroskedasticity

Weighted least squares

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{1}$$

$$\frac{y_i}{\sqrt{h(x_i)}} = \beta_0 \frac{1}{\sqrt{h(x_i)}} + \beta_1 \frac{x_i}{\sqrt{h(x_i)}} + \frac{u_i}{\sqrt{h(x_i)}} \tag{2}$$

with $\mathbf{Var}(u_i|x_i) = \sigma^2 h(x_i)$.

Now let's check that (2) is indeed homoskedastic.

$$\mathbf{Var}\left(\frac{u_i}{\sqrt{h(x_i)}}\middle| x_i\right) = \frac{1}{h(x_i)}\,\mathbf{Var}(u_i|x_i) = \frac{1}{h(x_i)}\sigma^2 h(x_i) =$$

Weighted least squares

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{1}$$

$$\frac{y_i}{\sqrt{h(x_i)}} = \beta_0 \frac{1}{\sqrt{h(x_i)}} + \beta_1 \frac{x_i}{\sqrt{h(x_i)}} + \frac{u_i}{\sqrt{h(x_i)}} \tag{2}$$

with $\mathbf{Var}(u_i|x_i) = \sigma^2 h(x_i)$.

Now let's check that $(\mathbf{2})$ is indeed homoskedastic.

$$\mathbf{Var}\left(\frac{u_i}{\sqrt{h(x_i)}}\middle| x_i\right) = \frac{1}{h(x_i)}\mathbf{Var}(u_i|x_i) = \frac{1}{h(x_i)}\sigma^2 h(x_i) = \sigma^2$$

Homoskedasticity!

Weighted least squares

Weighted least squares (WLS) estimators are a special class of **generalized least squares** (GLS) estimators focused on heteroskedasticity.

Weighted least squares

Weighted least squares (WLS) estimators are a special class of **generalized least squares** (GLS) estimators focused on heteroskedasticity.

$$y_i = \beta_0 + \beta_1 x_{1i} + u_i \quad \text{vs.} \quad \frac{y_i}{\sigma_i} = \beta_0 \frac{1}{\sigma_i} + \beta_1 \frac{x_{1i}}{\sigma_i} + \frac{u_i}{\sigma_i}$$

Notes:

1. WLS **transforms** a heteroskedastic model into a homoskedastic model.
2. **Weighting:** WLS downweights observations with higher variance u_i's.
3. **Big requirement:** WLS requires that we *know* σ_i^2 for each observation.
4. WLS is generally **infeasible**. *Feasible* GLS (FGLS) offers a solution.
5. Under its assumptions: WLS is the **best linear unbiased estimator**.

Living with heteroskedasticity

Heteroskedasticity-robust standard errors

Response #3:

- Ignore OLS's inefficiency (in the presence of heteroskedasticity).
- Focus on **unbiased estimates for our standard errors**.
- In the process: Correct inference.

Living with heteroskedasticity

Heteroskedasticity-robust standard errors

Response #3:

- Ignore OLS's inefficiency (in the presence of heteroskedasticity).
- Focus on **unbiased estimates for our standard errors**.
- In the process: Correct inference.

Q: What is a standard error?

Living with heteroskedasticity

Heteroskedasticity-robust standard errors

Response #3:

- Ignore OLS's inefficiency (in the presence of heteroskedasticity).
- Focus on **unbiased estimates for our standard errors**.
- In the process: Correct inference.

Q: What is a standard error?

A: The **standard deviation of an estimator's distribution**.

Heteroskedasticity-robust standard errors

Response #3:

- Ignore OLS's inefficiency (in the presence of heteroskedasticity).
- Focus on **unbiased estimates for our standard errors**.
- In the process: Correct inference.

Q: What is a standard error?

A: The **standard deviation of an estimator's distribution**.

Estimators (like $\hat{\beta}_1$) are random variables, so they have distributions.

Standard errors give us a sense of how much variability is in our estimator.

Living with heteroskedasticity

Heteroskedasticity-robust standard errors

Recall: We can write the OLS estimator for β_1 as

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_i \left(x_i - \overline{x}\right) u_i}{\sum_i \left(x_i - \overline{x}\right)^2} = \beta_1 + \frac{\sum_i \left(x_i - \overline{x}\right) u_i}{\text{SST}_x} \tag{3}$$

Heteroskedasticity-robust standard errors

Recall: We can write the OLS estimator for β_1 as

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_i \left(x_i - \overline{x}\right) u_i}{\sum_i \left(x_i - \overline{x}\right)^2} = \beta_1 + \frac{\sum_i \left(x_i - \overline{x}\right) u_i}{\text{SST}_x} \tag{3}$$

Let $\mathbf{Var}(u_i|x_i) = \sigma_i^2$.

Heteroskedasticity-robust standard errors

Recall: We can write the OLS estimator for β_1 as

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_i \left(x_i - \overline{x}\right) u_i}{\sum_i \left(x_i - \overline{x}\right)^2} = \beta_1 + \frac{\sum_i \left(x_i - \overline{x}\right) u_i}{\text{SST}_x} \tag{3}$$

Let $\mathop{\text{Var}}(u_i|x_i) = \sigma_i^2$.

We can use (3) to write the variance of $\hat{\beta}_1$, *i.e.*,

$$\mathop{\text{Var}}\left(\hat{\beta}_1 \middle| x_i\right) = \frac{\sum_i \left(x_i - \overline{x}\right)^2 \sigma_i^2}{\text{SST}_x^2} \tag{4}$$

Living with heteroskedasticity

Heteroskedasticity-robust standard errors

If we want unbiased estimates for our standard errors, we need an unbiased estimate for

$$\frac{\sum_i \left(x_i - \overline{x}\right)^2 \sigma_i^2}{\text{SST}_x^2}$$

Our old friend Hal White provided such an estimator:[†]

$$\widehat{\text{Var}}\left(\hat{\beta}_1\right) = \frac{\sum_i \left(x_i - \overline{x}\right)^2 e_i^2}{\text{SST}_x^2}$$

where the e_i comes from the OLS regression of interest.

† This specific equation is for simple linear regression.

Heteroskedasticity-robust standard errors

Our heteroskedasticity-robust estimators for the standard error of β_j.

Case 1 Simple linear regression, $y_i = \beta_0 + \beta_1 x_i + u_i$

$$\widehat{\text{Var}}\left(\hat{\beta}_1\right) = \frac{\sum_i \left(x_i - \overline{x}\right)^2 e_i^2}{\text{SST}_x^2}$$

Case 2 Multiple (linear) regression, $y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + u_i$

$$\widehat{\text{Var}}\left(\hat{\beta}_j\right) = \frac{\sum_i \hat{r}_{ij}^2 e_i^2}{\text{SST}_{x_j^2}}$$

where $\hat{r}_{ij}$ denotes the i$^{\text{th}}$ residual from regressing x_j on all other explanatory variables.

Living with heteroskedasticity

Heteroskedasticity-robust standard errors

With these standard errors, we can return to correct statistical inferencel

E.g., we can update our previous t statistic formula with our new heteroskedasticity-robust standard erros.

$$t = \frac{\text{Estimate} - \text{Hypothesized value}}{\text{Standard error}}$$

Heteroskedasticity-robust standard errors

Notes

- We are still using **OLS estimates for β_j**
- Our het.-robust standard errors use a **different estimator**.
- Homoskedasticity
  - Plain OLS variance estimator is more efficient.
  - Het.-robust is still unbiased.
- Heteroskedasticity
  - Plain OLS variance estimator is biased.
  - Het.-robust variance estimator is unbiased.

Heteroskedasticity-robust standard errors

These standard errors go by many names

- Heteroskedasticity-robust standard errors
- Het.-robust standard errors
- White standard errors
- Eicker-White standard errors
- Huber standard errors
- Eicker-Huber-White standards errors
- (some other combination of Eicker, Huber, and White)

Do not say: "Robust standard errors". The problem: "robust" to what?

Living with heteroskedasticity

Examples

Living with heteroskedasticity

Examples

Back to our test-scores dataset…

```
# Load packages
library(pacman)
p_load(tidyverse, Ecdat)
# Select and rename desired variables; assign to new dataset; format as tibble
test_df <- Caschool %>% select(
  test_score = testscr, ratio = str, income = avginc, enrollment = enrltot
) %>% as_tibble()
# View first 2 rows of the dataset
head(test_df, 2)
```

```
#> # A tibble: 2 × 4
#>   test_score ratio income enrollment
#>        <dbl> <dbl>  <dbl>      <int>
#> 1       691.  17.9  22.7         195
#> 2       661.  21.5   9.82        240
```

Living with heteroskedasticity

Example: Model specification

We found significant evidence of heteroskedasticity.

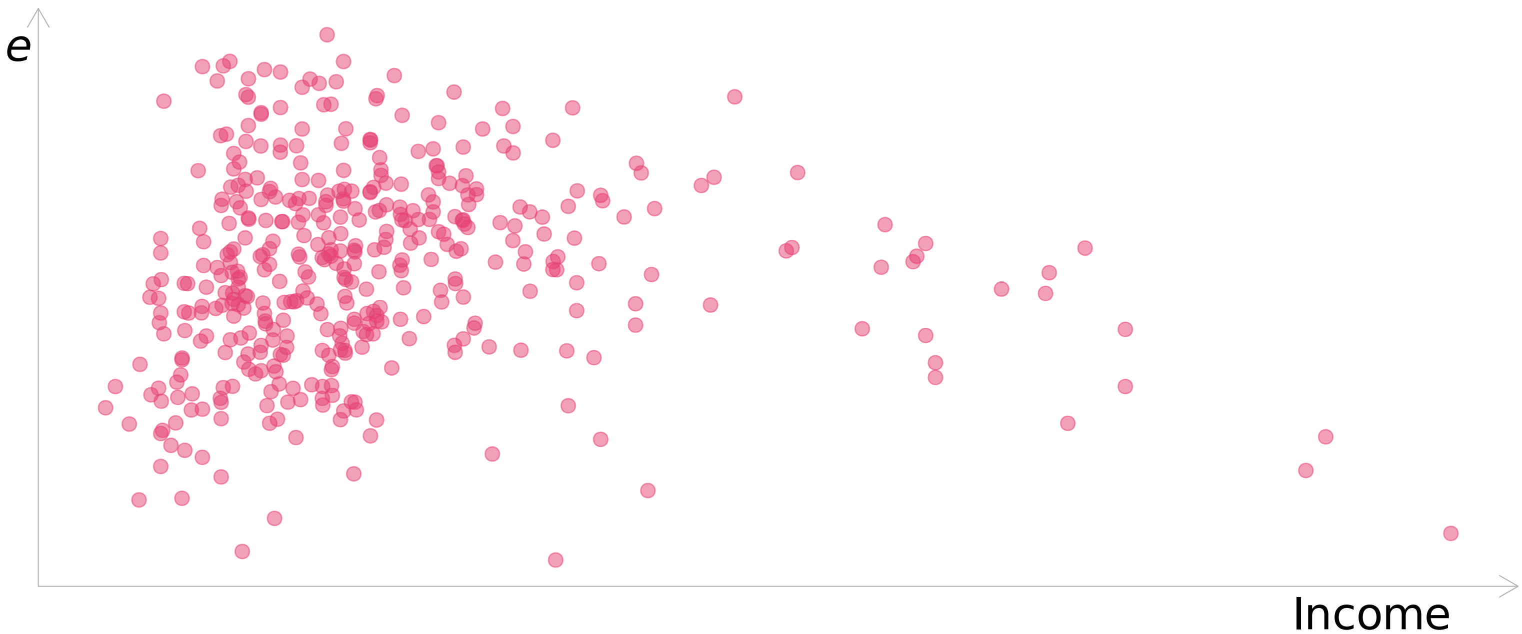
Let's check if it was due to misspecifying our model.

Living with heteroskedasticity

Example: Model specification

Model_1: $\text{Score}_i = \beta_0 + \beta_1 \text{Ratio}_i + \beta_2 \text{Income}_i + u_i$

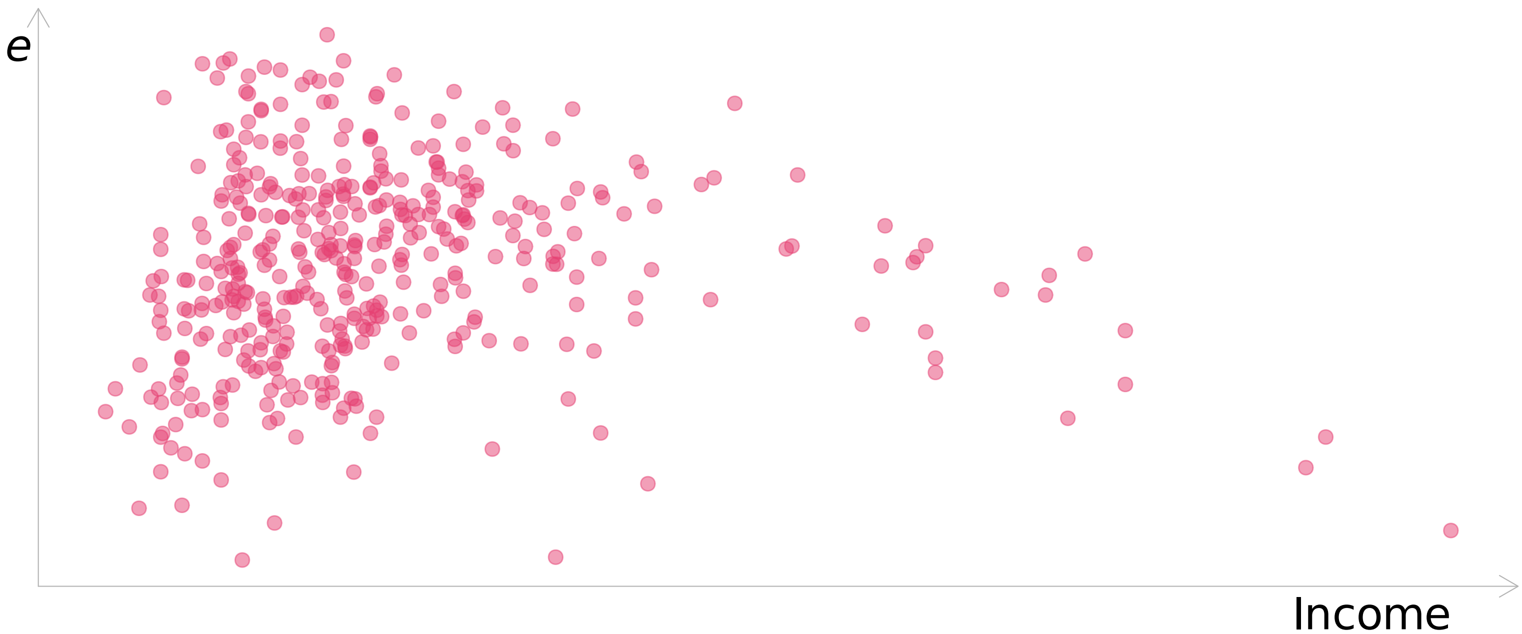
```
lm(test_score ~ ratio + income, data = test_df)
```



Example: Model specification

Model$_2$: $\log(\text{Score}_i) = \beta_0 + \beta_1 \text{Ratio}_i + \beta_2 \text{Income}_i + u_i$

```
lm(log(test_score) ~ ratio + income, data = test_df)
```

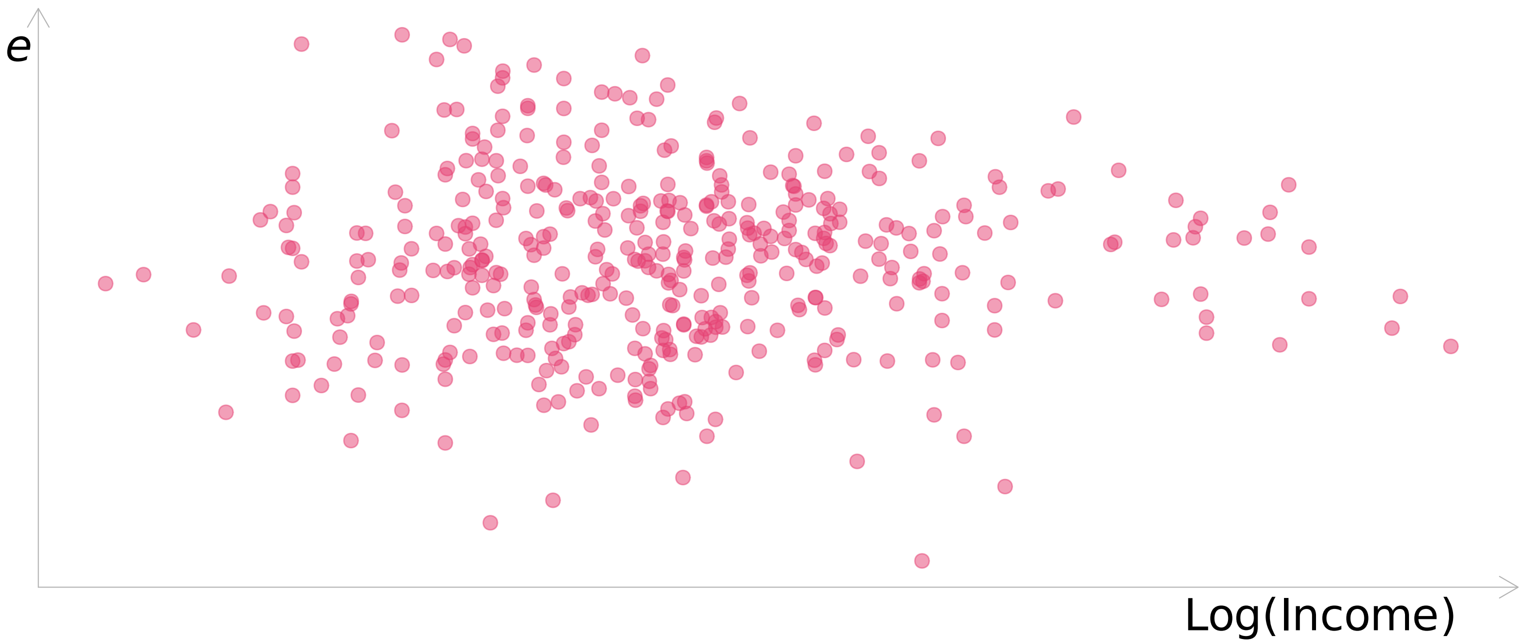


Living with heteroskedasticity

Example: Model specification

Model$_3$: $\log(\text{Score}_i) = \beta_0 + \beta_1 \text{Ratio}_i + \beta_2 \log(\text{Income}_i) + u_i$

```
lm(log(test_score) ~ ratio + log(income), data = test_df)
```



Example: Model specification

Let's test this new specification with the White test for heteroskedasticity.

$$\text{Model}_3\text{: } \log(\text{Score}_i) = \beta_0 + \beta_1 \text{Ratio}_i + \beta_2 \log(\text{Income}_i) + u_i$$

Example: Model specification

Let's test this new specification with the White test for heteroskedasticity.

$$\text{Model}_3\text{: } \log(\textbf{Score}_i) = \beta_0 + \beta_1 \textbf{Ratio}_i + \beta_2 \log(\textbf{Income}_i) + u_i$$

The regression for the White test

Example: Model specification

Let's test this new specification with the White test for heteroskedasticity.

$$\text{Model}_3\text{: } \log(\text{Score}_i) = \beta_0 + \beta_1 \text{Ratio}_i + \beta_2 \log(\text{Income}_i) + u_i$$

The regression for the White test

$$\begin{aligned} e_i^2 = &\alpha_0 + \alpha_1 \text{Ratio}_i + \alpha_2 \log(\text{Income}_i) + \alpha_3 \text{Ratio}_i^2 + \alpha_4 \left(\log(\text{Income}_i)\right)^2 \\ &+ \alpha_5 \left(\text{Ratio}_i \times \log(\text{Income}_i)\right) + v_i \end{aligned}$$

Example: Model specification

Let's test this new specification with the White test for heteroskedasticity.

$$\text{Model}_3\text{: } \log(\text{Score}_i) = \beta_0 + \beta_1 \text{Ratio}_i + \beta_2 \log(\text{Income}_i) + u_i$$

The regression for the White test

$$\begin{aligned} e_i^2 = &\alpha_0 + \alpha_1 \text{Ratio}_i + \alpha_2 \log(\text{Income}_i) + \alpha_3 \text{Ratio}_i^2 + \alpha_4 \left(\log(\text{Income}_i)\right)^2 \\ &+ \alpha_5 \left(\text{Ratio}_i \times \log(\text{Income}_i)\right) + v_i \end{aligned}$$

yields $R_e^2 \approx 0.029$

Example: Model specification

Let's test this new specification with the White test for heteroskedasticity.

$$\text{Model}_3\text{: } \log(\text{Score}_i) = \beta_0 + \beta_1 \text{Ratio}_i + \beta_2 \log(\text{Income}_i) + u_i$$

The regression for the White test

$$\begin{aligned} e_i^2 =& \alpha_0 + \alpha_1 \text{Ratio}_i + \alpha_2 \log(\text{Income}_i) + \alpha_3 \text{Ratio}_i^2 + \alpha_4 (\log(\text{Income}_i))^2 \\ &+ \alpha_5 \left(\text{Ratio}_i \times \log(\text{Income}_i)\right) + v_i \end{aligned}$$

yields $R_e^2 \approx 0.029$ and test statistic of

Living with heteroskedasticity

Example: Model specification

Let's test this new specification with the White test for heteroskedasticity.

$$\text{Model}_3\text{: } \log(\text{Score}_i) = \beta_0 + \beta_1 \text{Ratio}_i + \beta_2 \log(\text{Income}_i) + u_i$$

The regression for the White test

$$\begin{aligned} e_i^2 = & \alpha_0 + \alpha_1 \text{Ratio}_i + \alpha_2 \log(\text{Income}_i) + \alpha_3 \text{Ratio}_i^2 + \alpha_4 (\log(\text{Income}_i))^2 \\ & + \alpha_5 \left(\text{Ratio}_i \times \log(\text{Income}_i)\right) + v_i \end{aligned}$$

yields $R_e^2 \approx 0.029$ and test statistic of $\widehat{\text{LM}} = n \times R_e^2 \approx 12.2$.

Example: Model specification

Let's test this new specification with the White test for heteroskedasticity.

$$\text{Model}_3\text{: } \log(\text{Score}_i) = \beta_0 + \beta_1 \text{Ratio}_i + \beta_2 \log(\text{Income}_i) + u_i$$

The regression for the White test

$$\begin{aligned} e_i^2 = & \alpha_0 + \alpha_1 \text{Ratio}_i + \alpha_2 \log(\text{Income}_i) + \alpha_3 \text{Ratio}_i^2 + \alpha_4 \left(\log(\text{Income}_i)\right)^2 \\ & + \alpha_5 \left(\text{Ratio}_i \times \log(\text{Income}_i)\right) + v_i \end{aligned}$$

yields $R_e^2 \approx 0.029$ and test statistic of $\widehat{\text{LM}} = n \times R_e^2 \approx 12.2$.

Under H_0, LM is distributed as

Example: Model specification

Let's test this new specification with the White test for heteroskedasticity.

$$\text{Model}_3\text{: } \log(\text{Score}_i) = \beta_0 + \beta_1 \text{Ratio}_i + \beta_2 \log(\text{Income}_i) + u_i$$

The regression for the White test

$$\begin{aligned} e_i^2 = & \alpha_0 + \alpha_1 \text{Ratio}_i + \alpha_2 \log(\text{Income}_i) + \alpha_3 \text{Ratio}_i^2 + \alpha_4 (\log(\text{Income}_i))^2 \\ & + \alpha_5 (\text{Ratio}_i \times \log(\text{Income}_i)) + v_i \end{aligned}$$

yields $R_e^2 \approx 0.029$ and test statistic of $\widehat{\text{LM}} = n \times R_e^2 \approx 12.2$.

Under H_0, LM is distributed as χ_5^2

Example: Model specification

Let's test this new specification with the White test for heteroskedasticity.

$$\text{Model}_3\text{: } \log(\text{Score}_i) = \beta_0 + \beta_1 \text{Ratio}_i + \beta_2 \log(\text{Income}_i) + u_i$$

The regression for the White test

$$\begin{aligned} e_i^2 = & \alpha_0 + \alpha_1 \text{Ratio}_i + \alpha_2 \log(\text{Income}_i) + \alpha_3 \text{Ratio}_i^2 + \alpha_4 \left(\log(\text{Income}_i)\right)^2 \\ & + \alpha_5 \left(\text{Ratio}_i \times \log(\text{Income}_i)\right) + v_i \end{aligned}$$

yields $R_e^2 \approx 0.029$ and test statistic of $\widehat{\text{LM}} = n \times R_e^2 \approx 12.2$.

Under H_0, LM is distributed as $\chi_5^2 \implies$ p-value ≈ 0.033.

Example: Model specification

Let's test this new specification with the White test for heteroskedasticity.

$$\text{Model}_3\text{: } \log(\text{Score}_i) = \beta_0 + \beta_1 \text{Ratio}_i + \beta_2 \log(\text{Income}_i) + u_i$$

The regression for the White test

$$\begin{aligned} e_i^2 = & \alpha_0 + \alpha_1 \text{Ratio}_i + \alpha_2 \log(\text{Income}_i) + \alpha_3 \text{Ratio}_i^2 + \alpha_4 \left(\log(\text{Income}_i)\right)^2 \\ & + \alpha_5 \left(\text{Ratio}_i \times \log(\text{Income}_i)\right) + v_i \end{aligned}$$

yields $R_e^2 \approx 0.029$ and test statistic of $\widehat{\text{LM}} = n \times R_e^2 \approx 12.2$.

Under H_0, LM is distributed as $\chi_5^2 \implies$ p-value $\approx$ 0.033.

$\therefore$

Example: Model specification

Let's test this new specification with the White test for heteroskedasticity.

$$\text{Model}_3\text{: } \log(\text{Score}_i) = \beta_0 + \beta_1 \text{Ratio}_i + \beta_2 \log(\text{Income}_i) + u_i$$

The regression for the White test

$$\begin{aligned} e_i^2 = & \alpha_0 + \alpha_1 \text{Ratio}_i + \alpha_2 \log(\text{Income}_i) + \alpha_3 \text{Ratio}_i^2 + \alpha_4 (\log(\text{Income}_i))^2 \\ & + \alpha_5 (\text{Ratio}_i \times \log(\text{Income}_i)) + v_i \end{aligned}$$

yields $R_e^2 \approx \mathbf{0.029}$ and test statistic of $\widehat{\text{LM}} = n \times R_e^2 \approx \mathbf{12.2}$.

Under H_0, LM is distributed as $\chi_5^2 \implies p\text{-value} \approx 0.033$.

$\therefore$ **Reject H_0.**

Living with heteroskedasticity

Example: Model specification

Let's test this new specification with the White test for heteroskedasticity.

$$\text{Model}_3\text{: } \log(\text{Score}_i) = \beta_0 + \beta_1 \text{Ratio}_i + \beta_2 \log(\text{Income}_i) + u_i$$

The regression for the White test

$$\begin{aligned} e_i^2 = &\alpha_0 + \alpha_1 \text{Ratio}_i + \alpha_2 \log(\text{Income}_i) + \alpha_3 \text{Ratio}_i^2 + \alpha_4 \left(\log(\text{Income}_i)\right)^2 \\ &+ \alpha_5 \left(\text{Ratio}_i \times \log(\text{Income}_i)\right) + v_i \end{aligned}$$

yields $R_e^2 \approx 0.029$ and test statistic of $\widehat{\text{LM}} = n \times R_e^2 \approx 12.2$.

Under H_0, LM is distributed as $\chi_5^2 \implies$ p-value ≈ 0.033.

$\therefore$ **Reject H_0. Conclusion:**

Example: Model specification

Let's test this new specification with the White test for heteroskedasticity.

$$\text{Model}_3\text{: } \log(\text{Score}_i) = \beta_0 + \beta_1 \text{Ratio}_i + \beta_2 \log(\text{Income}_i) + u_i$$

The regression for the White test

$$\begin{aligned} e_i^2 = & \alpha_0 + \alpha_1 \text{Ratio}_i + \alpha_2 \log(\text{Income}_i) + \alpha_3 \text{Ratio}_i^2 + \alpha_4 \left(\log(\text{Income}_i)\right)^2 \\ & + \alpha_5 \left(\text{Ratio}_i \times \log(\text{Income}_i)\right) + v_i \end{aligned}$$

yields $R_e^2 \approx 0.029$ and test statistic of $\widehat{\text{LM}} = n \times R_e^2 \approx 12.2$.

Under H_0, LM is distributed as $\chi_5^2 \implies$ p-value $\approx$ 0.033.

$\therefore$ **Reject H_0. Conclusion:** There is statistically significant evidence of heteroskedasticity at the five-percent level.

Example: Model specification

Okay, we tried adjusting our specification, but there is still evidence of heteroskedasticity.

Next: In general, you will turn to heteroskedasticity-robust standard errors.

- OLS is still unbiased for the **coefficients** (the β_j's)
- Heteroskedasticity-robust standard errors are unbiased for the **standard errors** of the $\hat{\beta}_j$'s, *i.e.*, $\sqrt{\mathrm{Var}\left(\hat{\beta}_j\right)}$.

Example: Het.-robust standard errors

Let's return to our model

$$\text{Score}_i = \beta_0 + \beta_1 \text{Ratio}_i + \beta_2 \text{Income}_i + u_i$$

We can use the `fixest` package in **R** to calculate standard errors.

Example: Het.-robust standard errors

$$\text{Score}_i = \beta_0 + \beta_1 \text{Ratio}_i + \beta_2 \text{Income}_i + u_i$$

1. Run the regression with `feols()` (instead of `lm()`)

```
# Load 'fixest' package
p_load(fixest)
# Regress log score on ratio and log income
test_reg ← feols(test_score ~ ratio + income, data = test_df)
```

Example: Het.-robust standard errors

$$\text{Score}_i = \beta_0 + \beta_1 \text{Ratio}_i + \beta_2 \text{Income}_i + u_i$$

1. Run the regression with `feols()` (instead of `lm()`)

```
# Load 'fixest' package
p_load(fixest)
# Regress log score on ratio and log income
test_reg <- feols(test_score ~ ratio + income, data = test_df)
```

Notice that `feols()` uses the same syntax as `lm()` for this regression.

Example: Het.-robust standard errors

$$\text{Score}_i = \beta_0 + \beta_1 \text{Ratio}_i + \beta_2 \text{Income}_i + u_i$$

2. Estimate het.-robust standard errors with `vcov = 'hetero'` option in `summary()`

```
# Het-robust standard errors with 'vcov = 'hetero''
summary(test_reg, vcov = 'hetero')
```

```
#>                Estimate Std. Error  t value  Pr(>|t|)
#> (Intercept) 638.729155   7.301234 87.48235 < 2.2e-16 ***
#> ratio        -0.648740   0.353340 -1.83602  0.067066 .
#> income        1.839112   0.114733 16.02949 < 2.2e-16 ***
```

Living with heteroskedasticity

Example: Het.-robust standard errors

Ceofficients and **heteroskedasticity-robust standard errors**:

```
summary(test_reg, vcov = 'hetero')
```

```
#>              Estimate Std. Error  t value  Pr(>|t|)
#>  (Intercept) 638.729155   7.301234 87.48235 < 2.2e-16 ***
#>  ratio        -0.648740   0.353340 -1.83602  0.067066 .
#>  income        1.839112   0.114733 16.02949 < 2.2e-16 ***
```

Ceofficients and **plain OLS standard errors** (assumes homoskedasticity):

```
summary(test_reg, vcov = 'iid')
```

```
#>              Estimate Std. Error  t value  Pr(>|t|)
#>  (Intercept) 638.729155   7.449077 85.74608 < 2.2e-16 ***
#>  ratio        -0.648740   0.354405 -1.83051  0.067888 .
#>  income        1.839112   0.092787 19.82083 < 2.2e-16 ***
```

Example: WLS

We mentioned that WLS is often not possible—we need to know the functional form of the heteroskedasticity—either

A. σ_i^2

or

B. $h(x_i)$, where $\sigma_i^2 = \sigma^2 h(x_i)$

Example: WLS

We mentioned that WLS is often not possible—we need to know the functional form of the heteroskedasticity—either

A. σ_i^2

or

B. $h(x_i)$, where $\sigma_i^2 = \sigma^2 h(x_i)$

There *are* occasions in which we can know $h(x_i)$.

Living with heteroskedasticity

Example: WLS

Imagine individuals in a population have homoskedastic disturbances.

However, instead of observing individuals' data, we observe (in data) groups' averages (*e.g.*, cities, counties, school districts).

If these groups have different sizes, then our dataset will be heteroskedastic—in a predictable fashion.

Recall: The variance of the sample mean depends upon the sample size,

$$\text{Var}\left(\overline{x}\right) = \frac{\sigma_x^2}{n}$$

Example: WLS

Imagine individuals in a population have homoskedastic disturbances.

However, instead of observing individuals' data, we observe (in data) groups' averages (*e.g.*, cities, counties, school districts).

If these groups have different sizes, then our dataset will be heteroskedastic—in a predictable fashion.

Recall: The variance of the sample mean depends upon the sample size,

$$\mathrm{Var}\left(\overline{x}\right) = \frac{\sigma_x^2}{n}$$

Example: Our school testing data is averaged at the school level.

Example: WLS

Example: Our school testing data is averaged at the school level.

Even if individual students have homoskedastic disturbances, the schools would have heteroskedastic disturbances, *i.e.*,

Individual-level model: $\text{Score}_i = \beta_0 + \beta_1 \text{Ratio}_i + \beta_2 \text{Income}_i + u_i$

School-level model: $\overline{\text{Score}}_s = \beta_0 + \beta_1 \overline{\text{Ratio}}_s + \beta_2 \overline{\text{Income}}_s + \overline{u}_s$

where the s subscript denotes an individual school (just as i indexes an individual person).

$$\text{Var}\left(\overline{u}_s\right) = \frac{\sigma^2}{n_s}$$

Living with heteroskedasticity

Example: WLS

For WLS, we're looking for a function $h(x_s)$ such that $\mathbf{Var}\big(\overline{u}_s|x_s\big) = \sigma^2 h(x_s)$.

Example: WLS

For WLS, we're looking for a function $h(x_s)$ such that $\mathbf{Var}\left(\overline{u}_s|x_s\right) = \sigma^2 h(x_s)$.

We just showed[†] that $\mathbf{Var}\left(\overline{u}_s|x_s\right) = \dfrac{\sigma^2}{n_s}$.

† Assuming the individuals' disturbances are homoskedastic.

Example: WLS

For WLS, we're looking for a function $h(x_s)$ such that $\mathbf{Var}\left(\overline{u}_s|x_s\right) = \sigma^2 h(x_s)$.

We just showed[†] that $\mathbf{Var}\left(\overline{u}_s|x_s\right) = \dfrac{\sigma^2}{n_s}$.

Thus, $h(x_s) = 1/n_s$, where n_s is the number of students in school s.

† Assuming the individuals' disturbances are homoskedastic.

Living with heteroskedasticity

Example: WLS

For WLS, we're looking for a function $h(x_s)$ such that $\text{Var}\left(\overline{u}_s|x_s\right) = \sigma^2 h(x_s)$.

We just showed[†] that $\text{Var}\left(\overline{u}_s|x_s\right) = \dfrac{\sigma^2}{n_s}$.

Thus, $h(x_s) = 1/n_s$, where n_s is the number of students in school s.

To implement WLS, we divide each observation's data by $1/\sqrt{h(x_s)}$, meaning we need to multiply each school's data by $\sqrt{n_s}$.

† Assuming the individuals' disturbances are homoskedastic.

Example: WLS

For WLS, we're looking for a function $h(x_s)$ such that $\text{Var}\left(\overline{u}_s|x_s\right) = \sigma^2 h(x_s)$.

We just showed† that $\text{Var}\left(\overline{u}_s|x_s\right) = \dfrac{\sigma^2}{n_s}$.

Thus, $h(x_s) = 1/n_s$, where n_s is the number of students in school s.

To implement WLS, we divide each observation's data by $1/\sqrt{h(x_s)}$, meaning we need to multiply each school's data by $\sqrt{n_s}$.

The variable `enrollment` in the `test_df` dataset is our n_s.

† Assuming the individuals' disturbances are homoskedastic.

Example: WLS

Using WLS to estimate $\textbf{Score}_i = \beta_0 + \beta_1 \textbf{Ratio}_i + \beta_2 \textbf{Income}_i + u_i$

Step 1: Multiply each variable by $1/\sqrt{h(x_i)} = \sqrt{\textbf{Enrollment}_i}$

```
# Create WLS transformed variables, multiplying by sqrt of 'pop'
test_df <- mutate(test_df,
  test_score_wls = test_score * sqrt(enrollment),
  ratio_wls      = ratio * sqrt(enrollment),
  income_wls     = income * sqrt(enrollment),
  intercept_wls  = 1 * sqrt(enrollment)
)
```

Notice that we are creating a transformed intercept.

Example: WLS

Using WLS to estimate $\text{Score}_i = \beta_0 + \beta_1 \text{Ratio}_i + \beta_2 \text{Income}_i + u_i$

Step 2: Run our WLS (transformed) regression

```
# WLS regression
wls_reg <- lm(
  test_score_wls ~ -1 + intercept_wls + ratio_wls + income_wls,
  data = test_df
)
```

Example: WLS

Using WLS to estimate $\text{Score}_i = \beta_0 + \beta_1 \text{Ratio}_i + \beta_2 \text{Income}_i + u_i$

Step 2: Run our WLS (transformed) regression

```
# WLS regression
wls_reg ← lm(
  test_score_wls ~ -1 + intercept_wls + ratio_wls + income_wls,
  data = test_df
)
```

Note: The `-1` in our regression tells **R** not to add an intercept, since we are adding a transformed intercept (`intercept_wls`).

Living with heteroskedasticity

Example: WLS

The **WLS estimates and standard errors:**

```
#>                 Estimate Std. Error t value Pr(>|t|)
#>  intercept_wls 618.78331    8.26929  74.829   <2e-16 ***
#>  ratio_wls      -0.21314    0.37676  -0.566    0.572
#>  income_wls      2.26493    0.09065  24.985   <2e-16 ***
```

Living with heteroskedasticity

Example: WLS

The **WLS estimates and standard errors:**

```
#>                 Estimate Std. Error t value Pr(>|t|)
#> intercept_wls 618.78331    8.26929  74.829   <2e-16 ***
#> ratio_wls      -0.21314    0.37676  -0.566    0.572
#> income_wls      2.26493    0.09065  24.985   <2e-16 ***
```

The **OLS estimates** and **het.-robust standard errors:**

```
#>                Estimate Std. Error  t value  Pr(>|t|)
#> (Intercept) 638.729155   7.301234 87.48235 < 2.2e-16 ***
#> ratio        -0.648740   0.353340 -1.83602  0.067066 .
#> income        1.839112   0.114733 16.02949 < 2.2e-16 ***
```

Example: WLS

Alternative to doing your own weighting: feed `lm()` some `weights`.

```
lm(test_score ~ ratio + income, data = test_df, weights = enrollment)
```

Living with heteroskedasticity

In this example

- **Heteroskedasticity-robust standard errors** did not change our standard errors very much (relative to plain OLS standard errors).
- **WLS** changed our answers a bit—coefficients and standard errors.

Living with heteroskedasticity

In this example

- **Heteroskedasticity-robust standard errors** did not change our standard errors very much (relative to plain OLS standard errors).
- **WLS** changed our answers a bit—coefficients and standard errors.

These examples highlighted a few things:

1. Using the correct estimator for your standard errors really matters.[†]
2. Econometrics doesn't always offer an obviously *correct* route.

† Sit in on an economics seminar, and you will see what I mean.

Living with heteroskedasticity

In this example

- **Heteroskedasticity-robust standard errors** did not change our standard errors very much (relative to plain OLS standard errors).
- **WLS** changed our answers a bit—coefficients and standard errors.

These examples highlighted a few things:

1. Using the correct estimator for your standard errors really matters.[†]
2. Econometrics doesn't always offer an obviously *correct* route.

To see #1, let's run a simulation.

† Sit in on an economics seminar, and you will see what I mean.

Simulation

Let's examine a simple linear regression model with heteroskedasticity.

$$y_i = \underbrace{\beta_0}_{=1} + \underbrace{\beta_1}_{=10} x_i + u_i$$

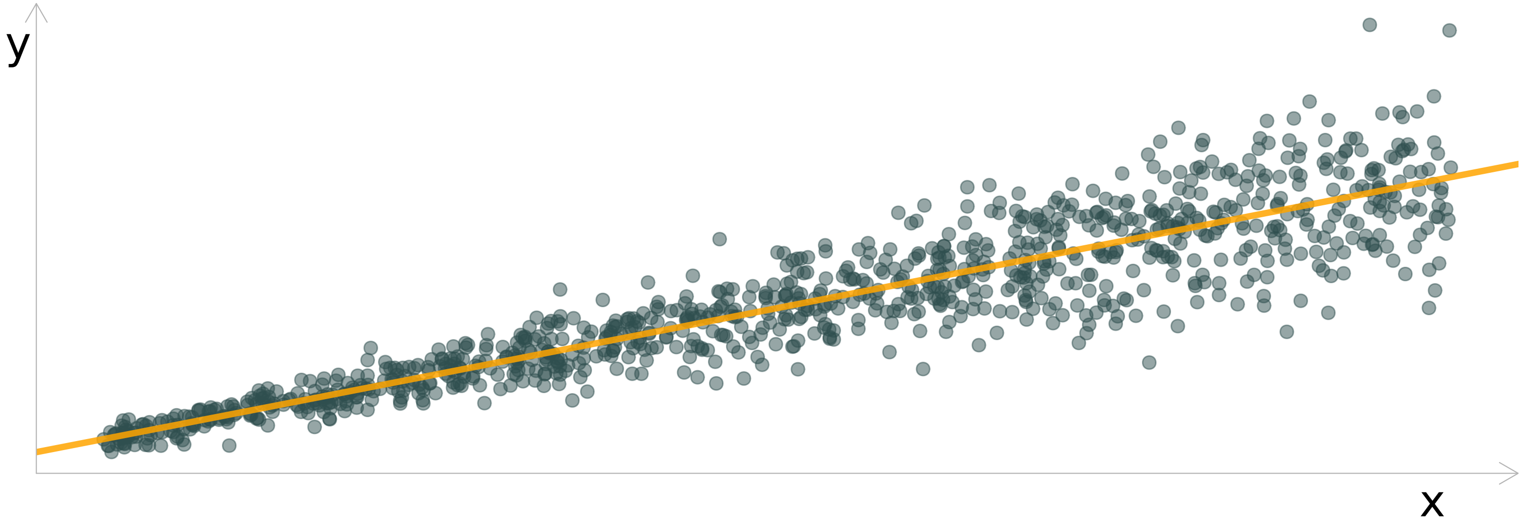
where $\mathbf{Var}(u_i|x_i) = \sigma_i^2 = \sigma^2 x_i^2$.

Simulation

Let's examine a simple linear regression model with heteroskedasticity.

$$y_i = \underbrace{\beta_0}_{=1} + \underbrace{\beta_1}_{=10} x_i + u_i$$

where $\mathbf{Var}(u_i|x_i) = \sigma_i^2 = \sigma^2 x_i^2$.

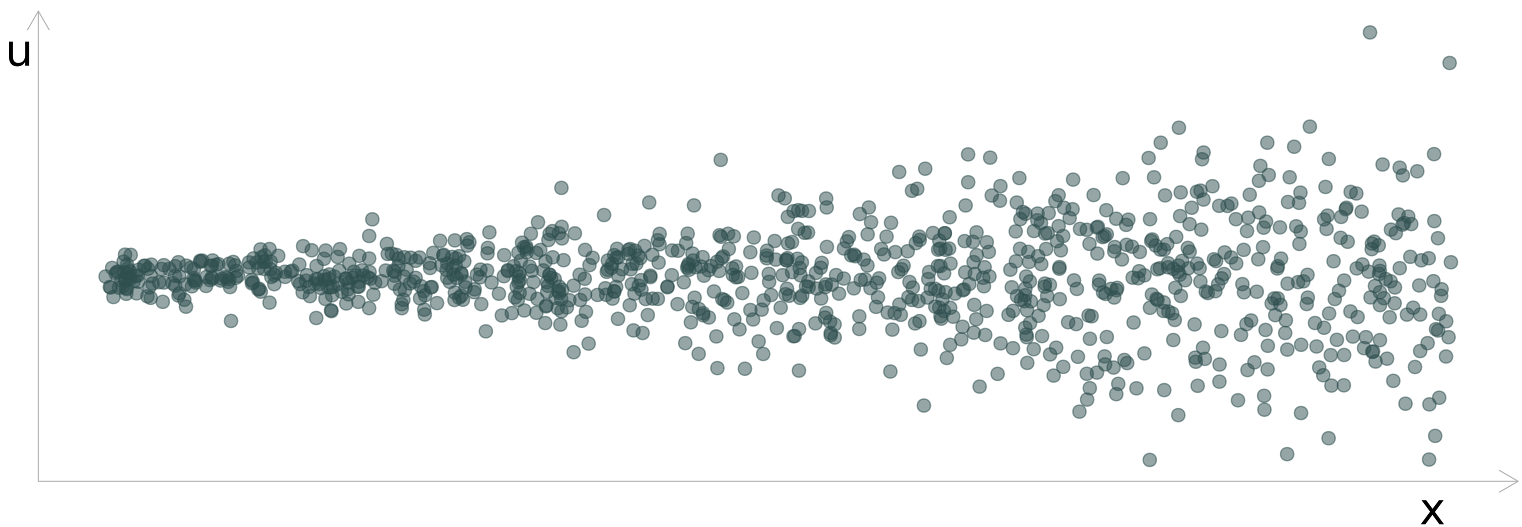


Simulation

Let's examine a simple linear regression model with heteroskedasticity.

$$y_i = \underbrace{\beta_0}_{=1} + \underbrace{\beta_1}_{=10} x_i + u_i$$

where $\mathbf{Var}(u_i | x_i) = \sigma_i^2 = \sigma^2 x_i^2$.



Simulation

Note regarding WLS:

Since $\mathbf{Var}(u_i|x_i) = \sigma^2 x_i^2$,

$$\mathbf{Var}(u_i|x_i) = \sigma^2 h(x_i) \implies h(x_i) = x_i^2$$

WLS multiplies each variable by $1/\sqrt{h(x_i)} = 1/x_i$.

Living with heteroskedasticity

Simulation

In this simulation, we want to compare

1. The **efficiency** of
   - OLS
   - WLS with correct weights: $h(x_i) = x_i$
   - WLS with incorrect weights: $h(x_i) = \sqrt{x_i}$
2. How well our **standard errors** perform (via confidence intervals) with
   - Plain OLS standard errors
   - Heteroskedasticity-robust standard errors
   - WLS standard errors

Simulation

The simulation plan:

```
Do 10,000 times:

 1. Generate a sample of size 30 from the population

 2. Calculate/save OLS and WLS (×2) estimates for β₁

 3. Calculate/save standard errors for β₁ using

      ◦ Plain OLS standard errors
      ◦ Heteroskedasticity-robust standard errors
      ◦ WLS (correct)
      ◦ WLS (incorrect)
```

Living with heteroskedasticity

Simulation

For one iteration of the simulation:

Code to generate the data...

```r
# Parameters
b0 <- 1
b1 <- 10
s2 <- 1
# Sample size
n <- 30
# Generate data
sample_df <- tibble(
  x = runif(n, 0.5, 1.5),
  y = b0 + b1 * x + rnorm(n, 0, sd = s2 * x^2)
)
```

Simulation

For one iteration of the simulation:

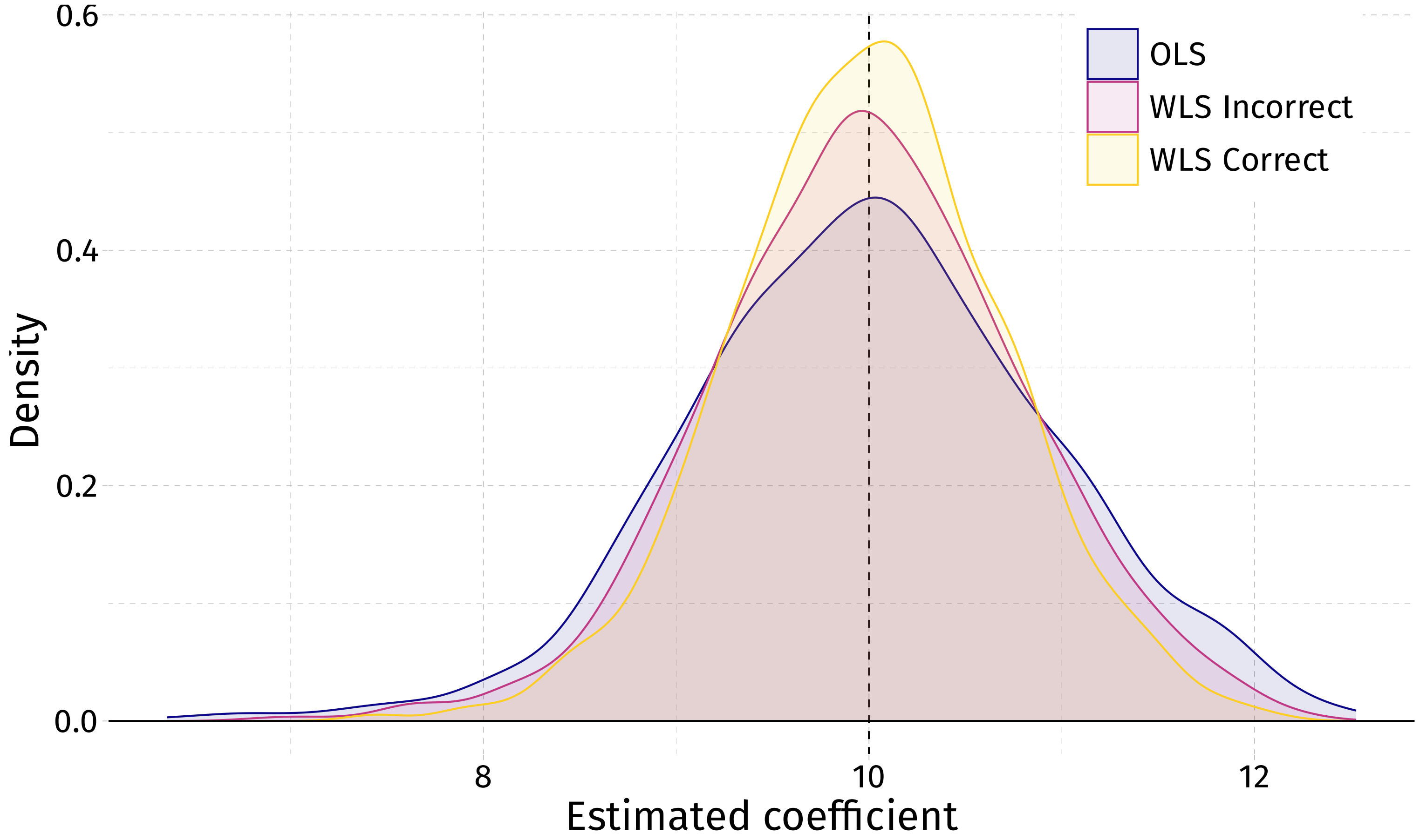
Code to estimate our coefficients and standard errors…

```
# OLS
ols ← feols(y ~ x, data = sample_df)
# WLS: Correct weights
wls_t ← lm(y ~ x, data = sample_df, weights = 1/x^2)
# WLS: Correct weights
wls_f ← lm(y ~ x, data = sample_df, weights = 1/x)
# Coefficients and standard errors
summary(ols, vcov = 'iid')
summary(ols, vcov = 'hetero')
summary(wls_t)
summary(wls_f)
```

Then save the results.

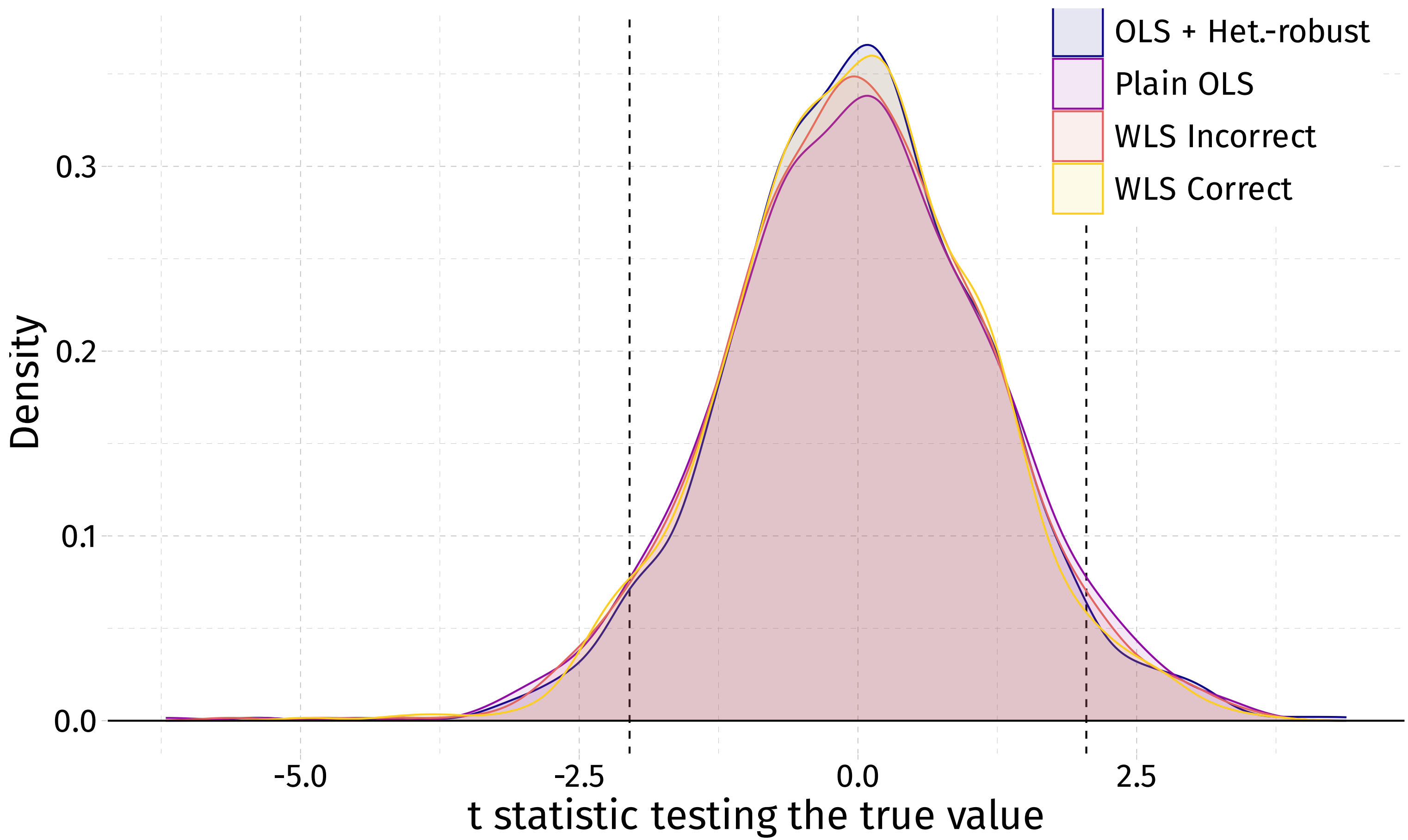
Living with heteroskedasticity

Simulation: Coefficients



Living with heteroskedasticity

Simulation: Inference



Simulation: Results

Summarizing our simulation results (10,000 iterations)

Estimation: Summary of $\hat{\beta}_1$'s

Estimator	**Mean**	**S.D.**
OLS	10.003	0.929
WLS Correct	9.990	0.702
WLS Incorrect	9.995	0.798

Simulation: Results

Summarizing our simulation results (10,000 iterations)

Estimation: Summary of $\hat{\beta}_1$'s

Estimator	Mean	S.D.
OLS	10.003	0.929
WLS Correct	9.990	0.702
WLS Incorrect	9.995	0.798

Inference: % of times we reject β_1

Estimators	% Reject
OLS + Het.-robust	7.4
OLS + Homosk.	9.1
WLS Correct	7.6
WLS Incorrect	8.3

Going further...

Similar violations

Assumptions

Recall our old assumption that led to this heteroskedasticity discussion:

> 5. The disurbances have **constant variance** σ^2
> and **zero covariance**, *i.e.*,
>
> - $\boldsymbol{E}\left[u_i^2 | X_i\right] = \text{Var}(u_i | X_i) = \sigma^2 \implies \text{Var}(u_i) = \sigma^2$
> - $\text{Cov}\left(u_i,\, u_j | X_i,\, X_j\right) = \boldsymbol{E}\left[u_i u_j | X_i,\, X_j\right] = 0$ for $i \neq j$

Similar violations

Assumptions

Recall our old assumption that led to this heteroskedasticity discussion:

> 5. The disurbances have **constant variance** σ^2
> and **zero covariance**, *i.e.*,
>
> - $\boldsymbol{E}\left[u_i^2 \middle| X_i\right] = \text{Var}(u_i|X_i) = \sigma^2 \implies \text{Var}(u_i) = \sigma^2$
> - $\text{Cov}\left(u_i,\, u_j \middle| X_i,\, X_j\right) = \boldsymbol{E}\left[u_i u_j \middle| X_i,\, X_j\right] = 0$ for $i \neq j$

Violation of constant variance $=$ **heteroskedasticity**

It's also possible (likely) that the **disturbances are correlated**.

Ignoring this correlation is even more problematic for inference.

Correlated disturbances

The problem

In many cases, **observations' disturbances** (u_i, u_j) **can be correlated**.

Remember

- The **disturbance** represents the un-included variables that affect y.
- Some **observations** in the sample may *relate* to other observations.

If these observation-level relationships extend to the variables in the disturbance, then disturbances can correlate.

$$\implies \text{Cov}\big(u_i, u_j | X_i, X_j\big) \neq 0.$$

Correlated disturbances

The problem

In many cases, **observations' disturbances** $(u_i,\, u_j)$ **can be correlated**.

Remember

- The **disturbance** represents the un-included variables that affect y.
- Some **observations** in the sample may *relate* to other observations.

If these observation-level relationships extend to the variables in the disturbance, then disturbances can correlate.

$$\implies \text{Cov}\big(u_i,\, u_j | X_i,\, X_j\big) \neq 0.$$

Ignoring this correlation can cause **big problems** in your inference.

Correlated disturbances

The intution

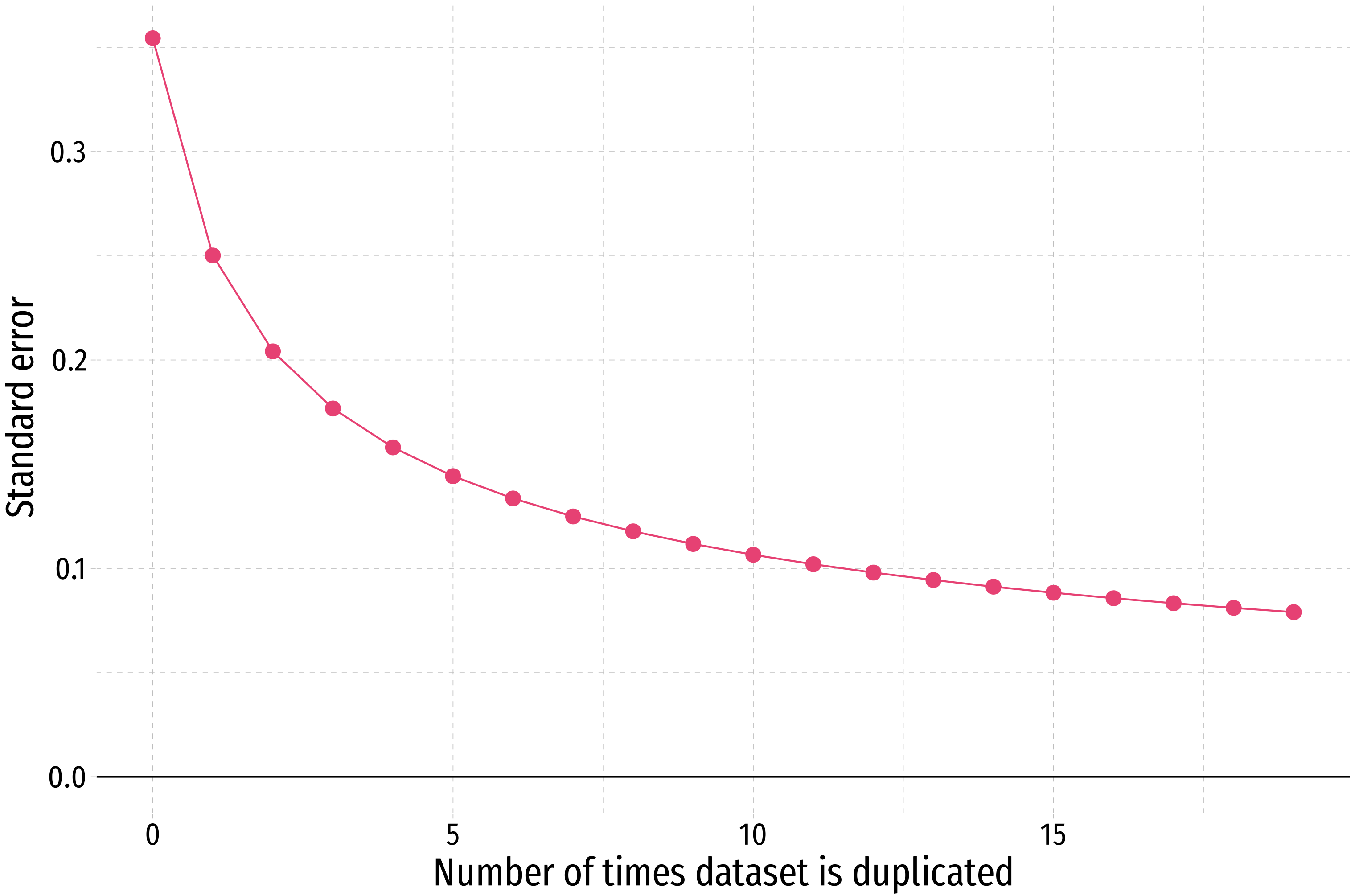
Why is ignoring this correlation problematic?

False precision: We can get "overconfident" in our knowledge.

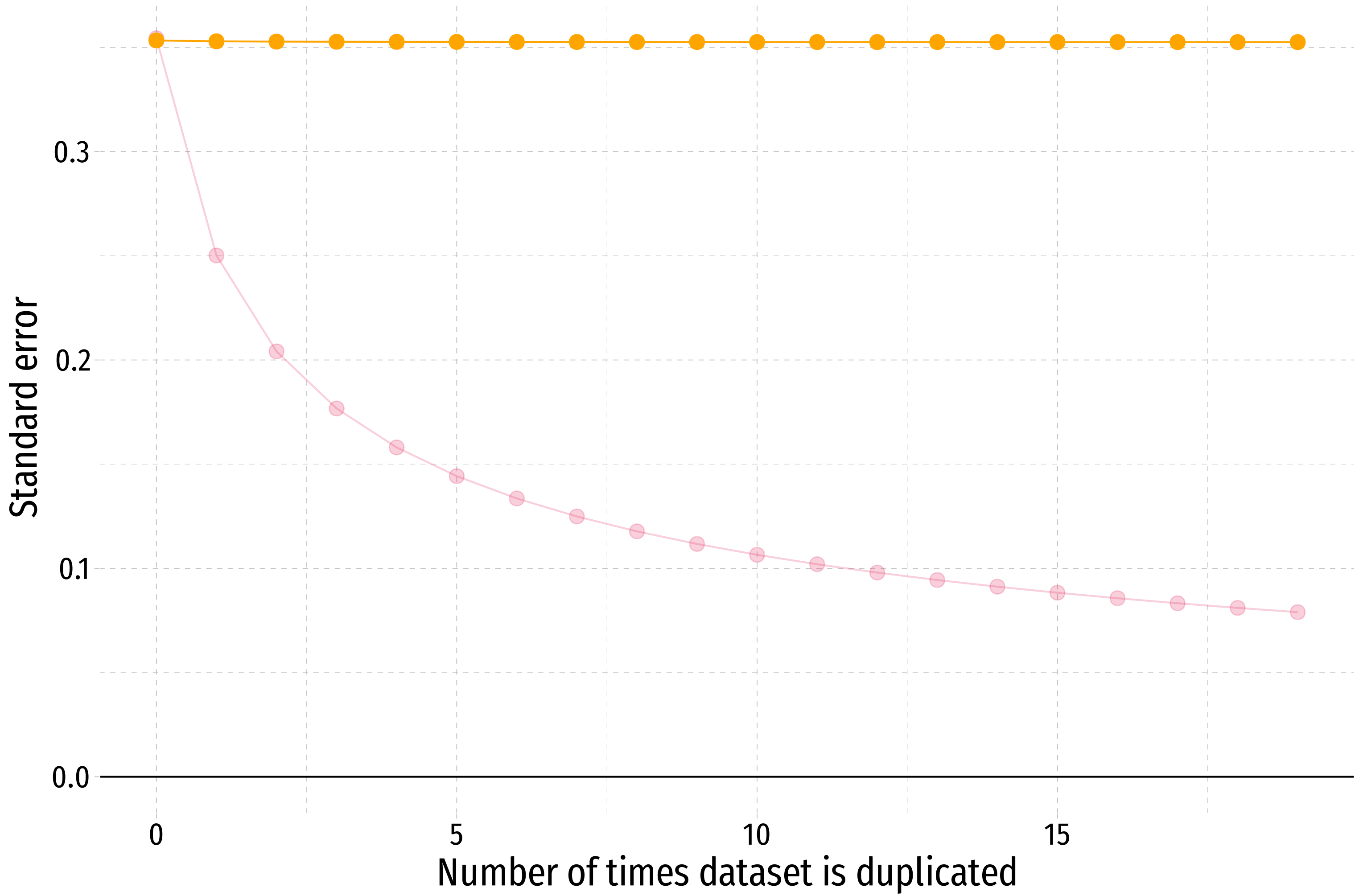
When we teating correlated observations as independent, we OLS thinks we're learning more than we actually are.

Extreme example: If duplicate your dataset (stack it on top of itself), plain OLS standard errors would decrease every time you duplicated the dataset.

The effect of duplicating our data on the OLS standard error of `ratio`.



Correcting our standard errors for clustering (observations' correlation).



Correlated disturbances

Examples

"Real" examples where disturbances might correlate:

- Students in a classroom (share teacher, curriculum, etc.)
- Counties in a state (share state-level policies/laws)
- Businesses in a city (share local economic shocks)
- Consecutive days in a sample (share events, weather, etc.)

Correlated disturbances

The solution

Just like we calculate *heteroskedasticity*-robust standard errors,
we can also calculate standard errors robust to correlated disturbances.

People call these *cluster-robust standard errors* (or just *clustered*).

From `fixest` package:

```
feols(y ~ x, data = fake_data, cluster = 'cluster_var')
```

or even

```
feols(y ~ x, data = fake_data, cluster = c('cluster1', 'cluster2'))
```

Final word

Better inference

1. You should default to assuming your data are heteroskedastic
2. Think about how your explanatory variables and/or disturbances correlate across observations.