Heteroskedasticity, Part 2

EC 421, Set 5

Edward Rubin
17 April 2019

Prologue

Schedule

Last Time

Heteroskedasticity: Issues and tests

Today

Living with heteroskedasticity

Upcoming

- First assignment! **Due at 11:59pm on Sunday** (4/21).
- No class/office hours on Monday (4/22).
- I do not have office hours tomorrow (4/18)
- GEs will send an announcement today about their office hours.

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?

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. e_i measures how far the i^{th} observation is from the **regression** line, *i.e.*,

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

Heteroskedasticity

Review

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

Answer 2:

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

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

By definition, $\mathrm{Var}(u_i|x_i) = \boldsymbol{E}\Big[u_i^2 - \underbrace{\boldsymbol{E}[u_i|x_i]^2}_{=0}\Big|x_i\Big] = \boldsymbol{E}\big[u_i^2\big|x_i\big]$

Thus, if we want to learn about the variance of u_i, we can focus on 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.

Heteroskedasticity

Review: Current assumptions

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

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

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

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

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

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

Heteroskedasticity

Review

Today we're focusing on assumption #5:

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

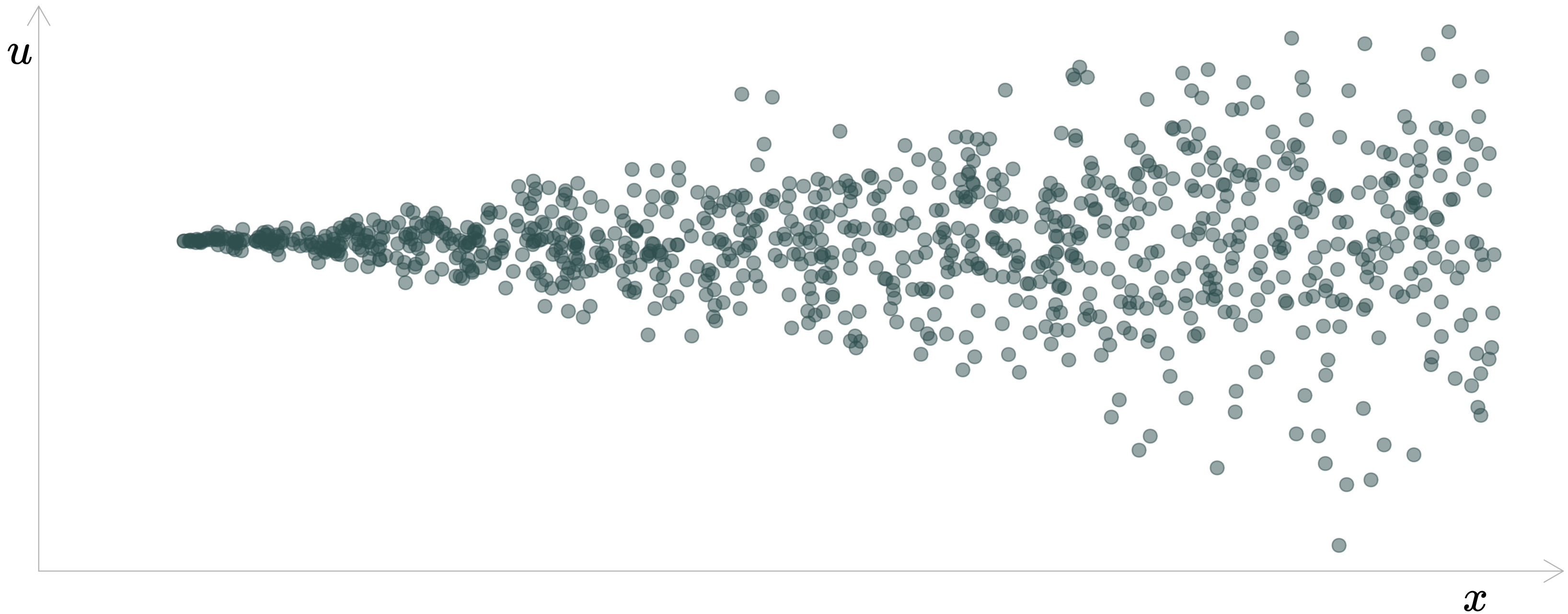
Heteroskedasticity: $\mathrm{Var}(u_i) = \sigma_i^2$ and $\sigma_i^2 \neq \sigma_j^2$ for some $i \neq j$.

Heteroskedasticity

Review

Classic example of heteroskedasticity: The funnel

Variance of u increases with x

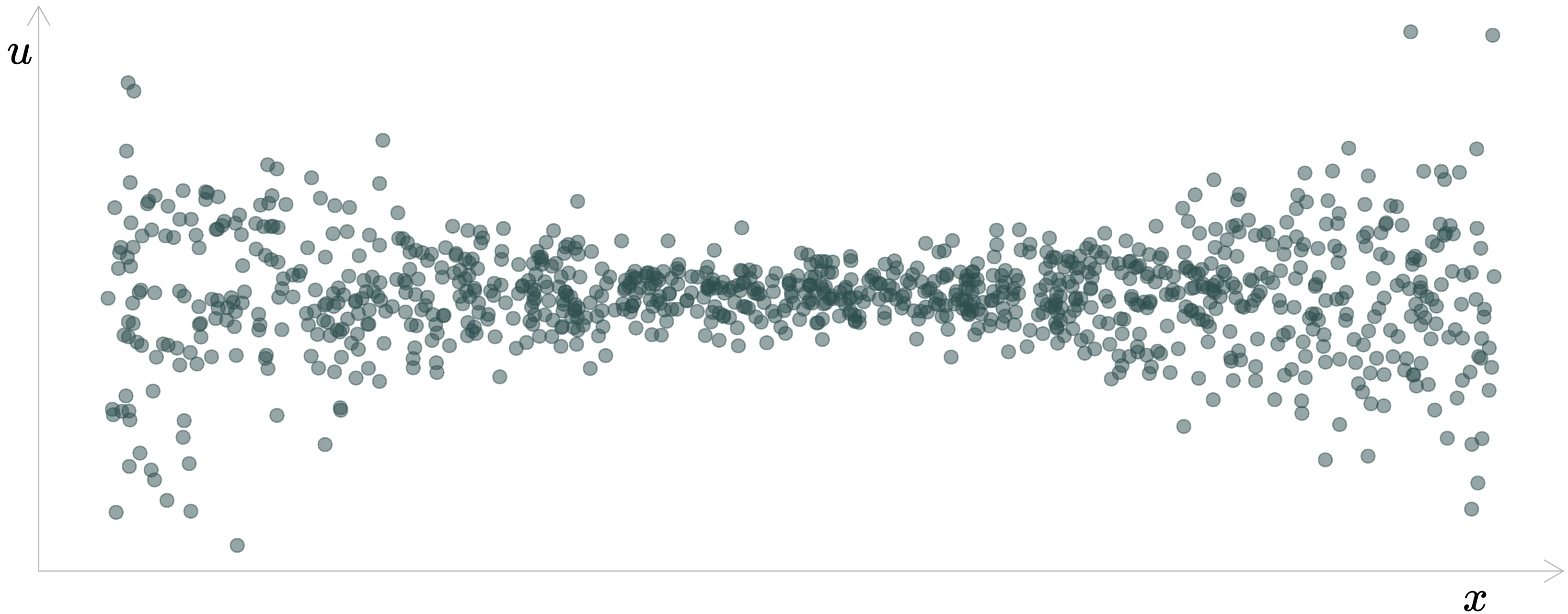


Heteroskedasticity

Review

Another example of heteroskedasticity: (double funnel?)

Variance of u increasing at the extremes of x

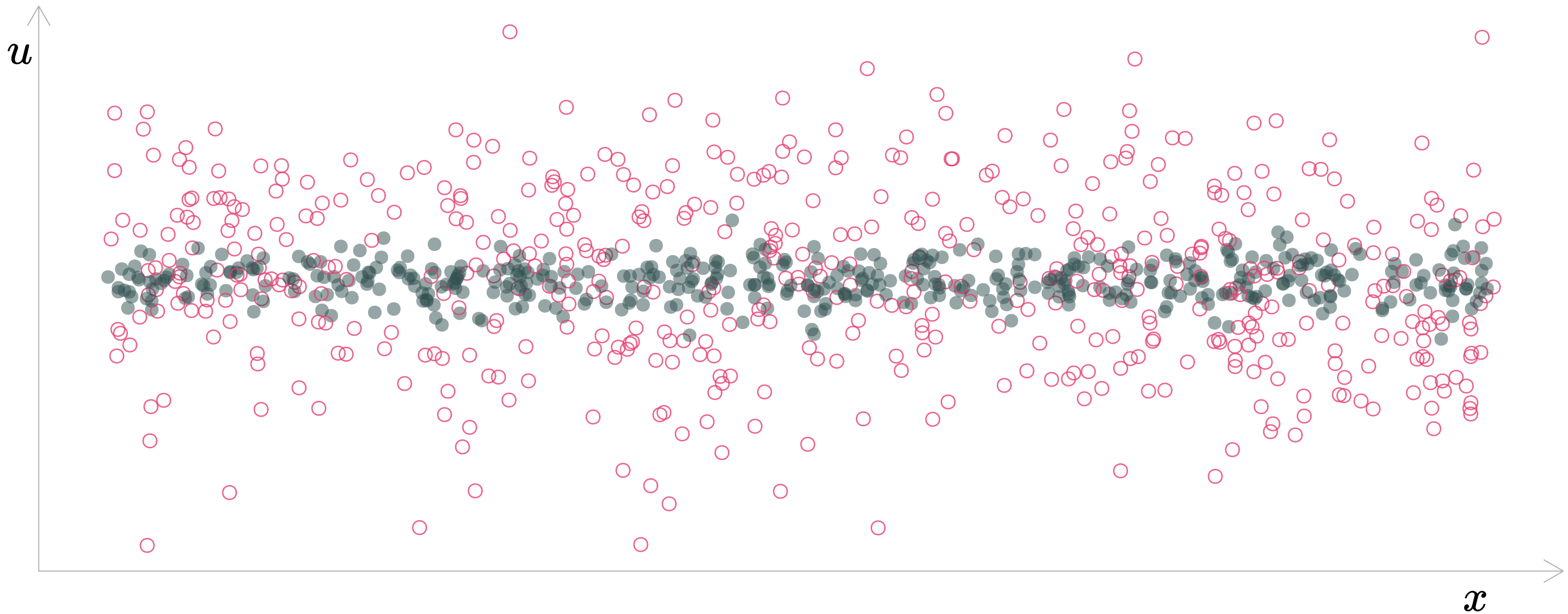


Heteroskedasticity

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
- Breusch-Pagan
- White

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.

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

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

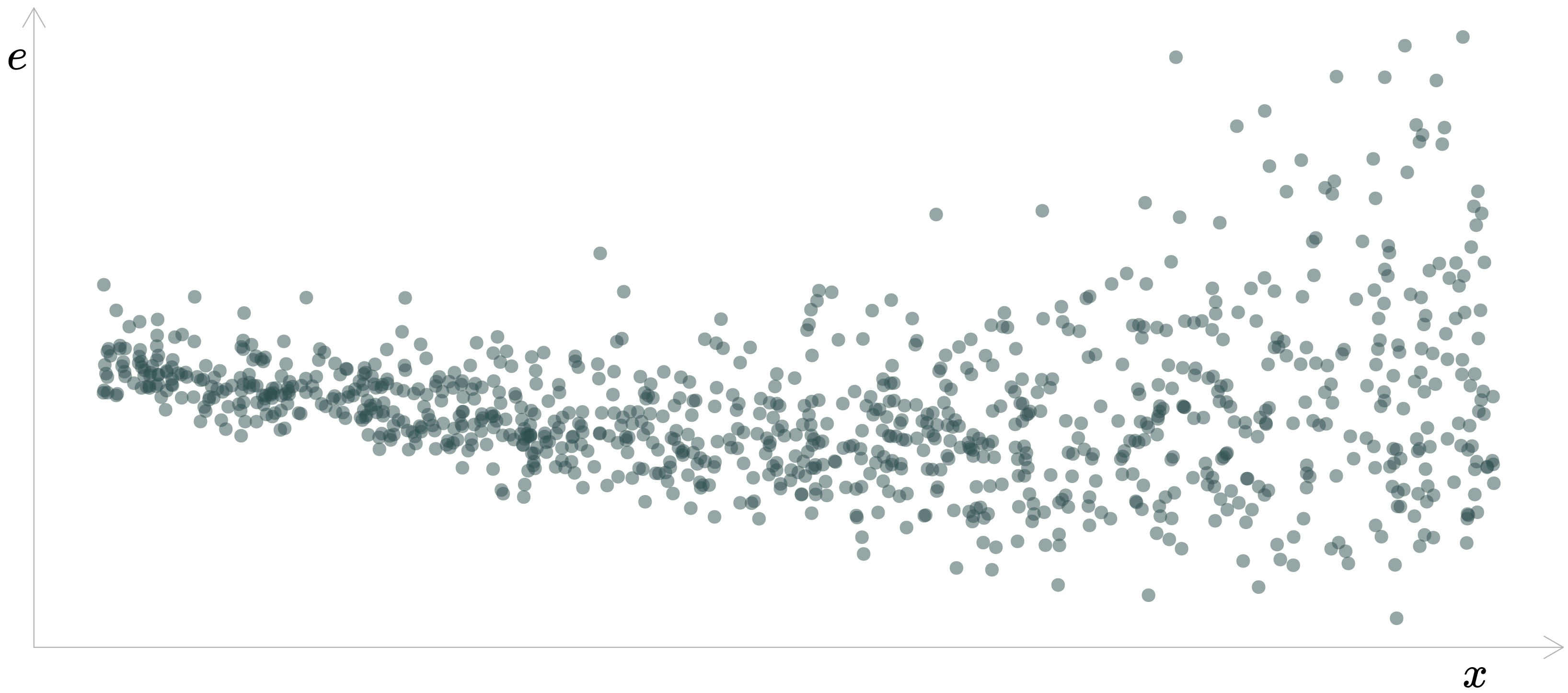
Then

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

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

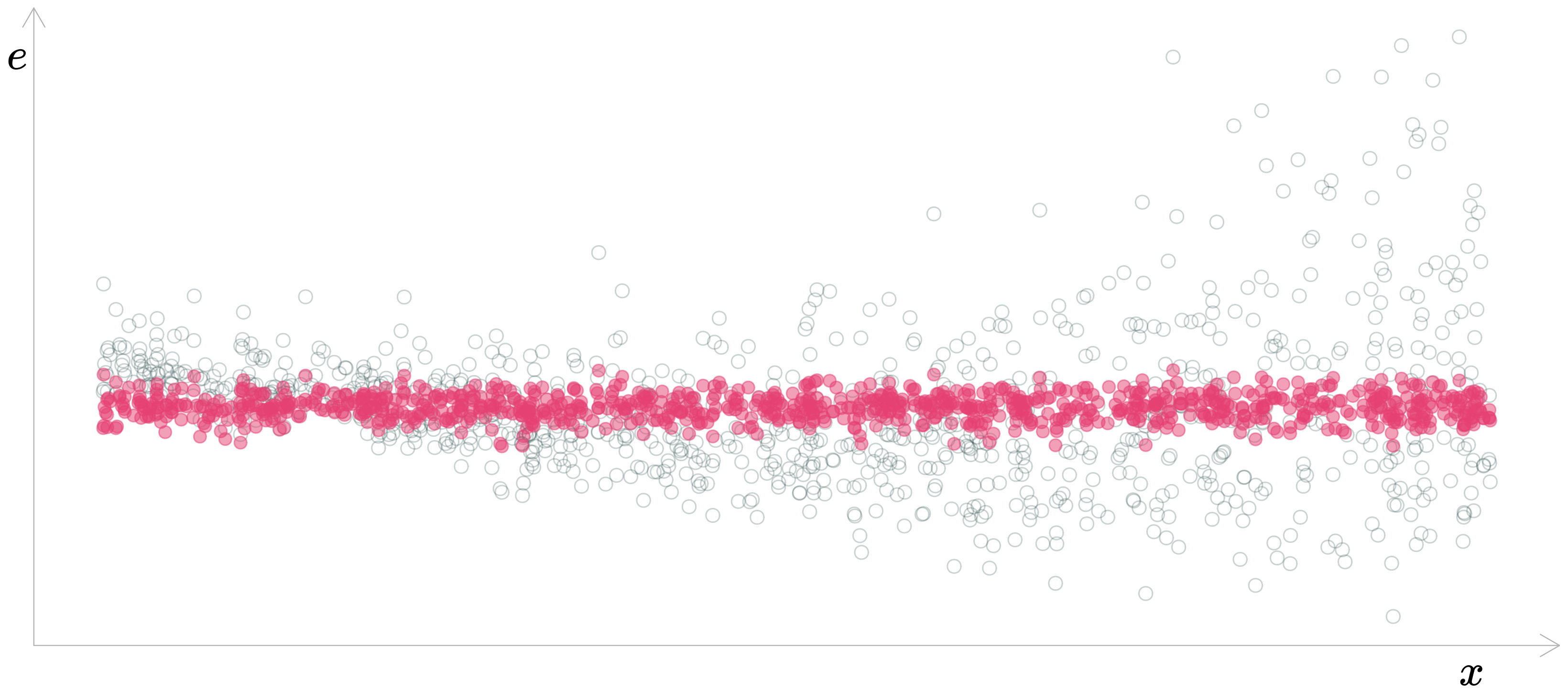
Misspecification

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



Misspecification

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



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.

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.*,

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

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 $h(x)$.

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.

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.

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

Homoskedasticity!

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.

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

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

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.

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

Examples

Back to our test-scores dataset…

```r
# 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 x 4
#>   test_score ratio income enrollment
#>        <dbl> <dbl>  <dbl>      <int>
#> 1       691.  17.9  22.7         195
#> 2       661.  21.5   9.82        240
```

Example: Model specification

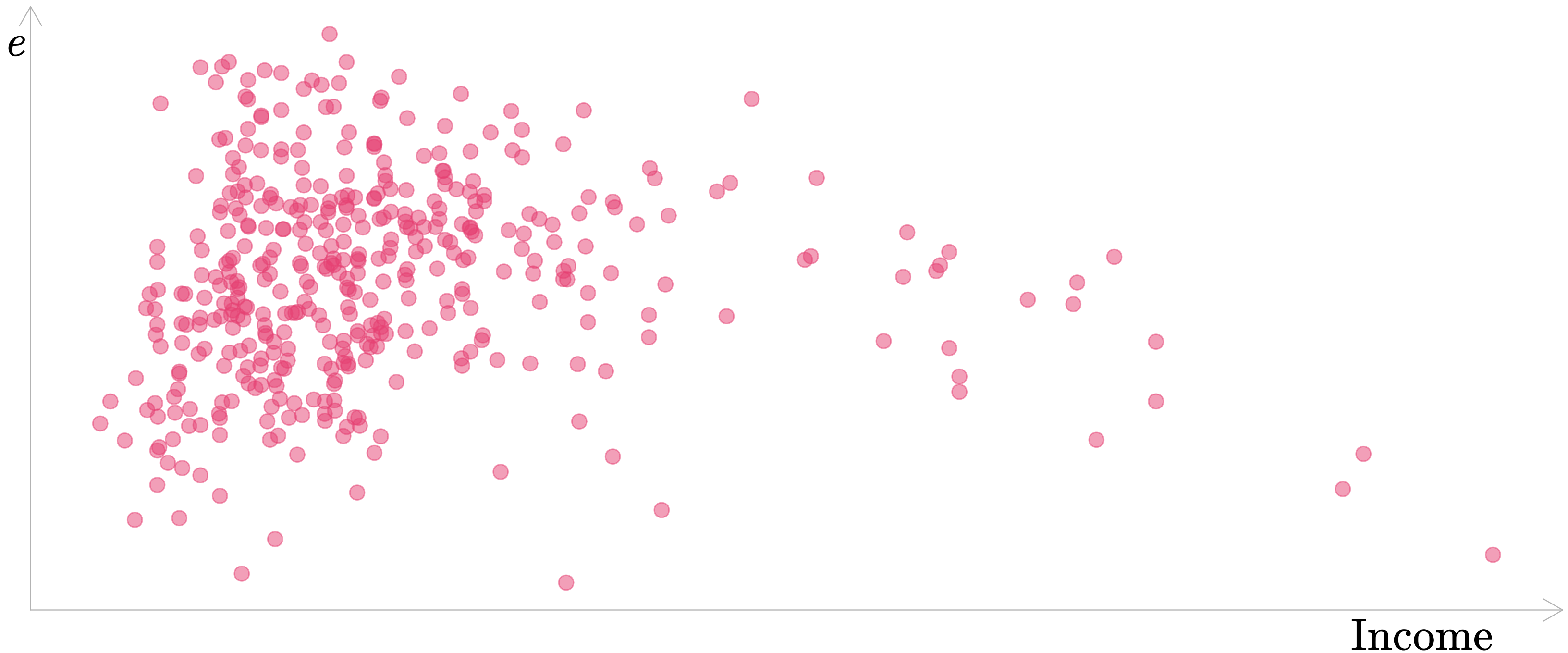
We found significant evidence of heteroskedasticity.

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

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)
```

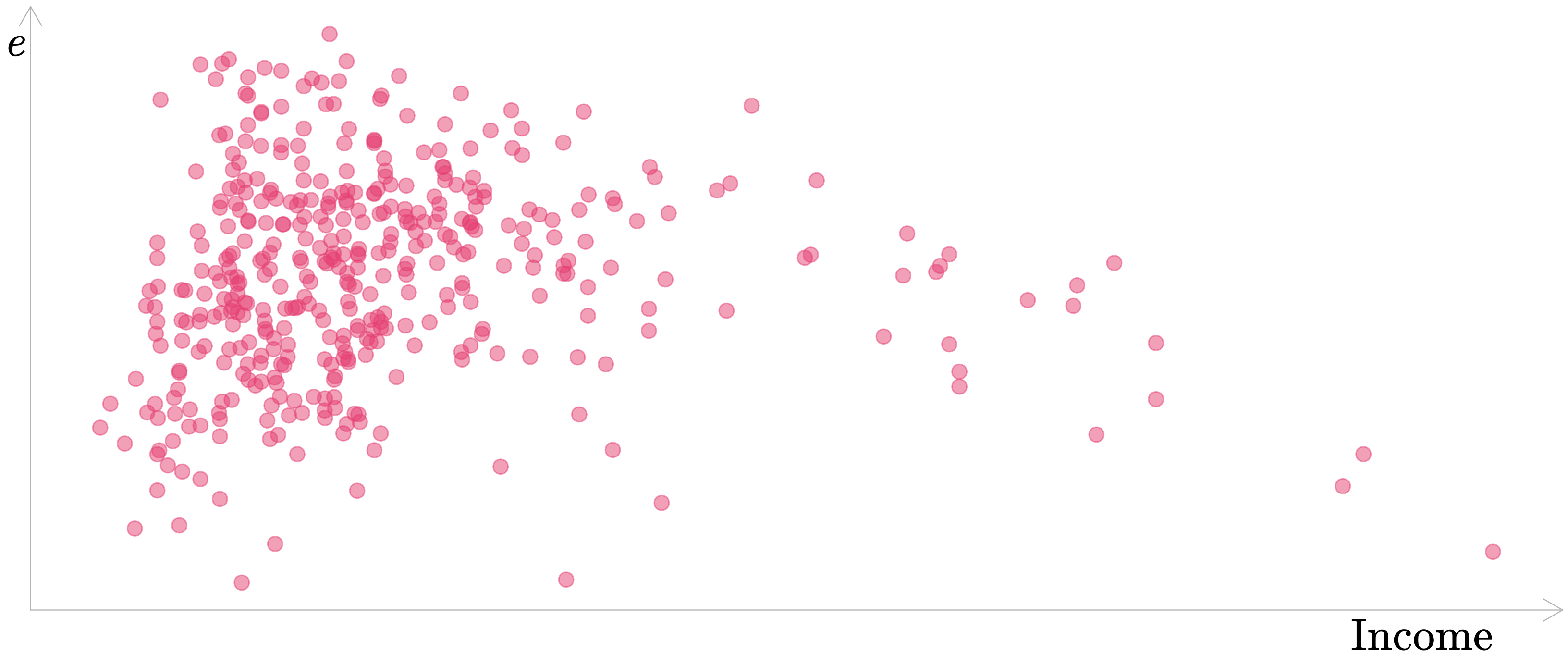


Living with heteroskedasticity

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)
```

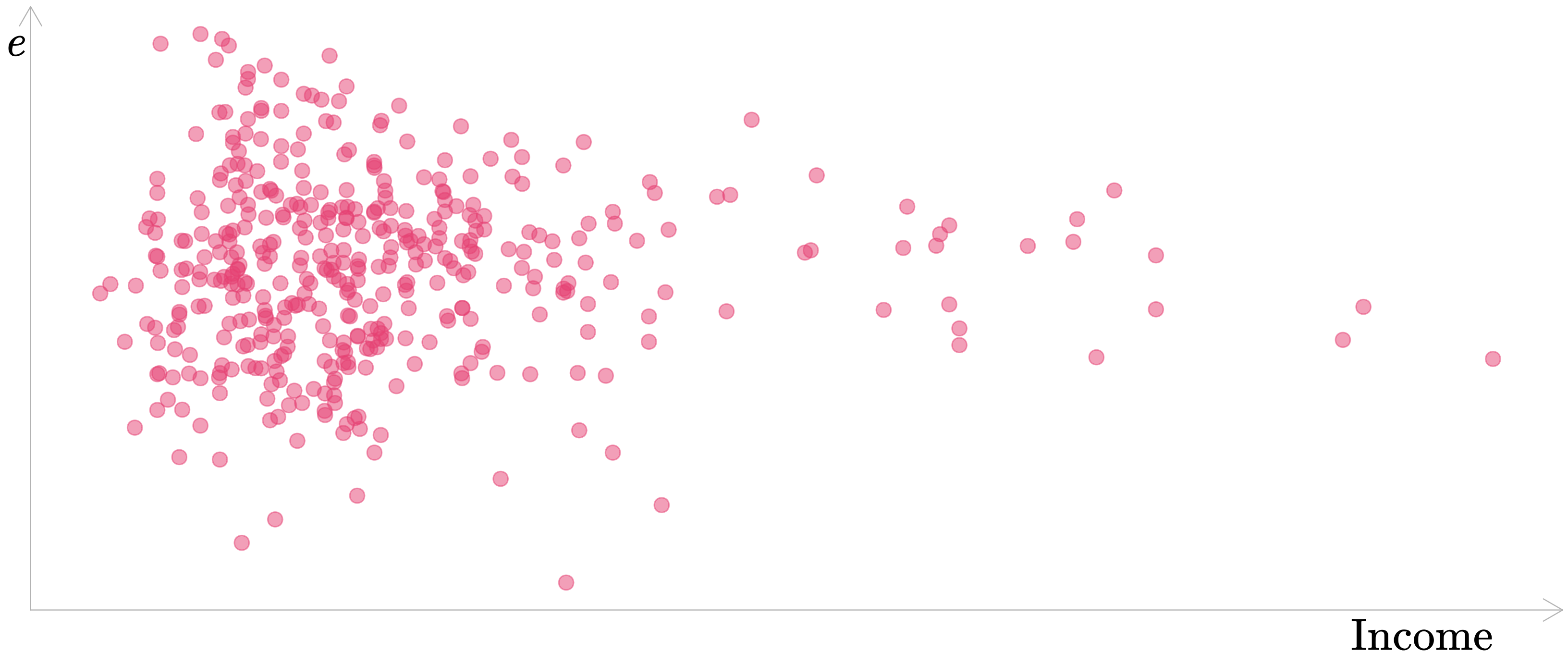


Living with heteroskedasticity

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.

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

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 `lfe` package in `R` to calculate standard errors.

Living with heteroskedasticity

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 `felm()` (instead of `lm()`)

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

Notice that `felm()` 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 `robust = T` option in `summary()`

```
# Het-robust standard errors with 'robust = T'
summary(test_reg, robust = T)
```

```
#>              Estimate Robust s.e t value Pr(>|t|)
#> (Intercept) 638.7292     7.3012  87.482   <2e-16 ***
#> ratio        -0.6487     0.3533  -1.836   0.0671 .
#> income        1.8391     0.1147  16.029   <2e-16 ***
```

Living with heteroskedasticity

Example: Het.-robust standard errors

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

```
summary(test_reg, robust = T)
```

```
#>              Estimate Robust s.e t value Pr(>|t|)
#> (Intercept) 638.7292      7.3012  87.482   <2e-16 ***
#> ratio        -0.6487      0.3533  -1.836   0.0671 .
#> income        1.8391      0.1147  16.029   <2e-16 ***
```

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

```
summary(test_reg, robust = F)
```

```
#>               Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 638.72915    7.44908  85.746   <2e-16 ***
#> ratio        -0.64874    0.35440  -1.831   0.0679 .
#> income        1.83911    0.09279  19.821   <2e-16 ***
```

Living with heteroskedasticity

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 $h(x_i)$.

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).

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

Example: WLS

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

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

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

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

```
# 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
)
```

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

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 Robust s.e t value Pr(>|t|)
#>  (Intercept) 638.7292     7.3012  87.482   <2e-16 ***
#>  ratio        -0.6487     0.3533  -1.836   0.0671 .
#>  income        1.8391     0.1147  16.029   <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.

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.

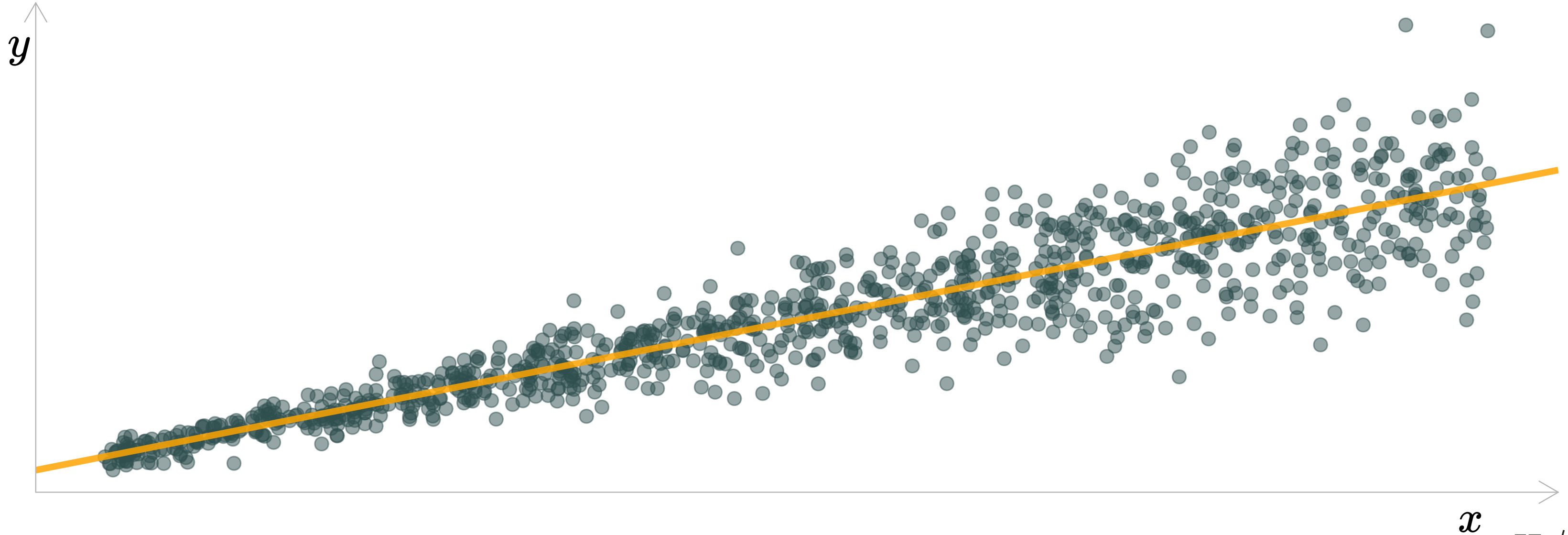
† Sit in on an economics seminar, and you will see what I mean.

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

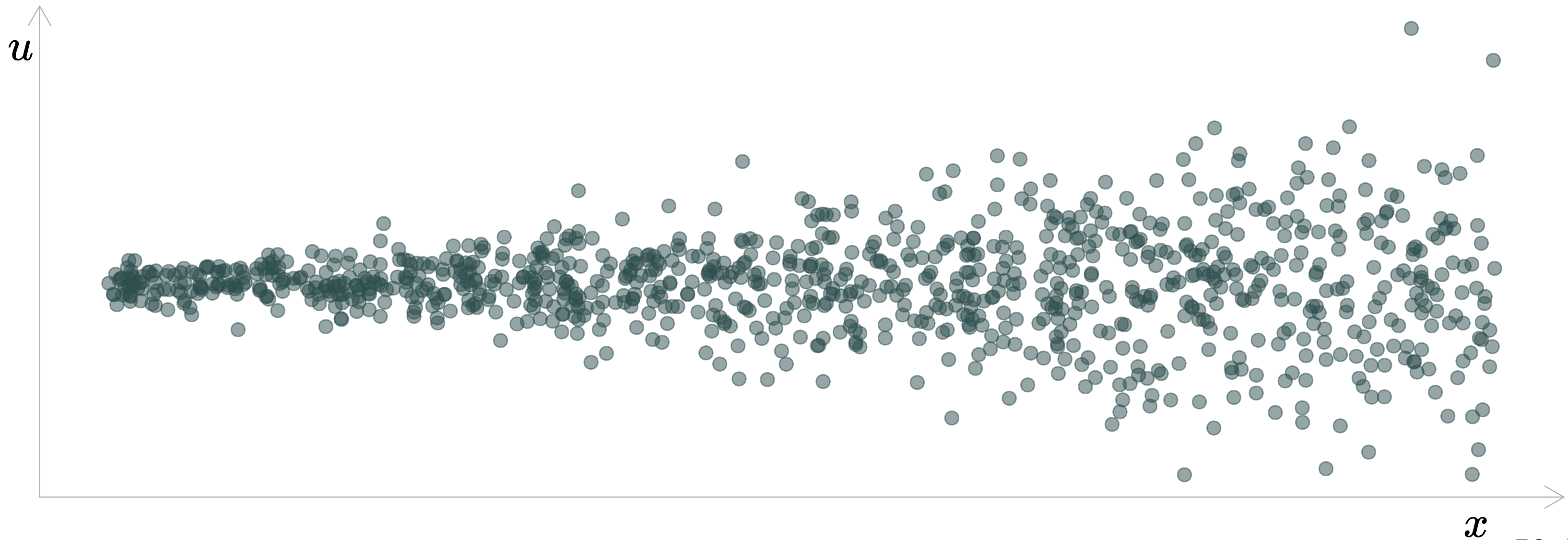


Simulation

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



Living with heteroskedasticity

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 $1/\sqrt{h(x_i)} = 1/x_i$.

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...

```
# 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:

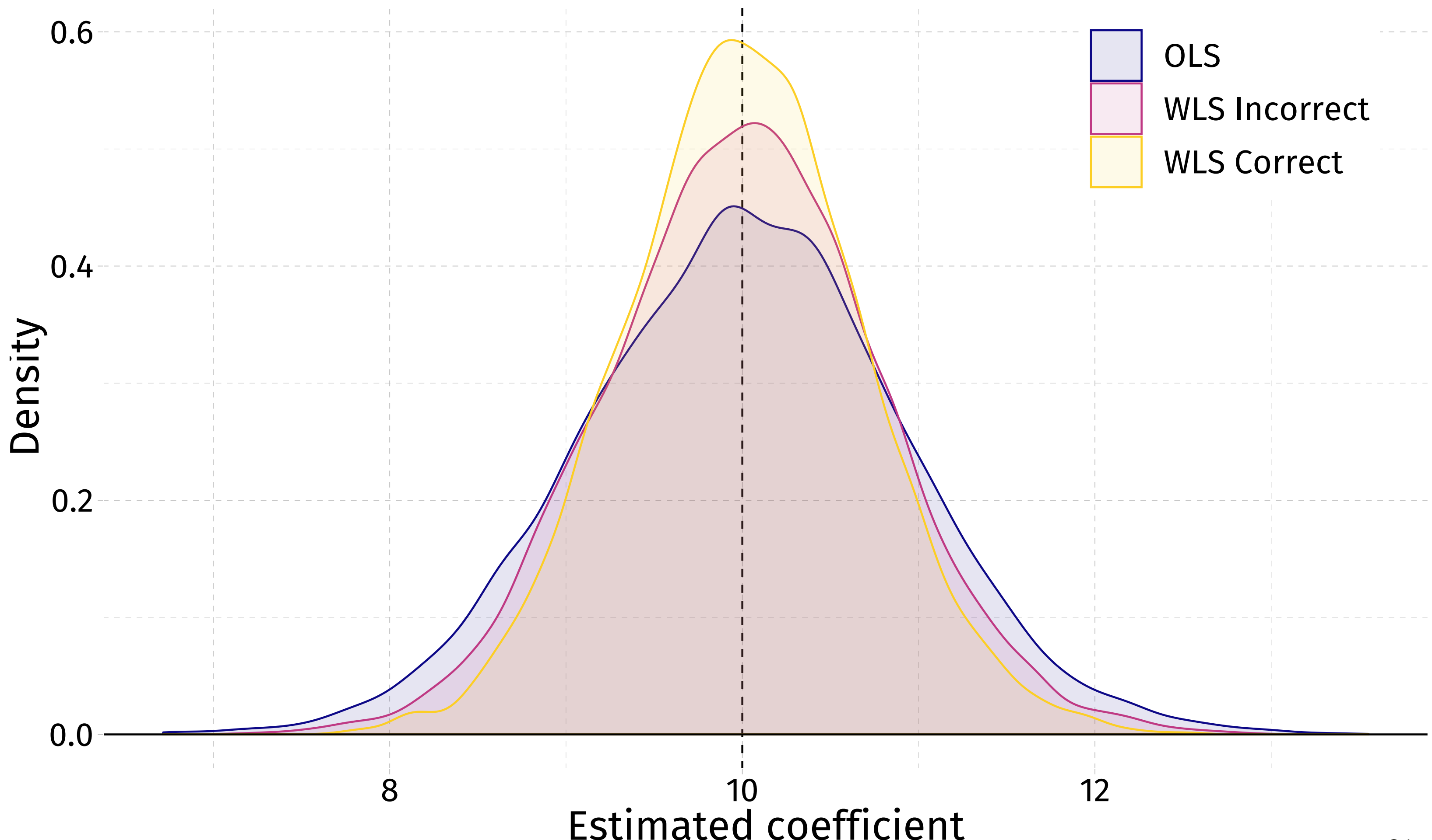
Code to estimate our coefficients and standard errors…

```
# OLS
ols ← felm(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, robust = F)
summary(ols, robust = T)
summary(wls_t)
summary(wls_f)
```

Then save the results.

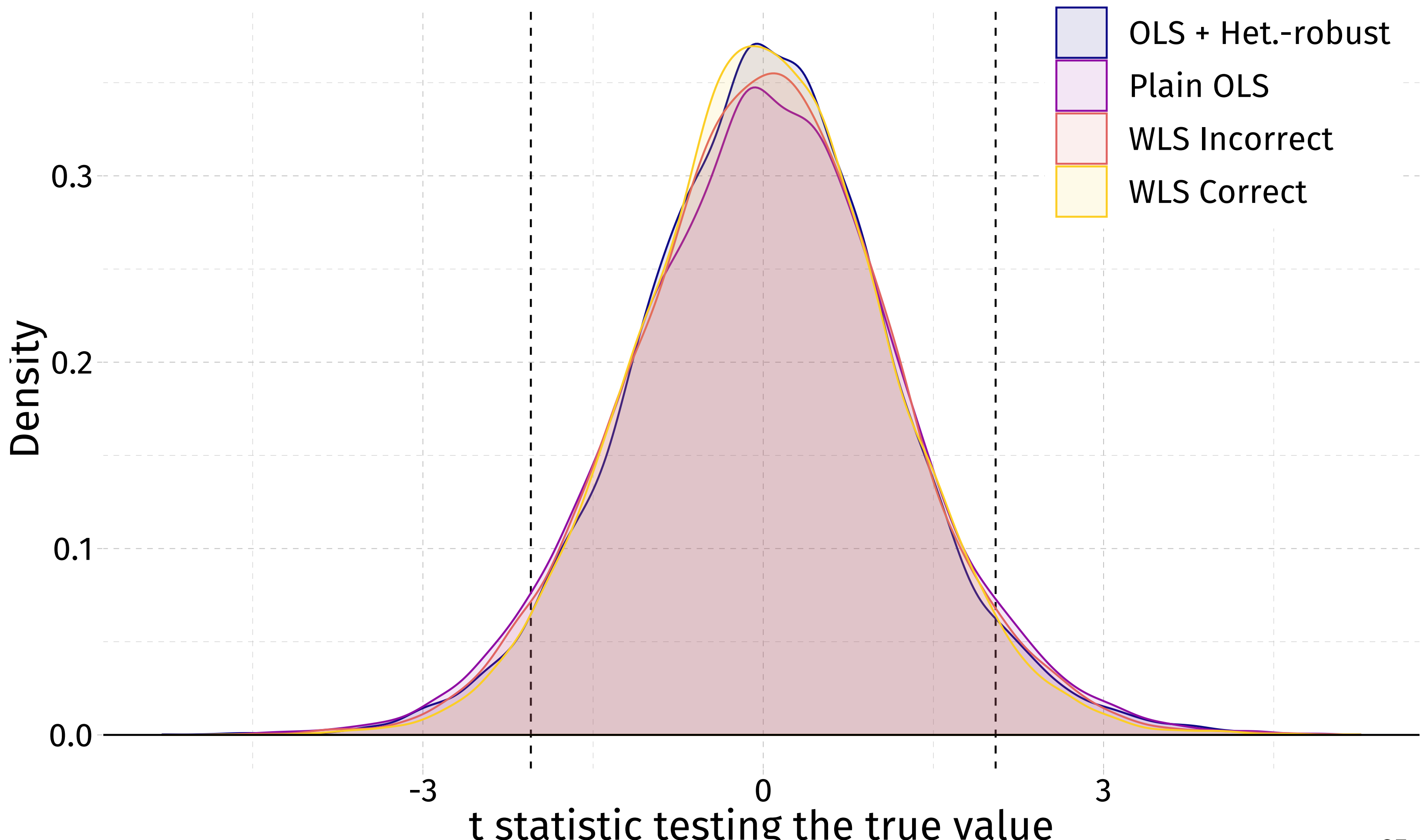
Living with heteroskedasticity

Simulation: Coefficients



Living with heteroskedasticity

Simulation: Inference



Living with heteroskedasticity

Simulation: Results

Summarizing our simulation results (10,000 iterations)

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

Estimator	Mean	S.D.
OLS	10.009	0.910
WLS Correct	10.005	0.682
WLS Incorrect	10.007	0.777

Inference: % of times we reject β_1

Estimators	% Reject
OLS + Het.-robust	7.6
OLS + Homosk.	9.0
WLS Correct	6.3
WLS Incorrect	7.9