Heteroskedasticity, Part 2

EC 421, Set 5

Edward Rubin
22 January 2019

Prologue

Schedule

Last Time

Heteroskedasticity: Issues and tests

Today

Living with heteroskedasticity

This week

First assignment! **Due at 11:59pm on 27 Jan. 2019.**

R showcase

R Markdown

- Simple mark-up language for for combining/creating documents, equations, figures, R, and more
- Basics of Markdown
- *E.g.*, `**I'm bold**`, `*I'm italic*`, `I ← "code"`

Econometrics with R

- (Currently) free, online textbook
- Written and published using R (and probably R Markdown)
- *Warning:* I haven't read this book yet.

Related: Tyler Ransom has a great cheatsheet for econometrics.

Heteroskedasticity: Review

Heteroskedasticity: Review

Let's recall our **current assumptions**

Heteroskedasticity: Review

Let's recall our **current assumptions**

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

Heteroskedasticity: Review

Let's recall our **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.

Heteroskedasticity: Review

Let's recall our **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

Let's recall our **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**: $\boldsymbol{E}[u|X] = 0$ ($\implies$ $\boldsymbol{E}[u] = 0$).

Heteroskedasticity: Review

Let's recall our **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 $i \neq j$

Heteroskedasticity: Review

Let's recall our **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.*,
>
> - $\boldsymbol{E}\left[u_i^2 | X_i\right] = \text{Var}(u_i | X) = \sigma^2 \implies \text{Var}(u_i) = \sigma^2$
> - $\text{Cov}\left(u_i,\, u_j | X_i,\, X_j\right) = \boldsymbol{E}\left[u_i u_j | X_i,\, X_j\right] = 0$ for $i \neq j$

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.*,
>
> - $\boldsymbol{E}\left[u_i^2 | X_i\right] = \text{Var}(u_i | X) = \sigma^2 \implies \text{Var}(u_i) = \sigma^2$
> - $\text{Cov}\left(u_i,\, u_j | X_i,\, X_j\right) = \boldsymbol{E}\left[u_i u_j | X_i,\, X_j\right] = 0$ for $i \neq j$

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

Violation of this assumption:

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

Heteroskedasticity: Review

Today we're focusing on assumption #5:

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

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

Violation of this assumption:

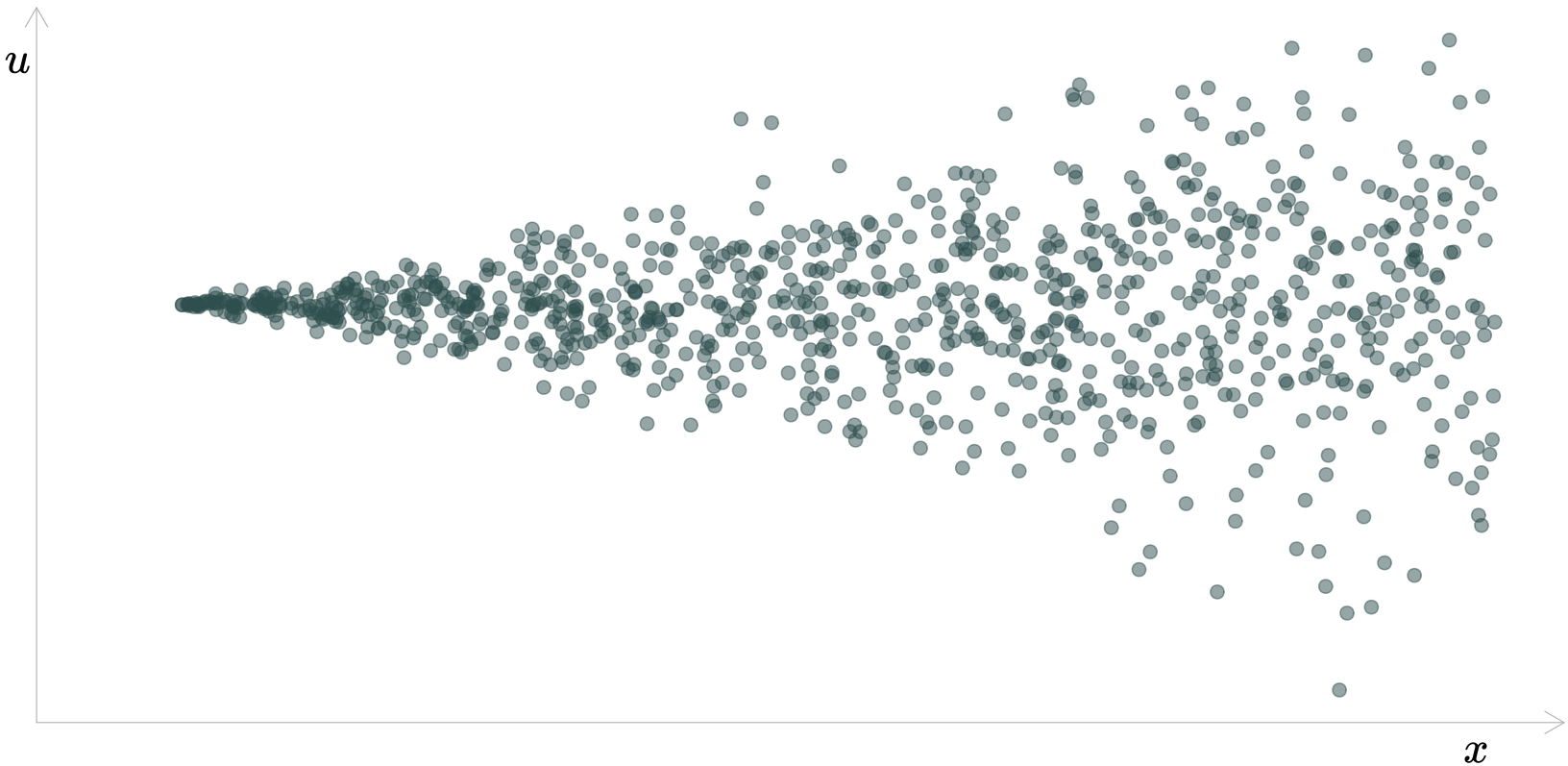
Heteroskedasticity: $\text{Var}(u_i) = \sigma_i^2$ and $\sigma_i^2 \neq \sigma_j^2$ for some $i \neq j$.

In other words: Our disturbances have different variances.

Heteroskedasticity: Review

Classic example of heteroskedasticity: The funnel

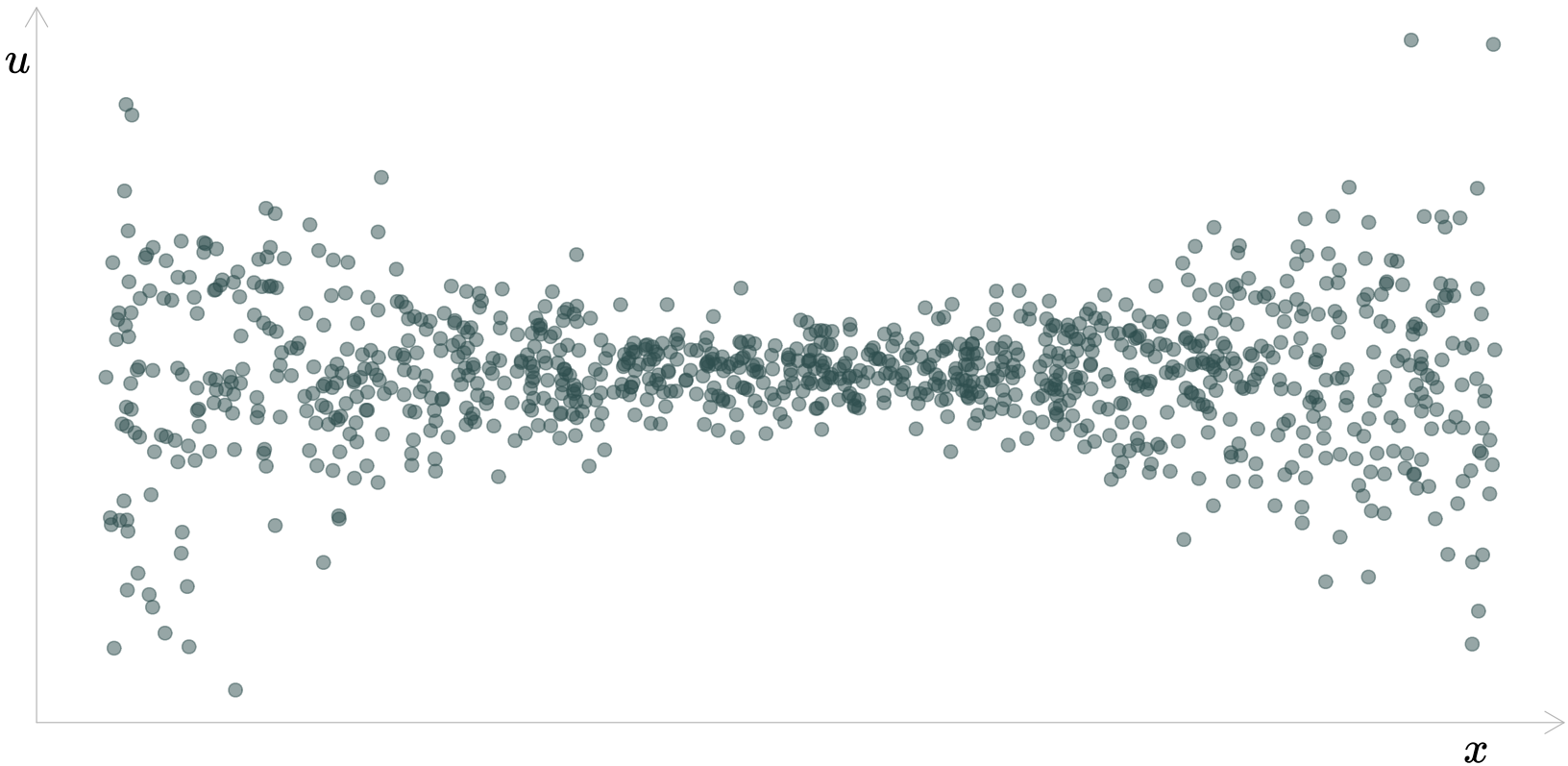
Variance of u increases with x



Heteroskedasticity: Review

Another example of heteroskedasticity: (double funnel?)

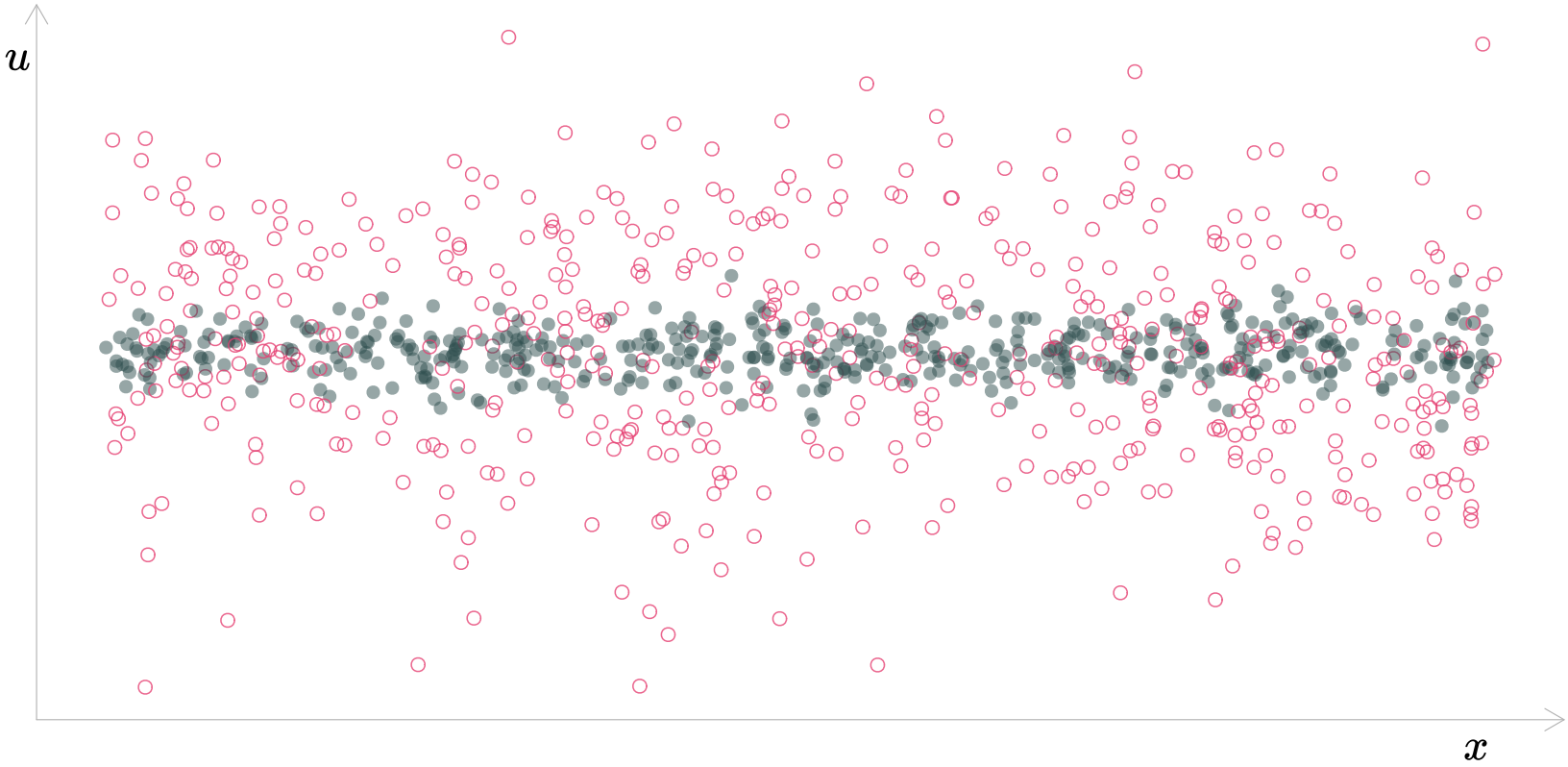
Variance of u increasing at the extremes of x



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 (henceforth: X).

Testing for heteroskedasticity

We have some tests that may help us detect heteroskedasticity.

- Goldfeld-Quandt
- Breusch-Pagan
- White

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).

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).

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.

Living with heteroskedasticity

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

However, we omit x^2 and estimate

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

Then

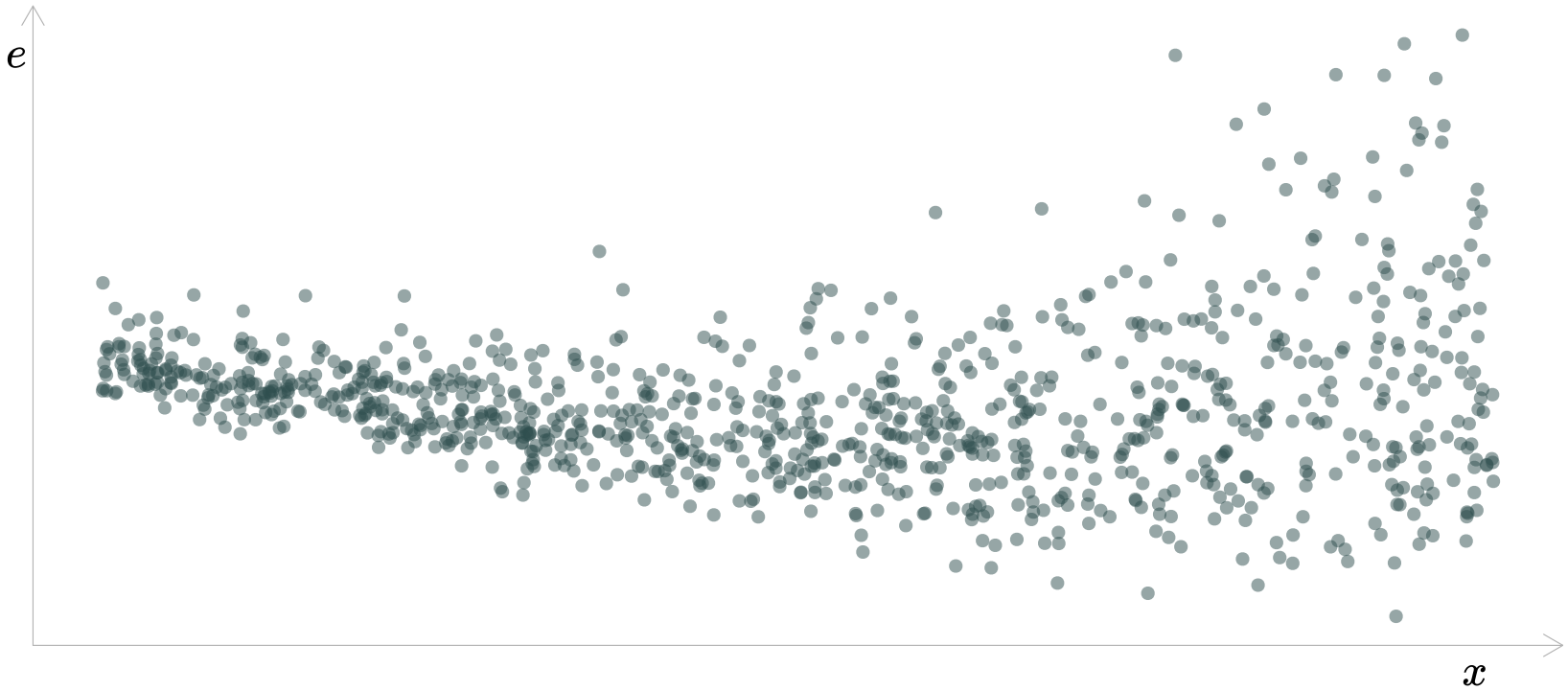
$$w_i = u_i + \beta_2 x_i^2 \implies \mathbf{Var}(w_i) = f(x_i)$$

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

Living with heteroskedasticity

Misspecification

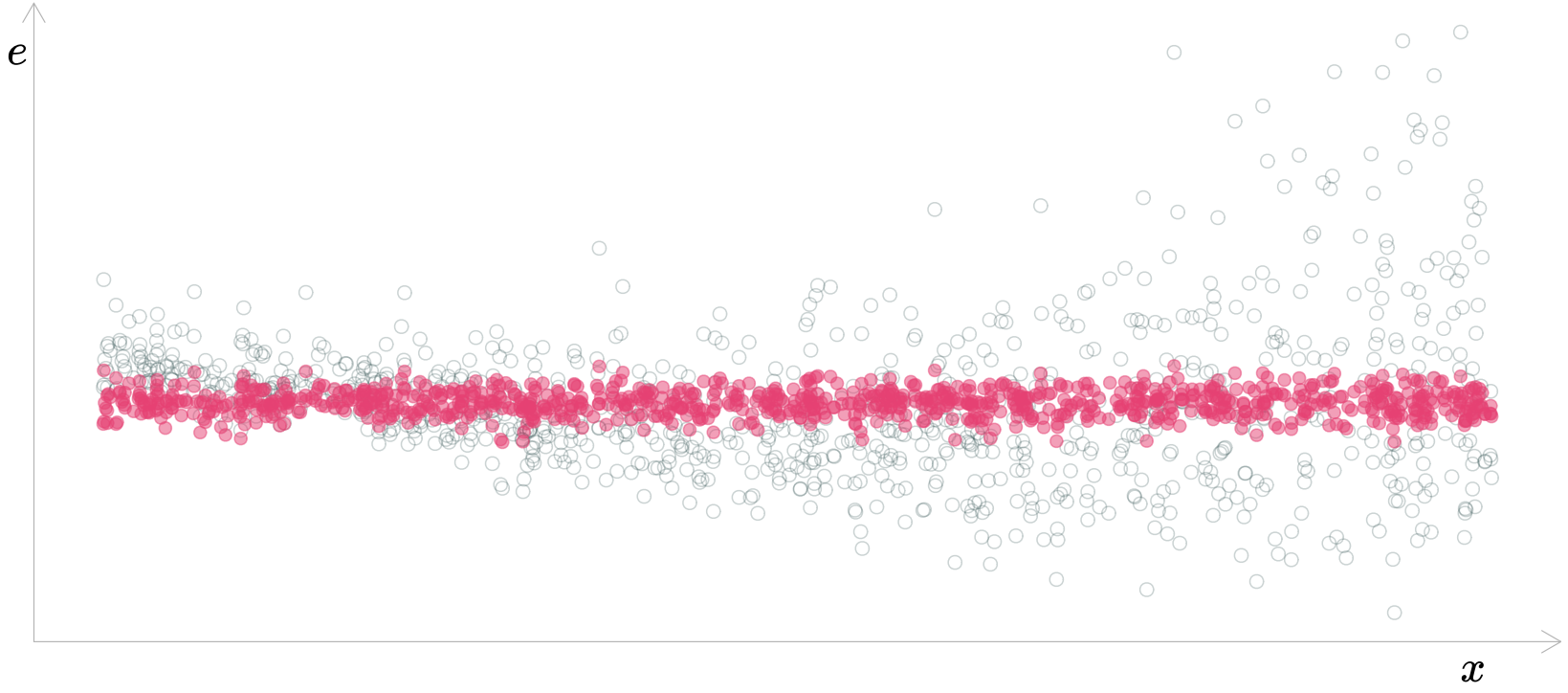
Truth: $\log(y_i) = \beta_0 + \beta_1 x_i + u_i$ **Misspecification:** $y_i = \beta_0 + \beta_1 x_i + v_i$



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.

Living with heteroskedasticity

Misspecification

More generally:

Misspecification problem: Incorrect specification of the regression model can cause heteroskedasticity.

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

Living with heteroskedasticity

Misspecification

More generally:

Misspecification problem: Incorrect specification of the regression model can cause heteroskedasticity.

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.

Living with heteroskedasticity

Misspecification

More generally:

Misspecification problem: Incorrect specification of the regression model can cause heteroskedasticity.

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: Adjusting the specification often doesn't solve the problem.

Living with heteroskedasticity

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

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

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

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

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

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

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

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

Living with heteroskedasticity

Weighted least squares

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

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

Whereas (1) is heteroskedastic,

Living with heteroskedasticity

Weighted least squares

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

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

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

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

Living with heteroskedasticity

Weighted least squares

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

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

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

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

Why is (2) homoskedastic?

Living with heteroskedasticity

Weighted least squares

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

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

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

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

Why is (2) homoskedastic?

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

Living with heteroskedasticity

Weighted least squares

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

Living with 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

Recall: We can write the OLS estimator for β_1 as

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

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

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

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

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

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.

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.

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

Living with heteroskedasticity

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.

Living with heteroskedasticity

Heteroskedasticity-robust standard errors

These standard errors go by many names

- Heteroskedasticity-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
test_df ← select(Caschool, test_score = testscr, ratio = str, income = avginc)
# Format as tibble
test_df ← as_tibble(test_df)
# View first 2 rows of the dataset
head(test_df, 2)
```

```
#> # A tibble: 2 x 3
#>   test_score ratio income
#>        <dbl> <dbl>  <dbl>
#> 1       691.  17.9  22.7
#> 2       661.  21.5   9.82
```

Example: Model specification

We found significant evidence of heteroskedasticity.

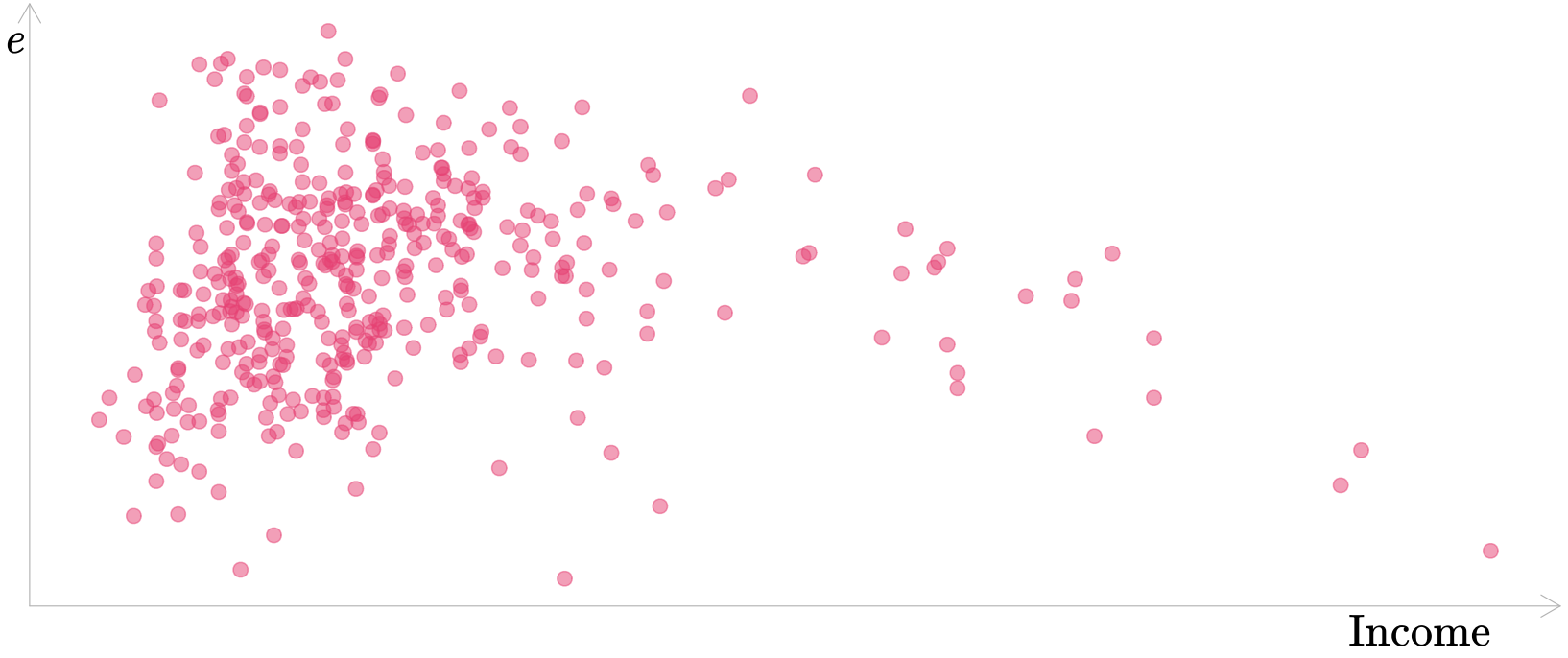
Let's check if it was due to misspecifying our model.

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)
```

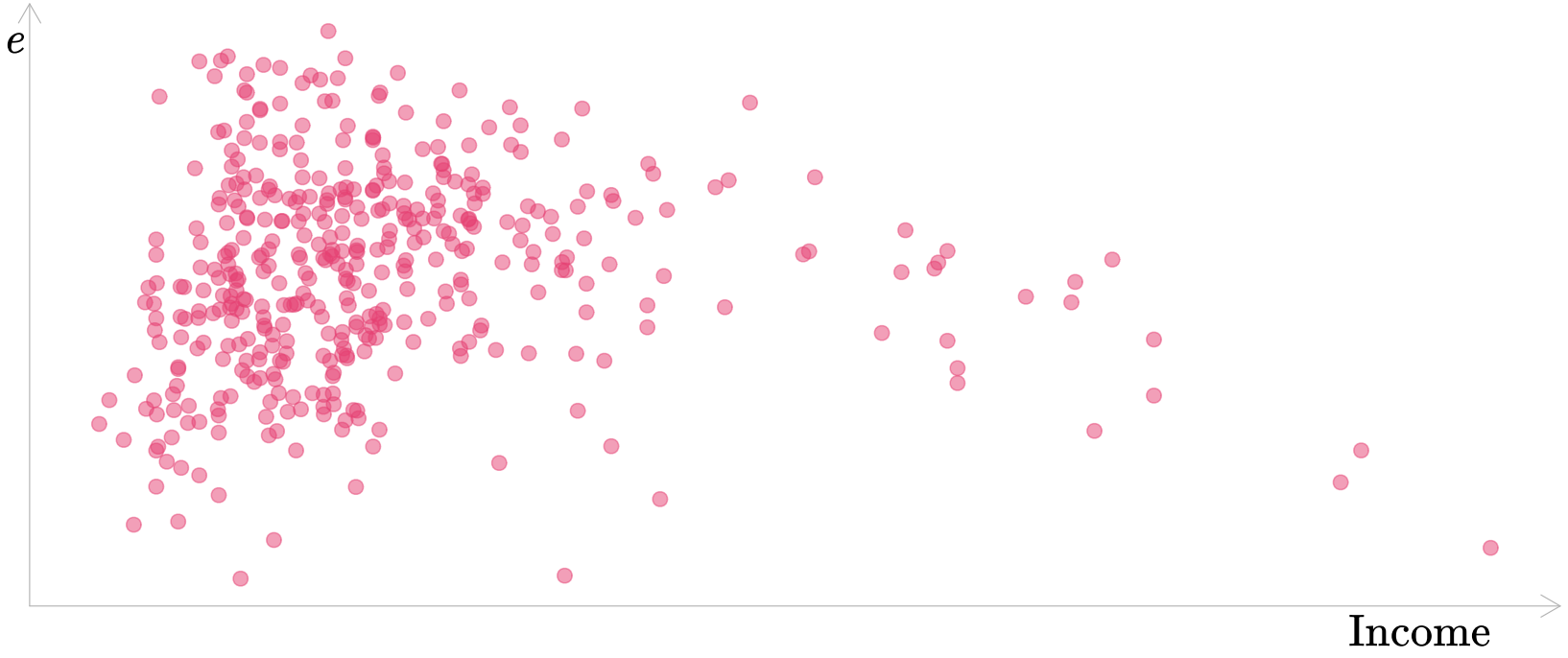


Living with heteroskedasticity

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)
```

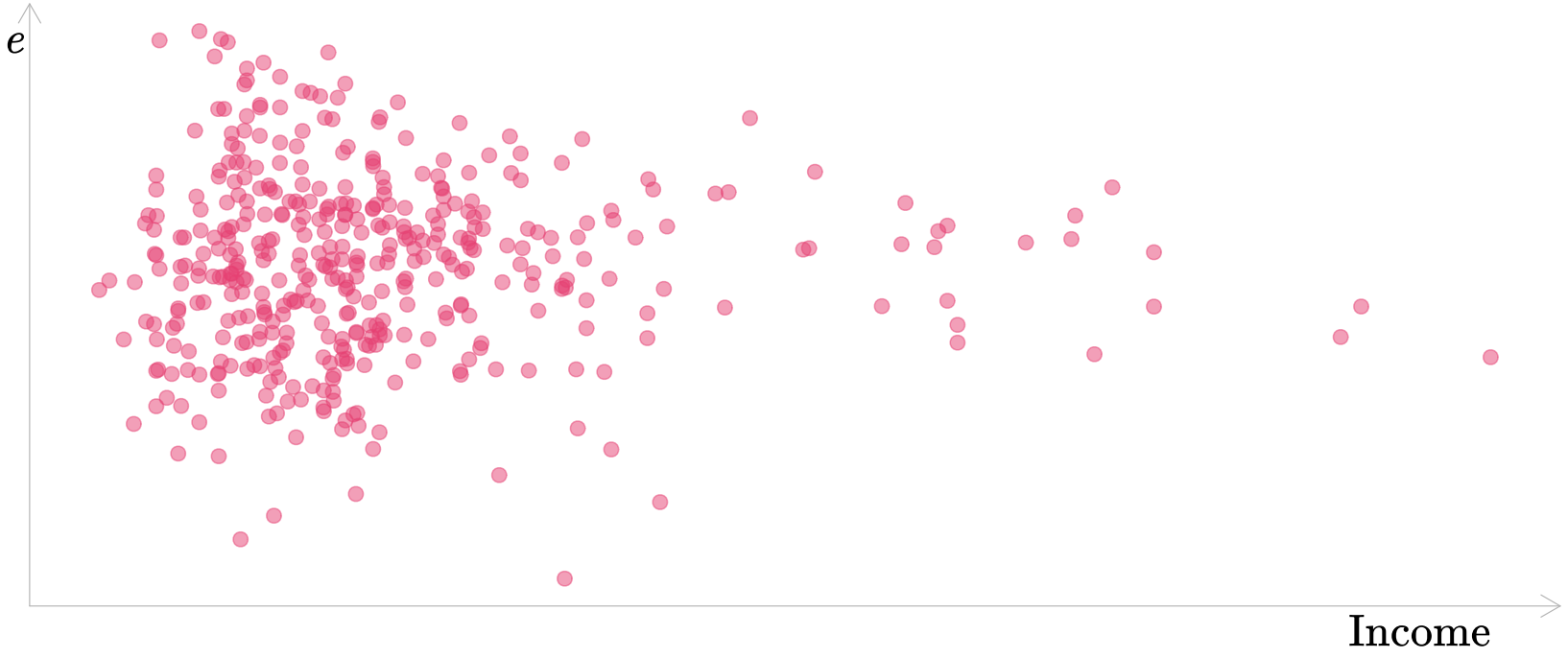


Living with heteroskedasticity

Example: Model specification

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

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

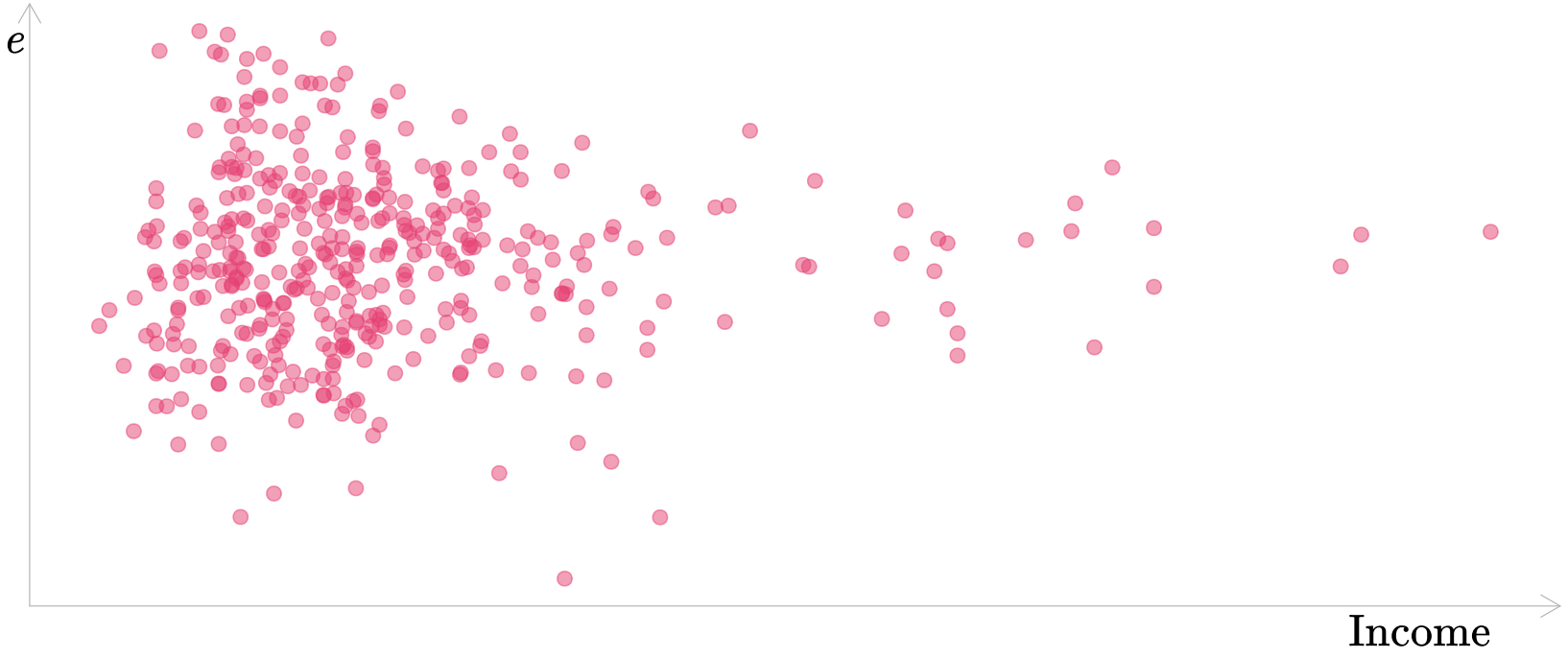


Living with heteroskedasticity

Example: Model specification

Model$_4$: Include interaction, first-order terms, and second-order terms

```
lm(test_score ~ I(ratio^2) + I(income^2) + income*ratio, data = test_df)
```



Living with heteroskedasticity

Example: Model specification