Heteroskedasticity

EC 421, Set 4

Edward Rubin
17 January 2019

Prologue

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.

Schedule

Last Time

We wrapped up our review.

Today

Heteroskedasticity

This week

First assignment! Due in a week.

Heteroskedasticity

Heteroskedasticity

Let's write down our **current assumptions**

Heteroskedasticity

Let's write down our **current assumptions**

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

Heteroskedasticity

Let's write down 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

Let's write down 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

Let's write down 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 dependent variables are **exogenous**: $E[u|X] = 0$ ($\implies$ $E[u] = 0$).

Heteroskedasticity

Let's write down 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 dependent variables are **exogenous**: $E[u|X] = 0$ ($\implies$ $E[u] = 0$).
5. The disurbances have **constant variance** σ^2 and **zero covariance**, *i.e.*,
   - $E\left[u_i^2|X\right] = \text{Var}(u_i|X) = \sigma^2 \implies \text{Var}(u_i) = \sigma^2$
   - $\text{Cov}\left(u_i,\, u_j|X\right) = E\left[u_i u_j|X\right] = 0$ for $i \neq j$

Heteroskedasticity

Let's write down 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 dependent variables are **exogenous**: $E[u|X] = 0$ ($\implies$ $E[u] = 0$).
5. The disurbances have **constant variance** σ^2 and **zero covariance**, *i.e.*,
   - $E\left[u_i^2|X\right] = \text{Var}(u_i|X) = \sigma^2 \implies \text{Var}(u_i) = \sigma^2$
   - $\text{Cov}\left(u_i,\, u_j|X\right) = E\left[u_i u_j|X\right] = 0$ 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

Today we're focusing on assumption #5:

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

Heteroskedasticity

Today we're focusing on assumption #5:

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

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

Heteroskedasticity

Today we're focusing on assumption #5:

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

Today we're focusing on assumption #5:

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

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

Violation of this assumption:

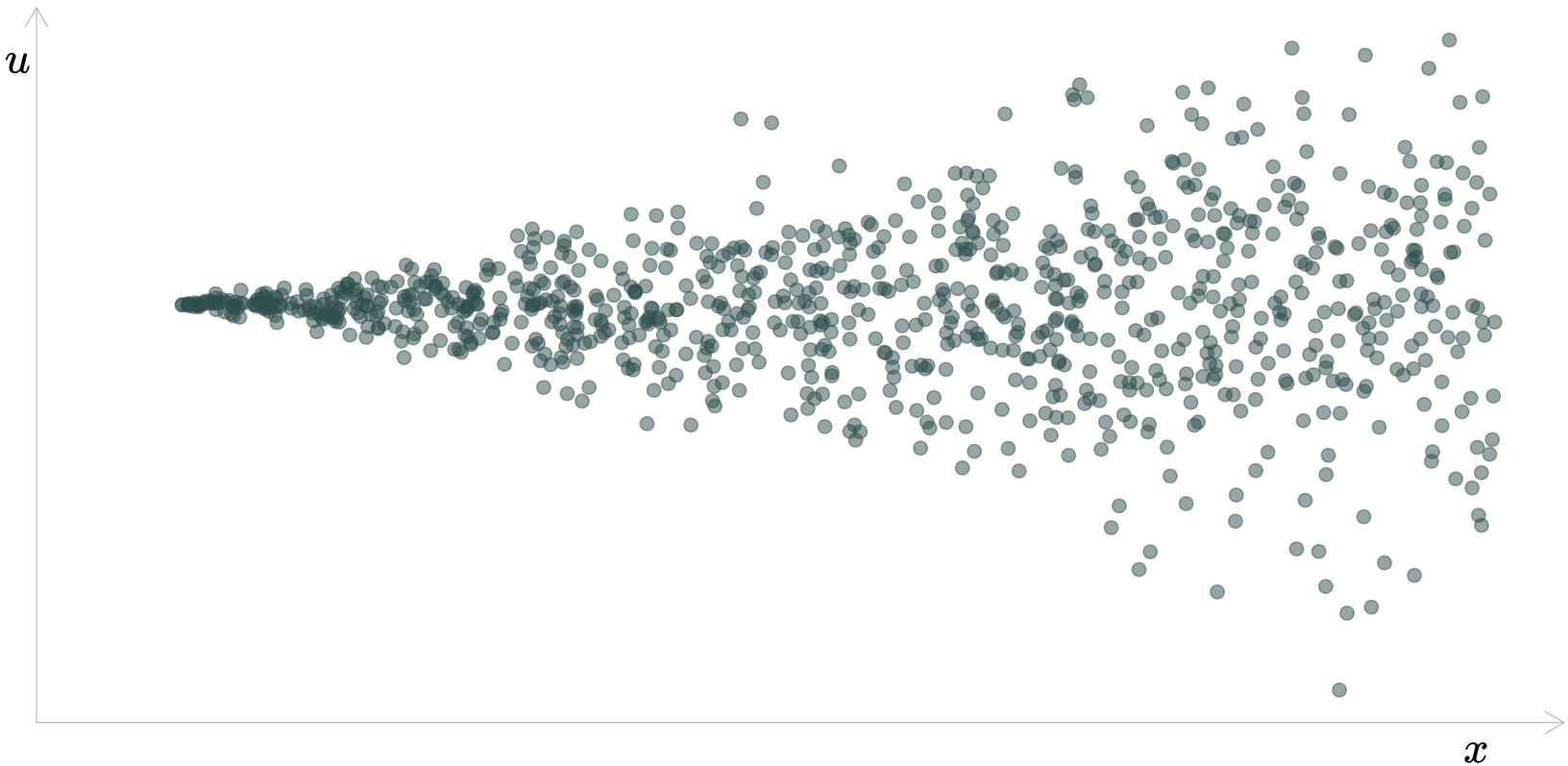
Heteroskedasticity: $\text{Var}(u_i) = \sigma_i^2$ and $\sigma_i^2 \neq \sigma_j^2$ for some $i \neq j$.

In other words: Our disturbances have different variances.

Heteroskedasticity

Classic example of heteroskedasticity: The funnel

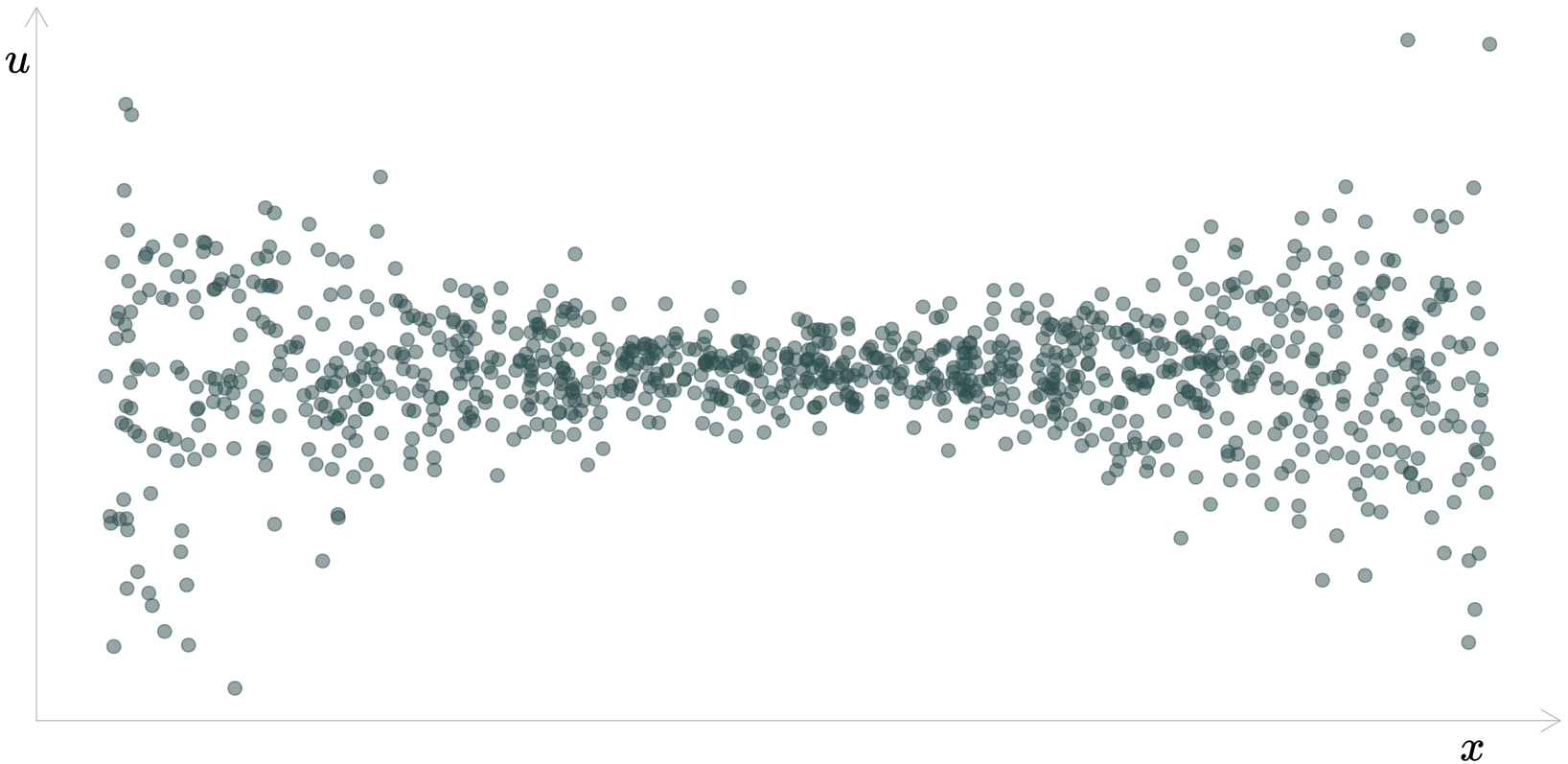
Variance of u increases with x



Heteroskedasticity

Another example of heteroskedasticity: (double funnel?)

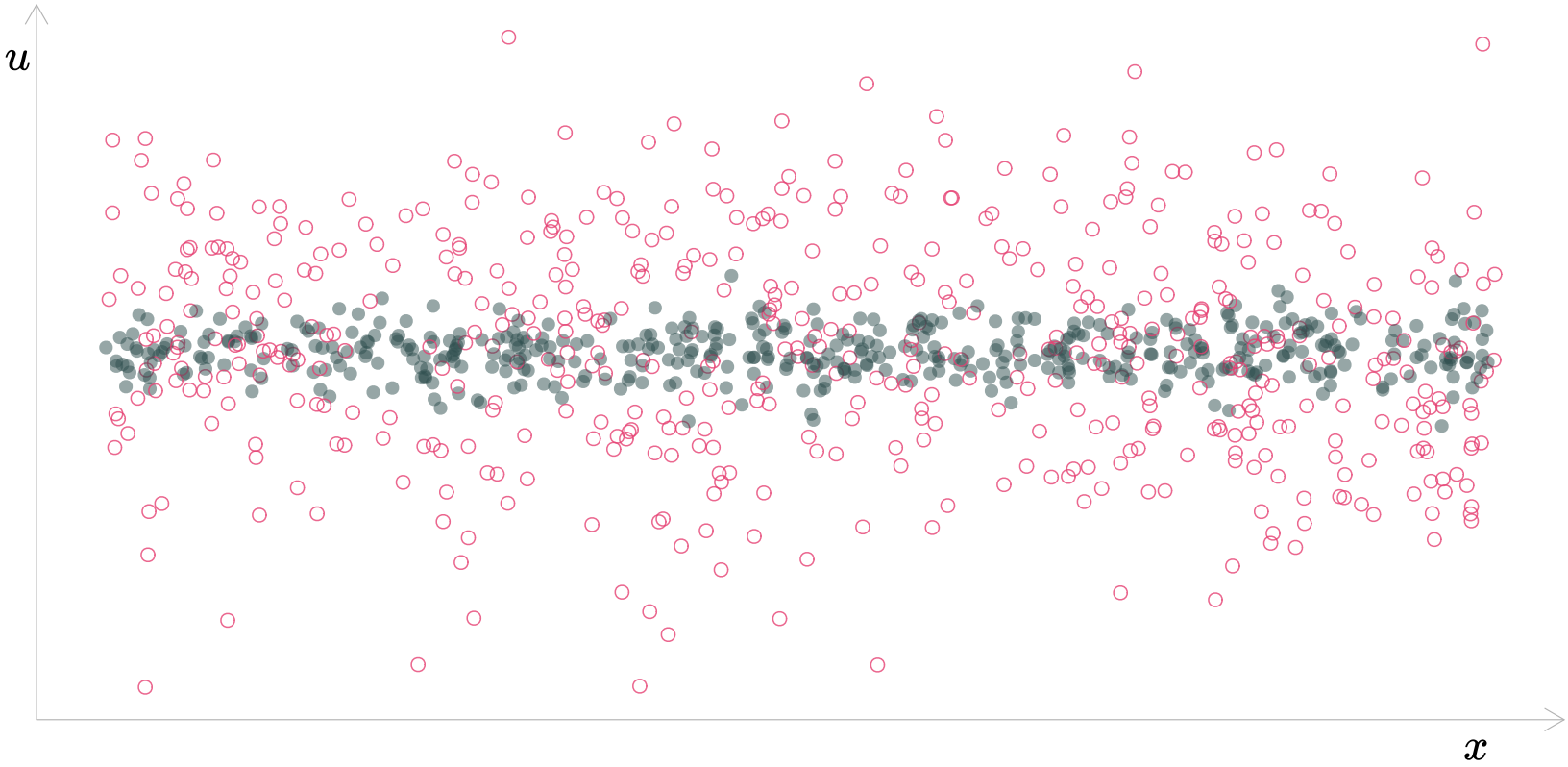
Variance of u increasing at the extremes of x



Heteroskedasticity

Another example of heteroskedasticity:

Differing variances of u by group



Heteroskedasticity

Heteroskedasticity is present when the variance of u changes with any combination of our explanatory variables x_1, through x_k (henceforth: X).

Heteroskedasticity

Heteroskedasticity is present when the variance of u changes with any combination of our explanatory variables x_1, through x_k (henceforth: X).

(Very common in practice)

Heteroskedasticity

Consequences

So what are the consquences of heteroskedasticity? Bias? Inefficiency?

First, let's check if it has consquences for the the unbiasedness of OLS.

Consequences

So what are the consquences of heteroskedasticity? Bias? Inefficiency?

First, let's check if it has consquences for the the unbiasedness of OLS.

Recall$_1$: OLS being unbiased means $E\left[\hat{\beta}_k \middle| X\right] = \beta_k$ for all k.

Consequences

So what are the consquences of heteroskedasticity? Bias? Inefficiency?

First, let's check if it has consquences for the the unbiasedness of OLS.

Recall$_1$: OLS being unbiased means $E\left[\hat{\beta}_k \middle| X\right] = \beta_k$ for all k.

Recall$_2$: We previously showed $\hat{\beta}_1 = \dfrac{\sum_i \left(y_i - \overline{y}\right)\left(x_i - \overline{x}\right)}{\sum_i \left(x_i - \overline{x}\right)^2}$

Heteroskedasticity

Consequences

So what are the consquences of heteroskedasticity? Bias? Inefficiency?

First, let's check if it has consquences for the the unbiasedness of OLS.

Recall$_1$: OLS being unbiased means $E\left[\hat{\beta}_k \middle| X\right] = \beta_k$ for all k.

Recall$_2$: We previously showed $\hat{\beta}_1 = \dfrac{\sum_i \left(y_i - \overline{y}\right)\left(x_i - \overline{x}\right)}{\sum_i \left(x_i - \overline{x}\right)^2}$

It will actually help us to rewrite this estimator as

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

Heteroskedasticity

Proof: Assuming $y_i = \beta_0 + \beta_1 x_i + u_i$

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_i \left(y_i - \overline{y}\right)\left(x_i - \overline{x}\right)}{\sum_i \left(x_i - \overline{x}\right)^2} \\
&= \frac{\sum_i \left(\left[\beta_0 + \beta_1 x_i + u_i\right] - \left[\beta_0 + \beta_1 \overline{x} + \overline{u}\right]\right)\left(x_i - \overline{x}\right)}{\sum_i \left(x_i - \overline{x}\right)^2} \\
&= \frac{\sum_i \left(\beta_1\left[x_i - \overline{x}\right] + \left[u_i - \overline{u}\right]\right)\left(x_i - \overline{x}\right)}{\sum_i \left(x_i - \overline{x}\right)^2} \\
&= \frac{\sum_i \left(\beta_1\left[x_i - \overline{x}\right]^2 + \left[x_i - \overline{x}\right]\left[u_i - \overline{u}\right]\right)}{\sum_i \left(x_i - \overline{x}\right)^2} \\
&= \beta_1 + \frac{\sum_i \left(x_i - \overline{x}\right)\left(u_i - \overline{u}\right)}{\sum_i \left(x_i - \overline{x}\right)^2}
\end{aligned}
$$

Heteroskedasticity

$$
\begin{aligned}
\hat{\beta}_1 = \cdots = \beta_1 + \frac{\sum_i \left(x_i - \overline{x}\right)\left(u_i - \overline{u}\right)}{\sum_i \left(x_i - \overline{x}\right)^2} \\
= \beta_1 + \frac{\sum_i \left(x_i - \overline{x}\right) u_i - \overline{u} \sum_i \left(x_i - \overline{x}\right)}{\sum_i \left(x_i - \overline{x}\right)^2} \\
= \beta_1 + \frac{\sum_i \left(x_i - \overline{x}\right) u_i - \overline{u} \left(\sum_i x_i - \sum_i \overline{x}\right)}{\sum_i \left(x_i - \overline{x}\right)^2} \\
= \beta_1 + \frac{\sum_i \left(x_i - \overline{x}\right) u_i - \overline{u} \left(\sum_i x_i - n\overline{x}\right)}{\sum_i \left(x_i - \overline{x}\right)^2} \\
= \beta_1 + \frac{\sum_i \left(x_i - \overline{x}\right) u_i - \overline{u} \left(\sum_i x_i - \sum_i x_i\right)}{\sum_i \left(x_i - \overline{x}\right)^2} \\
= \beta_1 + \frac{\sum_i \left(x_i - \overline{x}\right) u_i}{\sum_i \left(x_i - \overline{x}\right)^2}
\end{aligned}
$$

Consequences: Bias

We now want to see if heteroskedasticity biases the OLS estimator for β_1.

Consequences: Bias

We now want to see if heteroskedasticity biases the OLS estimator for β_1.

$$
\begin{aligned}
E\left[\hat{\beta}_1 \middle| X\right] &= E\left[\beta_1 + \frac{\sum_i \left(x_i - \overline{x}\right) u_i}{\sum_i \left(x_i - \overline{x}\right)^2} \middle| X\right] \\
&= \beta_1 + E\left[\frac{\sum_i \left(x_i - \overline{x}\right) u_i}{\sum_i \left(x_i - \overline{x}\right)^2} \middle| X\right] \\
&= \beta_1 + \frac{\sum_i \left(x_i - \overline{x}\right)}{\sum_i \left(x_i - \overline{x}\right)^2} \underbrace{E[u_i|X]}_{=0} \\
&= \beta_1 \quad 😹
\end{aligned}
$$

Consequences: Bias

We now want to see if heteroskedasticity biases the OLS estimator for β_1.

$$
\begin{aligned}
E\left[\hat{\beta}_1 \middle| X\right] &= E\left[\beta_1 + \frac{\sum_i \left(x_i - \overline{x}\right) u_i}{\sum_i \left(x_i - \overline{x}\right)^2} \middle| X\right] \\
&= \beta_1 + E\left[\frac{\sum_i \left(x_i - \overline{x}\right) u_i}{\sum_i \left(x_i - \overline{x}\right)^2} \middle| X\right] \\
&= \beta_1 + \frac{\sum_i \left(x_i - \overline{x}\right)}{\sum_i \left(x_i - \overline{x}\right)^2} \underbrace{E[u_i|X]}_{=0} \\
&= \beta_1 \quad 😹
\end{aligned}
$$

Phew. **OLS is still unbiased** for the β_k.

Consequences: Efficiency

OLS's efficiency and inference do not survive heteroskedasticity.

- In the presence of heteroskedasticity, OLS is **no longer the most efficient** (best) linear unbiased estimator.

Heteroskedasticity

Consequences: Efficiency

OLS's efficiency and inference do not survive heteroskedasticity.

- In the presence of heteroskedasticity, OLS is **no longer the most efficient** (best) linear unbiased estimator.
- It would be more informative (efficient) to **weight observations** inversely to their u_i's variance.
  - Downweight high-variance u_i's (too noisy to learn much).
  - Upweight observations with low-variance u_i's (more 'trustworthy').
  - Now you have the idea of weighted least squares (WLS)

Consequences: Inference

OLS **standard errors are biased** in the presence of heteroskedasticity.

- Wrong confidence intervals
- Problems for hypothesis testing (both t and F tests)

Heteroskedasticity

Consequences: Inference

OLS **standard errors are biased** in the presence of heteroskedasticity.

- Wrong confidence intervals
- Problems for hypothesis testing (both t and F tests)
- It's hard to learn much without sound inference.

Heteroskedasticity

Solutions

1. **Tests** to determine whether heteroskedasticity is present.

2. **Remedies** for (1) efficiency and (2) inference

Testing for heteroskedasticity

Testing for heteroskedasticity

While we *might* have solutions for heteroskedasticity, the efficiency of our estimators depends upon whether or not heteroskedasticity is present.

1. The **Goldfeld-Quandt test**
2. The **Breush-Pagan test**
3. The **White test**

Testing for heteroskedasticity

While we *might* have solutions for heteroskedasticity, the efficiency of our estimators depends upon whether or not heteroskedasticity is present.

1. The **Goldfeld-Quandt test**
2. The **Breush-Pagan test**
3. The **White test**

Each of these tests centers on the fact that we can **use the OLS residual e_i to estimate the population disturbance u_i**.

Testing for heteroskedasticity

The Goldfeld-Quandt test

Focuses on a specific type of heteroskedasticity: whether the variance of u_i differs **between two groups**.[†]

Remember how we used our residuals to estimate the σ^2?

$$s^2 = \frac{\text{SSE}}{n-1} = \frac{\sum_i e_i^2}{n-1}$$

We will use this same idea to determine whether there is evidence that our two groups differ in the variances of their disturbances, effectively comparing s_1^2 and s_2^2 from our two groups.

[†]: The G-Q test was one of the early tests of heteroskedasticity (1965).

Testing for heteroskedasticity

The Goldfeld-Quandt test

Operationally,

1. Order your the observations by x
2. Split the data into two groups of size n*
    - G_1: The first third
    - G_2: The last third
3. Run separate regressions of y on x for G_1 and G_2
4. Record SSE_1 and SSE_2
5. Calculate the G-Q test statistic

Testing for heteroskedasticity

The Goldfeld-Quandt test

The G-Q test statistic

$$F_{(n^\star - k,\, n^\star - k)} = \frac{\text{SSE}_2/(n^\star - k)}{\text{SSE}_1/(n^\star - k)} = \frac{\text{SSE}_2}{\text{SSE}_1}$$

follows an F distribution (under the null hypothesis) with $n^\star - k$ and $n^\star - k$ degrees of freedom.[†]

Testing for heteroskedasticity

The Goldfeld-Quandt test

The G-Q test statistic

$$F_{(n^\star - k,\, n^\star - k)} = \frac{\text{SSE}_2 / (n^\star - k)}{\text{SSE}_1 / (n^\star - k)} = \frac{\text{SSE}_2}{\text{SSE}_1}$$

follows an F distribution (under the null hypothesis) with $n^\star - k$ and $n^\star - k$ degrees of freedom.[†]

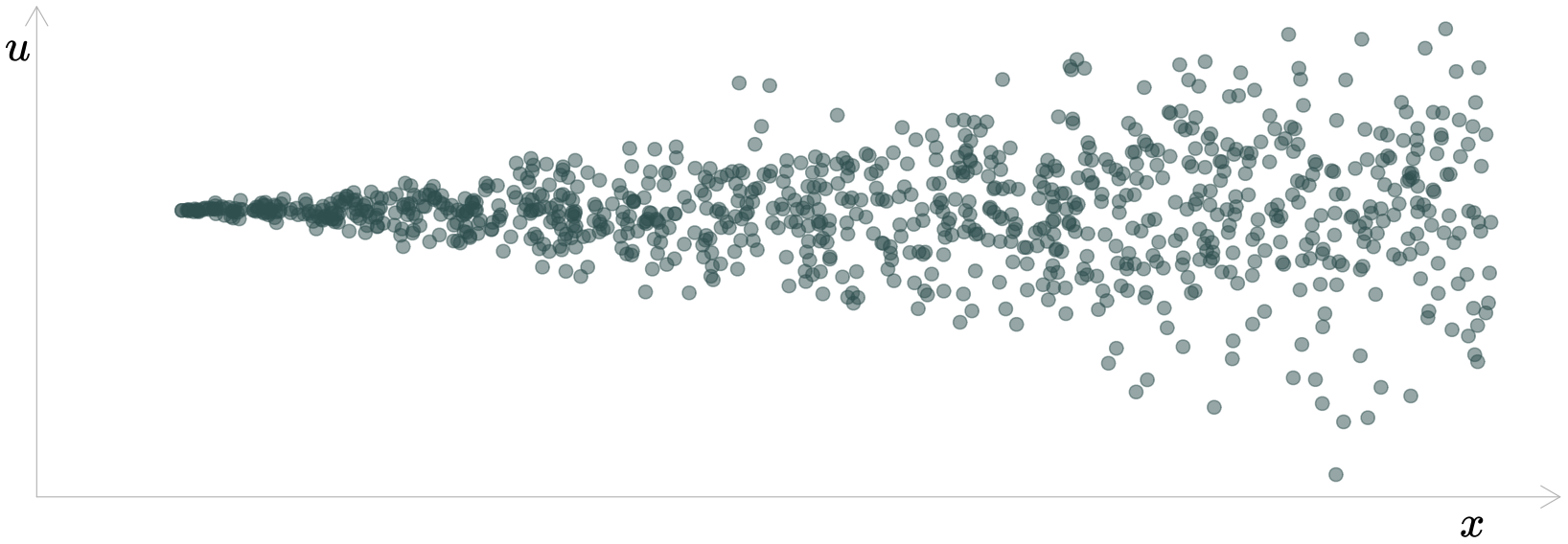
Notes

- The G-Q test requires the disturbances follow normal distributions.
- The G-Q assumes a very specific type/form of heteroskedasticity.
- Performs very well if we know the form of potentially heteroskedasticity.

[†]: Goldfelt and Quadt suggested $n^\star$ of $(3/8)n$.

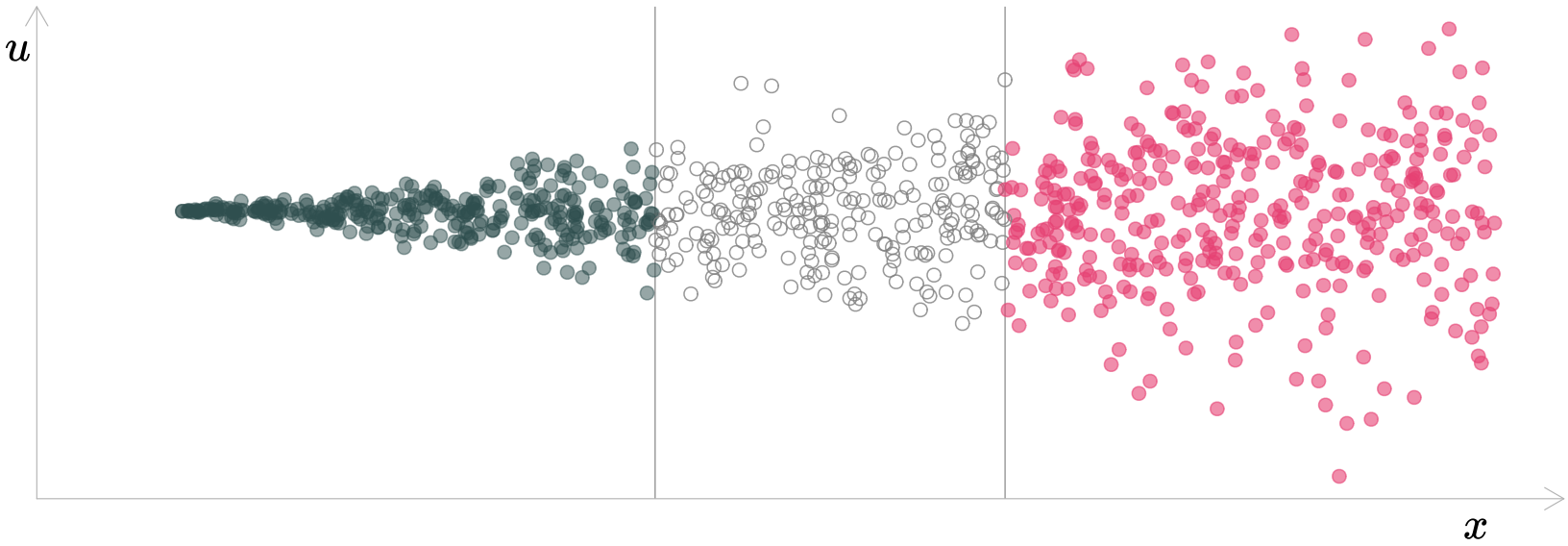
Testing for heteroskedasticity

The Goldfeld-Quandt test



Testing for heteroskedasticity

The Goldfeld-Quandt test



$$F_{375,\,375} = \frac{\text{SSE}_2 = 18,203.4}{\text{SSE}_1 = 1,039.5} \approx 17.5 \implies p\text{-value} < 0.001$$

$\therefore$ We reject H_0: $\sigma_1^2 = \sigma_2^2$ and conclude there is statistically significant evidence of heteroskedasticity.

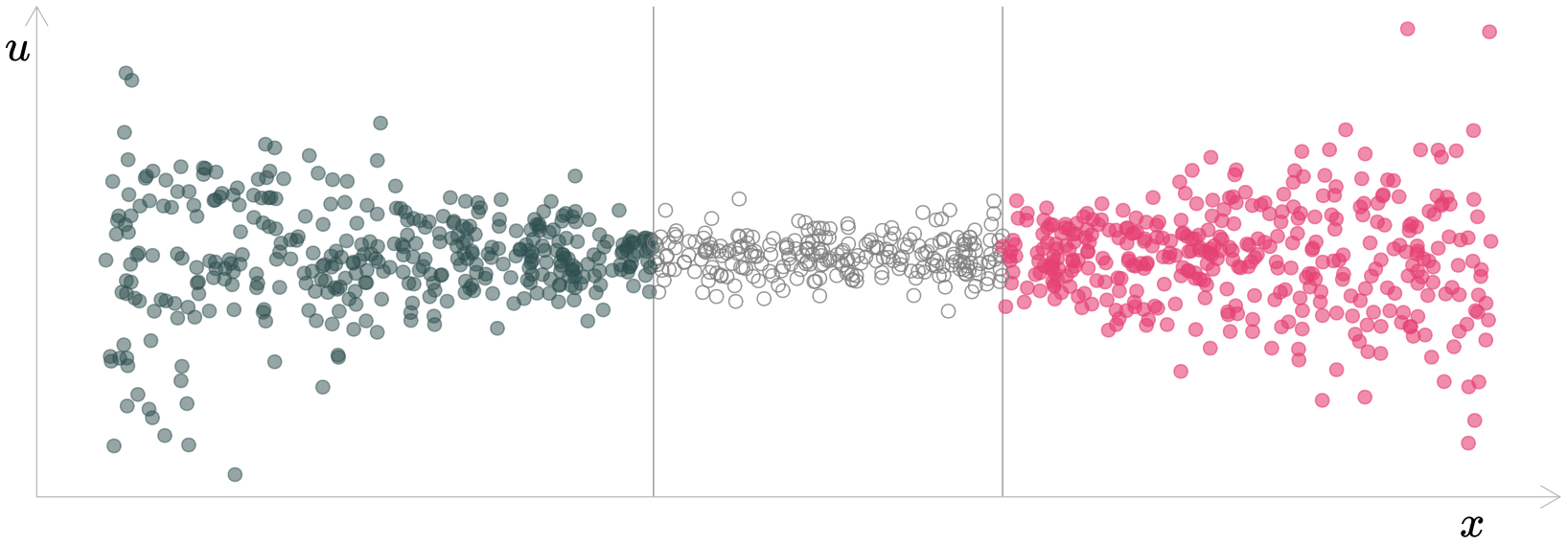
Testing for heteroskedasticity

The Goldfeld-Quandt test

The problem...

Testing for heteroskedasticity

The Goldfeld-Quandt test



$$F_{375,\,375} = \frac{\text{SSE}_2 = 14,516.8}{\text{SSE}_1 = 14,937.1} \approx 1 \implies p\text{-value} \approx 0.609$$

$\therefore$ We fail to reject H_0: $\sigma_1^2 = \sigma_2^2$ while heteroskedasticity is present.

Testing for heteroskedasticity

The Breush-Pagan test

Breusch and Pagan (1981) attempted to solve this issue of being too specific with the functional form of the heteroskedasticity.

- Allows the data to show if/how the variance of u_i correlates with X.
- If σ_i^2 correlates with X, then we have heteroskedasticity.
- Regresses e_i^2 on $X = [1, x_1, x_2, \ldots, x_k]$ and tests for joint significance.

Testing for heteroskedasticity

The Breush-Pagan test

How to implement:

1. Regress y on an intercept, x_1, x_2, …, x_k.
2. Record residuals e.
3. Regress e^2 on an intercept, x_1, x_2, …, x_k.

$$e_i^2 = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_k x_{ki} + v_i$$

4. Record R^2.
5. Test hypothesis H_0: $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$

Testing for heteroskedasticity

The Breush-Pagan test

The B-P test statistic[†] is

$$\mathrm{LM} = n \times R_e^2$$

where R_e^2 is the R^2 from the regression

$$e_i^2 = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_k x_{ki} + v_i$$

Under the null, LM is asymptotically distributed as χ_k^2.

Testing for heteroskedasticity

The Breush-Pagan test

The B-P test statistic[†] is

$$\mathrm{LM} = n \times R_e^2$$

where R_e^2 is the R^2 from the regression

$$e_i^2 = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_k x_{ki} + v_i$$

Under the null, LM is asymptotically distributed as χ_k^2.

This test statistic tests H_0: $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$.

Rejecting the null hypothesis implies evidence of heteroskedasticity.

[†]: This specific form of the test statistic actually comes form Koenker (1981).

Testing for heteroskedasticity

The Breush-Pagan test

Problem: We're still assuming a fairly restrictive **functional form** between our explanatory variables X and the variances of our disturbances σ_i^2.

Testing for heteroskedasticity

The Breush-Pagan test

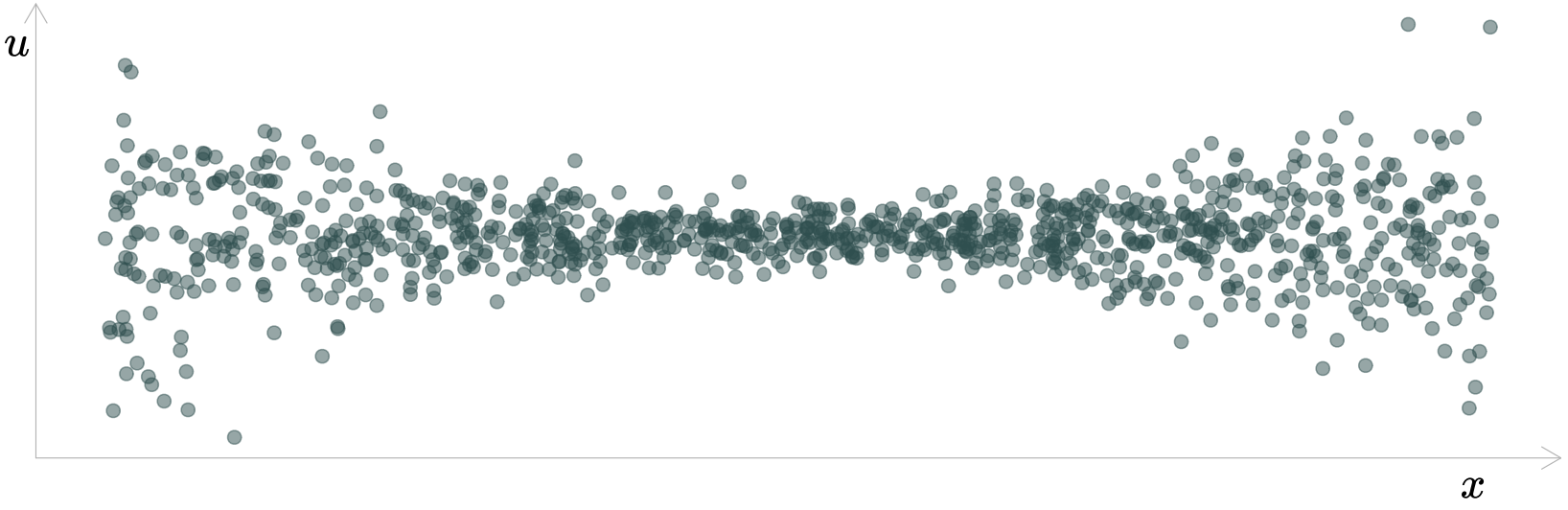
Problem: We're still assuming a fairly restrictive **functional form** between our explanatory variables X and the variances of our disturbances σ_i^2.

Result: B-P *may* still miss fairly simple forms of heteroskedasticity.

Testing for heteroskedasticity

The Breush-Pagan test

Breusch-Pagan tests are still **sensitive to functional form**.



$e_i^2 = \hat{\alpha}_0 + \hat{\alpha}_1 x_{1i}$	$\widehat{\text{LM}} = 1.26$	$\text{p-value} \approx 0.261$
$e_i^2 = \hat{\alpha}_0 + \hat{\alpha}_1 x_{1i} + \hat{\alpha}_2 x_{1i}^2$	$\widehat{\text{LM}} = 185.8$	$\text{p-value} < 0.001$

Testing for heteroskedasticity

The White test

So far we've been testing for specific relationships between our explanatory variables and the variances of the disturbances, *e.g.*,

- H_0: $\sigma_1^2 = \sigma_2^2$ for two groups based upon x_j (**G-Q**)
- H_0: $\alpha_1 = \cdots = \alpha_k = 0$ from $e_i^2 = \alpha_0 + \alpha_1 x_{1i} + \cdots + \alpha_k x_{ki} + v_i$ (**B-P**)

Testing for heteroskedasticity

The White test

So far we've been testing for specific relationships between our explanatory variables and the variances of the disturbances, *e.g.*,

- H_0: $\sigma_1^2 = \sigma_2^2$ for two groups based upon x_j (**G-Q**)
- H_0: $\alpha_1 = \cdots = \alpha_k = 0$ from $e_i^2 = \alpha_0 + \alpha_1 x_{1i} + \cdots + \alpha_k x_{ki} + v_i$ (**B-P**)

However, we actually want to know if

$$\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2$$

Q: Can't we just test this hypothesis?

Testing for heteroskedasticity

The White test

So far we've been testing for specific relationships between our explanatory variables and the variances of the disturbances, *e.g.*,

- H_0: $\sigma_1^2 = \sigma_2^2$ for two groups based upon x_j (**G-Q**)
- H_0: $\alpha_1 = \cdots = \alpha_k = 0$ from $e_i^2 = \alpha_0 + \alpha_1 x_{1i} + \cdots + \alpha_k x_{ki} + v_i$ (**B-P**)

However, we actually want to know if

$$\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2$$

Q: Can't we just test this hypothesis? **A:** Sort of.

Testing for heteroskedasticity

The White test

Toward this goal, Hal White took advantage of the fact that we can **replace the homoskedasticity requirement with a weaker assumption**:

- **Old:** $\text{Var}(u_i|X) = \sigma^2$
- **New:** u^2 is *uncorrelated* with the explanatory variables (*i.e.*, x_j for all j), their squares (*i.e.*, x_j^2), and the first-degree interactions (*i.e.*, $x_j x_h$).

Testing for heteroskedasticity

The White test

Toward this goal, Hal White took advantage of the fact that we can **replace the homoskedasticity requirement with a weaker assumption**:

- **Old:** $\text{Var}(u_i|X) = \sigma^2$
- **New:** u^2 is *uncorrelated* with the explanatory variables (*i.e.*, x_j for all j), their squares (*i.e.*, x_j^2), and the first-degree interactions (*i.e.*, $x_j x_h$).

This new assumption is easier to explicitly test (*hint:* regression).

Testing for heteroskedasticity

The White test

An outline of White's test for heteroskedasticity:

1. Regress y on x_1, x_2, …, x_k. Save residuals e.

2. Regress squared residuals on all explanatory variables, their squares, and interactions.

$$e^2 = \alpha_0 + \sum_{h=1}^{k} \alpha_h x_h + \sum_{j=1}^{k} \alpha_j x_j^2 + \sum_{\ell=1}^{k-1} \sum_{m=\ell+1}^{k} x_\ell x_m + v_i$$

3. Record R_e^2.

4. Calculate test statistic to test H_0: $\alpha_p = 0$ for all $p \neq 0$.

Testing for heteroskedasticity

The White test

Just as with the Bruesch-Pagan test, White's test statistic is

$$\mathrm{LM} = n \times R_e^2 \qquad \text{Under } \mathrm{H}_0,\ \mathrm{LM} \overset{\mathrm{d}}{\sim} \chi_k^2$$

but now the R_e^2 comes from the regression of e^2 on the explanatory variables, their squares, and their interactions.

$$e^2 = \alpha_0 + \underbrace{\sum_{h=1}^{k} \alpha_h x_h}_{\text{Expl. variables}} + \underbrace{\sum_{j=1}^{k} \alpha_j x_j^2}_{\text{Squared terms}} + \underbrace{\sum_{\ell=1}^{k-1} \sum_{m=\ell+1}^{k} x_\ell x_m}_{\text{Interactions}} + v_i$$

Testing for heteroskedasticity

The White test

Just as with the Bruesch-Pagan test, White's test statistic is

$$\text{LM} = n \times R_e^2 \qquad \text{Under H}_0\text{, LM} \overset{\text{d}}{\sim} \chi_k^2$$

but now the R_e^2 comes from the regression of e^2 on the explanatory variables, their squares, and their interactions.

$$e^2 = \alpha_0 + \underbrace{\sum_{h=1}^{k} \alpha_h x_h}_{\text{Expl. variables}} + \underbrace{\sum_{j=1}^{k} \alpha_j x_j^2}_{\text{Squared terms}} + \underbrace{\sum_{\ell=1}^{k-1} \sum_{m=\ell+1}^{k} x_\ell x_m}_{\text{Interactions}} + v_i$$

Note: If a variable is equal to its square (*e.g.*, binary variables), then you don't (can't) include it. THe same rule applies for interactions.

Testing for heteroskedasticity

The White test

Example: Consider the model[†] $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$

Step 1: Estimate the model; obtain residuals (e).

Step 2: Regress e^2 on explanatory variables, squares, and interactions.

$$\begin{aligned} e^2 =& \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_4 x_1^2 + \alpha_5 x_2^2 + \alpha_6 x_3^2 \\ &+ \alpha_7 x_1 x_2 + \alpha_8 x_1 x_3 + \alpha_9 x_2 x_3 + v \end{aligned}$$

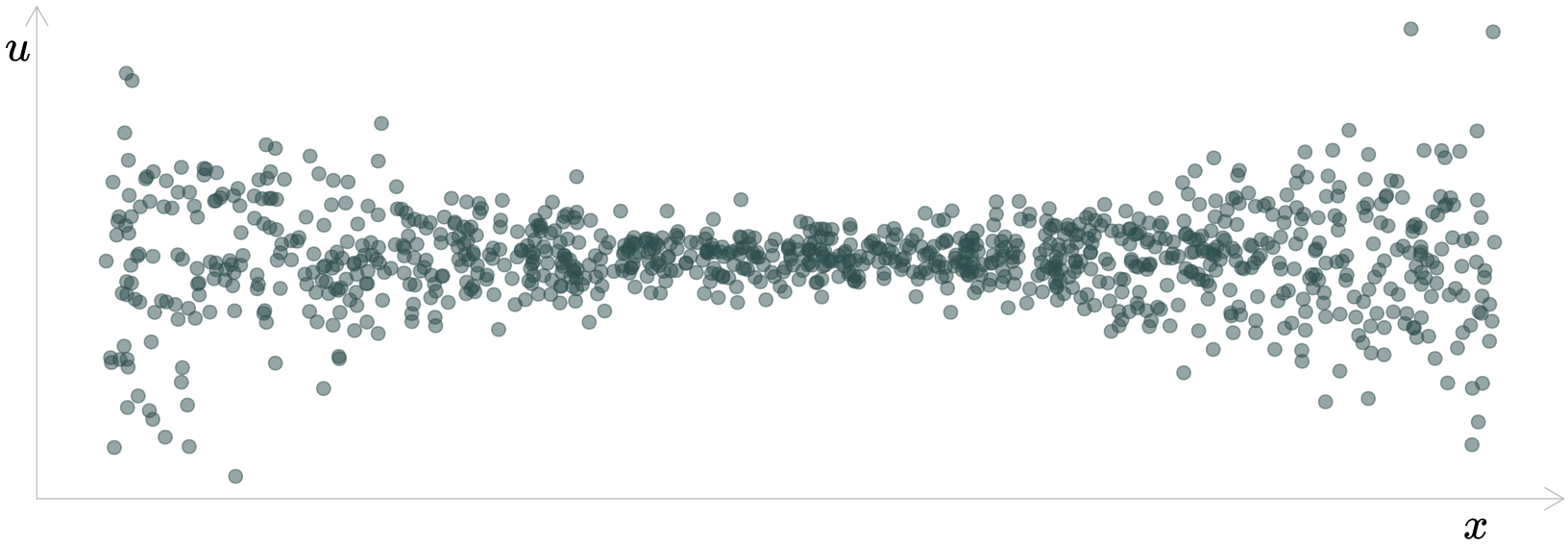
Record the R^2 from this equation (call it R_e^2).

Step 3: Test H_0: $\alpha_1 = \alpha_2 = \cdots = \alpha_9 = 0$ using $\text{LM} = nR_e^2 \overset{\text{d}}{\sim} \chi_3^2$.

[†]: To simplify notation here, I'm dropping the i subscripts.

Testing for heteroskedasticity

The White test



We've already done the White test for this simple linear regression.

$$e_i^2 = \hat{\alpha}_0 + \hat{\alpha}_1 x_{1i} + \hat{\alpha}_2 x_{1i}^2 \qquad \widehat{\text{LM}} = 185.8 \qquad \text{p-value} < 0.001$$

Testing for heteroskedasticity

Okay. We have some tests that may help us detect heteroskedasticity.

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**.
3. Live with OLS's inefficiency and make **corrections for inference**.
    - 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 function form doesn't cause heteroskedasticity.

[†]: *Specification:* Functional form and included variables.

Living with heteroskedasticity

Misspecification

Example: If the population relationship is

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$$

But you omit x^2 and estimate

$$y = \gamma_0 + \gamma_1 x + w$$

Then

$$w = u + \beta_2 x^2 \qquad \implies \qquad \mathbf{Var}(w) = f(x)$$

The variance of w changes systematically with x (heteroskedasticity).

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.

Living with heteroskedasticity

Weighted least squares

Weighted least squares (WLS) presents another approach.

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.

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?

$$\text{Var}\left(\frac{u_i}{\sigma_i}\right) = \frac{1}{\sigma_i^2}\text{Var}(u_i) = \frac{1}{\sigma_i^2}\sigma_i^2 = 1 \quad \forall i$$

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.

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

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

Notes:

1. **Inverse-variance weighting:** WLS downweights observations with higher variance in their errors.

2. **Big requirement:** WLS requires that we *know* σ_i^2 for each observation.