Heteroskedasticity

EC 421, Set 04

Edward Rubin

Winter 2022

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

Schedule

This week

First assignment!

Turn in **2 files**†

1. Your write up (*e.g.*, Word file).
2. The `R` script that generated your answers.

Important

- Your figures and regression results should be in the write up file.
- We should be able to easily find your answers for each question.
- **Do not copy.** (You will receive a zero.)

†: Unless you're using RMarkdown—then we need a PDF or HTML file.

Prerequisite material

The χ^2 distribution

Some test statistics are distributed as χ^2 random variables.

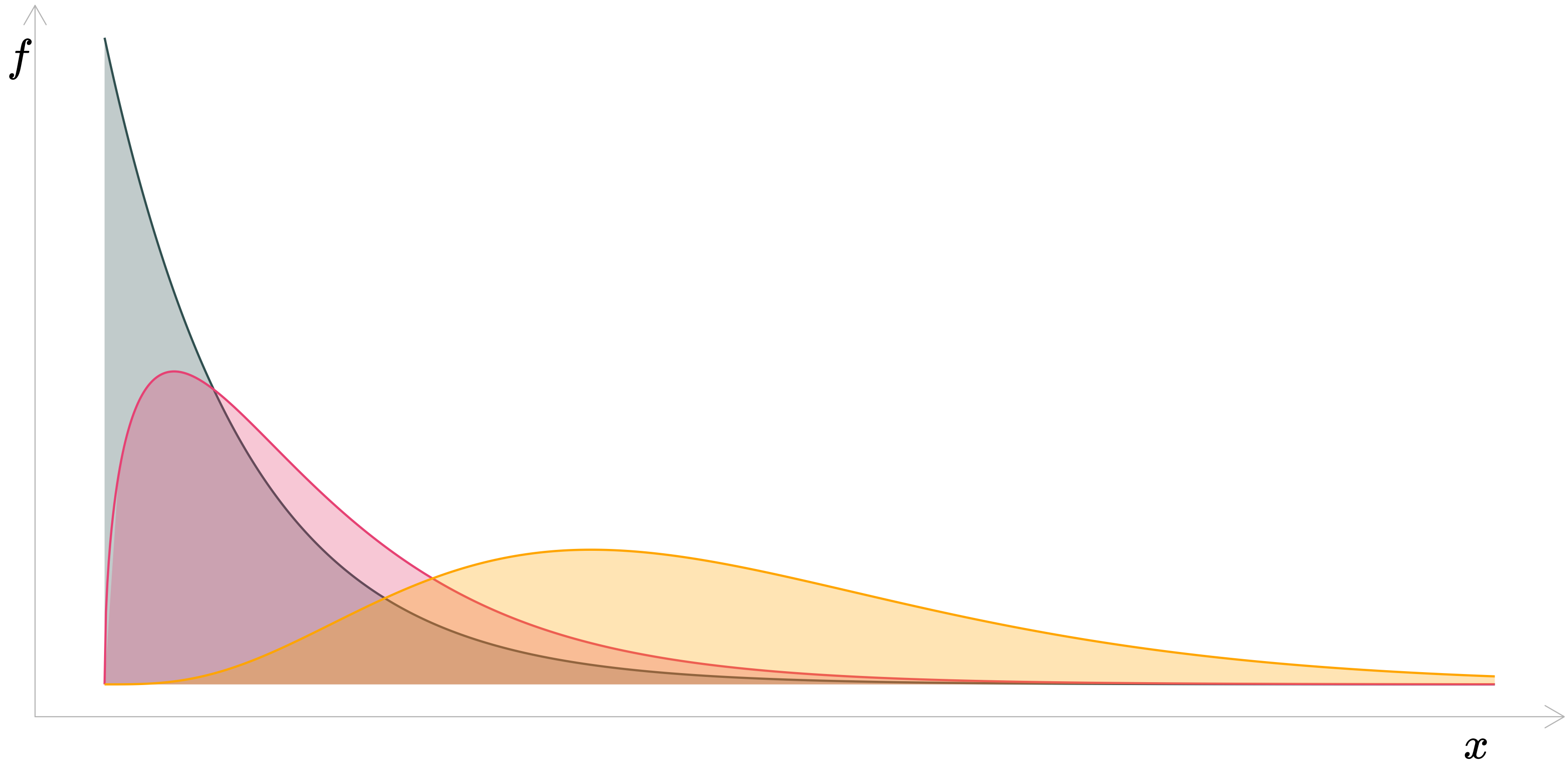
The χ^2 distribution is just another example of a common (named) distribution (like the Normal distribution, the t distribution, and the F).

The shape of the χ^2 distribution depends on a single parameter:

- We will call this parameter k
- Our test statistics will refer to k as *degrees of freedom*.

The χ^2 distribution

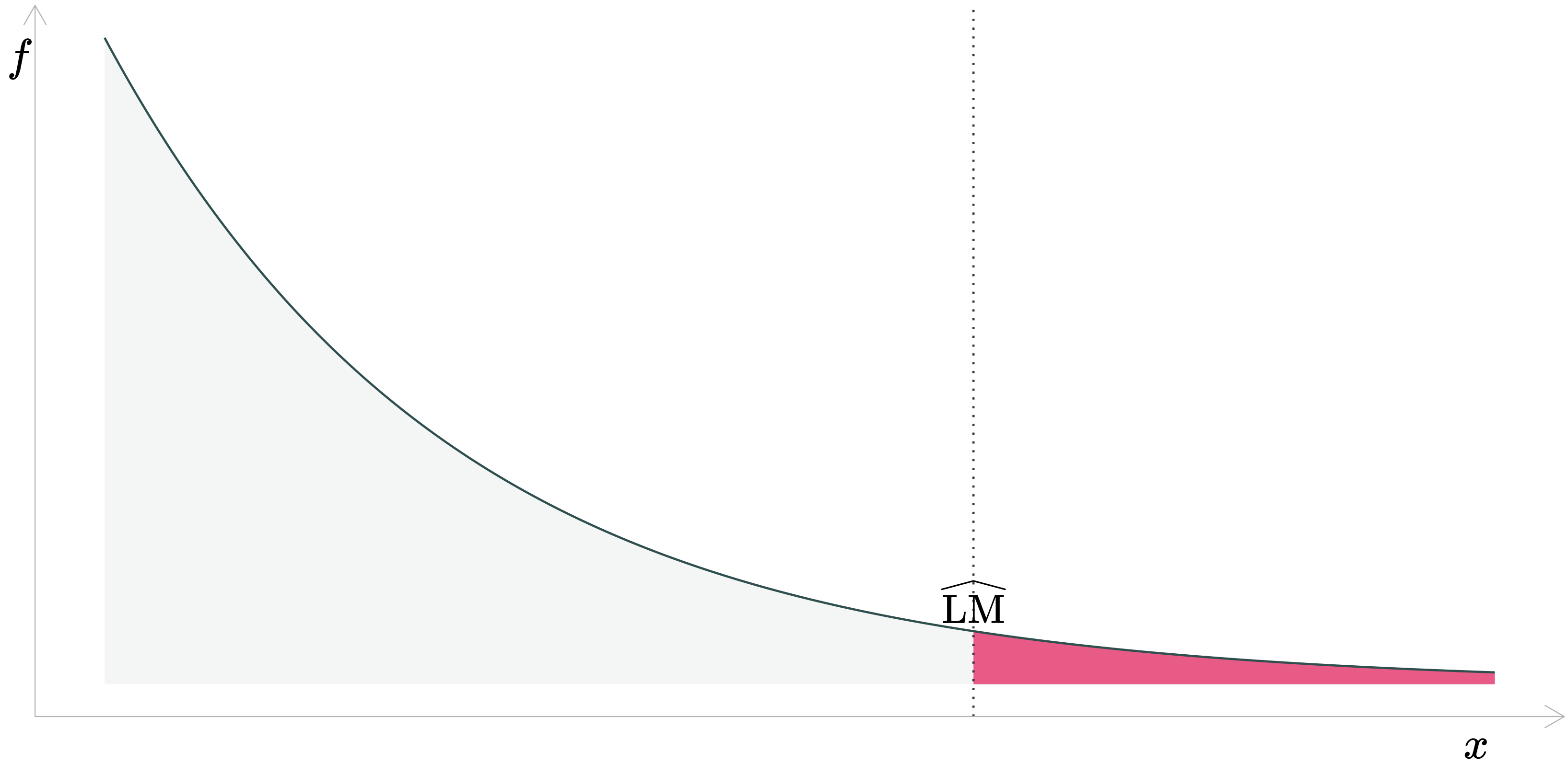
Three examples of χ_k^2: $k = 1$, $k = 2$, and $k = 9$



Prerequisite material

The χ^2 distribution

Probability of observing a more extreme test statistic $\widehat{\mathrm{LM}}$ under H_0



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

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 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.*,

    - $\boldsymbol{E}\left[u_i^2 \middle| X\right] = \mathrm{Var}(u_i|X) = \sigma^2 \implies \mathrm{Var}(u_i) = \sigma^2$
    - $\mathrm{Cov}\left(u_i,\, u_j \middle| X\right) = \boldsymbol{E}\left[u_i u_j \middle| X\right] = 0$ 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 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.*,
    - $\boldsymbol{E}\left[u_i^2|X\right] = \mathrm{Var}(u_i|X) = \sigma^2 \implies \mathrm{Var}(u_i) = \sigma^2$
    - $\mathrm{Cov}\left(u_i,\, u_j|X\right) = \boldsymbol{E}\left[u_i u_j|X\right] = 0$ for $i \neq j$
6. The disturbances come from a **Normal** distribution, *i.e.*, $u_i \overset{\text{iid}}{\sim} \mathrm{N}\left(0, \sigma^2\right)$.

Heteroskedasticity

Today we're focusing on assumption #5:

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

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

Violation of this assumption:

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

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

Violation of this assumption:

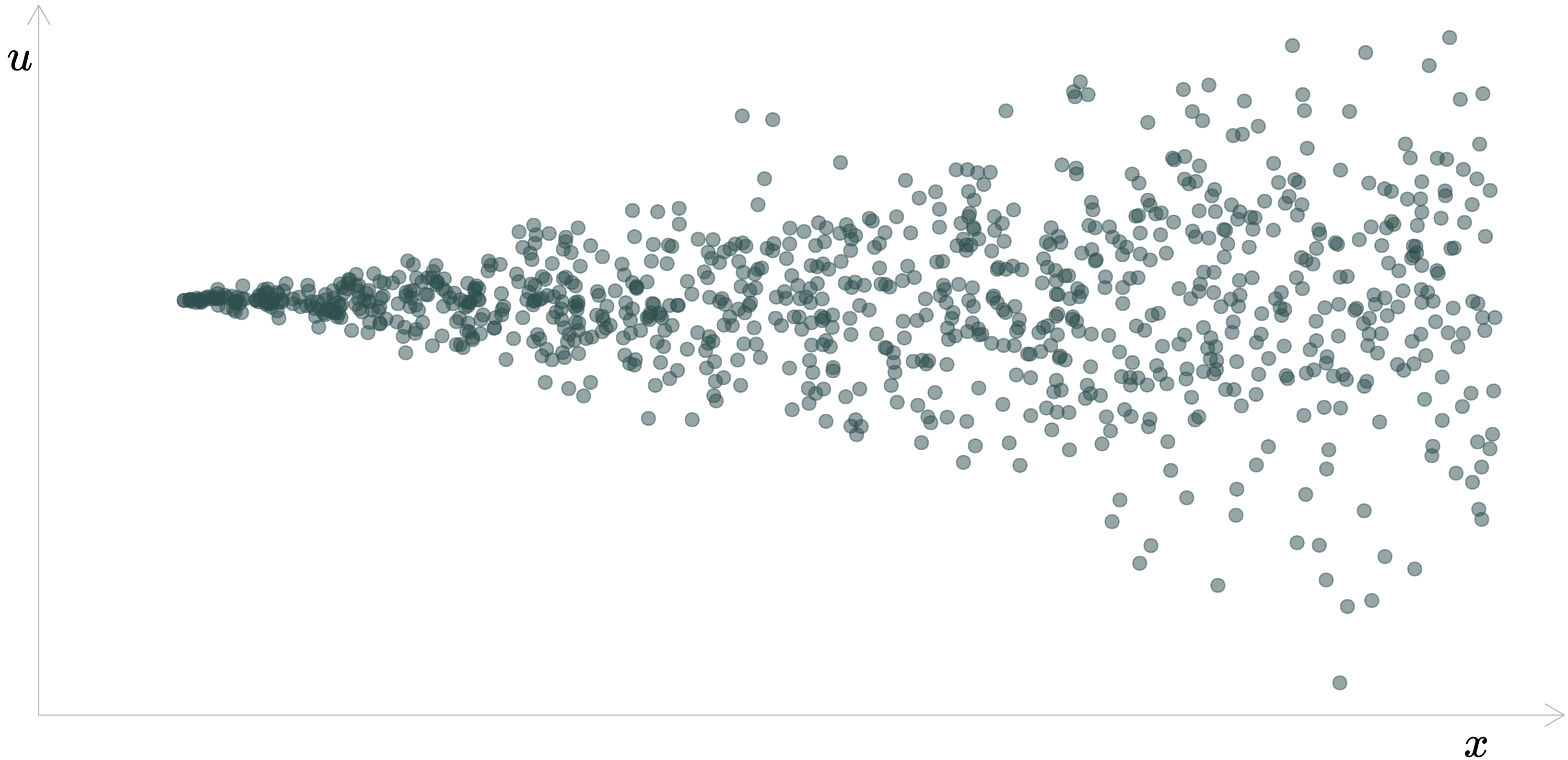
Heteroskedasticity: $\text{Var}(u_i) = \sigma_i^2$ and $\sigma_i^2 \neq \sigma_j^2$ for some $i \neq j$.

In other words: Our disturbances have different variances.

Heteroskedasticity

Classic example of heteroskedasticity: The funnel

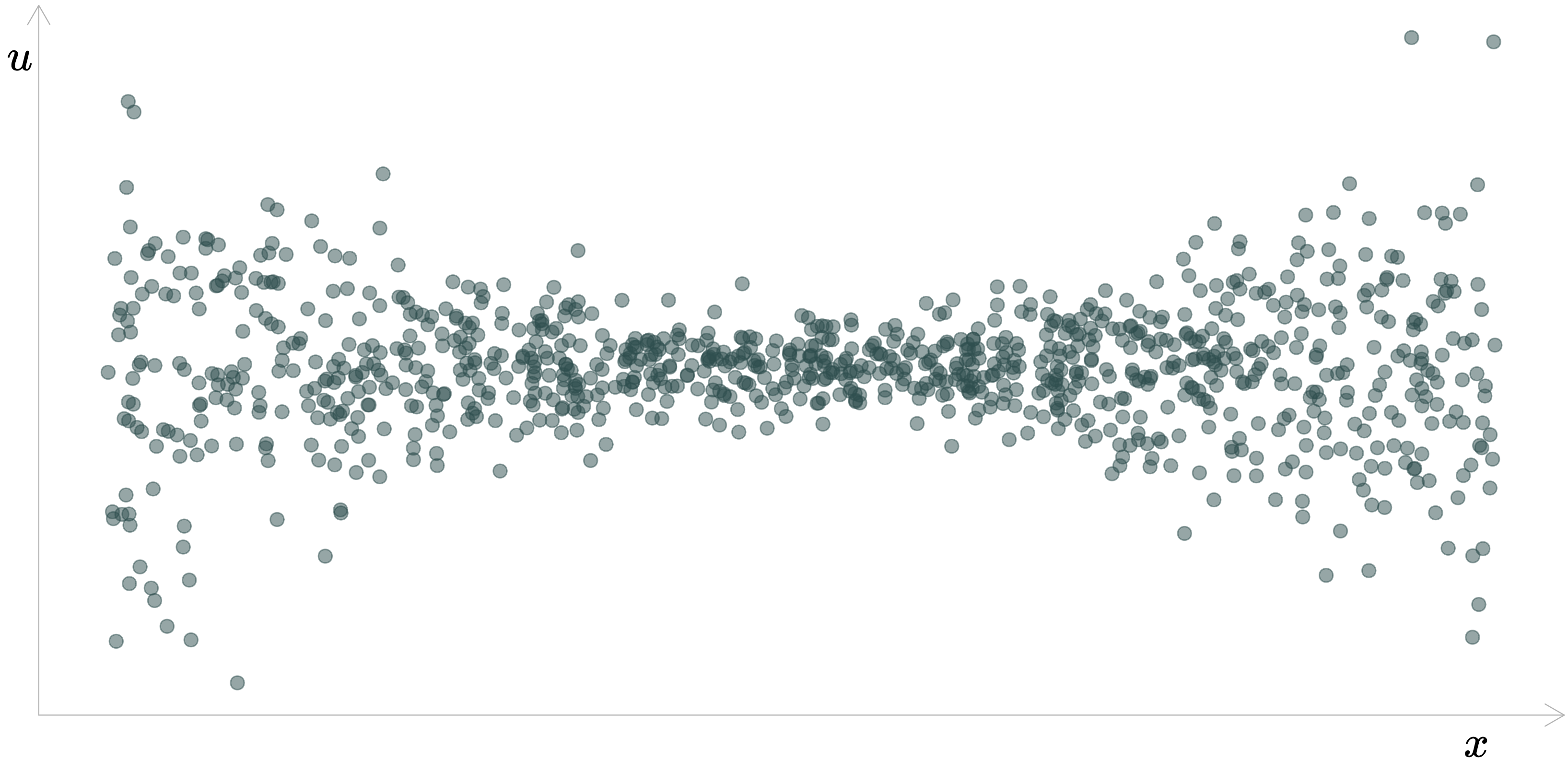
Variance of u increases with x



Heteroskedasticity

Another example of heteroskedasticity: (double funnel?)

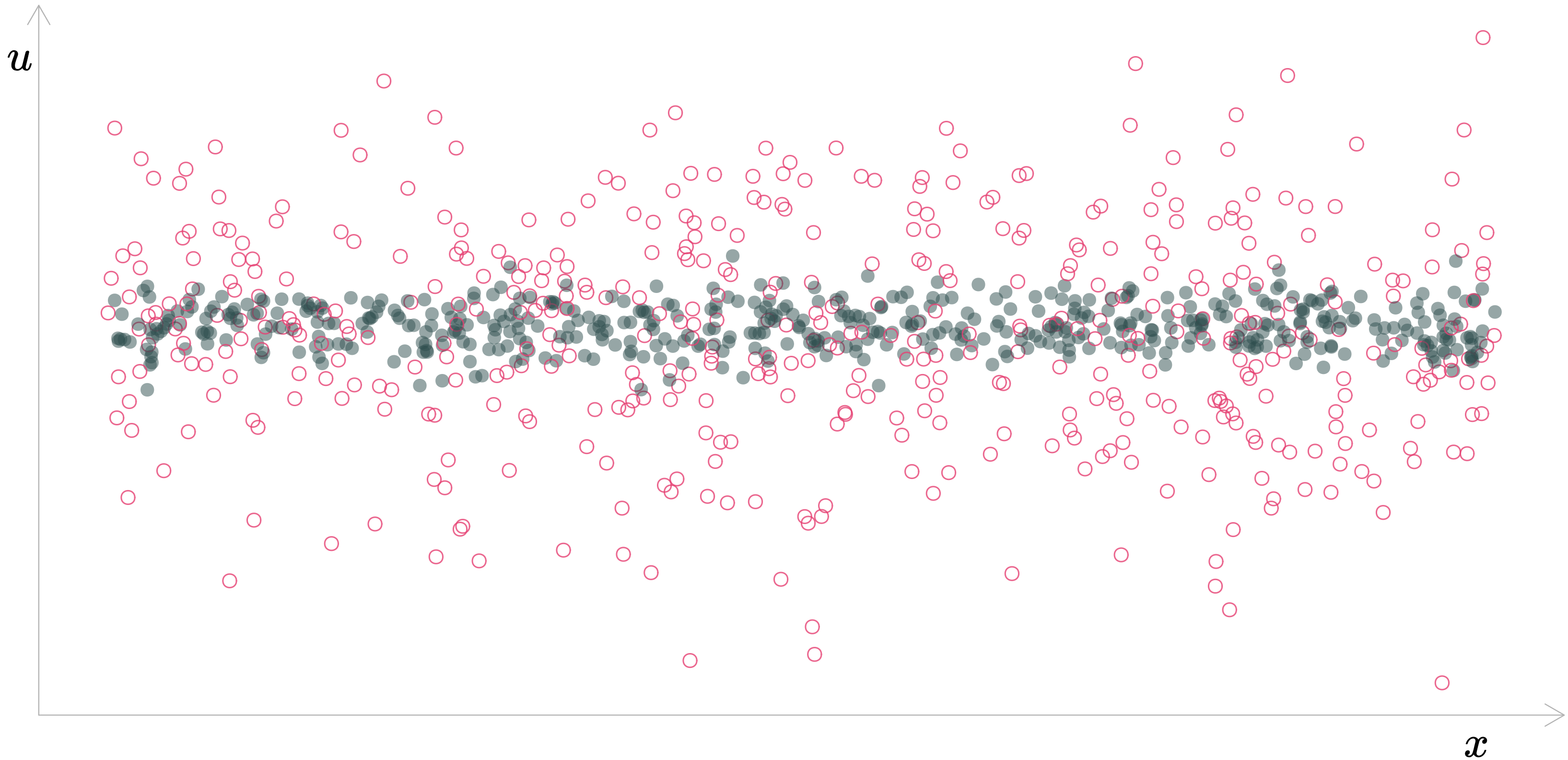
Variance of u increasing at the extremes of x



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).

It's very common in practice—and should probably be our default.

Heteroskedasticity

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

It's very common in practice—and should probably be our default.

Why we care: Heteroskedasticity shows us how small violations of our assumptions can affect OLS's performance.

Consequences

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

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

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 $\boldsymbol{E}\left[\hat{\beta}_k \middle| X\right] = \beta_k$ for all $\boldsymbol{k}$.

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 $\boldsymbol{E}\left[\hat{\beta}_k \middle| X\right] = \beta_k$ for all $\boldsymbol{k}$.

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 $\boldsymbol{E}\left[\hat{\beta}_k \middle| X\right] = \beta_k$ for all $\boldsymbol{k}$.

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.

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

Consequences: Bias

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

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

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.

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)

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 **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 **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**.

† There are many other options for testing, *e.g.*, the **Breusch-Pagan test**.

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).

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

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.[†]

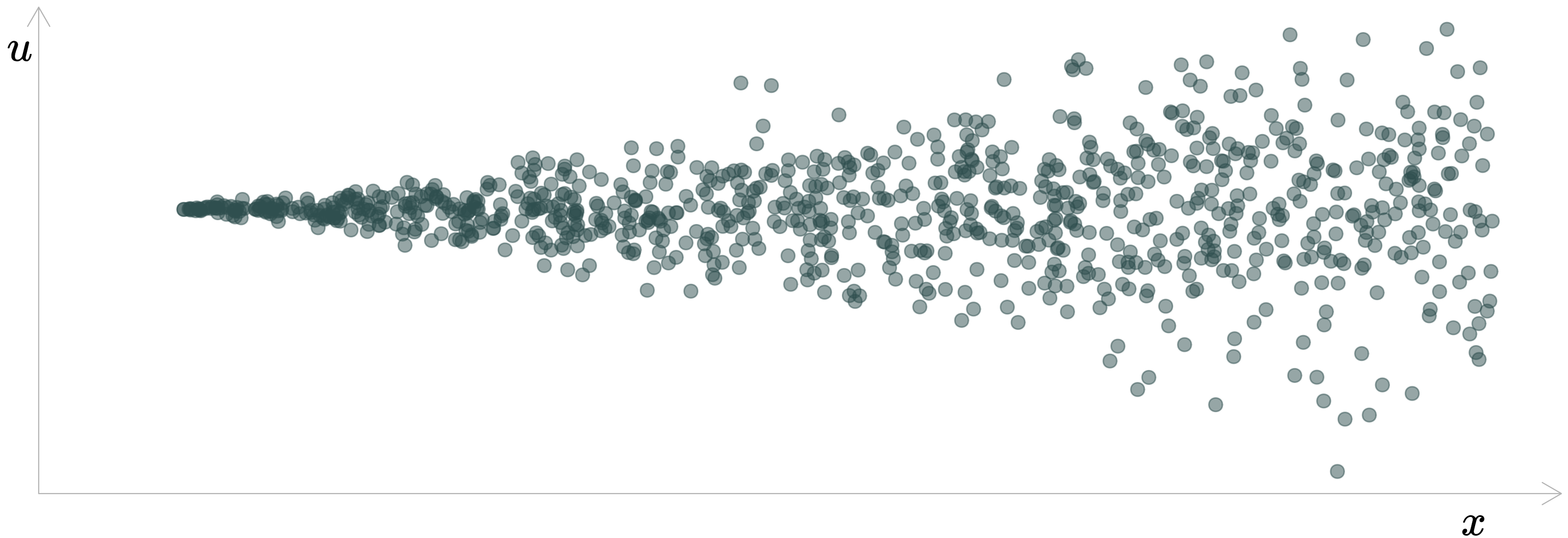
Notes

- 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.

[†]: Goldfeld and Quandt suggested $n^\star$ of $(3/8)n$. k gives number of estimated parameters (*i.e.*, $\hat{\beta}_j$'s).

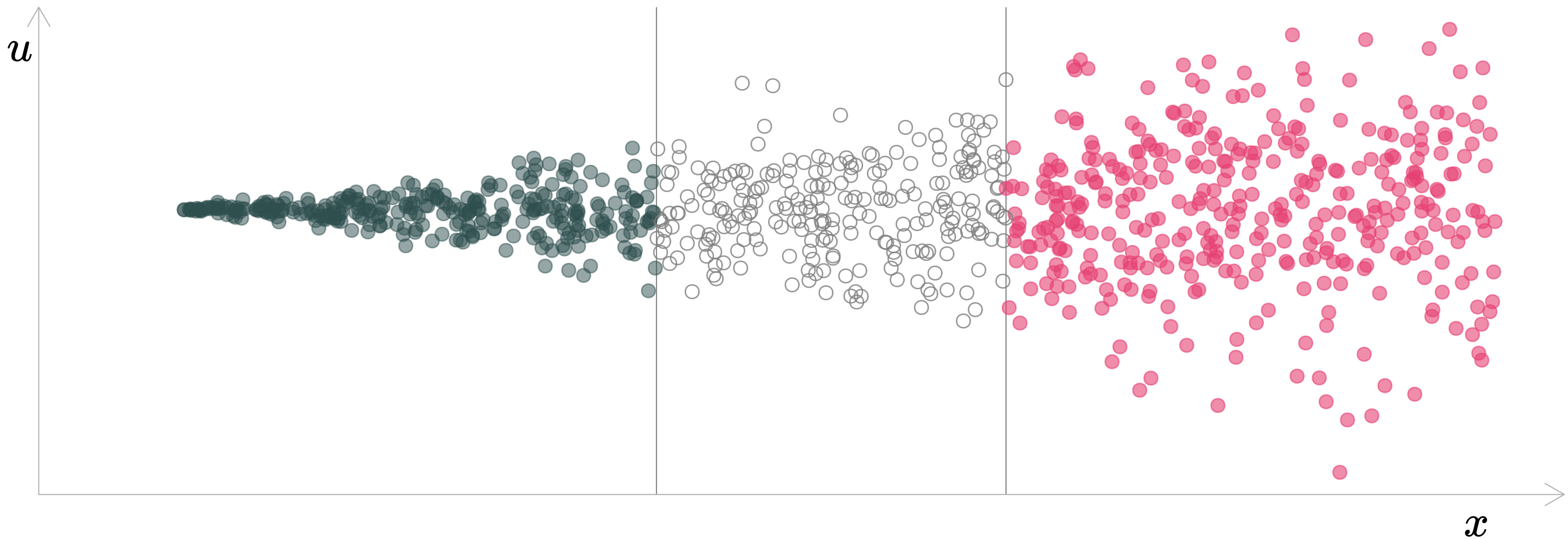
Testing for heteroskedasticity

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.

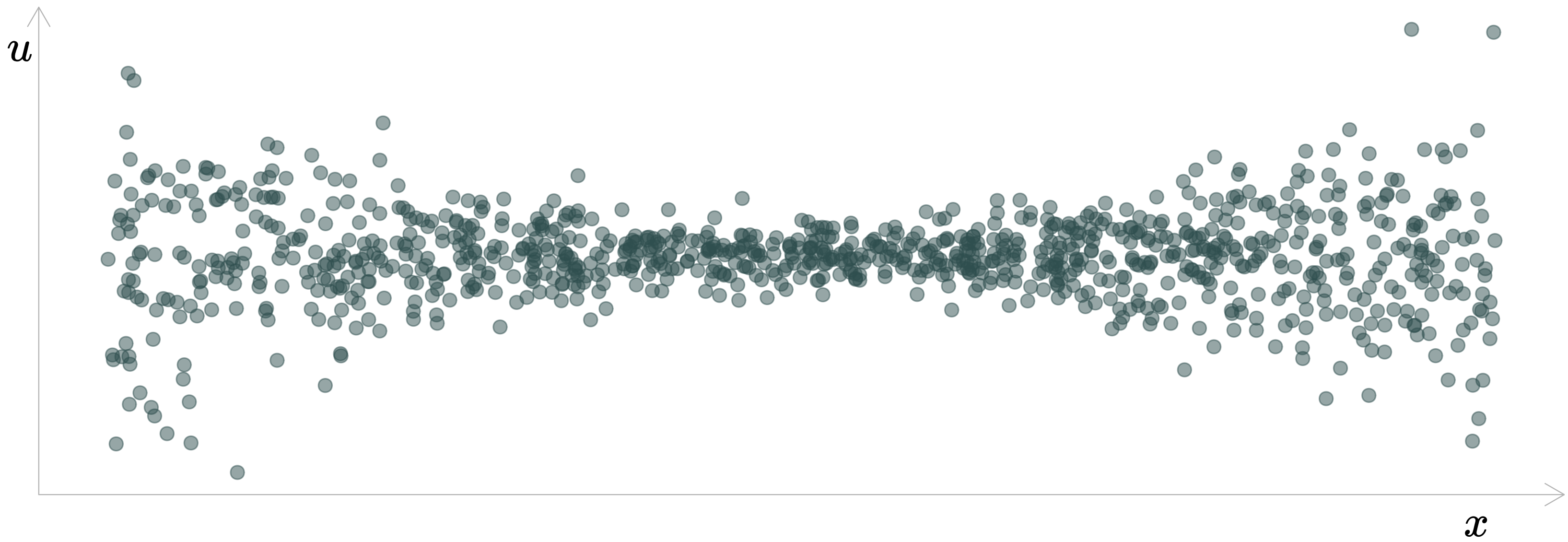
Testing for heteroskedasticity

The Goldfeld-Quandt test

The problem...

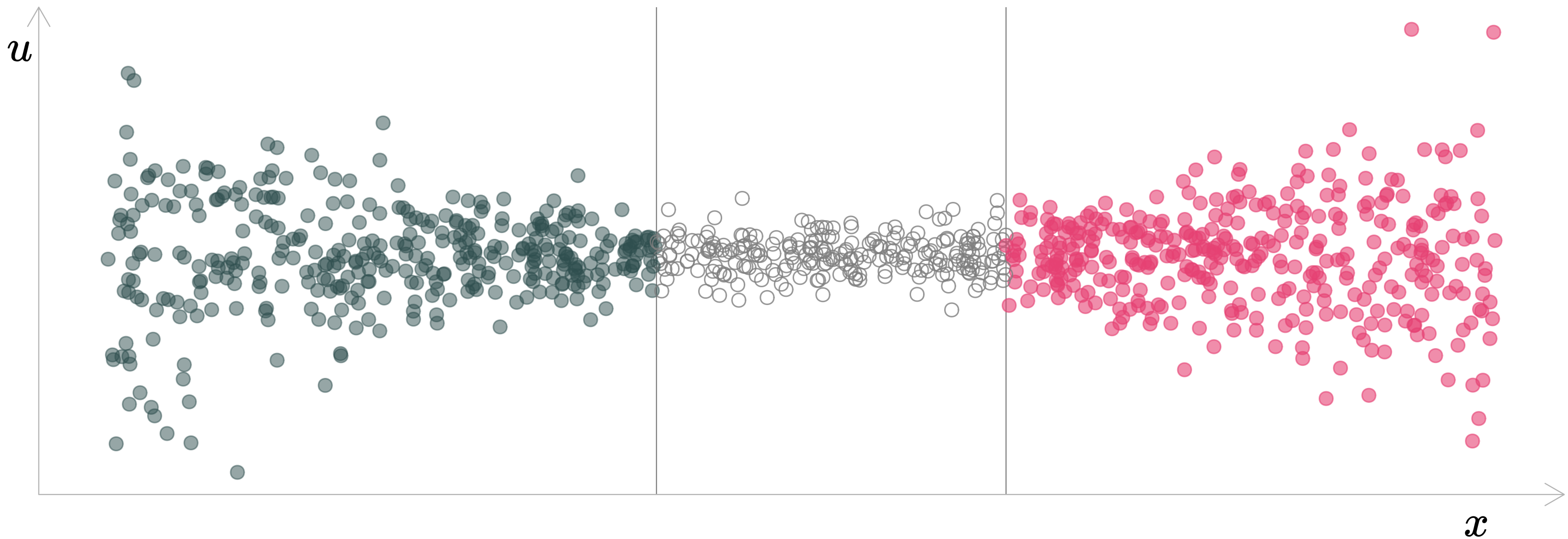
Testing for heteroskedasticity

The Goldfeld-Quandt test



Testing for heteroskedasticity

The Goldfeld-Quandt test



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

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

Testing for heteroskedasticity

The White test

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

- Regress e_i^2 on $X = [\mathbf{1},\, x_1,\, x_2,\, \ldots,\, x_k]$ and test for joint significance.
- 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.

The White test

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

- Regress e_i^2 on $X = [1,\ x_1,\ x_2,\ \ldots,\ x_k]$ and test for joint significance.
- 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.

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

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

- Regress e_i^2 on $X = [1,\, x_1,\, x_2,\, \ldots,\, x_k]$ and test for joint significance.
- 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.

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.

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:** $\mathrm{Var}(u_i|X) = \sigma^2$
- **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:** $\mathrm{Var}(u_i|X) = \sigma^2$

- **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).

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.

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

3. Record R_e^2.

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

The White test

White's test statistic is

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

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

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

Note: The k (for our χ_k^2) equals the number of estimated parameters in the regression above (the α_j), excluding the intercept (α_0).

Testing for heteroskedasticity

The White test

Practical 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.

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.

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

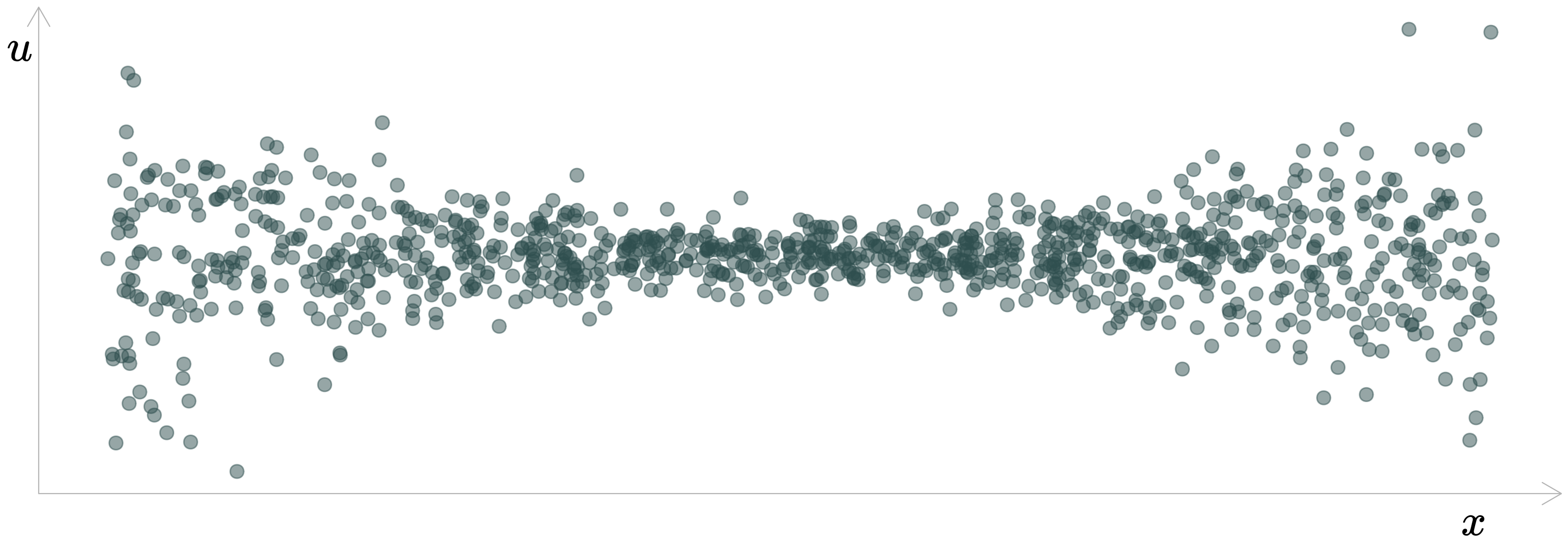
Record the R^2 from this equation (call it R_e^2).

Step 3: Test H_0: $\alpha_1 = \alpha_2 = \cdots = \alpha_9 = 0$ using $\text{LM} = nR_e^2 \overset{\text{d}}{\sim} \chi_9^2$.

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

Testing for heteroskedasticity

The White test



The White test for this simple linear regression.

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

Testing for Heteroskedasticity

Examples

Examples

Goal: Estimate the relationship between standardized test scores (outcome variable) and (1) student-teacher ratio and (2) income, *i.e.*,

$$\left(\textbf{Test score}\right)_i = \beta_0 + \beta_1 \textbf{Ratio}_i + \beta_2 \textbf{Income}_i + u_i \tag{1}$$

Potential issue: Heteroskedasticity... and we do not observe u_i.

Solution:

1. Estimate the relationship in (1) using OLS.
2. Test for heteroskedasticity.
   - Goldfeld-Quandt
   - White

Examples

We will use testing data from the dataset `Caschool` in the `Ecdat` R package.

```r
# 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 × 3
#>   test_score ratio income
#>        <dbl> <dbl>  <dbl>
#> 1       691.  17.9  22.7
#> 2       661.  21.5   9.82
```

Examples

Let's begin by estimating our model

$$(\textbf{Test score})_i = \beta_0 + \beta_1 \textbf{Ratio}_i + \beta_2 \textbf{Income}_i + u_i$$

```
# Estimate the model
est_model = lm(test_score ~ ratio + income, data = test_df)
# Summary of the estimate
tidy(est_model)
```

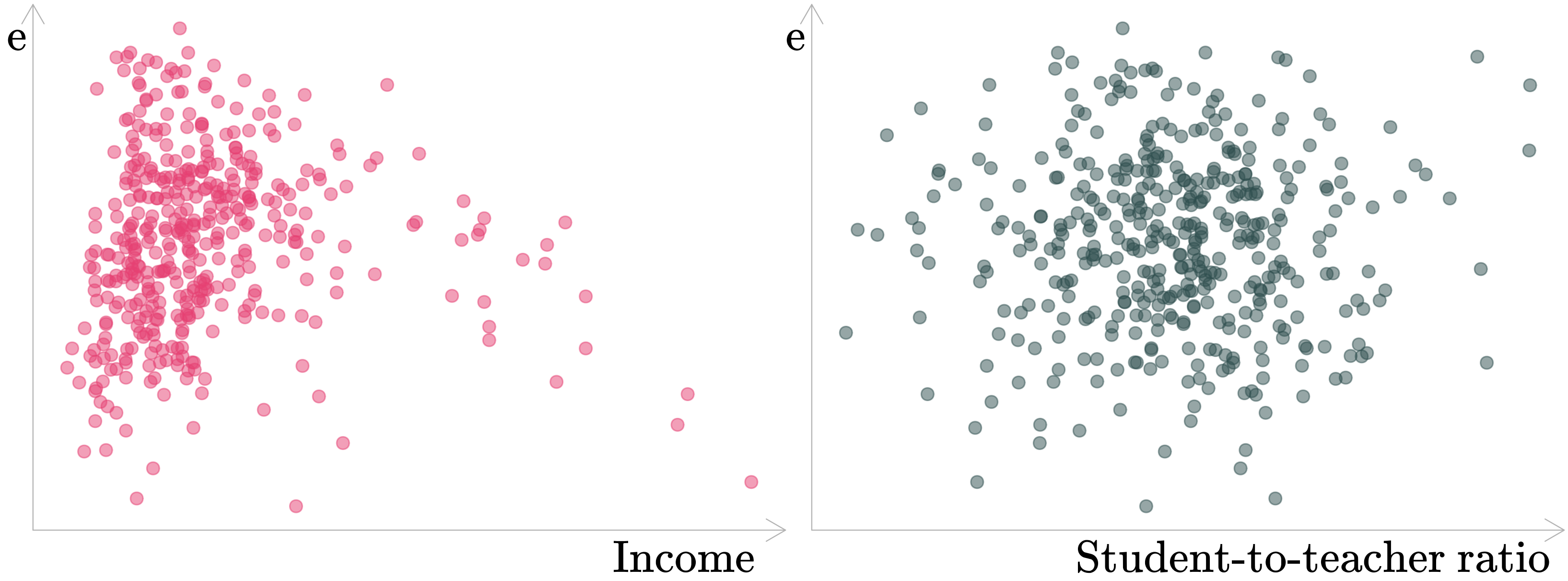
```
#> # A tibble: 3 × 5
#>   term        estimate std.error statistic   p.value
#>   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
#> 1 (Intercept)  639.       7.45        85.7  5.70e-267
#> 2 ratio         -0.649    0.354      -1.83 6.79e-  2
#> 3 income         1.84     0.0928      19.8  4.38e- 62
```

Testing for heteroskedasticity

Examples

Now, let's see what the residuals suggest about heteroskedasticity

```
# Add the residuals to our dataset
test_df$e = residuals(est_model)
```



Testing for heteroskedasticity

Example: Goldfeld-Quandt

Income looks potentially heteroskedastic; let's test via Goldfeld-Quandt.

```
# Arrange the data by income
test_df = arrange(test_df, income)
```

Testing for heteroskedasticity

Example: Goldfeld-Quandt

Income looks potentially heteroskedastic; let's test via Goldfeld-Quandt.

```
# Arrange the data by income
test_df = arrange(test_df, income)
# Re-estimate the model for the last and first 158 observations
est_model1 = lm(test_score ~ ratio + income, data = tail(test_df, 158))
est_model2 = lm(test_score ~ ratio + income, data = head(test_df, 158))
```

Testing for heteroskedasticity

Example: Goldfeld-Quandt

Income looks potentially heteroskedastic; let's test via Goldfeld-Quandt.

```
# Arrange the data by income
test_df = arrange(test_df, income)
# Re-estimate the model for the last and first 158 observations
est_model1 = lm(test_score ~ ratio + income, data = tail(test_df, 158))
est_model2 = lm(test_score ~ ratio + income, data = head(test_df, 158))
# Grab the residuals from each regression
e_model1 = residuals(est_model1)
e_model2 = residuals(est_model2)
```

Testing for heteroskedasticity

Example: Goldfeld-Quandt

Income looks potentially heteroskedastic; let's test via Goldfeld-Quandt.

```r
# Arrange the data by income
test_df = arrange(test_df, income)
# Re-estimate the model for the last and first 158 observations
est_model1 = lm(test_score ~ ratio + income, data = tail(test_df, 158))
est_model2 = lm(test_score ~ ratio + income, data = head(test_df, 158))
# Grab the residuals from each regression
e_model1 = residuals(est_model1)
e_model2 = residuals(est_model2)
# Calculate SSE for each regression
(sse_model1 = sum(e_model1^2))
```

```
#> [1] 19305.01
```

```r
(sse_model2 = sum(e_model2^2))
```

```
#> [1] 29537.83
```

Testing for heteroskedasticity

Example: Goldfeld-Quandt

Remember the Goldfeld-Quandt test statistic?

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

Example: Goldfeld-Quandt

Remember the Goldfeld-Quandt test statistic?

$$F_{n^\star-k,\, n^\star-k} = \frac{\text{SSE}_2}{\text{SSE}_1} \approx \frac{29,537.83}{19,305.01}$$

Example: Goldfeld-Quandt

Remember the Goldfeld-Quandt test statistic?

$$F_{n^\star-k,\, n^\star-k} = \frac{\text{SSE}_2}{\text{SSE}_1} \approx \frac{29,537.83}{19,305.01} \approx 1.53$$

Testing for heteroskedasticity

Example: Goldfeld-Quandt

Remember the Goldfeld-Quandt test statistic?

$F_{n^\star-k,\, n^\star-k} = \dfrac{\text{SSE}_2}{\text{SSE}_1} \approx \dfrac{29,537.83}{19,305.01} \approx 1.53$ Test via $F_{158-3,\, 158-3}$

Example: Goldfeld-Quandt

Remember the Goldfeld-Quandt test statistic?

$$F_{n^\star-k,\, n^\star-k} = \dfrac{\text{SSE}_2}{\text{SSE}_1} \approx \dfrac{29,537.83}{19,305.01} \approx 1.53 \quad \text{Test via } F_{158-3,\, 158-3}$$

```r
# G-Q test statistic
(f_gq = sse_model2/sse_model1)
```

```
#> [1] 1.530061
```

Testing for heteroskedasticity

Example: Goldfeld-Quandt

Remember the Goldfeld-Quandt test statistic?

$$F_{n^\star-k,\, n^\star-k} = \frac{\text{SSE}_2}{\text{SSE}_1} \approx \frac{29,537.83}{19,305.01} \approx 1.53 \quad \text{Test via } F_{158-3,\, 158-3}$$

```r
# G-Q test statistic
(f_gq = sse_model2/sse_model1)
```

```
#> [1] 1.530061
```

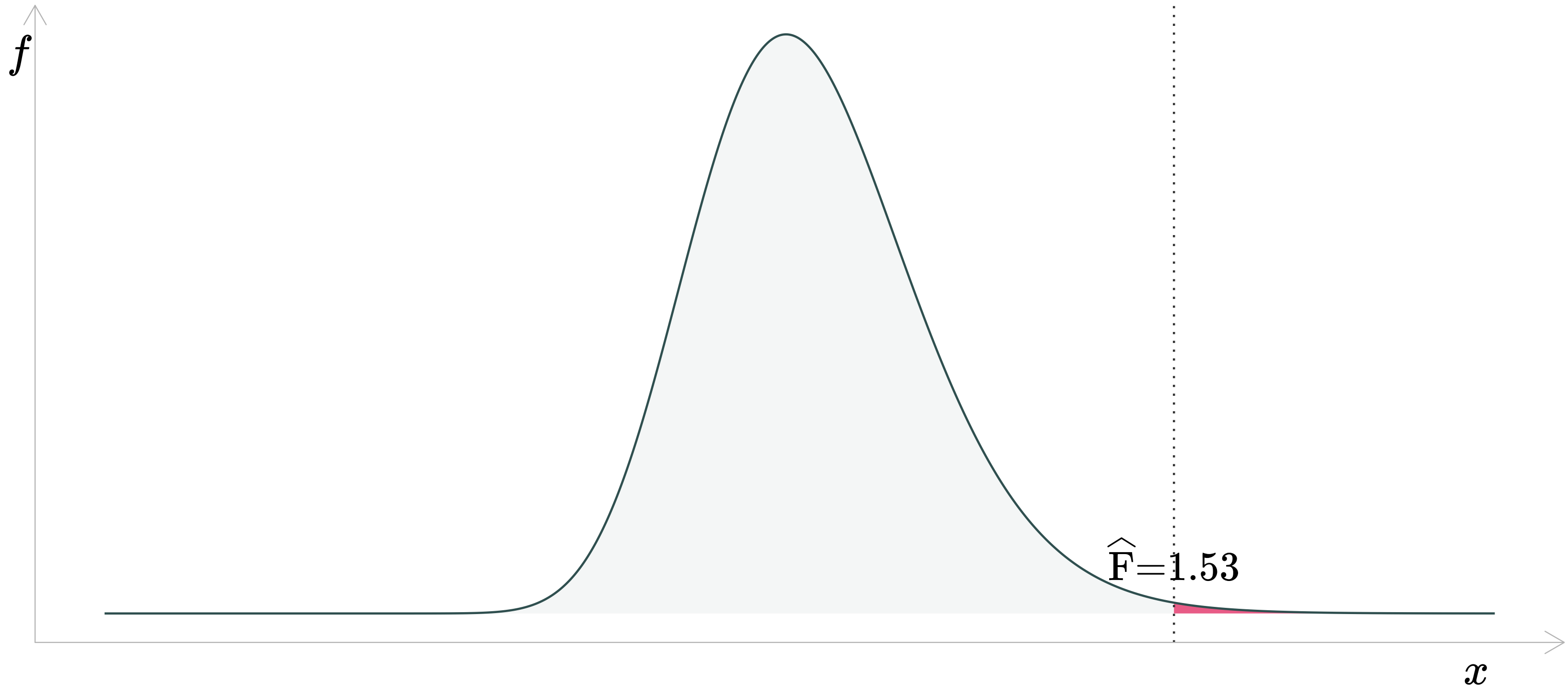
```r
# p-value
pf(q = f_gq, df1 = 158-3, df2 = 158-3, lower.tail = F)
```

```
#> [1] 0.004226666
```

Testing for heteroskedasticity

Example: Goldfeld-Quandt

The Goldfeld-Quandt test statistic and its null distribution



Testing for heteroskedasticity

Example: Goldfeld-Quandt

Putting it all together:

H_0: $\sigma_1^2 = \sigma_2^2$ *vs.* H_A: $\sigma_1^2 \neq \sigma_2^2$

Goldfeld-Quandt test statistic: $F \approx 1.53$

p-value ≈ 0.00423

$\therefore$ Reject H_0 (*p*-value is less than 0.05).

Conclusion: There is statistically significant evidence that $\sigma_1^2 \neq \sigma_2^2$. Therefore, we find statistically significant evidence of heteroskedasticity (at the 5-percent level).

Testing for heteroskedasticity

Example: Goldfeld-Quandt

What if we had chosen to focus on student-to-teacher ratio?

Testing for heteroskedasticity

Example: Goldfeld-Quandt

What if we had chosen to focus on student-to-teacher ratio?

```r
# Arrange the data by ratio
test_df = arrange(test_df, ratio)
# Re-estimate the model for the last and first 158 observations
est_model3 = lm(test_score ~ ratio + income, data = tail(test_df, 158))
est_model4 = lm(test_score ~ ratio + income, data = head(test_df, 158))
# Grab the residuals from each regression
e_model3 = residuals(est_model3)
e_model4 = residuals(est_model4)
# Calculate SSE for each regression
(sse_model3 = sum(e_model3^2))
```

```
#> [1] 26243.52
```

```r
(sse_model4 = sum(e_model4^2))
```

```
#> [1] 29101.52
```

Testing for heteroskedasticity

Example: Goldfeld-Quandt

$$F_{n^\star-k,\, n^\star-k} = \frac{\text{SSE}_4}{\text{SSE}_3} \approx \frac{29,101.52}{26,243.52} \approx 1.11$$

which has a *p*-value of approximately 0.2603.

Testing for heteroskedasticity

Example: Goldfeld-Quandt

$$F_{n^\star-k,\, n^\star-k} = \frac{\text{SSE}_4}{\text{SSE}_3} \approx \frac{29,101.52}{26,243.52} \approx 1.11$$

which has a p-value of approximately 0.2603.

$\therefore$ We would have failed to reject H_0, concluding that we failed to find statistically significant evidence of heteroskedasticity.

Testing for heteroskedasticity

Example: Goldfeld-Quandt

$$F_{n^\star - k,\, n^\star - k} = \frac{\text{SSE}_4}{\text{SSE}_3} \approx \frac{29,101.52}{26,243.52} \approx 1.11$$

which has a *p*-value of approximately 0.2603.

$\therefore$ We would have failed to reject H_0, concluding that we failed to find statistically significant evidence of heteroskedasticity.

Lesson: Understand the limitations of estimators, tests, *etc.*

Example: White

Let's test the same model and data with the White test.

Recall: We saved our residuals as `e` in our dataset, *i.e.*,

```
# Estimate the model
est_model = lm(test_score ~ ratio + income, data = test_df)
# Add the residuals to our dataset
test_df$e = residuals(est_model)
```

Example: White

The **White test** adds squared terms and interactions to initial regression specification (the right-hand side)

$$
\begin{aligned}
u_i^2 =& \alpha_0 + \alpha_1 \text{Ratio}_i + \alpha_2 \text{Income}_i \\
&+ \alpha_3 \text{Ratio}_i^2 + \alpha_4 \text{Income}_i^2 + \alpha_5 \text{Ratio}_i \times \text{Income}_i \\
&+ w_i
\end{aligned}
$$

The **White test** tests the null hypothesis

H_0: $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0$

We just need to write some `R` code to test H_0.

Example: White

Aside: `R` has funky notation for squared terms and interactions in `lm()`:

- **Squared terms** use `I()`, *e.g.*, `lm(y ~ I(x^2))`
- **Interactions** use `:` between the variables, *e.g.*, `lm(y ~ x1:x2)`

Example: Regress `y` on quadratic of `x1` and `x2`:

```
# Pretend quadratic regression w/ interactions
lm(y ~ x1 + x2 + I(x1^2) + I(x2^2) + x1:x2, data = pretend_df)
```

Example: White

Step 1: Regress e_i^2 on 1^{st} degree, 2^{nd} degree, and interactions

```r
# Regress squared residuals on quadratic of explanatory variables
white_model = lm(
  I(e^2) ~ ratio + income + I(ratio^2) + I(income^2) + ratio:income,
  data = test_df
)
# Grab the R-squared
(white_r2 = summary(white_model)$r.squared)
```

Example: White

Step 2: Collect R_e^2 from the regression.

```r
# Regress squared residuals on quadratic of explanatory variables
white_model = lm(
  I(e^2) ~ ratio + income + I(ratio^2) + I(income^2) + ratio:income,
  data = test_df
)
# Grab the R-squared
(white_r2 = summary(white_model)$r.squared)
```

```
#> [1] 0.07332222
```

Example: White

Step 3: Calculate White test statistic $\text{LM} = n \times R_e^2 \approx 420 \times 0.073$

```r
# Regress squared residuals on quadratic of explanatory variables
white_model = lm(
  I(e^2) ~ ratio + income + I(ratio^2) + I(income^2) + ratio:income,
  data = test_df
)
# Grab the R-squared
white_r2 = summary(white_model)$r.squared
# Calculate the White test statistic
(white_stat = 420 * white_r2)
```

```
#> [1] 30.79533
```

Example: White

Step 4: Calculate the associated *p*-value (where $\text{LM} \overset{d}{\sim} \chi^2_k$); here, $k = 5$

```
# Regress squared residuals on quadratic of explanatory variables
white_model = lm(
  I(e^2) ~ ratio + income + I(ratio^2) + I(income^2) + ratio:income,
  data = test_df
)
# Grab the R-squared
white_r2 = summary(white_model)$r.squared
# Calculate the White test statistic
white_stat = 420 * white_r2
# Calculate the p-value
pchisq(q = white_stat, df = 5, lower.tail = F)
```

```
#> [1] 1.028039e-05
```

Example: White

Putting everything together...

Example: White

Putting everything together...

H_0: $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0$

Heteroskedasticity

Example: White

Putting everything together...

H$_0$: $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0$ *vs.* H$_A$: $\alpha_i \neq 0$ for some $i \in \{1,\, 2,\, \ldots,\, 5\}$

Example: White

Putting everything together...

H_0: $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0$ *vs.* H_A: $\alpha_i \neq 0$ for some $i \in \{1,\, 2,\, \ldots,\, 5\}$

$$\begin{aligned} u_i^2 =& \alpha_0 + \alpha_1 \text{Ratio}_i + \alpha_2 \text{Income}_i \\ &+ \alpha_3 \text{Ratio}_i^2 + \alpha_4 \text{Income}_i^2 \\ &+ \alpha_5 \text{Ratio}_i \times \text{Income}_i + w_i \end{aligned}$$

Example: White

Putting everything together...

H$_0$: $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0$ *vs.* H$_A$: $\alpha_i \neq 0$ for some $i \in \{1,\, 2,\, \ldots,\, 5\}$

$$\begin{aligned} u_i^2 = &\alpha_0 + \alpha_1 \text{Ratio}_i + \alpha_2 \text{Income}_i \\ &+ \alpha_3 \text{Ratio}_i^2 + \alpha_4 \text{Income}_i^2 \\ &+ \alpha_5 \text{Ratio}_i \times \text{Income}_i + w_i \end{aligned}$$

Our White test statistic: $\text{LM} = n \times R_e^2 \approx 420 \times 0.073 \approx 30.8$

Example: White

Putting everything together...

H$_0$: $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0$ *vs.* H$_A$: $\alpha_i \neq 0$ for some $i \in \{1,\, 2,\, \ldots,\, 5\}$

$$\begin{aligned} u_i^2 =& \alpha_0 + \alpha_1 \text{Ratio}_i + \alpha_2 \text{Income}_i \\ &+ \alpha_3 \text{Ratio}_i^2 + \alpha_4 \text{Income}_i^2 \\ &+ \alpha_5 \text{Ratio}_i \times \text{Income}_i + w_i \end{aligned}$$

Our White test statistic: $\text{LM} = n \times R_e^2 \approx 420 \times 0.073 \approx 30.8$

Under the χ_5^2 distribution, this $\widehat{\text{LM}}$ has a *p*-value less than 0.001.

Example: White

Putting everything together...

H_0: $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0$ *vs.* H_A: $\alpha_i \neq 0$ for some $i \in \{1,\, 2,\, \dots,\, 5\}$

$$\begin{aligned} u_i^2 =& \alpha_0 + \alpha_1 \text{Ratio}_i + \alpha_2 \text{Income}_i \\ & + \alpha_3 \text{Ratio}_i^2 + \alpha_4 \text{Income}_i^2 \\ & + \alpha_5 \text{Ratio}_i \times \text{Income}_i + w_i \end{aligned}$$

Our White test statistic: $\text{LM} = n \times R_e^2 \approx 420 \times 0.073 \approx 30.8$

Under the χ_5^2 distribution, this $\widehat{\text{LM}}$ has a *p*-value less than 0.001.

$\therefore$ We **reject H_0**

Example: White

Putting everything together...

H_0: $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0$ *vs.* H_A: $\alpha_i \neq 0$ for some $i \in \{1,\ 2,\ \dots,\ 5\}$

$$\begin{aligned} u_i^2 = & \alpha_0 + \alpha_1 \text{Ratio}_i + \alpha_2 \text{Income}_i \\ & + \alpha_3 \text{Ratio}_i^2 + \alpha_4 \text{Income}_i^2 \\ & + \alpha_5 \text{Ratio}_i \times \text{Income}_i + w_i \end{aligned}$$

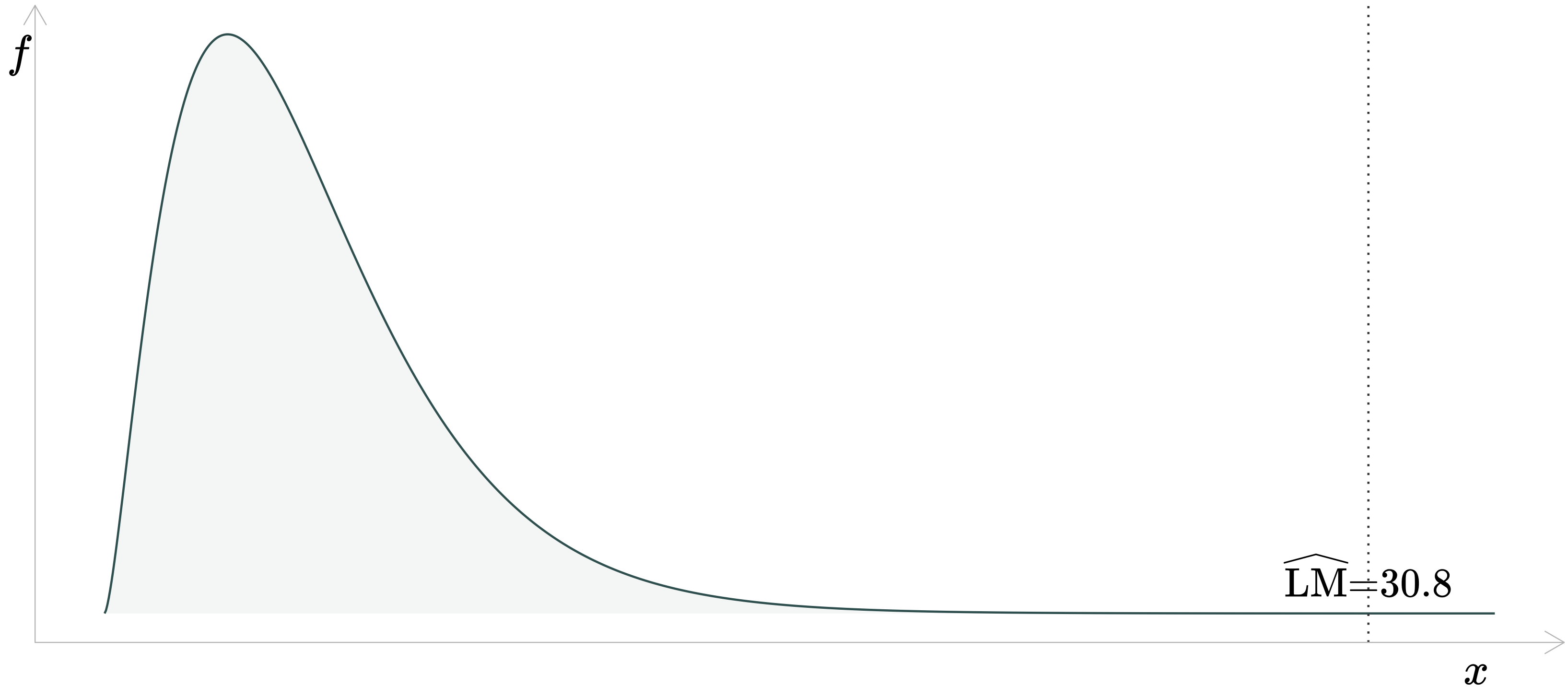
Our White test statistic: $\text{LM} = n \times R_e^2 \approx 420 \times 0.073 \approx 30.8$

Under the χ_5^2 distribution, this $\widehat{\text{LM}}$ has a *p*-value less than 0.001.

$\therefore$ We **reject H_0** and conclude there is **statistically significant evidence of heteroskedasticity** (at the 5-percent level).

Example: White

The White test statistic and its null distribution



Heteroskedasticity

Review questions

Review questions

- **Q:** What is the definition of heteroskedasticity?
- **Q:** Why are we concerned about heteroskedasticity?
- **Q:** Does plotting y against x, tell us anything about heteroskedasticity?
- **Q:** Does plotting e against x, tell us anything about heteroskedasticity?
- **Q:** Since we cannot observe the u_i's, what do we use to *learn about* heteroskedasticity?
- **Q:** Which test do you recommend to test for heteroskedasticity? Why?

Review questions

- **Q:** What is the definition of heteroskedasticity?

Review questions

- **Q:** What is the definition of heteroskedasticity?
- **A:**
  Math: $\mathrm{Var}(u_i|X) \neq \mathrm{Var}\big(u_j|X\big)$ for some $i \neq j$.
  Words: There is a systematic relationship between the variance of u_i and our explanatory variables.

Review questions

- **Q:** What is the definition of heteroskedasticity?
- **Q:** Why are we concerned about heteroskedasticity?

Review questions

- **Q:** What is the definition of heteroskedasticity?

- **Q:** Why are we concerned about heteroskedasticity?
- **A:** It biases our standard errors—wrecking our statistical tests and confidence intervals. Also: OLS is no longer the most efficient (best) linear unbiased estimator.

Heteroskedasticity

Review questions

- **Q:** What is the definition of heteroskedasticity?
- **Q:** Why are we concerned about heteroskedasticity?
- **Q:** Does plotting y against x, tell us anything about heteroskedasticity?

Review questions

- **Q:** What is the definition of heteroskedasticity?
- **Q:** Why are we concerned about heteroskedasticity?
- **Q:** Does plotting y against x, tell us anything about heteroskedasticity?
- **A:** It's not exactly what we want, but since y is a function of x and u, it can still be informative. If y becomes more/less disperse as x changes, we likely have heteroskedasticity.

Heteroskedasticity

Review questions

- **Q:** What is the definition of heteroskedasticity?
- **Q:** Why are we concerned about heteroskedasticity?
- **Q:** Does plotting y against x, tell us anything about heteroskedasticity?
- **Q:** Does plotting e against x, tell us anything about heteroskedasticity?

Heteroskedasticity

Review questions

- **Q:** What is the definition of heteroskedasticity?
- **Q:** Why are we concerned about heteroskedasticity?
- **Q:** Does plotting y against x, tell us anything about heteroskedasticity?
- **Q:** Does plotting e against x, tell us anything about heteroskedasticity?
- **A:** Yes. The spread of e depicts its variance—and tells us something about the variance of u. Trends in this variance, along x, suggest heteroskedasticity.

Heteroskedasticity

Review questions

- **Q:** What is the definition of heteroskedasticity?
- **Q:** Why are we concerned about heteroskedasticity?
- **Q:** Does plotting y against x, tell us anything about heteroskedasticity?
- **Q:** Does plotting e against x, tell us anything about heteroskedasticity?
- **Q:** Since we cannot observe the u_i's, what do we use to *learn about* heteroskedasticity?

Heteroskedasticity

Review questions

- **Q:** What is the definition of heteroskedasticity?
- **Q:** Why are we concerned about heteroskedasticity?
- **Q:** Does plotting y against x, tell us anything about heteroskedasticity?
- **Q:** Does plotting e against x, tell us anything about heteroskedasticity?
- **Q:** Since we cannot observe the u_i's, what do we use to *learn about* heteroskedasticity?
- **A:** We use the e_i's to predict/learn about the u_i's. This trick is key for almost everything we do with heteroskedasticity testing/correction.

Review questions

- **Q:** What is the definition of heteroskedasticity?
- **Q:** Why are we concerned about heteroskedasticity?
- **Q:** Does plotting y against x, tell us anything about heteroskedasticity?
- **Q:** Does plotting e against x, tell us anything about heteroskedasticity?
- **Q:** Since we cannot observe the u_i's, what do we use to *learn about* heteroskedasticity?
- **Q:** Which test do you recommend to test for heteroskedasticity? Why?

Heteroskedasticity

Review questions

- **Q:** What is the definition of heteroskedasticity?
- **Q:** Why are we concerned about heteroskedasticity?
- **Q:** Does plotting y against x, tell us anything about heteroskedasticity?
- **Q:** Does plotting e against x, tell us anything about heteroskedasticity?
- **Q:** Since we cannot observe the u_i's, what do we use to *learn about* heteroskedasticity?
- **Q:** Which test do you recommend to test for heteroskedasticity? Why?
- **A:** I like White. Fewer assumptions. Fewer issues.

Next time: Living/working with heteroskedasticity.

Appendix

One more test...

Testing for heteroskedasticity

The Breusch-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 Breusch-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: $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$

The Breusch-Pagan test

The B-P test statistic† is

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

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 Breusch-Pagan test

The B-P test statistic[†] is

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

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, $\mathbf{LM}$ 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 Breusch-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

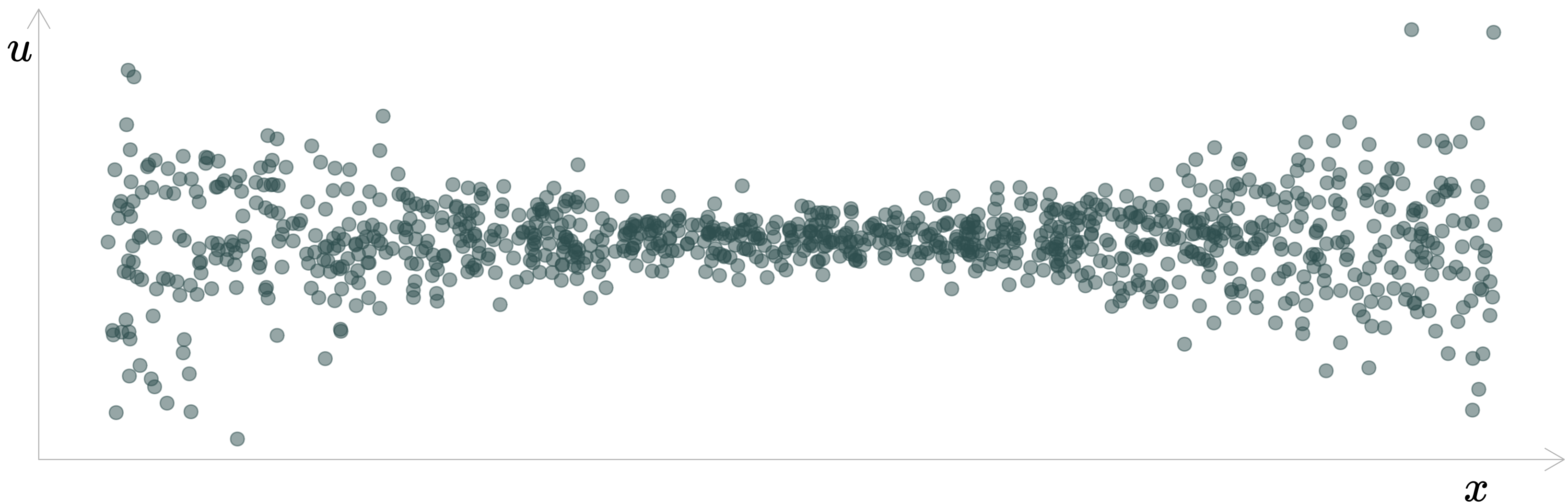
The Breusch-Pagan test

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.

The Breusch-Pagan test

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



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

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

Testing for heteroskedasticity

Example: Breusch-Pagan

Let's test the same model with the Breusch Pagan.

Recall: We saved our residuals as `e` in our dataset, *i.e.*,

```
test_df$e = residuals(est_model)
```

Testing for heteroskedasticity

Example: Breusch-Pagan

In B-P, we first regress e_i^2 on the explanatory variables,

```
# Regress squared residuals on explanatory variables
bp_model = lm(I(e^2) ~ ratio + income, data = test_df)
```

Testing for heteroskedasticity

Example: Breusch-Pagan

and use the resulting R^2 to calculate a test statistic.

```
# Regress squared residuals on explanatory variables
bp_model = lm(I(e^2) ~ ratio + income, data = test_df)
# Grab the R-squared
(bp_r2 = summary(bp_model)$r.squared)
```

```
#> [1] 3.23205e-05
```

Testing for heteroskedasticity

Example: Breusch-Pagan

The Breusch-Pagan test statistic is

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

Testing for heteroskedasticity

Example: Breusch-Pagan

The Breusch-Pagan test statistic is

$$\mathbf{LM} = \boldsymbol{n} \times \boldsymbol{R_e^2} \approx \mathbf{420} \times \mathbf{0.0000323}$$

Testing for heteroskedasticity

Example: Breusch-Pagan

The Breusch-Pagan test statistic is

$$\text{LM} = n \times R_e^2 \approx 420 \times 0.0000323 \approx 0.0136$$

which we test against a χ_k^2 distribution (here: $k = 2$).[†]

Testing for heteroskedasticity

Example: Breusch-Pagan

The Breusch-Pagan test statistic is

$$\text{LM} = n \times R_e^2 \approx 420 \times 0.0000323 \approx 0.0136$$

which we test against a χ_k^2 distribution (here: $k = 2$).[†]

```
# B-P test statistic
bp_stat = 420 * bp_r2
# Calculate the p-value
pchisq(q = bp_stat, df = 2, lower.tail = F)
```

```
#> [1] 0.9932357
```

[†]: k is the number of explanatory variables (excluding the intercept).

Testing for heteroskedasticity

Example: Breusch-Pagan

H_0: $\boldsymbol{\alpha_1 = \alpha_2 = 0}$ *vs.* H_A: $\boldsymbol{\alpha_1 \neq 0}$ and/or $\boldsymbol{\alpha_2 \neq 0}$

for the model $\boldsymbol{u_i^2 = \alpha_0 + \alpha_1 \textbf{Ratio}_i + \alpha_2 \textbf{Income}_i + w_i}$

Testing for heteroskedasticity

Example: Breusch-Pagan

H_0: $\boldsymbol{\alpha_1 = \alpha_2 = 0}$ *vs.* H_A: $\boldsymbol{\alpha_1 \neq 0}$ and/or $\boldsymbol{\alpha_2 \neq 0}$

for the model $\boldsymbol{u_i^2 = \alpha_0 + \alpha_1 \text{Ratio}_i + \alpha_2 \text{Income}_i + w_i}$

Breusch-Pagan test statistic: $\widehat{\textbf{LM}} \approx \mathbf{0.014}$

Testing for heteroskedasticity

Example: Breusch-Pagan

H_0: $\alpha_1 = \alpha_2 = 0$ *vs.* H_A: $\alpha_1 \neq 0$ and/or $\alpha_2 \neq 0$

for the model $u_i^2 = \alpha_0 + \alpha_1 \text{Ratio}_i + \alpha_2 \text{Income}_i + w_i$

Breusch-Pagan test statistic: $\widehat{\text{LM}} \approx 0.014$

p-value ≈ 0.993

Testing for heteroskedasticity

Example: Breusch-Pagan

H_0: $\boldsymbol{\alpha_1 = \alpha_2 = 0}$ *vs.* H_A: $\boldsymbol{\alpha_1 \neq 0}$ and/or $\boldsymbol{\alpha_2 \neq 0}$

for the model $\boldsymbol{u_i^2 = \alpha_0 + \alpha_1 \text{Ratio}_i + \alpha_2 \text{Income}_i + w_i}$

Breusch-Pagan test statistic: $\boldsymbol{\widehat{\text{LM}} \approx 0.014}$

p-value $\boldsymbol{\approx 0.993}$

$\therefore$ Fail to reject H_0 (the *p*-value is greater than 0.05)

Testing for heteroskedasticity

Example: Breusch-Pagan

H_0: $\alpha_1 = \alpha_2 = 0$ *vs.* H_A: $\alpha_1 \neq 0$ and/or $\alpha_2 \neq 0$

for the model $u_i^2 = \alpha_0 + \alpha_1 \text{Ratio}_i + \alpha_2 \text{Income}_i + w_i$

Breusch-Pagan test statistic: $\widehat{\text{LM}} \approx 0.014$

p-value ≈ 0.993

$\therefore$ Fail to reject H_0 (the *p*-value is greater than 0.05)

Conclusion: We do not find statistically significant evidence of heteroskedasticity at the 5-percent level.

Testing for heteroskedasticity

Example: Breusch-Pagan

H_0: $\alpha_1 = \alpha_2 = 0$ *vs.* H_A: $\alpha_1 \neq 0$ and/or $\alpha_2 \neq 0$

for the model $u_i^2 = \alpha_0 + \alpha_1 \text{Ratio}_i + \alpha_2 \text{Income}_i + w_i$

Breusch-Pagan test statistic: $\widehat{\text{LM}} \approx 0.014$

p-value ≈ 0.993

$\therefore$ Fail to reject H_0 (the *p*-value is greater than 0.05)

Conclusion: We do not find statistically significant evidence of heteroskedasticity at the 5-percent level. (We find no evidence of a *linear* relationship between u_i^2 and the explanatory variables.)

Example: Breusch-Pagan

The Breusch-Pagan test statistic and its null distribution