# Regression Analysis: Heteroskedasticity

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$$

Statistics and Econometrics

Jiahua Wu

382 Business School
j.wu@imperial.ac.uk

Imperial College Business School

Imperial means Intelligent Business

# Roadmap

- Regression analysis with cross-sectional data
  - Basics: estimation, inference, analysis with dummy variables
  - More involved: model specification and data issues
- Advanced topics
  - Binary dependent variable models
  - Panel data analysis
  - Time series analysis

# Outline (Wooldridge, Chap. 8.1 - 8.3)

- Consequences of heteroskedasticity
- Testing for heteroskedasticity
- Heteroskedasticity-robust inference

# Outline

- Consequences of heteroskedasticity
- Testing for heteroskedasticity
- Heteroskedasticity-robust inference

# Homoskedasticity vs Heteroskedasticity

- Recall that the variance of OLS estimator is given by

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}, \qquad j = 1, \ldots, k,$$

  where $SST_j = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$, $\bar{x}_j = n^{-1}\sum_{i=1}^{n} x_{ij}$, and $R_j^2$ is the R-squared from regressing $x_j$ on all other independent variables.

- For the variance formula to be valid, we need $u$ to be homoskedastic

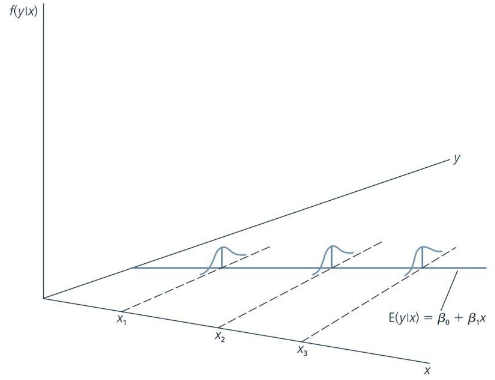
# Homoskedasticity vs Heteroskedasticity

## Assumption (homoskedasticity)

$Var(u_i|x_{i1}, \ldots, x_{ik}) = \sigma^2$ for $i = 1, 2, \ldots, n$. (It implies $Var(u_i) = \sigma^2$)

FIGURE 2.8

The simple regression model under homoskedasticity.



# Homoskedasticity vs Heteroskedasticity

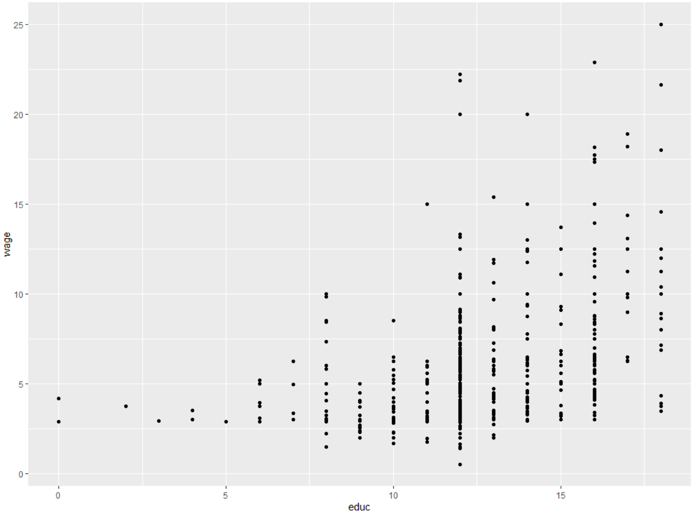



Figure: A scatter plot of *wage* and *educ* (wage1.RData)

# Consequences of Heteroskedasticity

- With heteroskedasticity,
  - OLS estimators are still unbiased
  - The standard errors of the estimates are biased if we ignore heteroskedasticity
  - We cannot use the usual $t$ statistic or $F$ statistic for drawing inferences

# Outline

- Consequences of heteroskedasticity
- Testing for heteroskedasticity
- Heteroskedasticity-robust inference

# Heteroskedasticity Tests

- Essentially want to test $H_0 : Var(u|x_1, x_2, \ldots, x_k) = \sigma^2$, which is equivalent to

$$H_0 : E(u^2|x_1, x_2, \ldots, x_k) = E(u^2) = \sigma^2$$

# Heteroskedasticity Tests: The Breusch-Pagan Test

- If assume a linear relationship between $u^2$ and $x_j$, i.e.,

$$u^2 = \delta_0 + \delta_1 x_1 + \cdots + \delta_k x_k + v,$$

  the null hypothesis of homoskedasticity is equivalent to $H_0 : \delta_1 = \delta_2 = \cdots = \delta_k = 0$.

- The Breusch-Pagan test
  - OLS $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$ and save the squared residuals $\hat{u}^2$
  - OLS $\hat{u}^2 = \delta_0 + \delta_1 x_1 + \cdots + \delta_k x_k + \textit{error}$ and save the R-squared $R^2_{\hat{u}^2}$
  - The test statistic

$$F = \frac{R^2_{\hat{u}^2}/k}{(1 - R^2_{\hat{u}^2})/(n-k-1)} \sim F_{k,n-k-1} \quad \text{under the null}$$

  - Reject the null if $F$ is too large (i.e., has a too-small $p$-value)

# Heteroskedasticity Tests: The White Test

- The Breusch-Pagan test will detect any linear forms of heteroskedasticity
- The White test allows for nonlinearities by using squares and crossproducts of all the $x$'s
- The White test
  - OLS $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$ and save the residuals and the fitted values, $\hat{u}$ and $\hat{y}$
  - OLS $\hat{u}^2 = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 + error$ and save the R-squared $R^2_{\hat{u}^2}$
  - The test statistic

$$F = \frac{R^2_{\hat{u}^2}/2}{(1 - R^2_{\hat{u}^2})/(n-3)} \sim F_{2,n-3} \quad \text{under the null}$$

  - Reject the null if $F$ is too large (i.e., has a too-small $p$-value)

# Heteroskedasticity Tests: An Example

- Eg. Wage model (wage1.RData)

$$\widehat{wage} = \underset{(.729)}{-2.873} + \underset{(.051)}{.599}\, educ + \underset{(.012)}{.022}\, exper + \underset{(.022)}{.169}\, tenure$$

$n = 526, R^2 = .306$

  - Breusch-Pagan test: $p$-value $= 2.349$e-09
  - White test: $p$-value $= 5.713$e-12

# Outline

- Consequences of heteroskedasticity
- Testing for heteroskedasticity
- Heteroskedasticity-robust inference

# Heteroskedasticity-Robust Inference

- It is possible to adjust the OLS standard errors to make the $t$ stat (or $F$ stat) valid in the presence of heteroskedasticity of unknown form
- The adjustment is called heteroskedasticity-robust procedure
- The procedure is "robust" because the adjusted $t$ stat (or $F$ stat) is valid regardless of the type of heteroskedasticity in the population (even if there is no heteroskedasticity)

# Robust Standard Errors

- Denote $r.se(\hat{\beta}_j)$ as robust standard error
- The robust $t$ stat is

$$t \text{ stat} = \frac{\hat{\beta}_j - a_j}{r.se(\hat{\beta}_j)}$$

- These robust standard errors only have asymptotic justification
  - With small sample sizes, robust $t$ stat will not have a distribution close to $t$, and inferences will not be correct
- The robust $F$ stat must be computed using a formula different from the original one

# Robust Standard Errors: An Example

- Eg. Wage model (wage1.RData)

$$\widehat{wage} = \underset{(.729)}{-2.873} + \underset{(.051)}{.599}\, educ + \underset{(.012)}{.022}\, exper + \underset{(.022)}{.169}\, tenure$$

The model with robust standard errors is

$$\widehat{wage} = \underset{[.807]}{-2.873} + \underset{[.061]}{.599}\, educ + \underset{[.011]}{.022}\, exper + \underset{[.029]}{.169}\, tenure$$

$n = 526, R^2 = .306$ (the robust results are in [ ])

- Hypotheses: $H_0 : \beta_{educ} - \beta_{exper} = 0$ vs $H_1 : \beta_{educ} - \beta_{exper} \neq 0$
  - $F$ stat: $(139.28)$, $[98.39]$

# When to Use Robust Standard Errors in Practice?

- Always report robust standard errors for models with non-constant variance, including linear probability model and panel data model
- For other models,
  - Test for heteroskedasticity
  - Report robust standard errors only if there is evidence of heteroskedasticity, because
    - usual standard errors are typically smaller
    - robust standard errors only have asymptotic justification