# LECTURE 8

# Introduction to Econometrics

# Multicollinearity & Heteroskedasticity

Hieu Nguyen

Fall semester, 2024

# On previous lectures

- We discussed the specification of a regression equation
- **Specification** consists of choosing:
  1. **correct independent variables** (lecture 7: omitted or irrelevant)
  2. **correct functional form** (lecture 6)
  3. correct form of the stochastic error term

# On today's lecture

- Today we will continue the discussion of choosing the correct independent variables by talking about **multicollinearity**

- But also, we will start the discussion of the correct form of the error term by talking about **heteroskedasticity**

- For both of these issues, we will learn
  - what is the nature of the problem
  - what are its consequences
  - how it is diagnosed
  - what are the remedies available

# Perfect multicollinearity

- Some explanatory variable is a perfect linear function of one or more other explanatory variables
- Violation of one of the classical assumptions
- OLS estimate cannot be found
- Intuitively: the estimator cannot distinguish which of the explanatory variables causes the change of the dependent variable if they move together
    - Technically: the matrix $\mathbf{X}'\mathbf{X}$ is singular (not invertible)
- Rare and easy to detect
    - Usually an obvious mistake: e. g. full set of dummies and a constant

# Examples of perfect multicollinearity

**Dummy variable trap**

- Inclusion of dummy variable for each category in the model with intercept
- Example: wage equation for sample of individuals who have high-school education or higher:

$$wage_i = \beta_1 + \beta_2 high\ school_i + \beta_3 university_i + \beta_4 phd_i + e_i$$

- Automatically detected by most statistical softwares

# Imperfect multicollinearity

- Two or more explanatory variables are highly correlated in the particular data set

- OLS estimate can be found, but it may be very imprecise

  Intuitively: the estimator can hardly distinguish the effects of the explanatory variables if they are highly correlated

  Technically: the matrix $\mathbf{X}^j\mathbf{X}$ is nearly singular and this causes the variance of the estimator $Var\left(\widehat{\beta}\right) = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$ to be very large

- Usually referred to simply as "multicollinearity"

# CONSEQUENCES OF MULTICOLLINEARITY

1. Estimates remain unbiased and consistent (estimated coefficients are not affected)

2. Standard errors of coefficients increase

    - Confidence intervals are very large - estimates are less reliable

    - $t$-statistics are smaller - variables may become insignificant

# DETECTION OF MULTICOLLINEARITY

- Some degree of multicollinearity exists in every equation - the aim is to recognize when it causes a severe problem

- Multicollinearity can be signaled by the underlying theory, but it is very sample depending

- We judge the severity of multicollinearity based on the properties of our sample and on the results we obtain

- One simple method: examine correlation coefficients between explanatory variables (run covariance)

  if some of them is too high, we may suspect that the coefficients of these variables can be affected by multicollinearity

# Example

- Estimating the demand for gasoline in the U.S.:

$$\widehat{PCON}_i = 389.6 - \underset{(13.2)}{36.5}\ TAX_i + \underset{(10.3)}{60.8}\ UHM_i - \underset{(0.043)}{0.061}\ REG_i$$

$$t = 5.92 \qquad -2.77 \qquad -1.43$$

$$R^2 = 0.924 \quad , \quad n = 50 \quad , \quad Corr(UHM, REG) = 0.978$$

| | | |
|---|---|---|
| $PCON_i$ | ... | petroleum consumption in the *i*-th state |
| $TAX_i$ | ... | the gasoline tax rate in the *i*-th state |
| $UHM_i$ | ... | urban highway miles within the *i*-th state |
| $REG_i$ | ... | motor vehicle registrations in the *i*-the state |

# EXAMPLE

- We suspect a multicollinearity between urban highway miles and motor vehicle registration across states, because those states that have a lot of highways might also have a lot of motor vehicles.
- Therefore, we might run into multicollinearity problems. How do we detect multicollinearity?
  - Look at correlation coefficient. It is indeed huge (0.978).
  - Look at the coefficients of the two variables. Are they both individually significant? *UHM* is significant, but *REG* is not. This further suggests a presence of multicollinearity.
- Remedy: try dropping one of the correlated variables.

# EXAMPLE

- Comparison:

$$\widehat{PCON_i} = 551.7 - \underset{(16.9)}{53.6}\, TAX_i + \underset{(0.012)}{0.186}\, REG_i$$

$$t = -3.18 \qquad 15.88$$

$$R^2 = 0.866 \quad , \quad n = 50$$

$$\widehat{PCON_i} = 410.0 - \underset{(13.1)}{39.6}\, TAX_i + \underset{(2.16)}{46.4}\, UHM_i$$

$$t = -3.02 \qquad 21.40$$

$$R^2 = 0.921 \quad , \quad n = 50$$

## REMEDIES FOR MULTICOLLINEARITY

- Drop a redundant variable
  - when the variable is not needed to represent the effect on the dependent variable
  - in case of severe multicollinearity, it makes no statistical difference which variable is dropped
  - theoretical underpinnings of the model should be the basis for such a decision

- Do nothing
  - when multicollinearity does not cause insignificant *t*-scores or unreliable estimated coefficients
  - deletion of collinear variable can cause specification bias

- Increase the size of the sample
  - the confidence intervals are narrower when we have more observations

# Remedies for multicollinearity

- Transform the multicollinear variables
  - in case when all variables are extremely important on theoretical grounds
  - we can try various transformations:
    1. Combination of multicollinear variables
    2. First differences (for time series)
    3. Increase the size of the sample (the confidence intervals are narrower when we have more observations)

# EXAMPLE

- **Estimating the demand equation for fish:**

$$\hat{F}_t = 7.96 + 0.03PF_t + 0.0047PB_t + 0.36 \ln YD_t$$
$$\qquad (0.03) \qquad (0.019) \qquad (1.15)$$

$$t = 0.98 \qquad 0.24 \qquad 0.31$$

$$R^2 = 0.667 \quad , \quad n{=}25$$

$F_t$ ... consumption of fish
$PF_t$ ... price of fish
$PB_t$ ... price of beef
$YD_t$ ... disposable income

# EXAMPLE

- If we transform as follows: $RP_t = \frac{PF_t}{PB_t}$
- We have:

$$\hat{F}_t = -5.17 - \underset{(1.43)}{1.93} RP_t + \underset{(0.66)}{0.36} \ln YD_t$$

$$t = -1.35 \qquad 4.13$$

$$R^2 = 0.588 \quad , \quad n=25$$

- It is better, but still not far from perfect. What we can do is to increase the number of observations.

# HETEROSKEDASTICITY

- Observations of the error term are drawn from a distribution that has no longer a constant variance

$$Var(\varepsilon_i) = \sigma_i^2 \quad , \quad i = 1, 2, \ldots, n$$

Note: constant variance means: $Var(\varepsilon_i) = \sigma^2 (i = 1, 2, \ldots, n)$

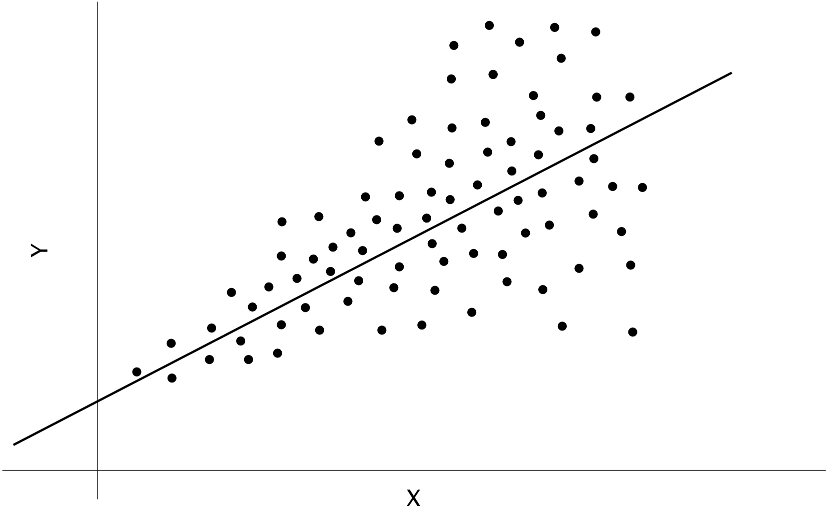
- Often occurs in data sets in which there is a wide disparity between the largest and smallest observed values

  Smaller values often connected to smaller variance and larger values to larger variance (e.g. consumption of households based on their income level)

- One particular form of heteroskedasticity (variance of the error term is a function of some observable variable):

$$Var(\varepsilon_i) = h(x_i) \quad , \quad i = 1, 2, \ldots, n$$

# HETEROSKEDASTICITY



# CONSEQUENCES OF HETEROSKEDASTICITY

- Violation of one of the classical assumptions

1. Estimates remain unbiased and consistent (estimated coefficients are not affected)
2. Estimated standard errors of the coefficients are biased
   - heteroskedastic error term causes the dependent variable to fluctuate in a way that the OLS estimation procedure attributes to the independent variable
   - heteroskedasticity biases *t* statistics, which leads to unreliable hypothesis testing
   - typically, we encounter underestimation of the standard errors, so the *t* scores are incorrectly too high

# VARIENCE OF OLS UNDER HETEROSKEDASTICITY

- Consider the model $\quad \boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$

- OLS estimate is $\quad \hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$

- Variance of the estimate is

$$Var(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

Where $\Omega = Var(\varepsilon) = \begin{pmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n^2 \end{pmatrix}$

- Compare to the variance under homoskedasticity:

$$Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

# Detection of heteroskedasticity

- There is a battery of tests for heteroskedasticity
  - Sometimes, simple visual analysis of residuals is sufficient to detect heteroskedasticity
- We will derive a test for the model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$$

- The test is based on analysis of residuals

$$e_i = y_i - \widehat{y}_i = y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i + \widehat{\beta}_2 z_i)$$

- The null hypothesis for the test is no heteroskedasticity: $E(e^2) = \sigma^2$
  - Therefore, we will analyse the relationship between $e^2$ and explanatory variables

# CHI SQUARED DISTRIBUTION

- Chi-squared distribution with k degrees of freedom: $\chi_k^2$
- Let $Z_i \sim N(0,1)$ for each $i$ and independent, then

$$X = \sum_{i=1}^{k} Z_i^2 \rightarrow X \sim \chi_k^2$$

# WHITE TEST FOR HETEROSKEDASTICITY

1. **Estimate the equation, get the residuals $e_i$**
2. **Regress the squared residuals on all explanatory variables and on squares and cross-products of all explanatory variables:**

$$e_i^2 = \alpha_0 + \alpha_1 x_i + \alpha_2 z_i + \alpha_3 x_i^2 + \alpha_4 z_i^2 + \alpha_5 x_i z_i + v_i \quad (1)$$

3. **Get the $R^2$ of this regression and the sample size $n$**

4. Test the joint significance of (2): test statistic $= nR^2 \sim \chi_k^2$, where $k$ is the number of slope coefficients in (2)

5. If $nR^2$ is larger than the $\chi_k^2$ critical value, then we have to reject $H_0$ of no heteroskedasticity

# BREUSCH PAGAN TEST FOR HETEROSKEDASTICITY

1. **Estimate the equation, get the residuals $e_i$**
2. **Standardize the squares of residuals:**

$$\tilde{e}_i^2 = \frac{e_i^2}{\tilde{\sigma}^2}, \quad \text{where} \quad \tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i$$

3. **Regress the squared residuals on all explanatory variables:**

$$e_i^2 = \alpha_0 + \alpha_1 x_i + \alpha_2 z_i + v_i \qquad (2)$$

4. Get the $R^2$ of this regression and the sample size $n$
5. Test the joint significance of (2): use F-test or alternative test statistic, LM = $nR^2 \sim \chi_k^2$, where $k$ is the number of slope coefficients in (2)

6. If $nR^2$ is larger than the $\chi^2$ critical value, then we have to reject $H_0$ of no heteroskedasticity

# REMEDIES FOR HETEROSKEDASTICITY

1. Redefining the variables

   in order to reduce the variance of observations with extreme values

   e.g. by taking logarithms or by scaling some variables

2. Weighted Least Squares (WLS)

   consider the model $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$

   suppose $Var(\varepsilon_i) = \sigma^2 z^2{}_i$

   it can be proved that if we redefine the model as

$$\frac{y_i}{z_i} = \beta_0 \frac{1}{z_i} + \beta_1 \frac{x_i}{z_i} + \beta_2 + \frac{\varepsilon_i}{z_i} \quad ,$$

   it becomes homoskedastic

3. **Heteroskedasticity-corrected robust standard errors**

## HETEROSKEDASTICITY-CORRECTED ROBUST ERRORS

- The logic behind:

  Since heteroskedasticity causes problems with the standard errors of OLS but not with the coefficients, it makes sense to improve the estimation of the standard errors in a way that does not alter the estimate of the coefficients (White, 1980)

- Heteroskedasticity-corrected standard errors are typically larger than OLS s.e., thus producing lower *t* scores

- In panel and cross-sectional data with group-level variables, the method of **clustering** the standard errors is the desired answer to heteroskedasticity

# SUMMARY

- Multicollinearity

  does not lead to inconsistent estimates, but it makes them lose significance

  if really necessary, can be remedied by dropping or transforming variables, or by getting more data

- Heteroskedasticity

  does not lead to inconsistent estimates, but invalidates inference

  can be simply remedied by the use of (clustered) robust standard errors

- Readings:
  Studenmund Chapter 8 and 10
  Wooldridge Chapter 8