Mon 2019-11-4

Chapter 19 → Term paper

Structural Multicollinearity
Data collinearity

Imperfect (near) Multicollinearity

$$X_{1i} = \alpha_0 + \alpha_1 X_{2i} + u_i$$

Where u_i can be regarded as an approximation ~~~/
Imperfect multicollinearity ~~can be viewed~~ is a strong linear relationship between variables. Multicollinearity is a sample phenomenon

i) If the goal is simply to predict Y from a set of X variables, then multicollinearity is not a problem
ii) If the goal is to understand how X influences Y, it is a problem
iii) P values can be misleading
iv) Confidence intervals are off

Correlation - linear strength between variables
Collinearity - One of the explanatories is linearly dependent of another

Sources of Multicollinearity
1) Data collection method
2) Constraints on model or population
3) Model specification
4) Over defined model

Consequences of Multicollinearity

- Estimates will remain unbiased
- Computed t will fall
- High R^2 value but few significant t's
- Wrong sign for regression coefficients

Auxiliary Regression
Regress each explanatory variable on the other explanatory

Variance inflation factor

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R_j^2}$$

Remedy

- Drop one of the variables
- Acquire more data
- Re-thinking the model: level-level, level-log, log-log

Wed 2019-11-06

Chapter 8 - Heteroscedasticity

HetSc is a violation of CLRM. u_i must have constant variance σ^2:

$$u \sim N(0, \sigma^2)$$
$$Var(u_i) = \sigma^2$$

If that is not the case:

$$Var(u_i) = \sigma_i^2 \quad \text{(non-constant variance)}$$

Types of HetSc:

- Conditional - nonconstant volatility when periods of high & low volatility cannot be IDed
- Unconditional - these periods can be IDed

Conditional is not predictable by nature

What is the nature of HetSc?
Omitted Independen Variables
How HetSc can be tested
What are consequences
Formal Method tests for HetSc
How do we detect it exists in a given situation
What do we do if Heteroscedasticity is observed

Homoscedasticity is when:

$$E(u_i^2) = \sigma^2 \text{ for all } i$$

As opposed to Het:

$$E(u_i^2) = \sigma_i^2$$

ie, conditional variances are not constant.

OR $E(u_i^2|X_i) \neq E(u^2|X_j)$ for some i, j

Basics of HetSc

Unconditional HetSc can be used when variables have identifiable seasonal variability, such as electricity usage

Why HetSc happens

i) Following error learning models: As people learn their errors of behavior become smaller, so σ_i^2 is expected to decrease

ii) Growth of incomes: In a regression of savings on income is likely to find σ_i^2 increasing with income

iii) Data-collecting technique: As techniques improve σ_i^2 is likely to decrease. eg. larger banks with better equipment are likely to commit fewer statement errors

iv) presence of outliers: including or excluding outliers is up to the discretion of the analyst

v) Misspecification of model: the wrong variables are included, or important ones excluded.

vi) Skewness in distr of one or more regressors

vii) Other sources:

1) incorrect data transformation form

2) incorrect functional form (eg. linear vs log)

Pure vs impure HetSc
i) Pure is caused by the error term of the correctly specified equation
ii) Impure is caused by specification error such as an omitted variable

Omitted variables: impure HetSc

Effects:
$y_i = \beta_0 + \beta_1 x + u$
$E(u_i) = 0$
$var(u_i) = \sigma_i^2$
$cov(u_i, u_j) = 0 \quad i \neq j$

Special considerations
Heteroscedasticity is frequent in cross-sectional data

Examples:
i) Operating income versus advertising expenses of 30 firms

How it can be tested:
looking at the plot

Calculating σ^2 under Hct Sc
OLS is unbiased under Hct Sc

$$Var(\hat{\beta}_i | X_i) = \sum Var($$

$$\downarrow$$

$$var(\hat{\beta} | X_i) = \sum (x_i - \bar{x})^2 \sigma_i^2$$

$$\downarrow$$

$$var(\hat{\beta}_i) = \sum (x - \bar{x}$$

$$t = \frac{\text{estimate} - \text{hypothesized value}}{\text{standard error}}$$

Only difference between it and the usual OLS t is how the SE is computed

Consequences of Hct Sc
a) OLS estimators are still linear
b) " " still unbiased
c) " " no longer have min variance
d) Usual OLS formulas to estimate are biased
e) This bias arises from the fact that $\hat{\sigma}^2$ is no longer unbiased for σ^2
f) CIs and hypoth test on t and f distr are unreliable
g) Loss of efficiency

There are several tests for Het Sc
We will do White test and Breusch-Pagan test
Htw
Chapter 4 questions 2 and 3 & 6

2b) i) $H_0: \beta_3 = 0$ $H_1: \beta_3 > 0$

ii) log-level $\%\Delta y = (100\beta_3)\Delta x$

$100 \cdot .00024 \cdot 50$

$= 1.2\%$

iii) $t_{\hat{\beta}_3} = \frac{.00024}{.00054} = 0.44$ $t_c = 1.282$

At the 90% significance level we fail to reject H_0.

iv) No, it fails to be significant and has only a marginal effect on the model

6) i) $H_0: \beta_0 = 0$ $\alpha = 0.05$ $df = 86$

$H_1: \beta_0 \neq 0$ $t_c = 1.988$ $t_{\hat{\beta}_0} = \frac{\hat{\beta}_0 - \beta_0}{Sc(\beta_0)} = \frac{-14.47}{16.67} = -.868$

Fail to reject H_0

$H_0: \beta_1 = 1$ $\alpha = 0.05$ $t_c = 1.988$

$H_1: \beta_1 \neq 1$ $df = 86$ $t_{\hat{\beta}_1} = \frac{.976 - 1}{.049} = -.490$

Fail to reject

Breusch-Pagan tests for condition HetSc
If the χ^2 value is significant with p-value below an appropriate threshold (e.g. $p < 0.05$), H_0: homoskedasticity is rejected, and HetSk is present

$$y_i = \beta_1 + \beta_2 x_{2i} + \dots \beta_k x_{ki} + u_i$$
$$\sigma_i^2 = \delta(\alpha_1 + \alpha_2 z_{2i} + \dots \alpha_m z_{mi})$$

assume

$$\sigma_i^2 = \alpha_1 + \alpha_2 z_{2i} + \dots \alpha_m z_{mi}$$

If $\alpha_2 = \alpha_3 = \dots = \alpha_m = 0$,
this implies $\sigma_i^2 = \alpha$, which is constant and the model is Homosk, so:

H_0: $\alpha_1 = \alpha_3 = \dots = \alpha_m = 0$, $\sigma_i^2 = \alpha_1$

In Stata: rvfplot, yline(0)

reg ...
estat hettest → gives p-value. If < 5%, then is HetSk

White test

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + u_i$$

$$\hat{y} = \hat{\beta}_1 + \hat{\beta} \dots$$

$$\hat{u}^2 = \alpha_1 + \alpha_2 x_{2i} + \alpha_3 x_{3i} + \alpha_4 x_{2i}^2 + \alpha_5 x_{3i}^2 + \alpha_6 x_{2i} x_{3i} + v_i$$

H_0: $\alpha_2 = \alpha_3 = \dots = \alpha_4 = 0$
H_1: $\alpha_2 \neq 0$, $\alpha_3 \neq 0$, ..., $\alpha \neq 0$

Lagrange Multiplier (LM) is R^2 times sample size n

$$LM = nR^2$$

If $nR^2 \geq \chi^2_{df}$

with $df = k - 1$

If $(nR^2 > \chi^2_{df})$ ε

there is heteroskedasticity

3

cross-sectional data : Het/Hk is very frequent

Panel data : combined cross-sectional and time-series data

Remedial measures

if variance is known

Divide both sides by σ_i

$$\frac{Y_i}{\sigma_i} = \beta_1\left(\frac{1}{\sigma_i}\right) + \beta_i\left(\frac{X_i}{\sigma_i}\right) + \frac{u_i}{\sigma_i}$$

let $v_i = \frac{u_i}{\sigma_i}$, the "transformed" error term

Now square v_i

$$v_i^2 = \frac{u_i^2}{\sigma_i^2}$$

$$E(v_i^2) = E\left(\frac{u_i^2}{\sigma_i^2}\right) = \frac{1}{\sigma_i^2} E(u_i^2)$$

$$\rightarrow E(u_i^2) = \sigma_i^2$$

$$E(v_i^2) = \left(\frac{1}{\sigma_i^2}\right) \sigma_i^2 = 1$$

↳ constant, therefore transformed model is homoskedastic

This is called method of weighted Least squares, or WLS

- lsrr package in R provides:
  - Bartlett Test
  - * Breusch-Pagan Test
  - Score test
  - * F test

If σ_i^2 is unknown

Two subcases:

1) Error variance is proportional X_i

Square root transformation

so

$$E(u_i^2) = \sigma^2 X_i$$

consider

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

$$\downarrow \frac{(\,)}{\sqrt{X_i}}$$

$$\frac{Y_i}{\sqrt{X_i}} = \frac{\beta_1}{\sqrt{X_i}} + \beta_2 \cdot \frac{X_i}{\sqrt{X_i}} + \frac{u_i}{\sqrt{X_i}}$$

$$\downarrow$$

$$\frac{Y_i}{\sqrt{X_i}} = \beta_1 \frac{1}{\sqrt{X_i}} + \beta \sqrt{X_i} + \frac{u_i}{\sqrt{X_i}}$$

let $V_i = \frac{u_i}{\sqrt{X_i}}$ $\quad V_i^2 = \frac{u_i^2}{X_i}$

$E(u_i^2) = \sigma^2 X_i$ $\quad E(V_i^2) = \frac{\sigma^2 X_i}{X_i} = \sigma^2$

so it's homoskedastic

Wed 2019-11-13

Situation 2: Error variance is proportional to mean of X_i^2

$$E(u^2) = \sigma^2 X_i^2 = \sigma^2 [E(X_i)]^2$$

Assume the model: mean transformation

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

↓ divide by X_i

$$\frac{Y_i}{X_i} = \frac{\beta_1}{X_i} + \frac{\beta_2 X_i}{X_i} + \frac{u_i}{X_i} = \frac{Y_i}{X_i} = \frac{\beta_1}{X_i} + \beta_2 + \frac{u_i}{X_i}$$

let $v_i = \frac{u_i}{X_i}$

$$\frac{Y_i}{X_i} = \beta_1 \frac{1}{X_i} + \beta_2 + v_i \qquad v_i^2 = \frac{u_i^2}{X_i^2}$$

$$E(u_i^2) = E\left(\frac{u_i^2}{X_i^2}\right) = \frac{1}{X^2} E(u_i^2)$$

$$E(u_i^2) = \frac{1}{X^2} \sigma^2 (X_i^2) = \sigma^2 \leftarrow$$ As a result, transformed model does not suffer from ~~heter~~ heteroskedasticity

Chapter 7 - Dummy variables

Qualitative variables as binary 1 or 0
Dummy variables use in regression to represent subgroups
Used in research design to distinguish treatment groups
Regression models containing only dummy explanatory variables are called analysis of variance (ANOVA)

Dummys are proxies for qualitative ~~thing~~ facts

Dummy variable trap
When dummys are multicollinear.

$Wage = \beta_0 + \delta_0 female + \beta_1 educ + u$

if $\delta_0 > 0$, women make more money
else if $\delta_0 < 0$, women make less money

Intercept for men is β_0
for women is $\beta_0 + \delta_0$

Mon 2019-11-18

Interaction:
interaction variables: formed by multiplication of two variables

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \delta x_1 x_2 + \varepsilon$$

ANOVA: used to analyze if there is a significant difference between groups (categorical variable)
ANOVA can show if survey or experiment results are significant

look at p-value of β_3

$$Y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

ANCOVA - Analysis of covariance

Wed 2018-11-20

Chapter 12 - Autocorrelation / serial correlation
•another violation of OLS
•frequent in time series data, as opposed to cross-sectional

The relationship between a given variable and itself over various time intervals. Often found in repeating patterns where the level of the variable affects its future level (such as stock bonds, GDP)

Occurs when the error term observations in a regression are correlated
Can be a prolem when analyzing historical data

Pure auto.: Model is correctly specified but errors are serial correlation
Most common is 1st order: in which error at time t is related to error @ previous time t-1:

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t \qquad -1 < \rho < 1$$

↑ population serial correlation

Occurs when assumption assuming uncorrelated observations of the error term is violated
No auto correlation means

$$E(u_i, u_j) = 0$$

pure serial correlation is the "default"

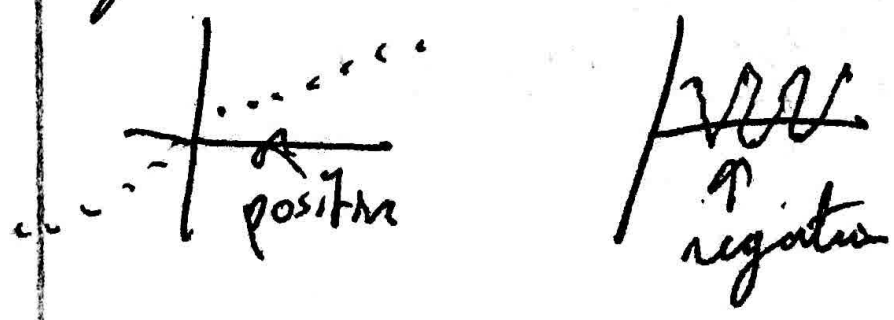
Impure serial correlation
Model is not well specified

Reasons for serial correlation
1) Inertia: (or sluggishness) frequent in time series data.
2) Model specification error.
3) Cobweb phenomenon: economic model that explain why prices might be subject to periodic fluctuations - based on time lag between supply and demand decisions

Data manipulation
In some cases reg. are involving quarterly data are often derived by averaging the data from 3 months

Spatial auto correlation:
occurs when two errors are spatially/geographically related

Negative auto correlation: switching of positive and negative errors:



$\varepsilon_t = \rho \varepsilon_{t-1} + u_t$ ← markov 1st order scheme

ε_t ↑ classical error term

error of eq in question

$\rho = 0 \Rightarrow$ no error term

$\rho \approx 0 \pm 1 \Rightarrow$ strong auto correlation

Most econometricians focus on positive serial correlation

1) Pure serial correlation does not cause bias in the coefficient estimates
2) They are not efficient i.e. not minimum bias
3) Estimated ~~variance~~ variance of OLS estimates are biased

d: Dublin watson statistic
always between 0 and 4
detects autocorrelation at lag 1

i) 2 - no autocorrelation
$0 < d < 2$ - positive serial correlation
$2 < d < 4$ - negative " "

1) Compute OLS
2) " $u_t = Y_t - \hat{\beta}_1 X_1 - \hat{\beta}_2 X_2 - \dots \hat{\beta}_k X_k$

3) $$d = \frac{\sum_{t=2}^{N} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum \hat{u}_t^2}$$

d is founded by 2 distributions: d_u and d_L, given by table

ε pop error term

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t$$

$\hat{\varepsilon}_t = \delta_0 + \delta_1 @ \hat{\varepsilon}_{t-1}$ — sample error term

δ_1 corresponds to ρ

Hence the test is

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

Another way to test for auto correlation

$$DW = \quad \Sigma$$