# Ch 7: Violations of regression assumptions – multicollinearity

# Overview of the lecture

- Not about difficult mathematics
- **It is about**
    - Communication of information
    - Critically evaluating computer output
- **Does your computer output and statistical model really do what you think it does?**

# Multicollinearity

• The name itself is revealing!

**• Multicollinearity=multiple (linear) relationships between the $X$-variables**

• It the $X$ variables are themselves collinear ("inter-related") it is hard to isolate the individual influence on $Y$

# Outline

1. Sources of multicollinearity
2. Theoretical consequences of multicollinearity
3. Practical consequences of multicollinearity
4. Detection of multicollinearity
5. Remedial measures
6. Worked Example

# 1.1 Nature of multicollinearity

- Multicollinearity refers to the situation where there is either an exact linear relationship or an approximate linear relationship amongst the $X$-variables
- Illustrative example data

| $X_2$ | $X_3$ | $X_3^*$ |
|---|---|---|
| 10 | 50 | 52 |
| 15 | 75 | 75 |
| 18 | 90 | 97 |
| 24 | 120 | 129 |
| 30 | 150 | 152 |

# 1.2 The nature of multicollinearity

- There is perfect collinearity between $X_2$ and $X_3$ since $X_3 = 5X_2$
- There is not perfect collinearity between $X_2$ and $X_3$ but the linear correlation between the two variables is very high (0.9959)
- **Both instances would qualify as multicollinearity**

# 1.3 Least Squares Estimation

- There is a rich mathematical theory – although that is a very different lecture!
- Least-squares estimator

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

- Theory implicitly assumes that the matrix $(X^T X)^{-1}$ can be inverted and the effect of individual $X$ variables upon $Y$ can be isolated
- **However this may not be the case**
- Exact linear relationships between the $X$-variables would mean that the matrix inverse $(X^T X)^{-1}$ does not exist
- In practice multicollinearity and approximate linear relationships between the $X$-variables may mean that $(X^T X)^{-1}$ can be calculated but may be numerically unstable
- **Mathematically we may not be able to trust the computer-generated output**

# 1.4 Sources of multicollinearity

• Encountered multicollinearity in my own work as a statistician – not just a purely theoretical problem

- Numerical examples when writing a statistics textbook
- Statistical analysis of survey data – regression problems with large number of redundant $X$ variables

• **Often "know" in advance when you might experience multicollinearity**

# 1.5 Sources of multicollinearity

1. Data collection method employed
2. Constraints on the model or the population
3. Model specification
4. An over-determined model
5. Common trends

# 1.6 Sources of multicollinearity

- **The data collection method employed**
  - Sampling over a limited range of values taken by the regressors in the population
- **Constraints on the model or population**
  - E.g. variables such as income and house size may be interrelated
- **Model Specification**
  - E.g. adding polynomial terms to a model when the range of the $X$-variables is small

# 1.7 Sources of multicollinearity in finance and accounting problems

- **An over-determined model**

  - May be caused by having many explanatory variables compared to the number of observations or by the sheer number of $X$ variables in large problems

  - Multicollinearity means it is difficult to isolate the effects of individual $X$ variables

  - Often occurs in statistical problems associated with the analysis of survey data

- **Common trends**

  - E.g. variables such as consumption, income, wealth, population etc may be correlated due to a dependence upon general economic trends and cycles

  - See e.g. the tutorial exercise

# 2.1 Theoretical consequences of multicollinearity

• Multicollinearity is essentially a sampling phenomenon associated with each particular data set – always present to a greater or lesser degree

• Theoretical effects are in some sense muted – but multicollinearity may remain an important practical problem

**• In the case of (approximate) multicollinearity OLS estimators retain desirable theoretical properties**

- Unbiased. Under repeated sampling the average of the sample values will converge to the true underlying population value
- Best Linear Unbiased Predictor

# 2.2 Theoretical consequences of multicollinearity

- Consider the OLS estimator $\hat{\beta}$
- If you have exact multicollinearity (so that the matrix $(X^TX)$ is not invertible) parameter estimates have infinite variances and covariances
- **In practical situations we only have approximate multicollinearity so in practice we can calculate** $(X^TX)^{-1}$
- Nonetheless under (approximate) multicollinearity parameter estimates may ...
  - Have large estimated standard errors
  - Be highly correlated

# 2.3 The curse of multicollinearity ...

- **Why is multicollinearity bad news?**
  1. **Large estimated standard errors**

     - Lack of precision associated with parameter estimtes

     - Wider confidence intervals

     - Large estimated standard errors may affect hypothesis tests
  2. **Correlated parameter estimates**

     - Another potential source of errors

     - Is suggestive of numerical problems with computational routines

# 2.4 The effect of multicollinearity upon hypothesis tests

- **It becomes harder to reject incorrect hypotheses**
- To test the null hypothesis that, say, $\beta_2 = 0$ we use a $t$-ratio:

$$t := \frac{\hat{\beta}_2}{\text{e.s.e.}(\hat{\beta}_2)} \tag{1}$$

- But multicollinearity increases the e.s.e so that the $t$-ratio reduces in size and its statistical significance becomes reduced
- **The net result is that it becomes harder to reject incorrect hypotheses**

# 3.1 Practical consequences of multicollinearity

- In practical problems look for something that "does not look quite right"
- **In my own experience you often see high $R^2$ values coupled with insignificant $t$-ratios**
- **This is contradictory**

 - The high $R^2$ suggests that the model is good and explains a lot of the variation in $Y$

 - But if individual $t$-ratios are non-significant this suggests that individual $X$-variables do not affect $Y$

# 3.2 Practical consequences of multicollinearity

- **Look for high $R^2$ values coupled with insignificant $t$-ratios but there are several related effects**

1. Multicollinearity results in large estimated standard errors
2. Larger estimated standard errors result in non-significant $t$-ratios
3. Larger estimated standard errors also results in wider confidence intervals since

$$\begin{aligned} 95\ \%\ \text{Confidence Interval} &= \hat{\beta}_2 \pm t_{n-p}(0.025)\text{e.s.e.}(\hat{\beta}_2) \\ p &= \text{No. of estimated parameters including the constant} \\ n-p &= \text{Residual d.f.} \end{aligned}$$

- **The width of the confidence interval increases as the estimated standard error increases**
- **High $R^2$ values coupled with low $t$-values ...**

# 3.3 Practical consequences of multicollinearity

- **Numerical instabilities**
  1. Parameter estimates and their associated estimated standard errors become very sensitive to small changes in the data
  2. Regression coefficients may “take the wrong sign” or otherwise “look strange”
  3. It may be difficult to assess the individual contributions of explanatory variables to the the regression sum of squares or to the $R^2$ statistic
  4. Parameter estimates may be highly correlated

# 3.4 Illustrative example of multicollinearity

- Consider the following illustrative example
- Want to explain expenditure $Y$ in terms of income $X_2$ and wealth $X_3$
- **To detect multicollinearity look out for a high $R^2$ value combined with low $t$-values ...**

# 3.5 Example data set

| Expenditure $Y$ | Income $X_2$ | Wealth $X_3$ |
|---|---|---|
| 70 | 80 | 810 |
| 65 | 100 | 1009 |
| 90 | 120 | 1273 |
| 95 | 140 | 1425 |
| 110 | 160 | 1633 |
| 115 | 180 | 1876 |
| 120 | 200 | 2052 |
| 140 | 220 | 2201 |
| 155 | 240 | 2435 |
| 150 | 260 | 2686 |

# 3.6 Entering the data into R

```
y<-c(70, 65, 90, 95, 110, 115, 120, 140, 155, 150)
x2<-c(80, 100, 120, 140, 160, 180, 200, 220, 240,
260)
x3<-c(810, 1009, 1273, 1425, 1633, 1876, 2052, 2201,
2435, 2686)
```

# 3.7 Analysis of regression example

- **Three basic statistics**
  - $R^2$, $F$-statistic, $t$-statistics
- **Using R**
  1. $R^2 = 0.9635$.
     The model explains a substantial amount of the variation (96.35% of the variation in the data)
  2. $F = 92.40196$
     From tables $F_{2,7} = 4.74$ so some evidence $p < 0.05$ that at least one of income and wealth affect expenditure
  3. $t$-statistics
     From tables $t_7(0.025) = 2.365$. The $t$-statistics are 1.144 for income and -0.526 for wealth
     Neither income nor wealth are individually statistically significant ($p > 0.05$)

# 3.8 Analysis of the regression example – continued

- We have a few of the tell-tale signs of multicollinearity
  1. High $R^2$ values low $t$-values
  2. The wealth variable has the "wrong sign" – it is likely that expenditure will increase as wealth increases
  3. The variables $X_2$ and $X_3$ are very highly correlated
- **The variables income and wealth are so highly correlated that it is impossible to isolate the individual impact of either income or wealth upon consumption**
- The example motivates the questions of how to detect multicollinearity in practice

# 4.1 Detection of multicollinearity

- Kmenta's warning

  - **The distinction is not between the presence and the absence of multicollinearity but between its various degrees**

  - Multicollinearity is a function of the sample and not of the population

  - **Multicollinearity is always liable to be present in a given dataset to a greater or lesser degree**

# 4.2 Detection of multicollinearity

1. High $R^2$ low $t$-values
2. High pairwise correlations between explanatory variables

- If the correlation coefficient of regressors is high, say greater than 0.8, then this indicates that multicollinearity might be a problem

- In the lecture example the correlation between income and wealth is 0.99896. This is very high suggesting that we might have a problem with multicollinearity

- In models involving two or more explanatory variables pairwise correlation will not provide a fool-proof guide to the presence of multicollinearity

# 4.3 Detection of multicollinearity

3. Examination of partial correlations

- Partial correlations may provide a better measure of multicollinearity than the simlpe pairwise correlations

- E.g. with three variables $X_2$, $X_3$, $X_4$

- The partial correlation coefficient $r_{23,4}$ would measure the correlation between $X_2$ and $X_3$ independent of $X_4$

# 4.4 Detection of multicollinearity – subsidiary or auxilliary regression

4. Subsidiary or Auxilliary regression

• Multicollinearity arises because one or more of the $X$ variables are linearly related or approximately linearly related to some combination of the other explanatory variables

• Can perform a subsidiary or auxilliary regression of $X_2$, $X_3$ etc against the other explanatory variables

• These secondary regressions are then subsidiary or auxilliary to the main regression for $Y$

# 4.5 Detection of multicollinearity – Klein's rule of thumb

- Klein's rule of thumb is that multicollinearity may only be a troublesome problem if the $R^2$ obtained from an auxilliary regression is greater than the overall $R^2$ obtained by regressing $Y$ against all the $X$ variables
- For the lecture example using R
  - Regressing $Y$ against $X_2$ and $X_3$ gives $R^2 = 0.963304$
  - Regressing $X_2$ against $X_3$ gives $R^2 = 0.997926$
- **Thus for the lecture example Klein's rule of thumb suggests that multicollinearity will be a serious problem**

# 5.1 Remedial measures – do nothing

- One possibility is to simply do nothing
- The renowned economist Blanchard describes multicollinearity as being “God’s will” and as not being a problem with Ordinary Least Squares or any other statistical technique
- When viewed from this perspective multicollinearity is essentially a data-deficiency problem and sometimes we have no control over the data we have available for analysis
- **Note that there is an important distinction here in terms of the quality of the available data in quantitative social science (finance and economics) and scientific experiments in physics and chemistry**

# 5.2 Remedial measures – use prior information about some parameters

- Textbooks often give some fairly extreme examples to illustrate this
- Incorporating prior information about parameters sensibly would require more advanced Bayesian statistical methods and lies far outside the scope of this course

# 5.3 Remedial measures – dropping variables

• The simplest approach to multicollinearity is to drop one or more of the collinear variables

- E.g. in the lecture example we could exclude wealth from the model

• Some economists might express concerns over specification bias if variables identified by economic theory are not included in the model. Whilst this may be a valid concern it does depend to some extent how much trust you want to place upon economic theory

• **Sometimes useful to use stepwise regression/similar techniques related to computer science to choose a model – see the worked example**

- May provide an “objective” approach to messy problems

# 5.4 Remedial measures – transformation of variables

• The problem of multicollinearity may be reduced by transforming variables

• This may be possible in various different ways

• E.g. if you have time series data one might consider forming a new model by taking first differences.

• This approach may reduce the problem of multicollinearity – see Gujarati and Porter Ch. 10.

# 5.5 Remedial measures – acquiring new data

- Acquire additional data or a new sample?
- **Multicollinearity is a sample feature**
- It is possible that in another sample involving the same variables the multicollinearity will not be as serious a problem as in the first sample
- Sometimes just acquiring more data – either increasing the sample size or including additional variables – can reduce the severity of the multicollinearity problem

# 5.6 Remedial measures – rethinking the model

- Sometimes a model chosen for empirical analysis is not carefully thought out
  - Some important variables may be omitted
  - The functional form of the model may have been incorrectly chosen
- It is also possible that using more advanced statistical techniques may be required. Possible examples include
  - Factor Analysis
  - Principal Components Analysis
  - Ridge Regression

# 5.7 Micronumerosity

- Multicollinearity has arguably received excessive attention in the literature
- Micronumerosity refers to the smallness of the sample size
- Micronumerosity is thought by some to be equally as important as multicollinearity
- **There are no “right” answers and no substitute for common sense and critical thinking**

# 6.1 Example: Longley data

- Want to illuatrate stepwise regression approaches
- Stepwise regression approaches represent an “objective” way of approaching messy problems
- Famous Longley dataset. Want to explain the number employed $Y$ in terms of

1. $X_2$ GNP
2. $X_3$ Number of unemployed people
3. $X_4$ Number of people in the armed services
4. $X_5$ Noninstitutionalised population over the age of 14
5. $X_6$ Time in years

- Fitting the full model gives high $R^2$ values and low $t$-values and suggests that we might have a problem with multicollinearity

# 6.2 R commands for reading in the data

- Data in the file `longley.txt`

```
longley<-read.table(‘‘E:longley.txt")
x2<-longley[,1]
x3<-longley[,2]
x4<-longley[,3]
x5<-longley[,4]
x6<-longley[,5]
y<-longley[,6]
```

# 6.3 Stepwise methods – three basic approaches

1. Forward Selection

   - Start with the basic model $Y_i = \beta_1 + u_i$ and add successive terms until no more terms are statistically significant

2. Backward Selection

   - Start with the full model $Y_i = \beta_1 + \beta_2 X_{2,i} + \ldots + \beta_p X_{p,i} + u_i$

   - Delete terms until all the variables remaining in the model are statistically significant

3. Stepwise Selection

   - Start with the basic model $Y_i = \beta_1 + u_i$ and add successive terms until no more terms are statistically significant

   - However, each time a variable enters the model a backward regression deletion step is performed to check that all variables included in the model remain statistically significant throughout

# 6.4 Stepwise regression in R

- Begin by fitting the full regression model

```
a.lm<-lm(y~x2+x3+x4+x5+x6)
```

- Then need to fit the null model with just the intercept term

```
null.lm<-lm(y~1)
```

- There are two basic differences depending on whether you are doing

1. Forward selection or stepwise regression
   - Start with the null model
2. Backward selection - Start with the full model

# 6.5 Stepwise regression in R

1. Stepwise selection
   `step(null.lm, direction = "both", scope = formula(a.lm))`
2. Forward selection
   `step(null.lm, direction = "forward", scope = formula(a.lm))`
3. Backward selection
   `step(a.lm, direction = ‘‘backward")`

• In this simple example all three approaches agree and select a model with the variables $X_2$, $X_3$, $X_4$ and $X_6$ (no $X_5$ term)

# 6.6 Final solution and interpretation

- Multicollinearity and inter-relationships between the $X$-variables means that there is some redundancy and the $X_5$ variable is not needed in the model
- This means that the number employed just depends on

1. $X_2$ GNP
2. $X_3$ Number of unemployed people
3. $X_4$ Number of people in the armed services
4. $X_6$ Time in years

- **Re-fit the model with just these variables and interpret the results**

```
step.lm<-lm(y~x2+x3+x4+x6)
summary(step.lm)
```

# 6.7 R results obtained

```
 Estimate Std.  Error t value Pr(>|t|)
(Intercept) -3.599e+03 7.406e+02 -4.859 0.000503 ***
x2 -4.019e-02 1.647e-02 -2.440 0.032833 *
x3 -2.088e-02 2.900e-03 -7.202 1.75e-05 ***
x4 -1.015e-02 1.837e-03 -5.522 0.000180 ***
x6 1.887e+00 3.828e-01 4.931 0.000449 ***
```

# 6.8 Final interpretation

**• The coefficient of $X2$ is negative and statistically significant. As GNP increases the number employed decreases**

**• The coefficient of $X3$ is negative and statistically significant. As the number of unemployed increases the number employed decreases**

**• The coefficient of $X4$ is negative and statistically significant. As the number of people in the armed services increases the number employed decreases**

**• The coefficient of $X6$ is positive. This suggests that the number employed is generally increasing over time.**

# 6.9 A final last word

- Previous slide details all the interpretation elements that would naturally be expected of you as part of this module
- However, there is a sense that there may be a little extra needed here for full interpretation of this example
- Since in this example the interpretations for $X_2$ and $X_4$ appear contradictory the suggestion here is that the effect of these variables is dwarfed by the general increase in employment over time. One might ask how likely this trend is to continue into the future?