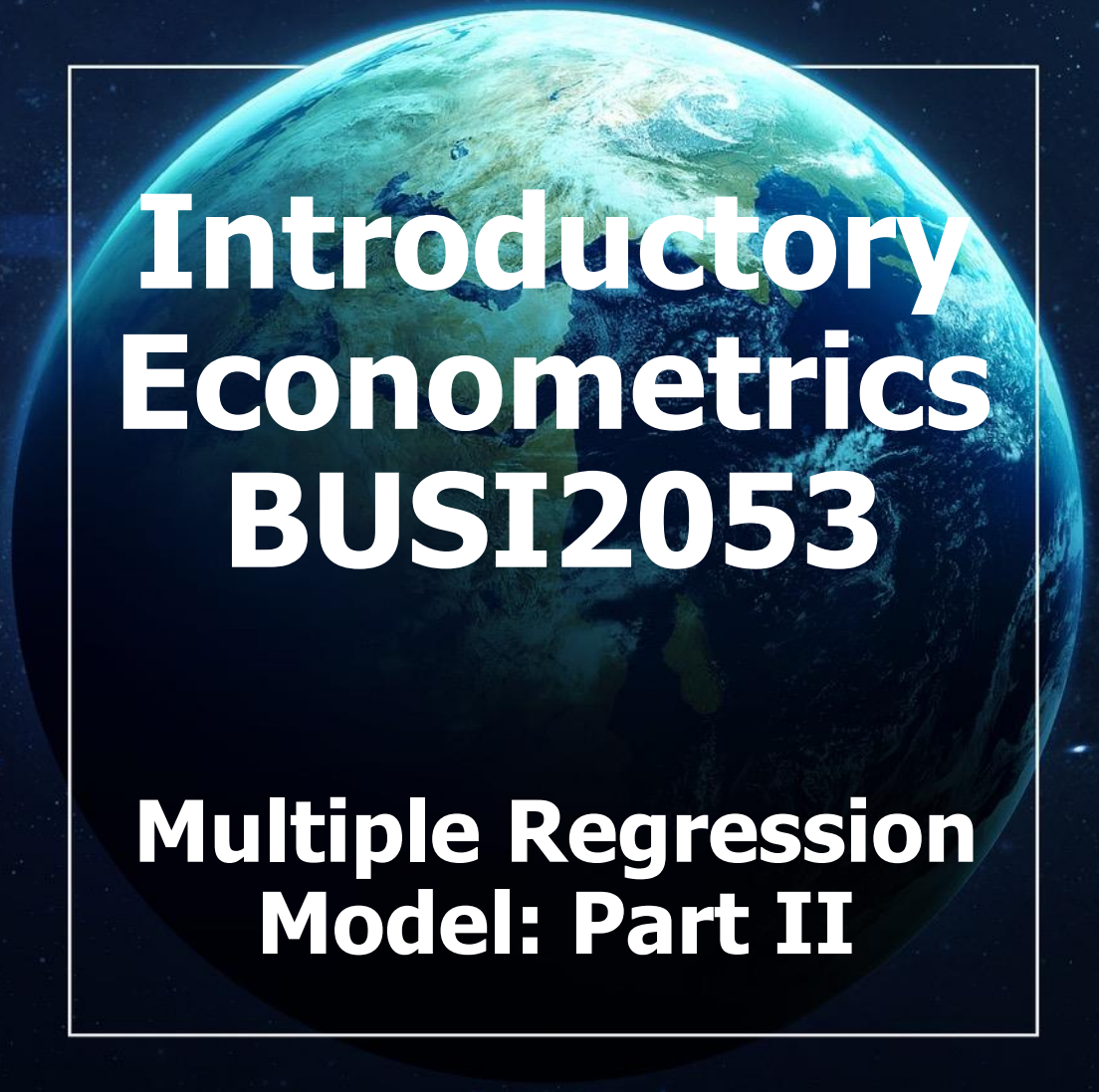
University of Nottingham
UK | CHINA | MALAYSIA



Introductory Econometrics BUSI2053

Multiple Regression Model: Part II

Lecture Outline

- Modelling specification and related issues
- F-test for Joint Significance
- F-test for Linear Restrictions
- Multicollinearity
- Reliability and Validity in Regression-Based Analytics

- **Suggested Reading:**

Chapter 5 & 6.1, 6.4: Hill, R.C., Griffiths W.E. and Lim, G.C. Principles of Econometrics, fourth edition, Wiley, 2012 (pp. 167-229; 240-242)

Chapter 7 & 8: Gujarati, D.N. and Porter D.C. Basic econometrics, 5th ed., McGraw-Hill, 2009 (pp. 188-252)

Chapter 3: Dougherty, Christopher. Introduction to econometrics. 4th ed. Oxford University Press, 2011 (pp.151-199)

Chapter 11, 14: Westhoff, F. An introduction to econometrics: a self-contained approach, MIT Press, 2013

Model Specification

- In any econometric investigation, choice of the model is one of the first steps
  - What are the important considerations when choosing a model?
  - What are the consequences of choosing the wrong model?
  - Are there ways of assessing whether a model is adequate?

- We may never know the true model. The most we can do is to find a "good" or "correct" model which does not commit the following "specification errors".
  1. Omission of a relevant variable(s)
  2. Inclusion of an unnecessary variable(s)
  3. Adopting an incorrect functional form
  4. Errors of measurement

Omitting a Relevant Variable

Assume the true model is: $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + e$

But for some reasons, the model estimated was: $\hat{y} = b_1 + b_2 x_2$

The consequences of omitting X_3 are as follows:

1. If the omitted variable x_3 is correlated with the included variable(s), x_2 in this case, then b_1 and b_2 are biased and inconsistent (i.e., $E(b_1) \neq \beta_1$ and $E(b_2) \neq \beta_2$)and the bias does not disappear no matter how large the sample size.
2. If x_3 and x_2 are uncorrelated, b_2 is unbiased, but b_1 is still biased.
3. The estimated standard errors are also inflated.
4. The estimated standard error of b_2 is a biased estimator of the variance of the true estimator β_2 .
5. As a consequence, the usual confidence intervals and hypothesis tests are likely to give misleading conclusions about the statistical significance of the estimated parameters. (why?).

Inclusion of an Irrelevant Variable

Assume the true model is: $y = \beta_1 + \beta_2 x_2 + e$

But for some reasons, the model estimated was: $\hat{y} = b_1 + b_2 x_2 + b_3 x_3$

The consequences of including X_3 are as follows:

1. The OLS estimators of the parameters are unbiased and consistent (i.e., $E(b_1) = \beta_1$ and $E(b_2) = \beta_2$)
2. The estimated standard errors are also correct.
3. The usual confidence intervals and hypothesis testing procedures remain valid.
4. The estimated *coefficients* will be inefficient (i.e., variances will be larger) compared to those of the estimated coefficients in the true model. Thus, probability inferences are less precise.

N.B. Don't be fooled into thinking more is better!

F-Test for the Joint Significance of variables

- A null hypothesis with multiple conjectures, expressed with more than one equal sign, is called a **joint hypothesis**
- Consider the following model that estimates icecream consumption:

$$cons = \beta_1 + \beta_2 price + \beta_3 income + \beta_4 temperature + e$$

Test whether or price and income have a joint effect on ice cream consumption

$H_0: \beta_2 = \beta_3 = 0$ (Price and Income will have no effect on sales)
$H_1: \beta_2 \neq 0$ or $\beta_3 \neq 0$ or both are nonzero (Price and Income have joint effect on sales)

Quiz: What is the difference between the test for overall significance and the test for joint significance?

F-Test for the Joint Significance of variables

Unrestricted and Restricted Models

- Relative to the null hypothesis $H_0 : \beta_2 = \beta_3 = 0$:
- **The unrestricted model** is:

$$cons = \beta_1 + \beta_2 price + \beta_3 income + \beta_4 temperature + e$$

- The restrictions in the null hypothesis have not been imposed on the model

- It contrasts with **the restricted model**, which is obtained by assuming the parameter restrictions in H_0 are true is:

$$cons = \beta_1 + \beta_4 temperature + e$$

F-Test for the Joint Significance of variables

Unrestricted and Restricted Models

- The F-test for the hypothesis $H_0 : \beta_2 = \beta_3 = 0$: is based on a comparison of the sum of squared errors or residuals (SSE) from the unrestricted model and the restricted model

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(n-k)}$$

where SSE_U = SSE (*Sum squared errors or residuals*) of the unrestricted model
SSE_R, = SSE (*Sum squared errors or residuals*) of the restricted model

J = the number of restrictions,

n = the number of observations and

k = the number of coefficients in the unrestricted model

F-Test for the Joint Significance of variables

$H_0: \beta_2 = \beta_3 = 0$
$H_1: \beta_2 \neq 0$ or $\beta_3 \neq 0$ or both are nonzero

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(n-k)}$$

F-test assess whether the reduction in the sum of squared errors is sufficiently large to be significant. (whether the addition of the variable to the model increases SSR "significantly" in relation to the SSE)

If adding the extra variables has little effect on reduction in the sum of squared errors, there is support for a null hypothesis that excludes them.

On the other hand, if adding the variables leads to a big reduction in the sum of squared errors, there is evidence against the null hypothesis.

F-Test for the Joint Significance of variables

Unrestricted model

Dataset: icecream (Verbeek)

```
Model 1: OLS, using observations 1-30
Dependent variable: cons

             coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const       0.1973       0.2702       0.7302   0.4718
  price      -1.044        0.8344      -1.252    0.2218
  income      0.003308     0.001171     2.824    0.0090   ***
  temp        0.003458     0.0004455    7.762    3.10e-08 ***

Mean dependent var   0.359433   S.D. dependent var   0.065791
Sum squared resid    0.035273   S.E. of regression   0.036833
R-squared            0.718994   Adjusted R-squared   0.686570
F(3, 26)             22.17489   P-value(F)           2.45e-07
```

Restricted model

```
Model 2: OLS, using observations 1-30
Dependent variable: cons

             coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const       0.206862     0.0247002     8.375    4.13e-09 ***
  temp        0.00310736   0.000477885   6.502    4.79e-07 ***

Mean dependent var   0.359433   S.D. dependent var   0.065791
Sum squared resid    0.050009   S.E. of regression   0.042262
R-squared            0.601593   Adjusted R-squared   0.587365
F(1, 28)             42.27997   P-value(F)           4.79e-07
```

$$H_0: \beta_2 = \beta_3 = 0$$

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(n-k)}$$

$$F = \frac{(0.050009 - 0.035273)/2}{0.035273/(30-4)} = 5.43$$

Since F_{test}=5.43>$F_{0.05,\,2,\,26}$=3.37, reject the null hypothesis that both $\beta_3 = 0$ and $\beta_4 = 0$, and conclude that at least one of them is not zero

Price and income do have a joint significant effect upon ice cream consumption

F Distribution: Critical Values of *F* (5% significance level)

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91	245.36	246.46	247.32	248.01
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.42	19.43	19.44	19.45
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.71	8.69	8.67	8.66
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.87	5.84	5.82	5.80
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.64	4.60	4.58	4.56
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.96	3.92	3.90	3.87
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.53	3.49	3.47	3.44
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.24	3.20	3.17	3.15
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.03	2.99	2.96	2.94
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.86	2.83	2.80	2.77
26	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.09	2.05	2.02	1.99
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.08	2.04	2.00	1.97
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.06	2.02	1.99	1.96
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.05	2.01	1.97	1.94
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.04	1.99	1.96	1.93
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11	2.04	1.99	1.94	1.91	1.88
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.95	1.90	1.87	1.84
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.95	1.89	1.85	1.81	1.78
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.86	1.82	1.78	1.75
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97	1.89	1.84	1.79	1.75	1.72

F-Test for the Linear restriction of variables

- We hypothesise that children educational attainment might be related to cognitive ability and family background (as represented by the mother's and father's educational attainment).
- Use EAFE01.gdt from Doherty:
- *S*=Year of schooling, *AVABC*=Cognitive ability; *SM*=Mother's education; *SF*=Father's education

```
Model 1: OLS, using observations 1-540
Dependent variable: S

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       4.37503      0.567874      7.704    6.41e-014 ***
  ASVABC      0.117100     0.0107325    10.91     3.62e-025 ***
  SM          0.122768     0.0447583     2.743    0.0063    ***
  SF          0.152430     0.0340015     4.483    9.00e-06  ***

Mean dependent var   13.67222   S.D. dependent var   2.555863
Sum squared resid    2261.762   S.E. of regression   2.054192
R-squared            0.357633   Adjusted R-squared   0.354038
F(3, 536)            99.47147   P-value(F)           3.40e-51
```

F-Test for the Linear restriction of variables

It was suggested that the impact of parental education might be the same for both parents, that is, that β_3 and β_4 might be equal.

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + \beta_4 SF + e$$

$$H_0: \beta_4 = \beta_3;\ H_1: \beta_4 \neq \beta_3$$

If $\beta_3 = \beta_4$; we can specify the model as:

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + \beta_3 SF + e$$
$$S = \beta_1 + \beta_2 ASVABC + \beta_3 (SM + SF) + e$$
$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SP + e$$ (where parental education SP=SF+SM)

Here, we restricted the model assuming impact of parental education might be the same for both parents. We can use a F-test to check whether our hypothesis is correct or not.

F-Test for the Linear restriction of variables

Unrestricted regression

	coefficient	std. error	t-ratio	p-value	
const	4.37503	0.567874	7.704	6.41e-014	***
ASVABC	0.117100	0.0107325	10.91	3.62e-025	***
SM	0.122768	0.0447583	2.743	0.0063	***
SF	0.152430	0.0340015	4.483	9.00e-06	***

Mean dependent var	13.67222	S.D. dependent var	2.555863
Sum squared resid	2261.762	S.E. of regression	2.054192
R-squared	0.357633	Adjusted R-squared	0.354038

Restricted regression:

	coefficient	std. error	t-ratio	p-value	
const	4.31408	0.549510	7.851	2.26e-014	***
ASVABC	0.117120	0.0107243	10.92	3.27e-025	***
SP	0.140244	0.0188039	7.458	3.55e-013	***

Mean dependent var	13.67222	S.D. dependent var	2.555863
Sum squared resid	2262.544	S.E. of regression	2.052633
R-squared	0.357411	Adjusted R-squared	0.355018

$$H_0: \beta_4 = \beta_3; H_1: \beta_4 \neq \beta_3$$

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(n-k)}$$

$$= \frac{(2262.544 - 2261.762)/1}{2261.762/536} = 0.1853$$

Since the F-statistic=0.1853 < $F_{critical,5\%,1,536} = 3.86$, do not reject H_0.

Conclusion: There is no significance different between the impacts of father's education and mother's education on child's education.

F Distribution: Critical Values of *F* (5% significance level)

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20
1	61.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91	245.36	246.46	247.32	248.01
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.42	19.43	19.44	19.45
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.71	8.69	8.67	8.66
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.87	5.84	5.82	5.80
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.64	4.60	4.58	4.56
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.96	3.92	3.90	3.87
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.53	3.49	3.47	3.44
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.24	3.20	3.17	3.15
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94	1.86	1.80	1.76	1.72	1.69
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.85	1.79	1.75	1.71	1.68
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.78	1.73	1.69	1.66
150	3.90	3.06	2.66	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.82	1.76	1.71	1.67	1.64
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.80	1.74	1.69	1.66	1.62
250	3.88	3.03	2.64	2.41	2.25	2.13	2.05	1.98	1.92	1.87	1.79	1.73	1.68	1.65	1.61
300	3.87	3.03	2.63	2.40	2.24	2.13	2.04	1.97	1.91	1.86	1.78	1.72	1.68	1.64	1.61
400	3.86	3.02	2.63	2.39	2.24	2.12	2.03	1.96	1.90	1.85	1.78	1.72	1.67	1.63	1.60
500	3.86	3.01	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.77	1.71	1.66	1.62	1.59

MR Assumption (Multicollinearity)

MR assumption 5: No exact linear relationship between the explanatory variables . Perfect multicollinearity occurs when there is an exact linear relationship between 2 or more of the explanatory variables in a regression model

Consider the model: $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + e$

The variance of the least squares estimator for b_2 is:

$$\text{var}(b_2) = \frac{\sigma^2}{(1 - r_{23}^2)\sum_{i=1}^{n}(x_2 - \bar{x}_2)^2}$$

where r_{23}^2 is correlation coefficient between x_2 and x_3

Quiz: What will happen to var(b_2) if there is a perfect correlation between x_2 and x_3, i.e., $r_{23}^2 = 1$?

Possible Sources of Multicollinearity

Multicollinearity: Not perfect but high degree of correlations among independent variables.

1. The data collection method employed. For example, sampling over a limited range of values taken by the regressors in the population.

2. Constraints on the model or the population being sampled. For example, electricity consumption (Y) on income (X_1) and house size (X_2) [Why?]

3. Model specification. For example, adding polynomial terms especially when the range of the explanatory variable is small.

4. Overdetermined model i.e., the number of explanatory variables exceeds the number of observations.

Consequences of Multicollinearity

- Under perfect multicollinearity, it is impossible to calculate OLS estimates of parameters.
- High degree of (but not perfect) multicollinearity:
  1. The regression estimates are unbiased and consistent, and their standard errors are correctly estimated (BLUE), although these tend to be large.
  2. It is likely that the usual *t*-tests are not reliable
  3. Estimators may be very sensitive to the addition or deletion of a few observations
  4. There are clearly problems of interpretation of the contribution of individual explanatory variable.

Detecting Multicollinearity

- Acknowledge that some multicollinearity exists in every regression equation.
- To determine how much multicollinearity exists:
  - High R^2 but few significant t-ratios
  - High Pair-wise correlation coefficients among the explanatory variables
  - High variance inflation factor (VIF)

$$VIF = \frac{1}{(1 - R_i^2)}$$

where $R_i{}^2$= the unadjusted R^2 of the auxiliary regression of x_i on other explanatory variables.

Detecting Multicollinearity (example)

$q = \beta_1 + \beta_2 k + \beta_3 l + e$

q = output, k = capital, l = labour

Data from POE (cobb.gdt)

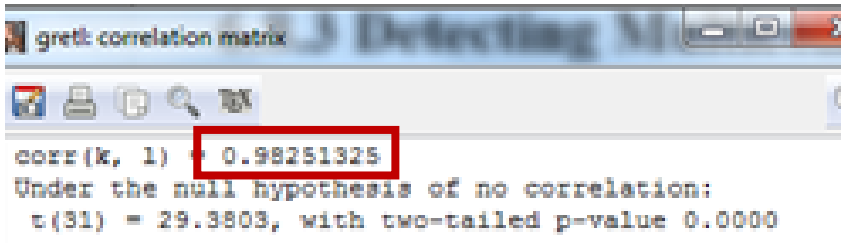
```
Model 1: OLS, using observations 1-33
Dependent variable: q

             coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const       -1.36303      3.54707     -0.3843   0.7035
  k            0.623191     0.741163     0.8408   0.4071
  l            0.484642     0.785943     0.6166   0.5421

Mean dependent var   21.43333   S.D. dependent var   7.628797
Sum squared resid    615.6514   S.E. of regression   4.530090
R-squared            0.669423   Adjusted R-squared   0.647384
F(2, 30)             30.37519   P-value(F)           6.15e-08
Log-likelihood      -95.10683   Akaike criterion     196.2137
Schwarz criterion    200.7032   Hannan-Quinn         197.7242
```

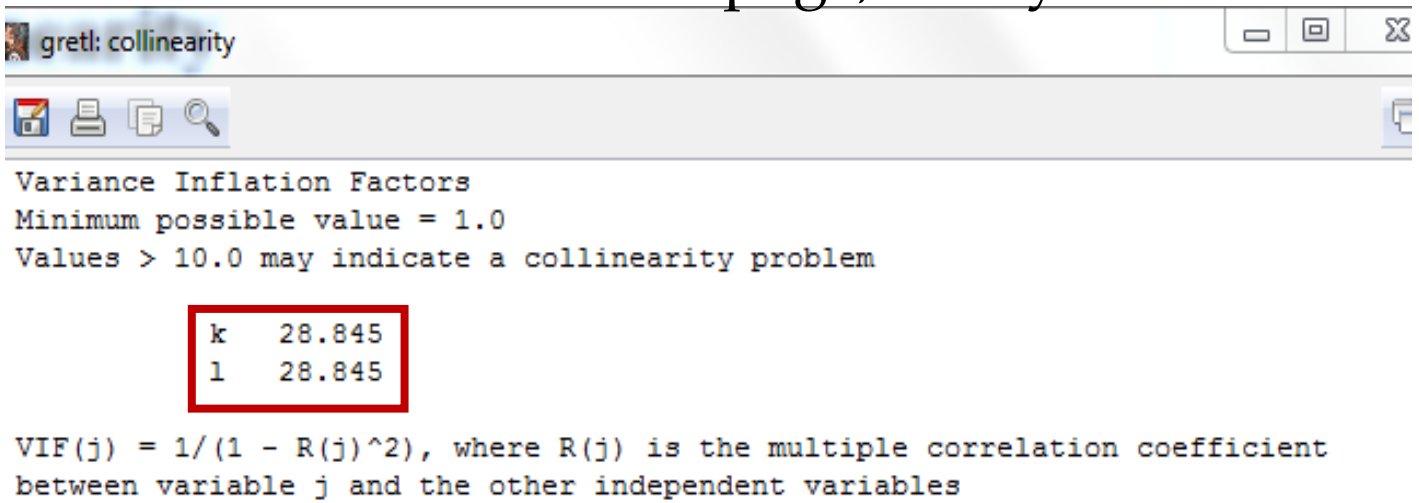
Despite a large F value and R-squared (indicating that the 2 explanatory variables jointly contribute significantly to the model), individually the explanatory variables are not significant. Severe multicollinearity causes such condition.

Detecting Multicollinearity (example)



The variable *k* and *l* are strongly correlated.

From the menu of model page, Analysis>>Collinearity



VIF >10, there is strong evidence of severe multicollinearity amongst the independent variables.

Remedies for Multicollinearity

- Do nothing (OLS estimators are still BLUE)
- Drop a redundant variable(s) (but be mindful of model specification error if you drop a relevant variable)
- Increase the sample size
- Transform the variables

Important Notes

Reliability and Validity in Regression-Based Analytics

- Data quality: Ensure completeness, accuracy, and cleanness of data
- Model is correctly specified with reasonable explanatory power (goodness of fit)
- Fulfillment of core assumptions so that Least Squares Estimators are BLUE (Best Linear Unbiased Estimators) and they are statistically significant
- Additional Considerations for Prescriptive Analytics: Causality (establishing cause-effect relationships) and actionable insights for decision-making

Summary

After this lecture, you should be able to:

- Explain the issues in model specification and identify the consequences of a mis-specified model
- Conduct F-tests for joint significance and linear restriction
- Explain the sources, consequences and remedies for multicollinearity problem

University of Nottingham
UK | CHINA | MALAYSIA