University of Nottingham
UK | CHINA | MALAYSIA

Introductory Econometrics BUSI2053

Multiple Regression Model: Part I

Lecture Outline

- Estimating the Parameters of the Multiple Regression Model
- Interpretation and Prediction
- Sampling Properties of the Least Squares Estimators
- Measuring goodness of fit in multiple regression
- F-test for overall significance

- **Suggested Reading:**

Chapter 5 & 6: Hill, R.C., Griffiths W.E. and Lim, G.C. Principles of Econometrics, fourth edition, Wiley, 2012 (pp. 167-229)

Chapter 10, 11: Westhoff (2013) An introduction to econometrics: a self-contained approach, MIT Press, 2013

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)

Why Multiple Regression?

- In practice, simple univariate regression is rather limited.
- It would be unrealistic to assume that only one independent (explanatory) variable affects the dependent variable in most situations.
    - For example, although the price of iPhone is a major determinant of its demand, there are other factors (such as, consumer's income, the price of other smart phones, preference, etc.) influencing the demand.
- In such situation, we can use **multiple regression** which allows several explanatory variables.
- An explanatory variable's coefficient estimates the change in the dependent variable resulting from a change in that particular explanatory variable while all other explanatory variables remain constant.

Economic and Econometric Models

- Let's set up an economic model in which sales revenue is hypothesised to depend on two explanatory variables, price and advertising expenditure

- The economic model is:

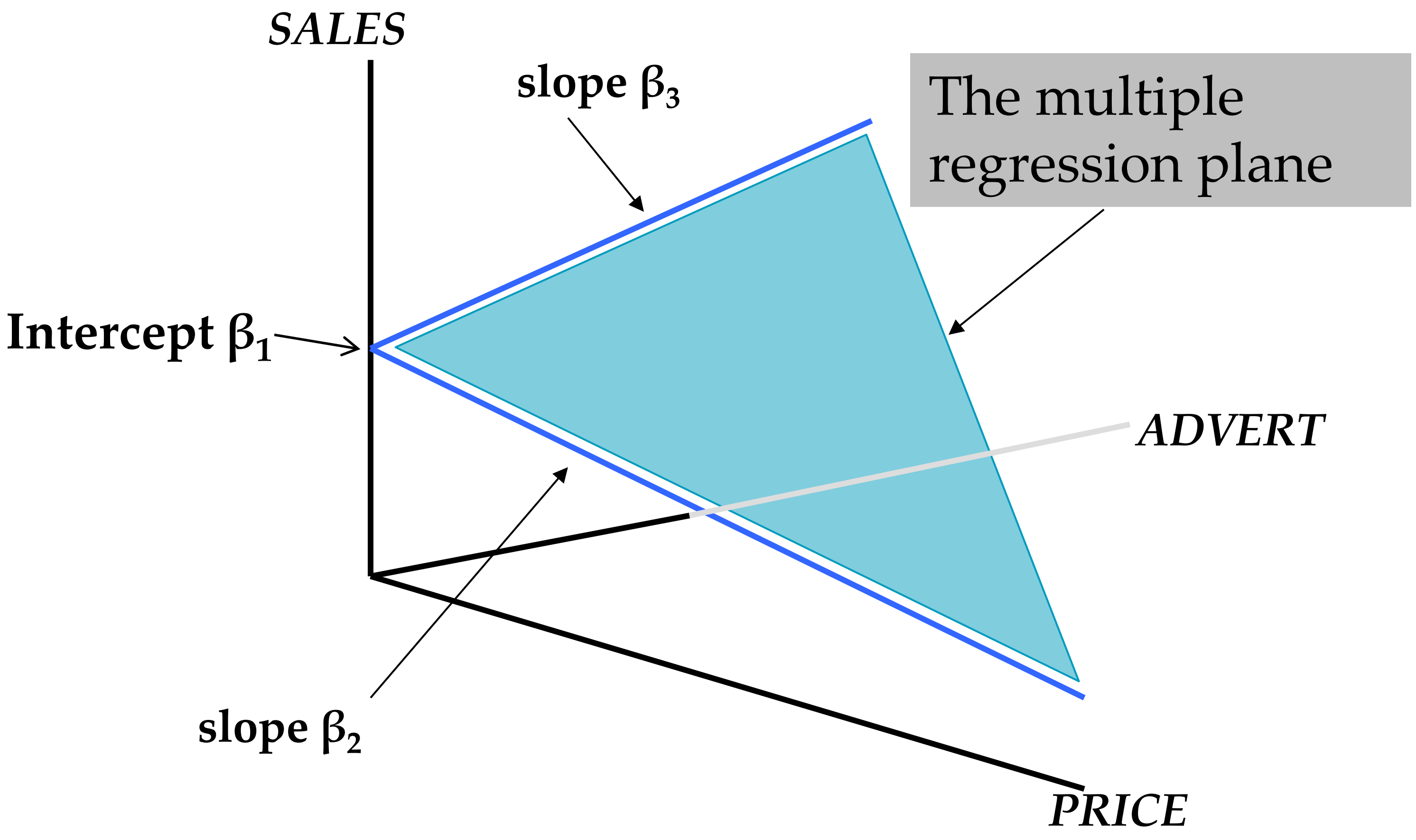
$$SALES = f(PRICE, ADVERT)$$

- Multiple regression model is:

$$E(SALES) = \beta_1 + \beta_2 PRICE + \beta_3\ ADVERT$$

The Econometric Model

$$E(SALES) = \beta_1 + \beta_2 PRICE + \beta_3\ ADVERT$$



The Econometric Model

$$SALES = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT + e$$

- β_2 is the change in monthly sales (*SALES*) when the price index (*PRICE*) is increased by one unit, and advertising expenditure (*ADVERT*) is held constant
- Similarly, β_3 is the change in monthly sales (*SALES*) when the advertising expenditure (*ADVERT*) is increased by one unit, and the price index (*PRICE*) is held constant

$$\beta_2 = \frac{\Delta SALES}{\Delta PRICE (ADVERT \text{ held constant})} = \frac{\partial SALES}{\partial PRICE}$$

$$\beta_3 = \frac{\Delta SALES}{\Delta ADVERT (PRICE \text{ held constant})} = \frac{\partial SALES}{\partial ADVERT}$$

A General Econometric Model

- In a general multiple regression model, a dependent variable y is related to a number of explanatory variables x_2, x_3, …, x_k through a linear equation that can be written:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + e$$

- The equation of or sales revenue can be viewed as

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

y = *SALES*, x_2 = *PRICE* and x_3 = *ADVERT,*

k (number of parameters) = 3

Assumptions of MR & Properties of Estimators

MR1. $y_i = \beta_1 + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + e_i, \qquad i = 1, \ldots, n$

MR2. $E(y_i) = \beta_1 + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} \Leftrightarrow E(e_i) = 0$

MR3. $\text{var}(y_i) = \text{var}(e_i) = \sigma^2$

MR4. $\text{cov}(y_i, y_j) = \text{cov}(e_i, e_j) = 0$

MR5. The values of each x_{ik} are not random and are **not exact linear functions of the other explanatory variables**

MR6. $y_i \sim N[(\beta_1 + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}), \sigma^2] \Leftrightarrow e_i \sim N(0, \sigma^2)$

Gauss-Markov Theorem: For the multiple regression model, if assumptions MR1–MR5 hold, then the least squares estimators are the best linear unbiased estimators (***BLUE***) of the parameters.

Estimating the parameters

Least Squares Estimation Procedure

- Sample regression model is:

$$y_i = b_1 + b_2 x_{2i} + b_3 x_{3i} + \hat{e}_i$$

- Mathematically we minimize the sum of squares of residuals to obtain estimators b_1, b_2, and b_3,

$$Minimise \sum \hat{e}_i^2 = \sum (y_i - b_1 - b_2 x_{2i} - b_3 x_{3i})^2$$

Estimating the parameters

Estimation with gretl

City	*SALES* $1,000 units	*PRICE* $1 units	*ADVERT* $1,000 units
1	73.2	5.69	1.3
2	71.8	6.49	2.9
3	62.4	5.63	0.8
4	67.4	6.22	0.7
5	89.3	5.02	1.5
.	.	.	.
.	.	.	.
.	.	.	.
73	75.4	5.71	0.7
74	81.3	5.45	2.0
75	75.0	6.05	2.2
Summary statistics			
Sample mean	77.37	5.69	1.84
Median	76.50	5.69	1.80
Maximum	91.20	6.49	3.10
Minimum	62.40	4.83	0.50
Std. Dev.	6.49	0.52	0.83

Monthly Sales, Price, and Advertising in Big Andy's Burger Barn (andy.gdt in gretl)

Estimating the parameters (GRETL)

```
Model 1: OLS, using observations 1-75
Dependent variable: sales

             coefficient   std. error   t-ratio    p-value
  --------------------------------------------------------
  const       118.914       6.35164      18.72     2.21e-029 ***
  price        -7.90785     1.09599      -7.215    4.42e-010 ***
  advert        1.86258     0.683195      2.726    0.0080    ***

Mean dependent var   77.37467   S.D. dependent var   6.488537
Sum squared resid    1718.943   S.E. of regression   4.886124
R-squared            0.448258   Adjusted R-squared   0.432932
F(2, 72)             29.24786   P-value(F)           5.04e-10
Log-likelihood      -223.8695   Akaike criterion     453.7390
Schwarz criterion    460.6915   Hannan-Quinn         456.5151
```

D #	Variable name	Descriptive label
0	const	
2	price	A price index for all products sold in a given month.
3	advert	Expenditure on advertising ($1000s)
1	sales	Monthly sales revenue ($1000s)

$$\widehat{sales} = 118.91 - 7.908 price + 1.863 advert$$

s.e. $(1.096)^{***}$ $(0.683)^{***}$ $R^2 = 0.448$

Estimating the parameters (SAS®)

The REG Procedure
Model: MODEL1
Dependent Variable: sales Monthly sales revenue ($1000s)

Number of Observations Read	75
Number of Observations Used	75

Analysis of Variance

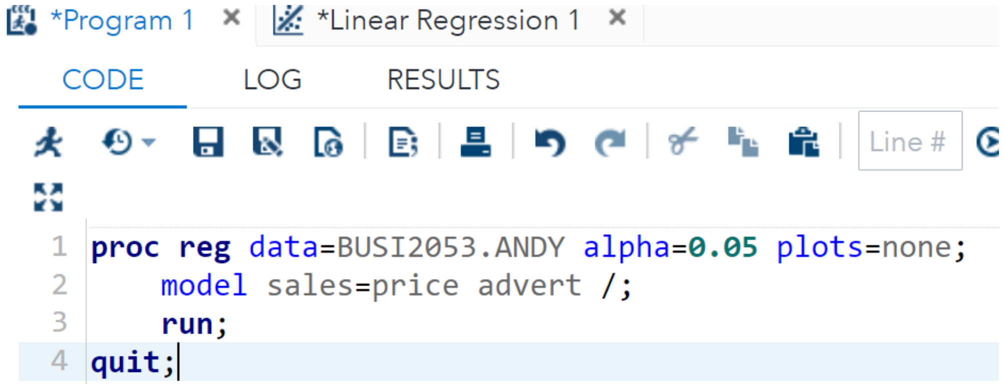
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1396.53893	698.26946	29.25	<.0001
Error	72	1718.94294	23.87421		
Corrected Total	74	3115.48187			

Root MSE	4.88612	R-Square	0.4483
Dependent Mean	77.37467	Adj R-Sq	0.4329
Coeff Var	6.31489		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	118.91361	6.35164	18.72	<.0001
price	A price index for all products sold in a given month.	1	-7.90785	1.09599	-7.22	<.0001
advert	Expenditure on advertising ($1000s)	1	1.86258	0.68320	2.73	0.0080

Use "Linear Regression" task or enter the following code in "Program 1" tab and click (assuming the dataset is saved in BUSI2053 folder)




```
proc reg data=busi2053.andy alpha=0.5 plots=none;
model sales=price advert /;
run;
quit;
```

$$\hat{sales} = 118.91 - 7.908price + 1.863advert$$

s.e. $\quad (1.096)^{***} \quad (0.683)^{***} \quad R^2 = 0.448$

Interpretating estimated coefficients

$$\hat{sales} = 118.91 - 7.908price + 1.863advert$$

s.e. (1.096)*** (0.683)*** $R^2 = 0.448$

Interpretations of the results:

1. **Holding advertising constant**, an increase in price index by $1 will lead to a fall in average (expected) monthly revenue of $7,908
2. Quiz**:** The estimated coefficient on advertising is positive; we estimate that $1000 increase in expenditure on advertising will lead to an increase in average (expected) monthly revenue by $1863, holding price index constant.

Interpretating estimated coefficients

$$\hat{sales} = 118.91 - 7.908 price + 1.863 advert$$

s.e. $(1.096)^{***}$ $(0.683)^{***}$ $R^2 = 0.448$

Interpretations of the results (Continued):

3. The estimated intercept implies that if both price and advertising expenditure were zero the expected sales revenue would be $118,914
   - Clearly, this outcome is not possible; a zero price implies zero sales revenue
   - it is important to recognize that the model is an approximation to reality **in the range of data we have** (the range of price index is between $4.83 and $6.49)
   - Including an intercept improves this approximation even when it is not directly interpretable

Predictive & Prescriptive Analytics

- **Predictive Analytics**: using the model to predict sales if price index is \$5.50 and advertising expenditure is \$1,200:

$$
\begin{aligned}
SALES &= 118.91 - 7.908PRICE + 1.863ADVERT \\
&= 118.91 - 7.908 \times 5.5 + 1.863 \times 1.2 \\
&= 77.652
\end{aligned}
$$

  - The predicted value of sales revenue for *PRICE* = 5.5 and *ADVERT* =1.2 is \$77,652.

- **Prescriptive analytics**: If Andy's Burger Barn wants to achieve a sales target of \$100,000 by setting price index of \$6, how much they have to spend on advertising?
- Solve: $100 = 118.91 - (7.908 \times 6) + 1.863ADVERT$

$$ADVERT = 15.3183$$

By setting price index of \$6 Andy's Burger Barn must spend \$15318.30 on advertising to achieve a sales target of \$100,000.

Statistical inference (Hypothesis testing)

Testing the Significance of a Single Coefficient

- Need to ask whether the sample data provide any evidence to suggest that y is related to each of the explanatory variables x_k, that is testing whether individual parameter β_k=0
- Testing this null hypothesis is sometimes called **a test of significance** for the explanatory variable x_k
- For our hamburger example, we can conduct a test that sales revenue is affected by price:
- The null and alternative hypotheses are: $H_0\colon\ \beta_2 = 0$ and $H_1\colon\ \beta_2 \neq 0$

Statistical inference (Hypothesis testing)

Testing the Significance of a Single Coefficient

We can use *p*-value to make conclusion

	coefficient	std. error	t-ratio	p-value	
const	118.914	6.35164	18.72	2.21e-029	***
price	-7.90785	1.09599	-7.215	4.42e-010	***
advert	1.86258	0.683195	2.726	0.0080	***

To test that sales revenue is dependent on price: $H_0\colon\ \beta_2 = 0$ and $H_1\colon\ \beta_2 \neq 0$

Reject H_0 at the 1% significance level because p=0.000 < **0.01**

To test that a higher spending on advertising leads to an increase in sales revenue:

$$H_0\colon\ \beta_3 = 0 \text{ and } H_1\colon\ \beta_3 > 0$$

What is p-value and what is your conclusion at 5% significance level?

Measuring Goodness-of-fit

Adjusted R²

- In the multiple regression model, the R^2 is still relevant, but it **can be made large by adding more and more variables**, even if the variables added have no justification
  - Algebraically, it is a fact that as variables are added the sum of squared errors SSE goes down, and thus R^2 goes up
  - An alternative measure of goodness of fit called the adjusted-R^2, denoted as $\overline{R^2}$

$$\bar{R}^2 = 1 - \frac{SSE/(n-k)}{SST/(n-1)}$$

- For the hamburger example, $\overline{R^2}$=0.433
  - After adjusting for the number of explanatory variables, **43.3%** of the variation in sales revenue is explained by the variation in price and the variation in the level of advertising expenditure.

Testing the Overall Significance of the Model

- Consider again the general multiple regression model with $(k - 1)$ explanatory variables and k unknown coefficients

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + e$$

- To examine whether we have a viable explanatory model, we set up the following null and alternative hypotheses:

$$H_0 : \beta_2 = \beta_3 = \ldots \beta_k = 0$$
$$H_1 : \text{ At least one of the } \beta_i \text{ is not equal to zero}$$

Testing the Overall Significance of the Model

- The test is sometimes referred to as a test of **the overall significance of the regression model.**

- To test the overall significance of a model, the F-test statistic can be written as:

$$F = \frac{(SST - SSE)/(k-1)}{SSE/(n-k)} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}$$

- This F-test evaluate the weighted variation in the dependent variable explained by the model against the weighted variation in the dependent variable not explained by the model

Testing the Overall Significance of the Model

Our Andy's burger barn example (R^2=0.4483)

$$SALES = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT + e$$
$$n = 75$$

H_0: $\beta_2 = \beta_3 = 0$

H_1: At least one of the β_i is not equal to zero

Test statistic: $F = \dfrac{R^2/(k-1)}{(1-R^2)/(n-k)} = \dfrac{0.4483/(3-1)}{(1-0.4483)/(75-3)} = 29.25$

at 5% significance level, the critical value for the F-statistic with (2, 72) degrees of freedom is F_c = **3.124**.

Since 29.25 > 3.124, we reject H_0 and conclude that the estimated relationship in our explanatory model is a significant one

(Note: we use R^2, not adjusted R^2 to calculate F-statistic)

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
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
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95	1.88	1.82	1.77	1.73	1.70
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

Testing the Overall Significance of the Model

```
Model 1: OLS, using observations 1-75
Dependent variable: sales

             coefficient   std. error   t-ratio    p-value
  --------------------------------------------------------
  const        118.914      6.35164     18.72     2.21e-029 ***
  price         -7.90785    1.09599     -7.215    4.42e-010 ***
  advert         1.86258    0.683195     2.726    0.0080    ***

Mean dependent var   77.37467   S.D. dependent var   6.488537
Sum squared resid    1718.943   S.E. of regression   4.886124
R-squared            0.448258   Adjusted R-squared   0.432932
F(2, 72)             29.24786   P-value(F)           5.04e-10
Log-likelihood      -223.8695   Akaike criterion     453.7390
Schwarz criterion    460.6915   Hannan-Quinn         456.5151
```

The REG Procedure
Model: MODEL1
Dependent Variable: sales Monthly sales revenue ($1000s)

Number of Observations Read	75
Number of Observations Used	75

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1396.53893	698.26946	29.25	<.0001
Error	72	1718.94294	23.87421		
Corrected Total	74	3115.48187			

Root MSE	4.88612	R-Square	0.4483
Dependent Mean	77.37467	Adj R-Sq	0.4329
Coeff Var	6.31489		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	118.91361	6.35164	18.72	<.0001
price	A price index for all products sold in a given month.	1	-7.90785	1.09599	-7.22	<.0001
advert	Expenditure on advertising ($1000s)	1	1.86258	0.68320	2.73	0.0080

Dataset: andy

- Since p-value $= P(F \geq 29.25) = 0.0000<0.05$, we reject H_0 and conclude that the estimated relationship is significant (The regression model as a whole is significant)

Summary

After this lecture, you should be able to:

- Explain the rationale of multiple regression
- Interpret the value of estimated coefficients
- Conduct the test for significance of a single coefficient
- Explain the rationale for using adjusting R-square and interpret it
- F-test for overall Significance of the regression model

University of Nottingham
UK | CHINA | MALAYSIA

Thank you!
Any question?