

**MULTIPLE REGRESSION ANALYSIS MODEL  
BUILDING AND  
ADVANCED TOPICS  
PART A**

# LEARNING OBJECTIVES

**Upon completing part A of this session, you should be able to do the following:**

- Explain model building using multiple regression analysis
- Apply multiple regression analysis to business decision-making situations
- Analyse and interpret the computer output for a multiple regression model
- Test the significance of the independent variables in a multiple regression model
- Recognize potential problems in multiple regression analysis and take steps to correct the problems

# MULTIPLE REGRESSION MODEL

The equation that describes how the dependent variable is related to the independent variables  $X_1, X_2, \dots, X_k$  and an error term is called the **multiple regression model**.

**Population model:**

The diagram shows the population regression model equation:  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon$ . Above the equation, three pink boxes with arrows point to specific parts: 'Y-intercept' points to  $\beta_0$ , 'Population slopes' points to the  $\beta$  coefficients ( $\beta_1, \beta_2, \dots, \beta_k$ ), and 'Random Error' points to  $\varepsilon$ .

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon$$

# MULTIPLE REGRESSION MODEL

A simple random sample is used to compute sample statistics  $b_0, b_1, b_2, \dots, b_k$  that are used as the point estimators of the parameters  $b_0, b_1, b_2, \dots, b_k$ .

**Estimated multiple regression model:**

The diagram shows the multiple regression equation  $\hat{y} = b_0 + b_1x_1 + b_2x_2 + \dots + b_kx_k$ . Three orange boxes with arrows point to specific parts of the equation: 'Estimated (or predicted) value of y' points to  $\hat{y}$ ; 'Estimated intercept' points to  $b_0$ ; and 'Estimated slope coefficients' points to the terms  $b_1x_1, b_2x_2, \dots, b_kx_k$ .

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \dots + b_kx_k$$

# MULTIPLE REGRESSION ASSUMPTIONS

Errors (residuals) from the regression model:

$$e = (y - \hat{y})$$

1. The model errors are independent and random
2. The errors are normally distributed
3. The mean of the errors is zero
4. Errors have a constant variance

# BASIC MODEL BUILDING CONCEPTS

1. Model Specification
2. Model Building
3. Model Diagnosis



# MODEL SPECIFICATION

1. Decide what you want to do and select the dependent variable
2. Determine the potential independent variables for your model
3. Gather sample data (observations) for all variables

# MODEL BUILDING

1. The process of constructing a mathematical equation that explains the relationship between independent variables and the dependent variable.



# MODEL DIAGNOSIS

1. Is the overall model significant?
2. Are the individual variables significant?
3. Is the standard deviation of the model error too large to provide meaningful results?
4. Have the regression analysis assumptions been satisfied?

# EXAMPLE

## BLITZ MANAGEMENT SALARY SURVEY

A sample of 20 management staff members are randomly selected. A suggestion was made that regression analysis could be used to determine if salary was related to the years of experience and the score on the BLITZ aptitude.

The years of experience, score on the aptitude test, and corresponding annual salary (\$'000s) for a sample of 20 management staff members is shown on the next slide.



# BLITZ MANAGEMENT SALARY SURVEY

## SAMPLE DATA

Exper.	Score	Salary	Exper.	Score	Salary
4	78	76.8	9	88	121.6
7	100	137.6	2	73	85.12
1	86	75.84	10	75	115.84
5	82	109.76	5	81	101.12
8	86	114.56	6	74	92.8
10	84	121.6	8	87	108.8
0	75	71.04	4	79	96.32
1	80	73.92	6	94	108.48
6	83	96	3	70	90.24
6	91	105.6	3	89	96

# BLITZ MANAGEMENT SALARY SURVEY

## ANNUAL SALARY MODEL

$Y$  = annual salary (\$'000s)

$X_1$  = years of experience

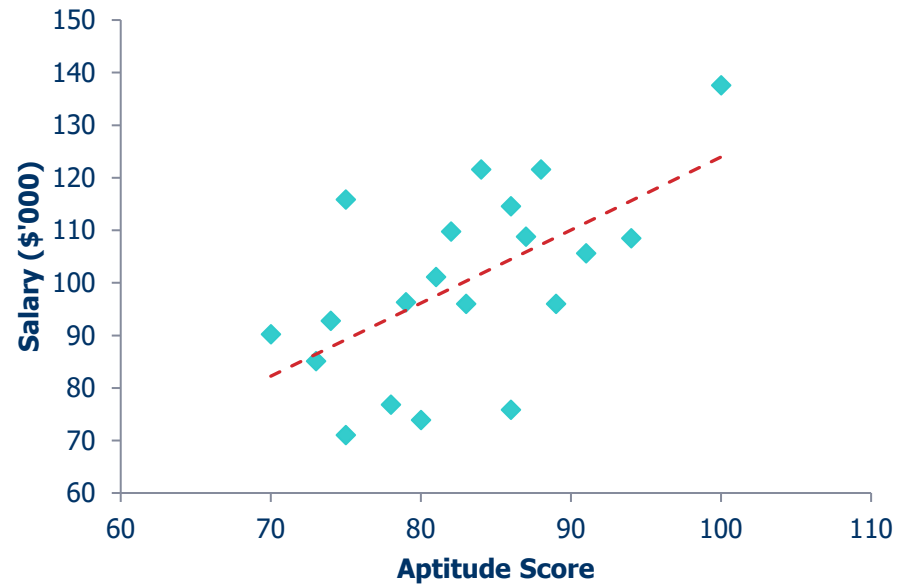
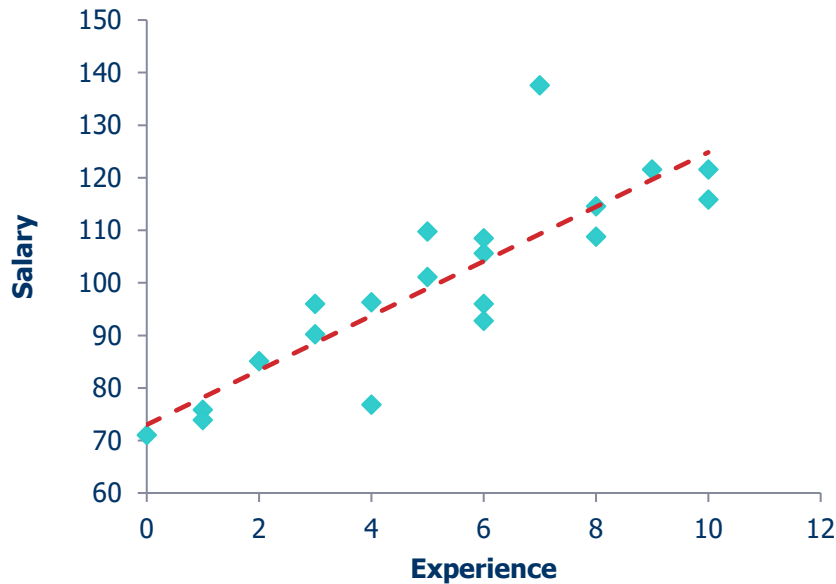
$X_2$  = score on aptitude test (out of 100)

Multiple regression model:

$$\text{Annual Salary} = b_0 + b_1 (\text{Years of Exp.}) + b_2 (\text{Aptitude Score})$$

# BLITZ MANAGEMENT SALARY SURVEY

## SCATTER PLOTS



Linear Associations

# BLITZ MANAGEMENT SALARY SURVEY CORRELATION MATRIX

	Annual Salary	Experience	Aptitude
Annual Salary	1		
Experience	0.855	1	
Aptitude	0.589	0.336	1

There appears to be a linear association between Salary v. Experience and Salary v. Score. In line with **Scatterplots**.

# ESTIMATING A MLR EQUATION

Computer software is generally used to generate the coefficients and measures of goodness of fit for multiple regression.

- Excel:
  - Data → Data Analysis → Regression

# BLITZ MANAGEMENT SALARY SURVEY

## MULTIPLE REGRESSION OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.913
R Square	0.834
Adjusted R Square	0.815
Standard Error	7.740
Observations	20

### ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	2	5123.364	2561.682	42.760	0.000
Residual	17	1018.439	59.908		
Total	19	6141.804			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	10.157	19.699	0.516	0.613	-31.406	51.719
Experience (Years)	4.492	0.635	7.070	0.000	3.152	5.833
Aptitude Score	0.803	0.248	3.243	0.005	0.281	1.325

**WE WILL NOW DISCUSS KEY ASPECTS OF THIS REGRESSION OUTPUT**



# MULTIPLE COEFFICIENT OF DETERMINATION ( $R^2$ )

- $R^2$  has a similar meaning and interpretation to  $R^2$  from simple linear regression.

$$R^2 = \frac{SSR}{SST} = \frac{\text{Sum of squares regression}}{\text{Total sum of squares}}$$

- The difference this time is that all of the independent variables ( $X$ 's) are used in the calculation of  $R^2$
- $R^2$  the proportion of variability in the dependent ( $Y$ ) which can be explained by the joint variation in the dependent ( $X$ ) variables.
- **$1 - R^2$**  is the proportion not explained by the regression model and would be accounted for by variables not included in the model.

# BLITZ MANAGEMENT SALARY SURVEY

## R<sup>2</sup> EXPLAINED

- In our case R<sup>2</sup> is 83.42%, which means that 83.4% of the variation in management staff salaries can be explained by the variation in 'Experience' and 'Aptitude Score'.
- The remaining  $1 - R^2 = 16.6\%$  of variation in salaries would be explained by other factors (e.g. Education, gender etc.) not included in the model.
- The value of R<sup>2</sup> is quite high, indicating a regression model that fit the data well (strong predictive power for the model).

# BLITZ MANAGEMENT SALARY SURVEY

## R<sup>2</sup>

Regression Statistics	
Multiple R	0.913
R Square	0.834
Adjusted R Square	0.815
Standard Error	7.740
Observations	20

$$R^2 = \frac{SSR}{SST} = \frac{5123.34}{6141.80} = 0.834$$

### ANOVA

	df	SS	MS	F	Significance F
Regression	2	5123.364	2561.682	42.760	0.000
Residual	17	1018.439	59.908		
Total	19	6141.804			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	10.157	19.699	0.516	0.613	-31.406	51.719
Experience (Years)	4.492	0.635	7.070	0.000	3.152	5.833
Aptitude Score	0.803	0.248	3.243	0.005	0.281	1.325

# ADJUSTED $R^2$

- $R^2$  never decreases when a new  $X$  variable is added to the model!
  - ✓ This can be a disadvantage when comparing models
- What is the net effect of adding a new variable?
  - ✓ We lose a degree of freedom when a new  $X$  variable is added
  - ✓ Did the new  $X$  variable add enough explanatory power to offset the loss of one degree of freedom?

# ADJUSTED R<sup>2</sup>

Shows the proportion of variation in  $y$  explained by all  $x$  variables **adjusted for the number of  $x$  variables used**.

$$R_A^2 = 1 - (1 - R^2) \left( \frac{n - 1}{n - k - 1} \right)$$

(where  $n$  = sample size,  $k$  = number of independent variables)

- Penalise excessive use of unimportant independent variables
- Smaller than  $R^2$
- Useful in comparing among models

# BLITZ MANAGEMENT SALARY SURVEY

## ADJUSTED R<sup>2</sup>

Regression Statistics	
Multiple R	0.913
R Square	0.834
Adjusted R Square	0.815
Standard Error	7.740
Observations	20

$$R^2_{Adj} = .815$$

81.5 percent of the variation in Salary is explained by variation in Experience and Aptitude, taking into account the sample size, and the number of independent variables.

### ANOVA

	df	SS	MS	F	Significance F
Regression	2	5123.364	2561.682	42.760	0.000
Residual	17	1018.439	59.908		
Total	19	6141.804			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	10.157	19.699	0.516	0.613	-31.406	51.719
Experience (Years)	4.492	0.635	7.070	0.000	3.152	5.833
Aptitude Score	0.803	0.248	3.243	0.005	0.281	1.325

# IS THE MODEL SIGNIFICANT?

## F-Test for Overall Significance of the Model:

- Shows if there is a linear relationship between all of the x variables considered together and y.
- Use F-test statistic
- Hypotheses:  
 $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$  (no linear relationship)  
 $H_A: \text{At least one } \beta_i \neq 0$  (at least one independent variable affects y)

# BLITZ MANAGEMENT SALARY SURVEY

## F-TEST

<i>Regression Statistics</i>	
Multiple R	0.913
R Square	0.834
Adjusted R Square	0.815
Standard Error	7.740
Observations	20

<i>ANOVA</i>					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	2	5123.364	2561.682	42.760	0.000
Residual	17	1018.439	59.908		
Total	19	6141.804			

*p* -value for  
the F-Test

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	10.157	19.699	0.516	0.613	-31.406	51.719
Experience (Years)	4.492	0.635	7.070	0.000	3.152	5.833
Aptitude Score	0.803	0.248	3.243	0.005	0.281	1.325



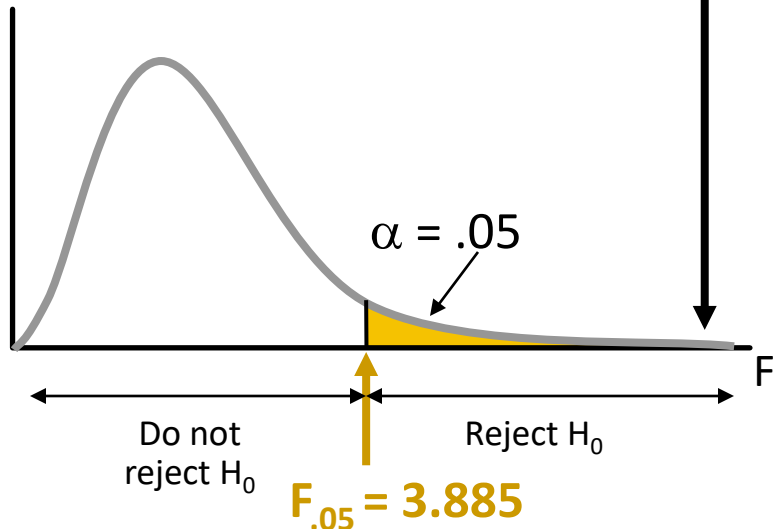
# F-TEST FOR OVERALL SIGNIFICANCE

$$H_0: \beta_1 = \beta_2 = 0$$

$$H_A: \beta_1 \text{ and } \beta_2 \text{ not both zero}$$

$$\alpha = .05$$

$$df_1 = 2 \quad df_2 = 12$$



## Test Statistic:

$$F_{sample} = \frac{2561.68}{59.90} = 42.76$$

## Decision:

Reject  $H_0$  at  $\alpha = 0.05$

## Conclusion:

The regression model does explain a significant portion of the variation in Salary.

(In other words, there is evidence that at least one independent variable affects  $y$ ).

# ARE INDIVIDUAL VARIABLES SIGNIFICANT?

Even if the F-test shows the model overall is significant, we still need to check if all of independent variables individually are significant.

- A separate  $t$ -test is conducted for each of the independent variables in the model.
- We refer to each of these  $t$ -tests as a test for individual significance.

For all independent variables we test:

$H_0 : \beta_j = 0$  (no linear relationship)

$H_1 : \beta_j \neq 0$  (linear relationship does exist between  $x_i$  and  $y$ )

# BLITZ MANAGEMENT SALARY SURVEY

## T-TESTS

Regression Statistics	
Multiple R	0.913
R Square	0.834
Adjusted R Square	0.815
Standard Error	7.740
Observations	20

T-value for experience is  $t = 7.070$ , with  $p$ -value of 0.000.

T-value for aptitude is  $t = 3.243$ , with  $p$ -value of 0.005.

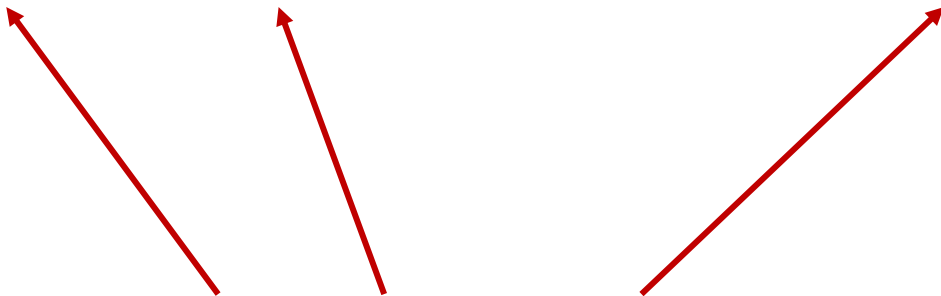
ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	2	5123.364	2561.682	42.760	0.000
Residual	17	1018.439	59.908		
Total	19	6141.804			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	10.157	19.699	0.516	0.613	-31.406	51.719
Experience (Years)	4.492	0.635	7.070	0.000	3.152	5.833
Aptitude Score	0.803	0.248	3.243	0.005	0.281	1.325

# BLITZ MANAGEMENT SALARY SURVEY

## ESTIMATED REGRESSION EQUATION

$$\text{Salary} = 10.157 + 4.492 \times \text{Experience} + 0.803 \times \text{Aptitude}$$

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2$$


# BLITZ MANAGEMENT SALARY SURVEY

## INTERPRETATION OF $b_0$

- $b_0$  is the 'intercept' term of 10.157.
- It is the value of  $y$  when **all**  $x$  values are 0.
- It tells us that **on average**, salaries of management staff with no experience and with a score of 0 on aptitude test would earn \$10,157.
- **Does not have a practical interpretation** as a management staff would NOT be employed with an aptitude score of 0.

# BLITZ MANAGEMENT SALARY SURVEY

## INTERPRETATION OF $b_1$

- $b_1$  is the coefficient of 'Experience' = 4.492
- It tells us that, assuming **no change** in the aptitude score, a management staff member with an extra year's work experience would earn, **on average**, an extra \$4,492 in annual salary.

# BLITZ MANAGEMENT SALARY SURVEY

## INTERPRETATION OF $b_2$

- $b_2$  is the coefficient of 'Score' = 0.803
- It tells us that, assuming **no change** in experience, every extra point scored on the aptitude test, results in an extra \$803 in annual salary, **on average**.

# CONFIDENCE INTERVALS FOR $\beta_j$

- Assuming our residual analysis does not highlight any problems, we can use a regression model to:
- Construct confidence intervals for the coefficients  $\beta_j$  and interpret them from a practical point of view.

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	10.157	19.699	0.516	0.613	-31.406	51.719
Experience (Years)	4.492	0.635	7.070	0.000	3.152	5.833
Aptitude Score	0.803	0.248	3.243	0.005	0.281	1.325

- ✓ We are 95% confident that an extra year of experience is worth, on average, between \$3,152 and \$5,833 more (assuming no change in the aptitude score).
- ✓ At 95% confidence level, salary is estimated to increase by between \$281 to \$1,325 for each unit of increase in aptitude score, on average.



# USING THE REGRESSION MODEL PREDICTION

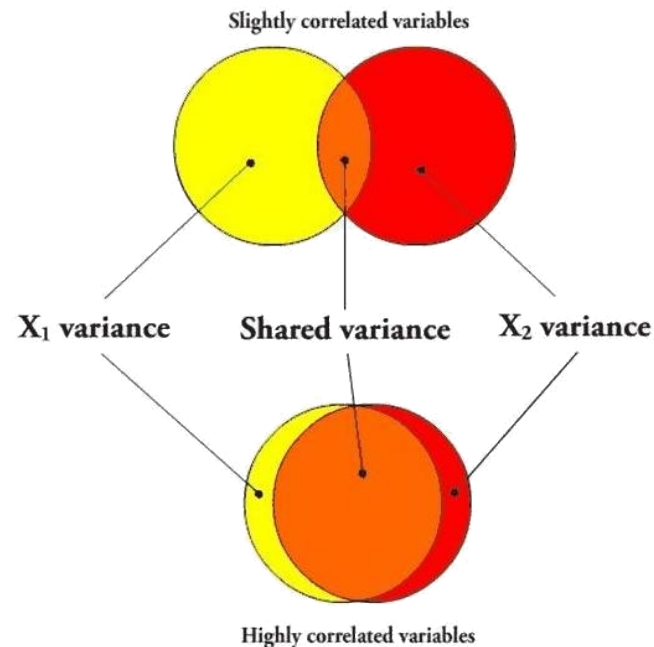
Determine estimates for a management staff member with 5 years experience and a score of 80 on the aptitude test.

$$\begin{aligned}\text{Salary} &= 10.157 + 4.492 \times \text{Experience} + 0.803 \times \text{Aptitude} \\ &= 10.157 + 4.492 \times 5 + 0.803 \times 80 \\ &= 96.857 \\ &= \$96,857\end{aligned}$$

# MULTICOLLINEARITY

**Multi-collinearity:** High correlation exists between two independent variables

It is not a mistake in the model specification, but due to the nature of the data at hand.



# MULTICOLLINEARITY

## WHY IMPORTANT?

Including two highly correlated independent variables can adversely affect the regression results:

- No new information provided
- Can lead to unstable coefficients (large standard error and low t-values)
- Coefficient signs may not match prior expectations

# DETECTING MULTICOLLINEARITY

## CORRELATION MATRIX

- The preliminary **correlation analysis** is one way of minimising the chance of this occurring.
- When the independent variables are highly correlated (say,  $|r| > 0.8$ ), it is not possible to determine the separate effect of any particular independent variable on the dependent variable.
- Generally we would **remove one** of the offending variables.
- If the regression equation is to be used **only for predictive purposes**, multi-collinearity is usually not a serious problem.

# DETECTING MULTICOLLINEARITY (VARIANCE INFLATIONARY FACTOR)

$VIF_j$  is used to measure collinearity:

$$VIF_j = \frac{1}{1 - R_j^2}$$

$R_j^2$  is the coefficient of determination when the  $j^{\text{th}}$  independent variable is regressed on the remaining  $k - 1$  independent variables.

If  $VIF_j \geq 5$ ,  $x_j$  is highly correlated with the other explanatory variables

