DR. ALVIN'S PUBLICATIONS

# MULTIPLE REGRESSION

## DR. ALVIN ANG



1 | PAGE

## CONTENTS

Ι.	What is Multiple Regression (MR)?4
<i>II</i> .	MR Example5
Α	. Using Excel6
B	. Question 1: Formulate the MR Equation7
С.	. Question 2: Interpret the MR Equation8
D	. Question 3: What is the Coefficient of Multiple Determination (Mulitple R <sup>2</sup> )?9
E. N	. Question 4: What is the Adjusted Coefficient of Multiple Determination (Adjusted Iulitple R <sup>2</sup> )?10
F.	Question 4: What is the Multiple Standard Error of Estimate (syx)?11
G	. Question 5: What is the ANOVA table?13
н	. Question 6: What is the Significance F?15
	1. The Global F Test15
١.	Question 7: What are the Individual P Values?18
	1. Individual F Test
J.	Question 8: What are the Individual t stats?21
K	. Question 9: What are the Lower and Upper 95%?22
<i>III</i> .	MR Assumptions24
Α	. Assumption 1: Linearity24
B	. Assumption 2: Homoscedasticity28
С.	Assumption 3: Normally Distributed33
D	Assumption 4: Multicollinearity35
	1. Obtaining the R2 for Population
	2. Obtaining the R2 for Percent unemployed
Ε.	Assumption 5: Autocorrelation
IV.	Forward Selection40
Α.	Step 1: Check out individual R <sup>2</sup> and adjusted R <sup>2</sup> 40
В.	Step 2: Check out alternative R <sup>2</sup> and adjusted R <sup>2</sup> 43

## 2 | PAGE

C.	Step 3: Check out the Final R <sup>2</sup> and adjusted R <sup>2</sup>	43
V.	References	45
VI.	About Dr. Alvin Ang	46

#### I. WHAT IS MULTIPLE REGRESSION (MR)?

- 1. Kindly refer to Ang (2019a) How to Perform Simple Linear Regression using Excel
- 2. In that article, we talked about  $Y = mX + c \rightarrow$  How to linearly regress scattered points onto a straight line.
- 3. That is 1 Y (dependent variable) and 1 X (independent variable).
- 4. In this article, we expand that to Multiple Variables, namely, X1, X2, X3...etc... but still 1 Y.
- 5. Which makes it impossible to draw on a graph because there are many dimensions.
- 6. But we are still trying to "linearly regress" every Variable unto Y.

## 4 | P A G E

#### II. MR EXAMPLE

#### Given:

Sales			Advertising	
(000)	Population	Percent	Expense	Mall
	(000,000)	Unemployed	(000)	Location
5.17	7.50	5.1	59.0	0
5.78	8.71	6.3	62.5	0
4.84	10.00	4.7	61.0	0
6.00	7.45	5.4	61.0	1
6.00	8.67	5.4	61.0	1
6.12	11.00	7.2	12.5	0
6.40	13.18	5.8	35.8	0
7.10	13.81	5.8	59.9	0
8.50	14.43	6.2	57.2	1
7.50	10.00	5.5	35.8	0
9.30	13.21	6.8	27.9	0
8.80	17.10	6.2	24.1	1
9.96	15.12	6.3	27.7	1
9.83	18.70	0.5	24.0	0
10.12	20.20	5.5	57.2	1
10.70	15.00	5.8	44.3	0
10.45	17.60	7.1	49.2	0
11.32	19.80	7.5	23.0	0
11.87	14.40	8.2	62.7	1
11.91	20.35	7.8	55.8	0
12.60	18.90	6.2	50.0	0
12.60	21.60	7.1	47.6	1
14.24	25.25	0.4	43.5	0
14.41	27.50	4.2	55.9	0
13.73	21.00	0.7	51.2	1
13.73	19.70	6.4	76.6	1
13.80	24.15	0.5	63.0	1
14.92	17.65	8.5	68.1	0
15.28	22.30	7.1	74.4	1
14.41	24.00	0.8	70.1	0

## 5 | PAGE

#### A. USING EXCEL

		DATA REVIEW	V VIEW							
		ar 🔗 🗄	🖥 Flash Fill	- □ Consolidate	e	0 Group	• +=	洛 Solver		
			* Remove Duplica	tes 🗄 What-If Ana	alvsis -	2 Unaro	up [	- Data An	alvsis	
		Text to	Z Data Validation	<ul> <li>Relationship</li> </ul>	nc	E Subtot	al			
		Columns =			ps	E Subioi				
			Data I	0015		Outlin	ne la l	Analys	IS	
			r.	Г (		1		V		-
		tising Expense (000)	Mall Location	r G	п	1	,	N	L	
		59	9 0	Data Analysis				?	X	
		62.5	5 0	Analysis Tools						
		6:	1 0	E-Test Two-Sampl	le for Varia	ances			ОК	
		6:	1 1	Fourier Analysis				Ca	incel	
		6	1 1	Histogram Moving Average						
		12.5		Random Number	Generatio	in			lelp	
		59.0		Regression						
		57.2	2 1	Sampling t-Test: Paired Two	Sample f	or Means				
		35.8	во	t-Test: Two-Samp	le Assumir	ng Equal Va	ariances	¥		
		27.9	ə 0							
			OUT FORMULAS		105140					
FI		Connections	OUT FORMULAS		VIEVV	Ban Car		É Graup -	+3 2.4	Selver.
		Properties			ash Fill Imove Dunlir	cates 📰 Wh	at-lf Δnalvsis τ		′ '⊒ i≱3 	Data Analysis
Get E	xternal Refres	h Z S	Sort Filter	Text to	inove bupit	cutes un vin	at it Analysis	- ongroup		Julu Analysis
D	ata 👻 🛛 🖂 🗸	🖧 Edit Links 🛛 🗛	🌄 Adv	anced Columns 🗟 Da	ata Validation	n ▼ ⊏a Rela	ationships	F Subtotal		
D	ata • All ∘	Connections	Sort & Filter	anced Columns 😽 Da	ata Validation Data	n ▼ ⊏a Rela Tools	ationships	문國 Subtotal Outline	Es.	Analysis
D B1	ata × All ×	Edit Links Connections	Sort & Filter	anced Columns 🗟 Da	ata Validation Data	n ▼ Rela Tools	ationships	E Subtotal Outline	G	Analysis
D B1	All -	$\begin{array}{c} \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \\ \end{array} \end{array} \xrightarrow{P} \begin{array}{c} A^{P} \end{array}$	Sort & Filter	Columns 🗟 Da	ata Validation Data F	n v 🕰 Rela Tools	ationships	EIII Subtotal Outline	G.	Analysis
D: B1	All - All - A Sales (000)	Connections	Sort & Filter 17.1 C It Unemployed - Advert	D D D D D D D Ma	E E E	n - Regression	ationships	E Subtotal Outline	F <sub>24</sub>	Analysis ? ×
D	All - All - A Sales (000) Po 5.17 5.78	Edit Links Connections      Connections      Connection      fx      pulation (000,000) Percen      7.5      8.71	Sort & Filter 17.1 C it Unemployed Advect 5.1 6.3	D D D Columns D D S S S S S D Columns S D Columns S D D Columns S D Columns S D D Columns S D S D S D S D S D S D S D S D S S D S S D S S D S S D S S D S S D S S D S S S D S S S S S S S S S S S S S	E Il Location	Regression	ationships	E Subtotal Outline SAS1:SAS	31	Analysis ? X OK
D B1 1 2 3 4	All - All - A Sales (000) 5.17 5.78 4.84	Edit Links Connections      Connections      Connection	C Advert	D Columns D Lising Expense (000) Ma 59 62.5 61	E E E E C C C C C C C C	Regression	ationships ange: ange:	Subtotal Outline SAS1:SAS SBS1:SES	G 31 31	Analysis       ?     ×       OK     Cancel
D B1 1 2 3 4 5 6	All -	Edit Links Connections :	Adv Sort & Filter 17.1 C tt Unemployed Advact 5.1 6.3 4.7 5.4 5.4	D Columns S Da Lising Expense (000) Ma 59 62. 61 61 61	E E Il Location 0 0 1 1	Regression Input Input Regression	ationships ange: ange:	Subtotal Outline SAS1:SAS SBS1:SES Constant is	5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5	Analysis       ?     ×       OK     Cancel       Help
D B1 1 2 3 4 5 6 7	All ~ All ~ Sales (000) Po 5.17 5.78 4.84 6 6 6 6 6.12	Edit Links Connections :	Adv Sort & Filter 17.1 C tt Unemployed Advect 6.3 4.7 5.4 5.4 7.2	D Columns S Da Lising Expense (000) Ma 59 62.5 61 61 61 61 61 12.5	E E Il Location 0 0 1 1 1 0	Regression Input Input Y Ra Date States Confi	ange: ange: ange:	SAS1:SAS SBS1:SES Constant is 95 %	r⊊ 31 ■ 31 ■ 2ero	Analysis
D B1 1 2 3 4 5 6 7 8 9	3 • All • Al	Edit Links Connections :	Adv Sort & Filter 17.1 C t Unemployed Advect 5.1 6.3 4.7 5.4 7.2 5.8	D Columns S Da Columns S Da Columns S Da Sep 62.5 61 61 61 61 12.5 35.8 59.9	E B B B B C C C C C C C C C C C C C C C	Regression Input Input Y Ra Input X Ra Input X Ra Input X Confi	ange: ange: 5] dence Level: ttions	Subtotal Outline SAS1:SAS SBS1:SES Constant is 95 %	r₂ 31 ■ 2ero	Analysis  ? ×  OK  Cancel  Help
D B1 1 2 3 4 5 6 7 8 9 10	All - All - Sales (000) Po 5.17 5.78 4.84 6 6 6 6.12 6.4 6.4 6.5 8.5	Edit Links Connections pulation (000,000) Percen 7.5 8.71 10 7.45 8.67 11 13.18 13.81 14.43	Adv Sort & Filter 17.1 C tt Unemployed 6.3 4.7 5.4 7.2 5.8 5.8 6.2	D Columns S Da Columns D Da Columna D Da Col	E B B B C C C C C C C C C C C C C C C C	Regression Input Input Y Ri Input X Ri Confi Output op	ange: ange: 5] dence Level: ttions ut Range:	Subtotal Outline SAS1:SAS SBS1:SES Constant is 95 %	31 ES	Analysis  ? ×  OK  Cancel  Help
D. B1 1 2 3 4 5 6 7 8 9 10 11	3 Sales (000) Po 5.17 5.78 4.84 6 6 6.12 6.4 7.1 8.5 7.5 9 3	Edit Links         A*           Connections	Adv Sort & Filter 17.1 C tt Unemployed 6.3 4.7 5.4 7.2 5.8 5.8 6.2 5.5 6.8	D Columns S Da Columns Da Columna Col	E B Il Location 0 0 0 1 1 1 0 0 0 0 1 0 0 0 0 1 0	Regression Input Input Input Y Ri Input X Input Y Ri Confi Output op Output op	ange: ange: 5] dence Level: ttions tt Range: Vorksheet <u>P</u> ly:	Ell Subtotal Outline	75 31 (%) 31 (%) Zero	Analysis  ? ×  OK  Cancel  Help
D B1 1 2 3 4 5 6 7 7 8 9 10 11 12 13	All - All - Sales (000) Po 5.17 5.78 4.84 6 6 6.12 6.4 7.1 8.5 7.5 9.3 8.8	Lefit Links Connections pulation (000,000) Percent 7.5 8.71 10 7.45 8.67 11 13.18 13.81 14.43 10 13.21 17.1	Adv Sort & Filter 17.1 C tt Unemployed 6.3 4.7 5.4 7.2 5.8 5.8 6.2 5.5 6.8 6.2	D Columns is Da Columns is Columns is Column Columns is Column Column Columns is Column Col	E Ull Location 0 0 0 1 1 0 0 0 0 1 1 0 0 0 0 1 1 0 0 0 0 1 1	Regression Input Input Y Ri Input X Ri Confi Output op Qutpu Regiduation New Y Residuation	ange: ange: s] dence Level: tions tt Range: Vorksheet <u>P</u> ly: <u>Vorksook</u>	Ell Subtotal Outline	75 31 56 31 56 Zero	Analysis  ? ×  OK  Cancel  Help
D B1 1 2 3 4 5 6 7 8 9 10 11 11 12 13 14 15	All - All - Sales (000) Po 5.17 5.78 4.84 6 6 6.12 6.4 7.1 8.5 7.5 9.3 8.8 9.96 0.00	Edit Links         A*           Connections         Factor           pulation (000,000)         Percent           7.5         8.71           10         7.45           8.67         11           13.18         13.81           14.43         10           13.21         17.1           15.12         49.7	Adv Sort & Filter 17.1 C tt Unemployed 6.3 4.7 5.4 7.2 5.8 5.8 6.2 5.5 6.8 6.2 6.3 0.5 6.3	D Columns is Da Columns is Columns is Column Columns is Column Column Columns is Column Co	E Ull Location 0 0 0 1 1 1 0 0 0 0 0 1 1 0 0 0 0 1 1 1 0 0 0 0 1 1 1 0 0 0 0 0 0 0 1 1 1 1 0 0 0 0 1	Regression Input Input Input ¥ R: Confi Output op Qutpu Residuals Residuals	ange: ange: ange: dence Level: tions tt Range: Vorksheet <u>P</u> ly: Vorkbook	Ell Subtotal Outline	rs 31 TS Zero	Analysis  ? ×  OK  Cancel  Help
D. B1 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16	All - All - Sales (000) Po 5.17 5.78 4.84 6 6 6.12 6.4 7.1 8.5 7.5 9.3 8.8 9.99 9.83 10.12	La Edit Links Connections pulation (000,000) Percen 7.5 8.71 10 7.45 8.67 11 13.18 13.81 14.43 10 13.21 17.1 15.12 18.7 20.2	Adv Sort & Filter 17.1 C tt Unemployed 6.3 4.7 5.4 7.2 5.8 6.2 5.5 6.8 6.2 6.3 0.5 5.5	D Columns is Da Columns is Columns is Columna Columns is Columna Columns is Columna Columns is Columna Columns is Columna Columna Columns is Columna Col	E Ull Location 0 0 0 1 1 1 0 0 0 0 1 1 1 0 0 0 0 1 1 1 1 0 0 0 0 0 1 1 1 1 0 0 0 0 1 1 1 1 0 0 0 1	Regression Input Input Y R: Input Y R: Confi Output op Qutp Residuals Residu Stand	ange: ange: 5] dence Level: tions tt Range: Vorksheet <u>P</u> ly: <u>V</u> orkbook jals ardized Residua	Ell Subtotal Outline	rs 31 TS 2ero dual Plots Fit Plots	Analysis       ?     ×       OK     Cancel       Help
D. B1 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 10	All - Sales (000) Po 5.17 5.78 4.84 6 6 6.12 6.4 7.1 8.5 7.5 9.3 8.8 9.96 9.83 10.12 10.7	Edit Links         A*           Connections         Factor           pulation (000,000)         Percent           7.5         8.71           10         7.45           8.67         11           13.18         13.81           14.43         10           13.21         17.1           15.12         18.7           20.2         15	Adv Sort & Filter 17.1 C t Unemployed 6.3 4.7 5.4 7.2 5.8 6.2 5.5 6.8 6.2 6.3 0.5 5.5 5.8 7.1	D Columns ⇒ Da Columns ⇒ Da Sy 62.5 61 61 61 61 61 61 61 61 61 61 61 72.5 858 59.9 57.2 858 27.9 24.1 27.7 24 4.3	E Il Location 0 0 0 0 1 1 0 0 0 0 1 1 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0	Regression Input Input y Ri Input y Ri Input y Ri Confi Output op Output Residual Stand	ationships ange: ange: ange: ange: ange: tions trange: Vorksheet <u>P</u> ly: <u>V</u> orkbook ardized Residua ardized Residua ardized sesidua	Ell Subtotal Outline SAS1:SAS SBS1:SES Constant is 95 % Is Is Resignation Is Resignation Is	rs 31 The second secon	Analysis  ? ×  OK Cancel Help
D B1 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19	A All - Sales (000) Po 5.17 5.78 4.84 6 6 6.12 6.4 7.1 8.5 7.5 9.3 8.8 9.96 9.83 10.12 10.7 10.45 11.32	Edit Links Connections Connec	Adv Sort & Filter 17.1 C t Unemployed 6.3 4.7 5.4 7.2 5.8 6.2 5.5 6.8 6.2 6.3 0.5 5.5 5.8 6.2 6.3 0.5 5.5 5.8 7.1 7.5	D Columns ⇒ Da Columns ⇒ Da Sy 62.5 61 61 61 61 61 61 61 61 61 61 61 72.5 85.8 59.9 57.2 55.8 27.9 24.1 27.7 24 57.2 44.3 49.2 23	E Il Location 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0	Regression Input Input y Ri Input y Ri Input y Ri Confi Output op Output op Output op Output op Residud Stand	ange: ange: ange: ange: ange: ange: tions at Range: Vorksheet <u>Ply</u> : <u>Vorksheet Ply</u> : <u>Vorksheet Ply</u> : <u>Vorksheet Ply</u> : <u>Vorksheet Ply</u> : <u>Vorksheet Ply</u> : <u>als</u> ardized Residua robability Plo	Ell Subtotal Outline SAS1:SAS SBS1:SES Constant is 95 % Is Is Resignation Is Is	IS IN INCLASSING AND	Analysis  ? ×  OK Cancel Help
D. B1 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20	A All - Sales (000) Po 5.17 5.78 4.84 6 6 6 6.12 6.4 7.1 8.5 7.5 9.3 8.8 9.96 9.83 10.12 10.7 10.45 11.32 11.87	Edit Links Connections Connec	Adv Sort & Filter 17.1 C t Unemployed 6.3 4.7 5.4 7.2 5.8 6.2 5.5 6.8 6.2 6.3 0.5 5.5 5.8 6.2 6.3 0.5 5.5 5.8 6.2 6.3 0.5 5.5 5.8 6.2 6.3 0.5 5.5 5.8 6.2 6.3 7.2 5.5 6.8 6.2 7.2 5.5 5.5 5.8 6.2 5.5 5.5 5.8 6.2 5.5 5.5 5.5 5.8 6.2 5.5 5.5 5.5 5.5 5.5 5.5 5.5 5	D Columns ⇒ Da Columns ⇒ Da Sy 62.5 61 61 61 61 61 61 61 61 61 61 61 72.5 35.8 59.9 57.2 35.8 27.9 24.1 27.7 24 44.3 49.2 23 62.7	E Il Location 0 0 0 1 1 0 0 0 0 1 1 0 0 0 0 1 1 0 0 0 0 1 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0	Regression Input Input y Ri Input y Ri Input y Ri Confi Output op Qutpu Output op Qutpu Output op New W Residuals Stand	ange: ange: ange: ange: ange: tions ut Range: Vorksheet <u>Ply:</u> <u>Vorksheet Ply:</u> <u>Vorksheet Ply:</u> <u>Vorksheet Ply:</u> <u>vorkbook</u> ardized Residua robability Plo	Ell Subtotal Outline SAS1:SAS SBS1:SES SBS1:SES Constant is 95 % 95 % Is Resigned Is	IS IN INCLASSING	Analysis
D. B1. 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 22	A All - Sales (000) Po 5.17 5.78 4.84 6 6 6.12 6.4 7.1 8.5 7.5 9.3 8.8 9.96 9.83 10.12 10.7 10.45 11.32 11.87 11.91 12.6	Edit Links Connections Connec	Adv Sort & Filter 17.1 C t Unemployed 6.3 4.7 5.4 7.2 5.8 6.2 5.5 6.8 6.2 6.3 0.5 5.5 5.8 6.2 6.3 0.5 5.5 5.8 6.2 6.3 0.5 5.5 5.8 6.2 6.3 0.5 5.5 5.8 6.2 6.3 6.3 6.2 6.3 6.2 6.3 6.2 6.3 6.3 6.2 6.3 6.3 6.2 6.3 6.2 6.3 6.3 6.2 6.3 6.2 6.3 6.3 6.2 6.3 6.2 6.3 6.3 6.2 6.3 6.2 6.3 6.3 6.2 6.3 6.2 6.3 6.2 6.3 6.3 6.2 6.3 6.3 6.3 6.2 6.3 6.2 6.3 6.2 6.3 6.3 6.5 5.5 5.5 5.5 5.5 5.5 5.5 5.5	D Columns ⇒ Da Columns ⇒ Da Sy 62.5 61 61 61 61 61 61 61 61 61 62 7.2 35.8 59.9 57.2 35.8 27.9 24.1 27.7 24 57.2 44.3 49.2 23 62.7 55.8 50	E Il Location 0 0 0 1 1 0 0 0 0 1 1 0 0 0 1 1 0 0 0 1 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0	Regression Input Input y Ri Input y Ri Input y Ri Confi Output op Output op Output op Output op New W Residuals Stand	ange: ange: ange: ange: ange: tions ut Range: Vorksheet <u>P</u> ly: Vorkbook ardized Residua robability Plo	Ell Subtotal Outline SAS1:SAS SBS1:SES SBS1:SES Constant is 95 % Is Is Results	IS IN INCLASSING AND INCLASSINA AND INCLASSINA AND	Analysis
D. B1. 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23	A All - Sales (000) Po 5.17 5.78 4.84 6 6 6.12 6.4 7.1 8.5 7.5 9.3 8.8 9.96 9.83 10.12 10.7 10.45 11.32 11.87 11.91 12.6 12.6	Edit Links Connections Connec	Adv Sort & Filter 17.1 C t Unemployed 6.3 4.7 5.4 7.2 5.8 6.2 5.5 6.8 6.2 5.5 6.8 6.2 6.3 0.5 5.5 5.8 6.2 6.3 0.5 5.5 5.8 6.2 6.3 0.5 5.5 5.8 6.2 6.3 0.5 5.5 5.8 6.2 6.3 6.2 7.1 7.2 7.2 7.2 7.2 7.2 7.2 7.2 7.2	D Columns ⇒ Da Columns ⇒ Da Sy 62.5 61 61 61 61 61 61 61 61 61 62 75.2 35.8 59.9 57.2 35.8 27.9 24.1 27.7 24 57.2 44.3 49.2 23 62.7 55.8 50 47.6	E E Il Location 0 0 0 1 1 0 0 0 0 1 1 0 0 0 1 1 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0	Regression Input Input y Ri Input y Ri Input y Ri Output op Output op Output op Output op Output op Output op Output op Nerway Residuals Stand	ange: ange: ange: i] dence Level: tions ut Range: Vorksheet <u>P</u> ly: Vorksheet <u>P</u> ly: Vorkbook i andized Residua ardized Residua	Ell Subtotal Outline SAS1:SAS SBS1:SES SBS1:SES Constant is 95 % Is Results Is Results Is	IS IN INCLASSING AND INCLASSINA AND INCLASSINA AND	Analysis
D. B1 1 2 3 4 5 6 6 7 8 9 10 11 12 13 14 15 16 17 18 9 9 20 21 22 23 24 25	A All - Sales (000) Po 5.17 5.78 4.84 6 6 6.12 6.4 7.1 8.5 7.5 9.3 8.8 9.96 9.83 10.12 10.7 10.45 11.87 11.8	Edit Links Connections Connec	Adv Sort & Filter 17.1 C t Unemployed 6.3 4.7 5.4 7.2 5.8 6.2 5.5 6.8 6.2 5.5 6.8 6.2 6.3 0.5 5.5 5.8 6.2 6.3 0.5 5.5 5.8 6.2 6.3 0.5 5.5 5.8 6.2 6.3 0.5 5.5 5.8 6.2 6.3 0.5 5.5 5.8 6.2 6.3 0.5 5.5 5.8 6.2 6.3 0.5 5.5 5.8 6.2 6.3 6.2 7.1 7.2 5.8 6.2 7.2 5.5 6.8 6.2 7.5 5.5 5.5 5.5 5.5 5.5 5.5 5.5	D Columns ⇒ Da Columns ⇒ Da Sy 62.5 61 61 61 61 61 61 61 61 61 61	E Il Location 0 0 1 1 0 0 0 0 1 1 0 0 0 0 1 1 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0	Regression Input Input y Ri Input y Ri Input y Ri Ultabets Confi Output op Output op Output op Output op Output op Nerwal Residuals Stand Normal P	ange: ange: ange: 3] dence Level: tions ut Range: Vorksheet <u>P</u> ly: Vorkbook ardized Residua ardized Residua ardized Residua	Ell Subtotal Outline	IS IN INCLASSING INCLA	Analysis
D. B1 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26	A All - Sales (000) Po 5.17 5.78 4.84 6 6 6.12 6.4 7.1 8.5 7.5 9.3 8.8 9.96 9.83 10.12 10.7 10.45 11.32 11.87 11.91 12.6 12.6 14.24 14.44 13.73	Edit Links         A*           Connections         7.5           mulation (000,000)         Percentric and a straight of the straightof the straight of the straight of the straight of the straight of	Adv Sort & Filter 17.1 C tt Unemployed 6.3 4.7 5.4 7.2 5.8 6.2 5.5 6.8 6.2 5.5 6.8 6.2 6.3 0.5 5.5 5.8 6.2 6.3 0.5 5.5 5.8 6.2 7.1 7.2 8.2 7.1 6.2 7.1 7.2 7.2 7.2 7.2 7.2 7.2 7.2 7.2	D Columns is Da Columns is Columns is Columns is Columns i	E Il Location 0 0 1 1 0 0 0 1 1 0 0 0 1 1 0 0 0 1 1 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0	Regression Input Input y Ri Input y Ri Input y Ri Ultabets Confi Output op Output op Output op Output op Output op Nerwal Residuals Stand	ange: ange: ange: i] dence Level: tions ut Range: Vorksheet <u>P</u> ly: Vorkbook j ardized Residua ardized Residua ardized Residua	Ell Subtotal Outline	I I I I I I I I I I I I I I I I I I I	Analysis
D. B11 2 3 4 5 6 7 7 8 9 9 10 11 12 13 14 15 16 17 18 9 20 21 22 23 22 22 22 22 22 22 22 22 22 22 22	A All - Sales (000) Po 5.17 5.78 4.84 6 6 6.12 6.4 7.1 8.5 7.5 9.3 8.8 9.96 9.83 10.12 10.7 10.45 11.32 11.87 11.91 12.6 12.6 14.24 14.44 13.73 13.73 13.73 13.73 13.87 13.9	Edit Links         A*           Connections         Far           pulation (000,000)         Percen           7.5         8.71           10         7.5           8.71         10           7.45         8.67           11         13.18           13.81         14.43           10         13.21           17.1         15.12           18.7         20.2           15         17.6           19.8         14.4           20.35         27.5           21.6         25.25           27.5         21           19.7         24.15	Adv Sort & Filter 17.1 C t Unemployed 6.3 4.7 5.4 7.2 5.8 6.2 5.5 6.8 6.2 5.5 6.8 6.2 6.3 0.5 5.5 5.8 6.2 6.3 0.5 5.5 5.8 6.2 7.1 7.2 8.2 7.1 8.2 7.1 8.2 7.1 8.2 7.2 5.5 5.5 5.5 5.8 6.2 7.1 7.5 8.2 7.1 7.2 7.2 7.5 7.5 7.5 7.5 7.5 7.5 7.5 7.5	D Columns is Da Columns is Columns is Columns is Columns is Columns is Columns is Columns is C	E E Il Location 0 0 1 1 0 0 0 1 1 0 0 0 0 1 1 0 0 0 0 1 1 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0	Regression Input Input y Ri Input y Ri Output op Output op Output op Residuals Residuals	ange: ange: ange: i] dence Level: tions ut Range: Vorksheet <u>P</u> ly: Vorkbook juals ardized Residua robability Plo	Ell Subtotal Outline	I I I I I I I I I I I I I I I I I I I	Analysis
D. B11 2 3 4 5 6 7 8 9 9 9 9 10 11 12 13 14 15 16 17 18 19 20 21 22 24 25 26 27 28 29 20 21 22 23 24 25 26 27 28 29 20 20 21 22 22 22 22 22 22 22 22 22	Atta * All * 3 Sales (000) Po 5.17 5.78 4.84 6 6 6.12 6.4 7.1 8.5 7.5 9.3 8.8 9.96 9.83 10.12 10.7 10.45 11.32 11.87 11.91 12.6 12.6 14.41 13.73 13.8 14.92	Edit Links         A*           Connections         7.5           mulation (000,000)         Percentric Arrows           7.5         8.71           10         7.45           8.67         11           13.18         13.81           14.43         10           13.21         17.1           15.12         18.7           20.2         15           17.6         19.8           14.4         20.35           18.9         21.6           25.25         27.5           21         19.7           24.15         17.65	Adv Sort & Filter 17.1 C t Unemployed 6.3 4.7 5.4 7.2 5.8 6.2 5.5 6.8 6.2 5.5 6.8 6.2 5.5 6.8 6.2 5.5 5.8 6.2 5.5 5.8 6.2 7.1 7.5 8.2 7.1 7.5 8.2 7.1 7.5 8.2 7.1 7.5 8.2 7.1 7.5 8.2 7.1 7.5 8.2 7.1 7.5 8.2 7.1 7.5 8.2 7.5 8.5 7.5 8.5 7.5 7.5 7.5 7.5 7.5 7.5 7.5 7	D Columns is Da Columns is Columns is Columns is Columns is Columns is Columns is Columns is Columns is Columns is Columns is Columns is C	E E Il Location 0 0 0 1 1 0 0 0 0 1 1 0 0 0 0 1 1 0 0 0 0 1 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0	Regression Input Input y Ri Input y Ri Output op Output op Output op Residuals Residuals	ange: ange: ange: i] dence Level: tions ut Range: Vorksheet <u>P</u> ly: Vorkbook juals ardized Residua robability Plo	Ell Subtotal Outline	I I I I I I I I I I I I I I I I I I I	Analysis
D. B11 2 3 4 5 6 7 8 9 9 9 9 10 11 12 13 14 15 16 17 18 19 20 21 22 3 24 25 26 27 28 29 30 20 21 22 3 24 25 26 27 28 29 30 20 21 22 22 22 22 22 22 22 22 22	Atta * All * 3 Sales (000) Po 5.17 5.78 4.84 6 6 6 6.12 6.4 7.1 8.5 7.5 9.3 8.8 9.96 9.83 10.12 10.7 10.45 11.32 11.87 11.91 12.6 12.6 14.41 13.73 13.73 13.8 14.92 15.28	Edit Links         A*           Connections         Far           pulation (000,000)         Percen           7.5         8.71           10         7.5           8.71         10           7.45         8.67           11         13.18           13.81         14.43           10         13.21           17.1         15.12           18.7         20.2           15         17.6           19.8         14.4           20.35         18.9           21.6         25.25           27.5         21           19.7         24.15           17.65         22.3	Adv Sort & Filter 17.1 C t Unemployed 6.3 4.7 5.4 7.2 5.8 6.2 5.5 6.8 6.2 5.5 6.8 6.2 6.3 0.5 5.5 5.8 6.2 6.3 0.5 5.5 6.8 6.2 7.1 7.5 8.2 7.1 0.4 4.2 7.1 0.4 4.2 7.1 0.5 8.5 5.5 5.5 5.5 5.5 5.5 5.5 5	D Columns is Da Columns is Columns is Columns is Columns is Columns is Columns is Columns is Col	E Il Location 0 0 0 1 1 0 0 0 0 1 1 0 0 0 0 1 1 0 0 0 0 1 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0	Regression Input Input y Ri Input y Ri Output op Output op Output op Residuals Residuals	ange: ange: ange: ange: i] dence Level: tions ut Range: Vorksheet <u>P</u> ly: Vorksheet <u>P</u> ly: Vorksheet <u>P</u> ly: andized Residua robability Plo andized Residua	Ell Subtotal Outline	I I I I I I I I I I I I I I I I I I I	Analysis

## 6 | PAGE

SUMMARY OUTPUT								
Regression Statist	tics							
Multiple R	0.9205436							
R Square	0.84740053							
Adjusted R Square	0.82298461							
Standard Error	1.39577331							
Observations	30							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	4	270.4615184	67.6153796	34.706891	7.22088E-10			
Residual	25	48.70457828	1.948183131					
Total	29	319.1660967						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-1.66853987	1.408315529	-1.184777016	0.24724684	-4.569019996	1.23194026	-4.569019996	1.231940255
	0.55400704	0.050630063	10 00000815	5.468E-11	0.447634803	0.65617981	0.447634803	0.656179811
Population (000,000) X1	0.55190/31	0.050629062	10.30033813	0.4000 11	01111001000		01117001000	
Population (000,000) X1 Percent Unemployed X2	0.55190731 0.20316264	0.030629062	1.735154932	0.09502688	-0.037980835	0.44430612	-0.037980835	0.444306121
Population (000,000) X1 Percent Unemployed X2 Advertising Expense (000) >	0.55190731 0.20316264 3 0.03135496	0.050629062 0.117086169 0.016062075	1.735154932 1.952111518	0.09502688	-0.037980835 -0.001725501	0.44430612 0.06443543	-0.037980835 -0.001725501	0.444306121 0.064435426

#### **B.** QUESTION 1: FORMULATE THE MR EQUATION

**Multiple Regression Equation**  $\hat{Y} = a + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k$ 

where

- X1 is one of the independent variables.
- X<sub>2</sub> is the second independent variable.
- $X_k$  is the  $k^{\text{th}}$  independent variable.
- a is the Y-intercept, the value of Y when all the X's are zero.
- $b_j$  is the net change in  $\hat{Y}$  for each unit change in  $X_{j}$ , holding all other X's constant.
- *j* the subscript can assume values between 1 and *k*, which is the number of independent variables.

Figure 1: MR Equation (SUSS, 2014)

## 7 | P A G E

#### The MR Equation is:

 $\hat{Y} = -1.669 + 0.552(Population) + 0.203(Unemployed) + 0.031(Advert) + 0.220(Mall)$ 

#### C. QUESTION 2: INTERPRET THE MR EQUATION

- For every additional 1 million increase in Population, the estimated mean Sales is increased by \$552.
- For every additional 1% in Unemployment, the estimated mean Sales is increased by \$203.
- For every additional \$1000 increase in Advertising Expense, the estimated mean Sales is increased by \$31.
- If the store is located in the mall location, the estimated mean Sales is increased by \$220.

## 8 | P A G E

Regression Statis	tics							
Multiple R	0.9205436							
R Square	0.84740053							
Adjusted R Square	0.82298461							
Standard Error	1.39577331							
Observations	30							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	4	270.4615184	67.6153796	34.706891	7.22088E-10			
Residual	25	48.70457828	1.94818313					
Total	29	319.1660967						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-1.6685399	1.408315529	-1.18477702	0.24724684	-4.56902	1.23194026	-4.56902	1.23194026
Population (000,000)	0.55190731	0.050629062	10.9009981	5.468E-11	0.447634803	0.65617981	0.4476348	0.65617981
Percent Unemployed	0.20316264	0.117086169	1.73515493	0.09502688	-0.03798084	0.44430612	-0.03798084	0.44430612
Advertising Expense (000)	0.03135496	0.016062075	1.95211152	0.06221113	-0.0017255	0.06443543	-0.0017255	0.06443543
Mall Location	0.21979032	0.540028513	0.40699763	0.68747343	-0.89241922	1.33199987	-0.89241922	1.33199987

## D. QUESTION 3: WHAT IS THE COEFFICIENT OF MULTIPLE DETERMINATION (MULITPLE R<sup>2</sup>)?

Coefficient of Multiple Determination	$R^2 = \frac{SSR}{SS \ total}$
---------------------------------------	--------------------------------

Figure 2: R<sup>2</sup> (SUSS, 2014)

- R<sup>2</sup> must always be between 0 and 1, inclusive.
- That is,  $0 \le \mathbb{R}^2 \le 1$ .
- The closer R<sup>2</sup> is to 1.0, the stronger the association between Y and the set of independent variables, X1, X2, X3.
- For example, if R<sup>2</sup> = 0.92 for the Y hat equation given above, that means that X1, X2 and X3 account for 92 percent of the variation of Y hat.

## 9 | P A G E

## E. QUESTION 4: WHAT IS THE ADJUSTED COEFFICIENT OF MULTIPLE DETERMINATION (ADJUSTED MULITPLE R<sup>2</sup>)?

SUMMARY OUTPUT								
Regression Statis	tics							
Multiple R	0.9205436							
R Square	0.84740053							
Adjusted R Square	0.82298461							
Standard Error	1.39577331							
Observations	30							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	4	270.4615184	67.6153796	34.706891	7.22088E-10			
Residual	25	48.70457828	1.94818313					
Total	29	319.1660967						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-1.6685399	1.408315529	-1.18477702	0.24724684	-4.56902	1.23194026	-4.56902	1.23194026
Population (000,000)	0.55190731	0.050629062	10.9009981	5.468E-11	0.447634803	0.65617981	0.4476348	0.65617981
Percent Unemployed	0.20316264	0.117086169	1.73515493	0.09502688	-0.03798084	0.44430612	-0.03798084	0.44430612
Advertising Expense (000)	0.03135496	0.016062075	1.95211152	0.06221113	-0.0017255	0.06443543	-0.0017255	0.06443543
Mall Location	0.21979032	0.540028513	0.40699763	0.68747343	-0.89241922	1.33199987	-0.89241922	1.33199987

Adjusted Coefficient of Determination	$R_{adj}^{2} = 1 - \frac{\frac{SSE}{n - (k+1)}}{\frac{SStotal}{n-1}}$
---------------------------------------	---

Figure 3: R<sup>2</sup>adj (SUSS, 2014)

- $R^2_{adj}$  is needed because  $R^2$  is not very accurate.
- $R^2_{adj}$  is more effective than  $R^2$ .
- That's because as the number of independent variables, X1, X2, X3 etc... increases, R<sup>2</sup> increases.
- But if the independent variable is not a good predictor, it still increases R<sup>2</sup>.
- This makes R<sup>2</sup> inaccurate.
- $R^{2}_{adj}$  will not necessarily increase when a new variable is added to the model.
- Here,  $R^2_{adj} = 0.822$

10 | PAGE

#### F. QUESTION 4: WHAT IS THE MULTIPLE STANDARD ERROR OF ESTIMATE (SYX)?

ics							
0.9205436							
0.84740053							
0.82298461							
1.39577331							
30							
df	SS	MS	F	Significance F			
4	270.4615184	67.6153796	34.706891	7.22088E-10			
25	48.70457828	1.94818313					
29	319.1660967						
Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
-1.6685399	1.408315529	-1.18477702	0.24724684	-4.56902	1.23194026	-4.56902	1.23194026
0.55190731	0.050629062	10.9009981	5.468E-11	0.447634803	0.65617981	0.4476348	0.65617981
0.20316264	0.117086169	1.73515493	0.09502688	-0.03798084	0.44430612	-0.03798084	0.44430612
0.03135496	0.016062075	1.95211152	0.06221113	-0.0017255	0.06443543	-0.0017255	0.06443543
0.21979032	0.540028513	0.40699763	0.68747343	-0.89241922	1.33199987	-0.89241922	1.33199987
	ics 0.9205436 0.84740053 0.82298461 1.39577331 30 <i>df</i> <i>df</i> 4 25 29 <i>Coefficients</i> -1.6685399 0.55190731 0.20316264 0.03135496 0.21979032	ics         Image: style sty	ics         Instant           0.9205436         Instant           0.84740053         Instant           0.82298461         Instant           1.39577331         Instant           300         Instant           319.166097         Instant           Coefficients         Standard Error         Instant           1.6685399         1.408315529         1.18477702           0.55190731         0.050629062         10.9009881           0.20316264         0.117086169         1.73515493           0.3135496         0.016062075         1.95211152           0.21979032         0.540028513         0.40699763	ics         Instant         Instant           0.9205436         Instant         Instant           0.84740053         Instant         Instant           0.82298461         Instant         Instant           0.80         Instant         Instant           0.80         Instant         Instant           0.80         Instant         Instant           0.81         Instant         Instant	icsIndexIndexIndexIndex0.9205436IndexIndexIndexIndex0.84740053IndexIndexIndexIndex0.82298461IndexIndexIndexIndex1.39577331IndexIndexIndexIndex30IndexIndexIndexIndex30IndexIndexIndexIndex30IndexIndexIndexIndex30IndexIndexIndexIndex30IndexIndexIndexIndex30IndexIndexIndexIndex30IndexIndexIndexIndex30IndexIndexIndexIndex319.166096IndexIndexIndexIndex319.166096IndexIndexIndexIndex319.166096IndexIndexIndexIndex319.166096IndexIndexIndexIndex319.166096IndexIndexIndexIndex319.166096IndexIndexIndexIndex319.166096IndexIndexIndexIndex319.166096IndexIndexIndexIndex319.166096IndexIndexIndexIndex319.166096IndexIndexIndexIndex319.166096IndexIndexIndexIndex319.166096IndexIndexIndexIndex31	icsIndexIndexIndexIndexIndex0.9205436IndexIndexIndexIndexIndex0.84740053IndexIndexIndexIndexIndex0.82298461IndexIndexIndexIndexIndex0.82298461IndexIndexIndexIndexIndex0.82298461IndexIndexIndexIndexIndex1.39577331IndexIndexIndexIndexIndex30IndexIndexIndexIndexIndexIndex30IndexIndexIndexIndexIndexIndex30IndexIndexIndexIndexIndexIndex30IndexIndexIndexIndexIndexIndex30IndexIndexIndexIndexIndexIndex30IndexIndexIndexIndexIndexIndex30IndexIndexIndexIndexIndexIndex30IndexIndexIndexIndexIndexIndex319IndexIndexIndexIndexIndexIndex319IndexIndexIndexIndexIndexIndex319IndexIndexIndexIndexIndexIndex319IndexIndexIndexIndexIndexIndex319IndexIndexIndexIndexIndexIndex <tr< td=""><td>icsIndexIndexIndexIndexIndexIndex0.9205436IndexIndexIndexIndexIndexIndex0.84740053IndexIndexIndexIndexIndexIndex0.82298461IndexIndexIndexIndexIndexIndex0.82298461IndexIndexIndexIndexIndexIndex0.82298461IndexIndexIndexIndexIndexIndex1.39577331IndexIndexIndexIndexIndexIndex30IndexIndexIndexIndexIndexIndex30IndexIndexIndexIndexIndexIndex30IndexIndexIndexIndexIndexIndex30IndexIndexIndexIndexIndexIndex30IndexIndexIndexIndexIndexIndex30IndexIndexIndexIndexIndexIndex30IndexIndexIndexIndexIndexIndex30IndexIndexIndexIndexIndexIndex30IndexIndexIndexIndexIndexIndex31IndexIndexIndexIndexIndexIndex31IndexIndexIndexIndexIndexIndex31IndexIndexIndexIndexIndexIndex31</td></tr<>	icsIndexIndexIndexIndexIndexIndex0.9205436IndexIndexIndexIndexIndexIndex0.84740053IndexIndexIndexIndexIndexIndex0.82298461IndexIndexIndexIndexIndexIndex0.82298461IndexIndexIndexIndexIndexIndex0.82298461IndexIndexIndexIndexIndexIndex1.39577331IndexIndexIndexIndexIndexIndex30IndexIndexIndexIndexIndexIndex30IndexIndexIndexIndexIndexIndex30IndexIndexIndexIndexIndexIndex30IndexIndexIndexIndexIndexIndex30IndexIndexIndexIndexIndexIndex30IndexIndexIndexIndexIndexIndex30IndexIndexIndexIndexIndexIndex30IndexIndexIndexIndexIndexIndex30IndexIndexIndexIndexIndexIndex31IndexIndexIndexIndexIndexIndex31IndexIndexIndexIndexIndexIndex31IndexIndexIndexIndexIndexIndex31

Multiple Standard Error of Estimate	$s_{y,12k} = \sqrt{\frac{(Y - \hat{Y})^2}{n - (k+1)}}$
-------------------------------------	--

where

- Y is the observation.
- $\hat{Y}$  is the value estimated from the regression equation.
- *n* is the number of observations in the sample.
- *k* is the number of independent variables.
- $s_{y\cdot 12\cdots k}$  is the standard error of estimate. The subscripts indicate the number of independent variables being used to estimate the value of *Y*.

$$s_{y.123..k} = \sqrt{\frac{SSE}{n - (k+1)}}$$

Figure 4: Multiple Standard Error of Estimate Equation (SUSS, 2014)

- sy.1.2...k Measures the error of Y hat.
- That is, it measures the error between Actual Y and Y hat.
- Y hat is the predicted value of the dependent variable.

## 11 | PAGE

- $S_{y,x}$ : Standard Error of Estimate = 1.396 x 10<sup>3</sup>
- $S_{y,x}$  is a measure of dispersion of values around the regression line
- $S_{y,x}$ : is the Standard Deviation of the Residuals away from the proposed line.
- Figure 5 shows what Residuals are.



Figure 5: What are Residuals?

- Small  $S_{y,x}$  = Lesser Scatter = Good predictor
- Big  $S_{y,x}$  = More Scatter = Bad predictor
- Similar to Multiple R or r, both measures strength of relationship between X and Y
- But  $S_{y.x}$  has same units as Y, Multiple R or r has range -1 to 1
- Since S<sub>y.x</sub> = 1.396, this shows that about 68% of the predictions should be within ±1.396 x 10<sup>3</sup> (±1σ) of the actual repair costs and about 95% should be within (1.396 x 10<sup>3</sup> x 2) = ±2.792 (±2σ) of actual repair costs.

#### G. QUESTION 5: WHAT IS THE ANOVA TABLE?

SOMMARY COTFOT								
Regression Statis	tics							
Multiple R	0.9205436							
R Square	0.84740053							
Adjusted R Square	0.82298461							
Standard Error	1.39577331							
Observations	30							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	4	270.4615184	67.6153796	34.706891	7.22088E-10			
Residual	25	48.70457828	1.94818313					
Total	29	319.1660967						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-1.6685399	1.408315529	-1.18477702	0.24724684	-4.56902	1.23194026	-4.56902	1.23194026
Population (000,000)	0.55190731	0.050629062	10.9009981	5.468E-11	0.447634803	0.65617981	0.4476348	0.65617981
Percent Unemployed	0.20316264	0.117086169	1.73515493	0.09502688	-0.03798084	0.44430612	-0.03798084	0.44430612
Advertising Expense (000)	0.03135496	0.016062075	1.95211152	0.06221113	-0.0017255	0.06443543	-0.0017255	0.06443543
Mall Location	0.21979032	0.540028513	0.40699763	0.68747343	-0.89241922	1.33199987	-0.89241922	1.33199987

#### Analysis of Variance

Source	df	SS	MS	F
Regression	k	SSR	MSR = SSR/k	MSR / MSE
Error	n - (k + 1)	SSE	MSE = SSE/[n - (k + 1)]	
Total	<i>n</i> – 1	SS total		

Figure 6: ANOVA (SUSS, 2014)

Total variation = SS total =  $\sum (Y - \overline{Y})^2$ Error variation = SSE =  $\sum (Y - \hat{Y})^2$ Regression variation = SSR =  $\sum (\hat{Y} - \overline{Y})^2$  = (SS total – SSE)

Figure 7: SS Total, SSE, SSR (SUSS, 2014)

## 13 | PAGE

Clobal Test	$E = \frac{MSR}{MSR} = \frac{SSR/k}{N}$
Global Test	$T = \frac{1}{MSE} = \frac{1}{SSE/[n-(k+1)]}$

#### where

SSR is the sum of the squares "explained by" the regression.

- *k* is the number of independent variables.
- SSE is the sum of squares error.
- *n* is the number of observations.

Figure 8: Global F Test Equation (SUSS, 2014)

- Key purpose of the ANOVA table is to calculate the F statistic.
- Here, F statistic = 34.7
- Comparing this to F critical (alpha = 5%; numerator (regression df) =4; denominator (residual df) = 25) → Referring to F table → F critical = 2.76
- Since F statistic > F critical  $\rightarrow$  Accept H1

#### H. QUESTION 6: WHAT IS THE SIGNIFICANCE F?

SUMMARY OUTPUT								
Regression Statist	tics							
Multiple R	0.9205436							
R Square	0.84740053							
Adjusted R Square	0.82298461							
Standard Error	1.39577331							
Observations	30							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	4	270.4615184	67.6153796	34.706891	7.22088E-10			
Residual	25	48.70457828	1.948183131					
Total	29	319.1660967						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-1.6685399	1.408315529	-1.18477702	0.24724684	-4.569019996	1.23194026	-4.56902	1.231940255
Population (000,000)	0.55190731	0.050629062	10.90099815	5.468E-11	0.447634803	0.65617981	0.447634803	0.656179811
Percent Unemployed	0.20316264	0.117086169	1.735154932	0.09502688	-0.037980835	0.44430612	-0.03798084	0.444306121
Advertising Expense (000)	0.03135496	0.016062075	1.952111518	0.06221113	-0.001725501	0.06443543	-0.0017255	0.064435426
Mall Location	0.21979032	0.540028513	0.406997626	0.68747343	-0.89241922	1.33199987	-0.89241922	1.331999866

- Significance  $F = 7.22 \times 10^{-10}$
- This is actually a p-value used for Hypothesis Testing.
- In Ang (2019b), I mentioned about the Z and t test for Hypothesis Testing, but left out the F test.
- I will mention it here.

#### 1. THE GLOBAL F TEST

- In Ang (2019b), there is a standard 5 step procedure for Hypothesis Testing.
- It applies here as well.
- Step 1: State the Null and Alternate Hypothesis

 $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$  $H_1:$  At least one of the  $\beta$ 's is not zero.

### 15 | PAGE

- ο β1 refers to X1 (population)
- $\circ$   $\beta$ 2 refers to X2 (percent unemployed)
  - o β3 refers to X3 (advertising expense)
  - ο β4 refers to X4 (mall location)
  - 0 Thus, what H0 means is that  $\beta 1$  /  $\beta 2$  /  $\beta 3$  /  $\beta 4$  are all not important.
  - In other words, Population / Percent Unemployed / Advertising Expense / Mall location all does not affect Sales and are insignificant.
  - Which also means that the MR Equation formulated above is USELESS  $\rightarrow$  Since all the factors can't affect sales at all.
  - While H1 represents that at LEAST ONE of the factor is important and will significiantly affect sales.
  - $\circ~$  We will not know which factor is important ( $\beta1~/~\beta2~/~\beta3~/~\beta4$ ), but we know that at LEAST one of them will be significant.
- Step 2: State the Level of Significance, Alpha,  $\alpha = 5\%$
- Step 3: State the Test Statistic  $\rightarrow$  Global F Test
- Step 4: Formulate the Decision Rule



Figure 9: F Distribution 5%

- P-value is like a "disease"... it is represented by the orange color area in Figure 9.
- The Alpha area is represented by the brown color area, which takes up 5%.
- O If the p-value area is very small < 5%, it will not infiltrate the H0 area, but stay in H1 area → Thus H1 is accepted.</li>
- But if the p-value is large > 5%, it will infiltrate the H0 area  $\rightarrow$  Thus H0 is accepted.
- Since the p-value i.e. Significance  $F = 7.22 \times 10^{-10} < 5\%$
- Thus we accept H1  $\rightarrow$  The equation is important!

SUMMARY OUTPUT								
Regression Statist	Regression Statistics							
Multiple R	0.9205436							
R Square	0.84740053							
Adjusted R Square	0.82298461							
Standard Error	1.39577331							
Observations	30							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	4	270.4615184	67.6153796	34.706891	7.22088E-10			
Residual	25	48.70457828	1.94818313					
Total	29	319.1660967						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-1.6685399	1.408315529	-1.18477702	0.24724684	-4.56902	1.23194026	-4.56902	1.23194026
Population (000,000)	0.55190731	0.050629062	10.9009981	5.468E-11	0.447634803	0.65617981	0.4476348	0.65617981
Percent Unemployed	0.20316264	0.117086169	1.73515493	0.09502688	-0.03798084	0.44430612	-0.03798084	0.44430612
Advertising Expense (000)	0.03135496	0.016062075	1.95211152	0.06221113	-0.0017255	0.06443543	-0.0017255	0.06443543
Mall Location	0.21979032	0.540028513	0.40699763	0.68747343	-0.89241922	1.33199987	-0.89241922	1.33199987

#### I. QUESTION 7: WHAT ARE THE INDIVIDUAL P VALUES?

#### 1. INDIVIDUAL F TEST

- In Ang (2019b), there is a standard 5 step procedure for Hypothesis Testing.
- It applies here as well.
- Step 1: State the Null and Alternate Hypothesis

For Population	For % Unemployed	For Advertising	For Mall
$H_0: \beta_1 = 0$	$H_0:\beta_2=0$	$H_0:\beta_3=0$	$H_0:\beta_4=0$
$H_1: \beta_1 \neq 0$	$H_1: \beta_2 \neq 0$	$H_1: \beta_3 \neq 0$	$H_1: \beta_4 \neq 0$

- ο  $\beta$ 1 refers to X1 (population)
- $\circ$   $\beta$ 2 refers to X2 (percent unemployed)

- o β3 refers to X3 (advertising expense)
- $\circ$   $\beta$ 4 refers to X4 (mall location)
- 0 Thus, what H0 means is that  $\beta 1 / \beta 2 / \beta 3 / \beta 4$  are individually not important.
  - In other words, Population / Percent Unemployed / Advertising Expense / Mall location, individually tested, does not affect Sales and is insignificant.
  - While H1 represents that the particular factor is important and will significiantly affect sales.
- Step 2: State the Level of Significance, Alpha,  $\alpha = 5\%$
- Step 3: State the Test Statistic  $\rightarrow$  Individual F Test
- Step 4: Formulate the Decision Rule



Figure 10: F Distribution 5%

- o If the p-value area is very small < 5%, it will not infiltrate the H0 area, but stay in H1 area → Then H1 is accepted.</li>
- Step 5: Make the Decision
  - P-value for population =  $5.5 \times 10^{-11} < 5\%$  → Accept H1 → Retain
  - P-value for percent unemployed =  $0.1 > 5\% \rightarrow$  Accept H0  $\rightarrow$  Drop off
  - P-value for advertising expense =  $0.062 > 5\% \rightarrow \text{Accept H0} \rightarrow \text{Drop off}$

- P-value for mall location =  $0.687 > 5\% \rightarrow$  Accept H0  $\rightarrow$  Drop off
- Conclusion:
  - Only population is important.
  - Unemployment, advertising expense and mall location are all not important and can be dropped off
  - However, only drop off 1 variable at a time.
  - This is because each time you drop off an insignificant variable, another variable may suddenly become important.
  - Also, re-run the model after dropping each insignificant variable one at a time.

SUMINARY OUTPUT								
Regression Statis	tics							
Multiple R	0.9205436							
R Square	0.84740053							
Adjusted R Square	0.82298461							
Standard Error	1.39577331							
Observations	30							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	4	270.4615184	67.6153796	34.706891	7.22088E-10			
Residual	25	48.70457828	1.94818313					
Total	29	319.1660967						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-1.6685399	1.408315529	-1.18477702	0.24724684	-4.56902	1.23194026	-4.56902	1.23194026
Population (000,000)	0.55190731	0.050629062	10.9009981	5.468E-11	0.447634803	0.65617981	0.4476348	0.65617981
Percent Unemployed	0.20316264	0.117086169	1.73515493	0.09502688	-0.03798084	0.44430612	-0.03798084	0.44430612
Advertising Expense (000)	0.03135496	0.016062075	1.95211152	0.06221113	-0.0017255	0.06443543	-0.0017255	0.06443543
Mall Location	0.21979032	0.540028513	0.40699763	0.68747343	-0.89241922	1.33199987	-0.89241922	1.33199987

#### J. QUESTION 8: WHAT ARE THE INDIVIDUAL T STATS?

- The t stat meant to be used for individual hypothesis testing
- Purpose is to test whether each variable is significant or not.
- However, since we already did the p-test in the previous section, we may skip using the t stat.

~ ~ ~	U	~	U	L.		J		1
SUMMARY OUTPUT								
Regression Statis	tics							
Multiple R	0.9205436							
R Square	0.84740053							
Adjusted R Square	0.82298461							
Standard Error	1.39577331							
Observations	30							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	4	270.4615184	67.6153796	34.706891	7.22088E-10			
Residual	25	48.70457828	1.94818313					
Total	29	319.1660967						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-1.6685399	1.408315529	-1.18477702	0.24724684	-4.56902	1.23194026	-4.56902	1.23194026
Population (000,000)	0.55190731	0.050629062	10.9009981	5.468E-11	0.447634803	0.65617981	0.4476348	0.65617981
Percent Unemployed	0.20316264	0.117086169	1.73515493	0.09502688	-0.03798084	0.44430612	-0.03798084	0.44430612
Advertising Expense (000)	0.03135496	0.016062075	1.95211152	0.06221113	-0.0017255	0.06443543	-0.0017255	0.06443543
Mall Location	0.21979032	0.540028513	0.40699763	0.68747343	-0.89241922	1.33199987	-0.89241922	1.33199987

#### K. QUESTION 9: WHAT ARE THE LOWER AND UPPER 95%?

- The given coefficients are:
  - $\circ$  Population = 0.55
  - $\circ$  Percent Unemployed = 0.2
  - $\circ$  Advertising Expense = 0.03
  - $\circ$  Mall Location = 0.22
- However, this is not 100% accurate. Thus, a 95% confidence interval is attached to them.
- This means that the lower and upper limits (95%) are:
  - Population = between 0.448 and 0.656
  - Percent Unemployed = between -0.04 and 0.444

22 | P A G E

- Advertising Expense = between -0.002 and 0.064
- Mall Location = between -0.892 and 1.33

#### III. MR ASSUMPTIONS

#### A. ASSUMPTION 1: LINEARITY

- LINEARITY: There must be a linear relationship between the dependent variable, Y, and each independent variable, X1, X2 etc..
- We check this by using a *Line Fit Plot:*

Flash Fill   Text to   Columns   Data Validation   Data Tools     Analysis     Plo   Histogram   Moving Average   Random Number Generation   Rank and Percentile   Regression     Lampling	FOR	RMULAS	DATA	REVIEW	VIEW			Sign
Data Analysis ? X Analysis Tools OK Histogram Moving Average Random Number Generation Bank and Percentile Regression Lampling	Text 1 Colum	nalysis vsis	~					
t-Test: Paired Two Sample for Means t-Test: Two-Sample Assuming Equal Variances t-Test: Two-Sample Assuming Unequal Variances z-Test: Two Sample for Means	plo	ata Analysi Analysis To Histogram Moving Av Random N Bank and Regression Sampling t-Test: Pain t-Test: Two t-Test: Two z-Test: Two z-Test: Two	s iols /erage lumber Gene Percentile n red Two Sam o-Sample As o-Sample As o Sample fo	eration nple for Mear suming Equa suming Uneo r Means	ns al Variances qual Varian	ces v	? OK Canc <u>H</u> elp	el

Sales (000)	Population (000,000)	Percent Unemployed	Advertising Expense (000)	Mall Location	
5.17	7.5	5.1	59	0	Regression ? X
5.78	8.71	6.3	62.5	0	dinput
4.84	10	4.7	61	0	Input Y Bange: SA\$1:\$A\$31
6	7.45	5.4	61	1	Cancel
6	8.67	5.4	61	1	Input X Range: SBS1:SBS31
6.12	11	7.2	12.5	0	Help
6.4	13.18	5.8	35.8	0	✓ Labels Constant is Zero
7.1	13.81	5.8	59.9	0	Confidence Level: 95 %
8.5	14.43	6.2	57.2	1	Output options
7.5	10	5.5	35.8	0	
9.3	13.21	6.8	27.9	0	O Output Range:
8.8	17.1	6.2	24.1	1	New Worksheet <u>Ply:</u>
9.96	15.12	6.3	27.7	1	O New Workbook
9.83	18.7	0.5	24	0	Residuals
10.12	20.2	5.5	57.2	1	Residuals Residual Plots
10.7	15	5.8	44.3	0	Standardized Residuals ✓ Line Fit Plots
10.45	17.6	7.1	49.2	0	Namel Brahability
11.32	19.8	7.5	23	0	
11.87	14.4	8.2	62.7	1	
11.91	20.35	7.8	55.8	0	
12.6	18.9	6.2	50	0	
12.6	21.6	7.1	47.6	1	
14.24	25.25	0.4	43.5	0	
14.41	27.5	4.2	55.9	0	
13.73	21	0.7	51.2	1	
13.73	19.7	6.4	76.6	1	
13.8	24.15	0.5	63	1	
14.92	17.65	8.5	68.1	0	
15.28	22.3	7.1	74.4	1	
14.41	24	0.8	70.1	0	









- We produce 4 Line Fit Plots for the 4 Variables.
- We see that for all of them, a linear assumption is valid.
- Except for Mall Location, because it is a 0/1, we simply assume its linear.

#### **B. ASSUMPTION 2: HOMOSCEDASTICITY**

• HOMOSCEDASTICITY: The residuals (Figure 11) must exhibit Homoscedasticity.





Figure 11: What are Residuals?

• We use a *Residual Plot* to check Homoscedacticity.

FO	RMULAS DATA	REVIEW	VIEW			Sign
Text Colu	Flash Fill Flash Fill Fto mns S Data Validati	licates <sup>II</sup> ? ▼ on ▼ ¤	Outline	Solver Data A	nalysis	
[	Data Tools Data Analysis			Analy	sis ?	×
plo	Analysis Tools Histogram Moving Average Random Number Ge Rank and Percentile Regression	^	OK Cance <u>H</u> elp			
	t-Test: Paired Two Si t-Test: Two-Sample t-Test: Two-Sample z-Test: Two Sample 5.8					

	А	В	С	D	E	F G H I J K L
1	Sales (000)	Population (000,000)	Percent Unemployed	Advertising Expense (000)	Mall Location	
2	5.17	7.5	5.1	59	0	Regression ? X
3	5.78	8.71	6.3	62.5	0	Input
4	4.84	10	4.7	61	0	Input Y Range: CAS1-SAS31
5	6	7.45	5.4	61	1	Cancel
6	6	8.67	5.4	61	1	Input X Range: \$B\$1:\$E\$31
7	6.12	11	7.2	12.5	0	Help
8	6.4	13.18	5.8	35.8	0	Constant is Zero
9	7.1	13.81	5.8	59.9	0	Confidence Level: 95 %
10	8.5	14.43	6.2	57.2	1	Output aptions
11	7.5	10	5.5	35.8	0	
12	9.3	13.21	6.8	27.9	0	O Output Range:
13	8.8	17.1	6.2	24.1	1	New Worksheet <u>Ply:</u>
14	9.96	15.12	6.3	27.7	1	🔿 New <u>W</u> orkbook
15	9.83	18.7	0.5	24	0	Residuals
16	10.12	20.2	5.5	57.2	1	Residuals
17	10.7	15	5.8	44.3	0	Standardized Residuals
18	10.45	17.6	7.1	49.2	0	
19	11.32	19.8	7.5	23	0	Normal Probability
20	11.87	14.4	8.2	62.7	1	
21	11.91	20.35	7.8	55.8	0	
22	12.6	18.9	6.2	50	0	
23	12.6	21.6	7.1	47.6	1	
24	14.24	25.25	0.4	43.5	0	
25	14.41	27.5	4.2	55.9	0	
26	13.73	21	0.7	51.2	1	
27	13.73	19.7	6.4	76.6	1	
28	13.8	24.15	0.5	63	1	
29	14.92	17.65	8.5	68.1	0	
30	15.28	22.3	7.1	74.4	1	
31	14.41	24	0.8	70.1	0	



Figure 12: Individual Residual Plots for Each Variable

- Figure 12 shows Individual Residual Plots for each Variable.
- We see that they are scattered randomly and uniformly around the 0 line.
- This means that individually, they are Homoscedastictiy.
- But we are not done. We need to check Overall Homoscedasticity.

_									DOORT	(
IN	SERT PA	AGE LAYOUT	FORM	JLAS DATA	REVIEW VIE	N				
nded les	Table Pic	tures Online Picture	Shapes S	martArt Screenshot	🗎 Store	Recommende Charts	• ∎∎ • ﷺ • ♦	■ • 索 • ≪ • ìì •	PivotChart	Line
			Illustratio	ns	Add-ins		Cha	Scatter		S
: 🕽	× ✓ c	fx D	E	F G	н	1	,	••••		<u>M</u>
								<b>X</b>	$\succ$	_
νUT								Bubble		
Sales	Residuals									
6837	-0.18684							Kil More	Scatter Char	te
78183	-0.59818							Pere More	Seated Char	com -
71000	1 07005									

30 | P A G E

RESIDUAL	OUTPUT		
bservatio.	cted Sales	Residuals	
1	5.356837	-0.18684	Chart Title
2	6.378183	-0.59818	4
3	6.71805	-1.87805	
4	5.672691	0.327309	3
5	6.346018	-0.34602	
6	6.257149	-0.13715	
7	7.906449	-1.50645	1
8	9.009806	-1.90981	
9	9.568385	-1.06839	
10	6.090435	1.409565	
11	7.878465	1.421535	
12	10.00413	-1.20413	-2
13	9.044546	0.915454	• • • •
14	9.506227	0.323773	-3
15	12.61068	-2.49068	
16	9.177438	1.522562	Edit Series ? X
17	11.03015	-0.58015	Series <u>n</u> ame:
18	11.50411	-0.18411	Residuals vs Predicted Sales 🔣 Select Range
19	10.13061	1.739395	Series <u>X</u> values:
20	12.89705	-0.98705	= Sheet4!\$B\$27:\$B\$57 💽 = Predicted Sale
21	11.58986	1.010135	Series <u>Y</u> values:
22	13.4074	-0.8074	=Sheet4!\$C\$27:\$C\$57
23	13.71233	0.527674	
24	16.11494	-1.70494	
25	11.88889	1.841108	
26	13.12586	0.604145	
27	13.95676	-0.15676	
28	11.93478	2.98522	
29	14.63405	0.645953	
30	13.93775	0.472252	



Figure 13: Overall Resdial Plot

- Figure 13 shows the Overall Residual Plot. The Residuals vs Predicted Sales.
- Figure 13 was gotten from drawing a scatter plot from the *Residual Output*.
- Since Figure 13 shows that the residuals are uniformly randomly scattered around 0, this means that we can assume Overall Homoscedasitity.

#### C. ASSUMPTION 3: NORMALLY DISTRIBUTED

- The Residuals need to follow the Normal Distribution.
- We plot a Histogram to check this.



## 33 | P A G E





- Figure 14 shows the final Histogram plotted of the residuals.
- It looks Normally Distributed, thus the Residuals are assumed Normally Distributed.

#### D. ASSUMPTION 4: MULTICOLLINEARITY

- Multicollinearity must NOT exist.
- Multicollinearity = Correlation between independent variables. Example, X1 is related to X2.
- Perhaps if X1 increase, X2 increases also.
- Then either X1 or X2 must be removed from the model.
- Variance Inflation Factor (VIF) is used to measure Multicollinearity.
- It is computed for each independent variable, Xi, i = 1, 2, etc...
- IF VIF > 10 OR R2 for that independent variable >= 0.9 --> UNSATISFACTORY. Independent varable must be removed.



Figure 15: VIF (SUSS, 2014)

• There is no short cut to measure VIF in excel, thus we need to get the R<sup>2</sup> for each and every variable in order to get their VIF.



35 | P A G E

#### 1. OBTAINING THE R2 FOR POPULATION

	-	-	-	-	
Sales (000)	Population (000,000)	Percent Unemployed	Advertising Expense (000)	Mall Location	
5.17	7.5	5.1	59	0	Regression ? X
5.78	8.71	6.3	62.5	0	Input
4.84	10	4.7	61	0	Input Y Range: SBS1:SBS31
6	5 7.45	5.4	61	1	Cancel
6	8.67	5.4	61	1	Input X Range: SC\$1:\$E\$31
6.12	11	7.2	12.5	0	Help
6.4	13.18	5.8	35.8	0	
7.1	. 13.81	5.8	59.9	0	Con <u>f</u> idence Level: 95 %
8.5	i 14.43	6.2	57.2	1	Output options
7.5	i 10	5.5	35.8	0	
9.3	13.21	6.8	27.9	0	
8.8	3 17.1	6.2	24.1	1	New Worksheet <u>Ply:</u>
9.96	5 15.12	6.3	27.7	1	O New Workbook
9.83	18.7	0.5	24	0	Residuals
10.12	20.2	5.5	57.2	1	Residuals Residual Plots
10.7	15	5.8	44.3	0	Standardized Residuals Line Fit Plots
10.45	5 17.6	7.1	49.2	0	Normal Probability
11.32	19.8	7.5	23	0	Normal Probability Plots
11.87	14.4	8.2	62.7	1	
11.91	. 20.35	7.8	55.8	0	
12.6	5 18.9	6.2	50	0	
12.6	5 21.6	7.1	47.6	1	
14.24	25.25	0.4	43.5	0	
14.41	. 27.5	4.2	55.9	0	
13.73	21	0.7	51.2	1	
13.73	19.7	6.4	76.6	1	
13.8	24.15	0.5	63	1	
14.92	17.65	8.5	68.1	0	
15.28	3 22.3	7.1	74.4	1	
14.41	. 24	0.8	70.1	0	

- Take note of something weird here!
- Input Y Range = Population
- Input X Range = Percent unemployed / Advertising / Mall Location!

SUMMARY OUTPUT	
Regression St	atistics
Multiple R	0.39026999
R Square	0.152310665
Adjusted R Square	0.054500357
Standard Error	5.406651004
Observations	30

## 36 | PAGE

#### 2. OBTAINING THE R2 FOR PERCENT UNEMPLOYED

		-	-	-	 -	
	Sales (000)	Population (000,000)	Advertising Expense (0	Mall Location	Percent Unemployed	
	5.17	7.5	59	0	5.1	Regression ? ×
	5.78	8.71	62.5	0	6.3	Input
	4.84	10	61	. 0	4.7	Input Y Range: SGS1:SGS31
	6	7.45	61	1	5.4	Cancel
	6	8.67	61	1	5.4	Input X Range: \$B\$1:\$D\$31
	6.12	11	12.5	0	7.2	Help
	6.4	13.18	35.8	0	5.8	
	7.1	13.81	59.9	0	5.8	Con <u>r</u> idence Level: 95 %
)	8.5	14.43	57.2	1	6.2	Output options
	7.5	10	35.8	0	5.5	O Output Bangar
!	9.3	13.21	27.9	0	6.8	
ł	8.8	17.1	24.1	1	 6.2	New Worksheet Ply:
ł	9.96	15.12	27.7	1	6.3	O New Workbook
į	9.83	18.7	24	0	0.5	Residuals
j	10.12	20.2	57.2	1	5.5	Residuals Residual Plots
1	10.7	15	44.3	0	5.8	Standardized Residuals
5	10.45	17.6	49.2	0	7.1	Normal Probability
)	11.32	19.8	23	0	7.5	Normal Probability Plots
)	11.87	14.4	62.7	1	8.2	
	11.91	20.35	55.8	0	7.8	
!	12.6	18.9	50	0	6.2	
1	12.6	21.6	47.6	1	7.1	
ł	14.24	25.25	43.5	0	0.4	
į	14.41	27.5	55.9	0	4.2	
į	13.73	21	51.2	1	0.7	
1	13.73	19.7	76.6	1	6.4	
1	13.8	24.15	63	1	0.5	
1	14.92	17.65	68.1	0	8.5	
)	15.28	22.3	74.4	1	7.1	
	14.41	24	70.1	0	0.8	

- Take note of something weird here!
- Input Y Range = Percent unemployed
- Input X Range = Population / Advertising / Mall Location!

<b>M</b>	U
SUMMARY OUTPUT	
Regression Stati	istics
Multiple R	0.370977
R Square	0.137624
Adjusted R Square	0.038119
Standard Error	2.337882
Observations	30

• And so we repeat the same for Advertising and Mall Location.

## 37 | PAGE

• We end up with this:

Variable	$R_j^2$	$VIF_j = \frac{1}{1 - R_j^2}$
Population (000,000)	0.152311	1.179677
	0.137624	1.159587
Unemployed (%)		
Advertising (\$'000)	0.07783	1.084398
Mall Location	0.07218	1.077795

• We see that the VIF is less than 10 for each  $\rightarrow$  No Multicollinearity!

## 38 | PAGE

#### E. ASSUMPTION 5: AUTOCORRELATION

- Autocorrelation must NOT exist.
- Autocorrelation = Correlation between residuals after a long period of time.
- Residual plot is used to detect Autocorrelation.
- IF long run plot shows a pattern of residuals occuring, e.g. constantly above or below the horizontal line of Y hat = 0, then Autocorrelation has existed.

## 39 | P A G E

#### IV. FORWARD SELECTION

- Forward Selection means to step by step include one variable at a time into the equation.
- Then use  $R^2$  and adjusted  $R^2$  as indicators to check which variables should be included / excluded.
- In other words, rather than using the Individual F test to seek out which are the insignificant variables, and then dropping them off one by one, Forward Selection is an alternative method to refine the best model.

Example, we are given this data:

price	sg metres	size of garage	size of bedroom
. 65	1	0	2
73	1.1	0	2
85	1.15	1	2
87	1.4	0	3
98	1.7	1	3
105	1.8	1	4
95	1.9	0	3
125	1.9	1	4
125	2.1	2	4
137	2.1	2	4
150	2.3	2	4

- Y: Price
- X1: Sq Metres
- X2: Size of Garage
- X3: Size of Bedroom

#### A. STEP 1: CHECK OUT INDIVIDUAL R<sup>2</sup> AND ADJUSTED R<sup>2</sup>

• There are 3 possible models:

### 40 | PAGE

$$Y = a + bX_1$$
$$Y = a + bX_2$$
$$Y = a + bX_3$$



SUMMARY OUTPUT		_						]
Regression S	Statistics		sq metres line fit plot					
Multiple R	0.932676854	200	' 1					
R Square	0.869886115	.ë 100			•••			
Adjusted R Square	0.855429016	۹.		- C.		<ul> <li>price</li> </ul>		
Standard Error	10.2990493	0	-			Predi	cted price	
Observations	11		0 1 2 3					
			sq metres					
ANOVA								
	df	SS	MS	F	gnificance	F		
Regression	1	6382.275	6382.275	60.17017	2.83E-05			
Residual	9	954.6337	106.0704					
Total	10	7336.909						
	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.09	pper 95.09
Intercept	9.502604757	12.5832	0.755182	0.46944	-18.9626	37.96777	-18.9626	37.96777
sq metres	56.39411099	7.270146	7.756944	2.83E-05	39.9479	72.84032	39.9479	72.84032

Figure 16: Output from Sq Meters Regression

- We see from Figure 16 that the assumption of linearity is OK from the Line Fit Plot.
- We see that the R2 is 0.869, and the adjusted R2 is 0.855.
- If we were to repeat this process for the other variables, we will obtain:

Independent Variable in the Model	R <sup>2</sup>	Adjusted R <sup>2</sup>	Se	Parameter Estimates
X <sub>1</sub>	0.879	0.855	10.299	<i>b</i> <sub>0</sub> =9.503 <i>b</i> <sub>1</sub> =56.394
X <sub>2</sub>	0.759	0.731	14.030	<i>b</i> <sub>0</sub> =78.290 <i>b</i> <sub>1</sub> =28.382
X3	0.793	0.770	12.982	<i>b</i> <sub>0</sub> =16.250 <i>b</i> <sub>1</sub> =27.607

• We see that X1 gives the highest R2 (which means X1 is the most significant in affecting price), thus we should use (currently)

$$Y = 9.503 + 56.394X_1$$

## 42 | PAGE

#### B. STEP 2: CHECK OUT ALTERNATIVE R<sup>2</sup> AND ADJUSTED R<sup>2</sup>

• Now we have 2 new possible models:

$$Y = a + b_1X_1 + b_2X_2$$
  
 $Y = a + b_1X_1 + b_2X_3$ 

• After regression analysis...

Independent	$\mathbb{R}^2$	Adjusted R <sup>2</sup>	Se	Parameter
Variable in the				Estimates
Model				
X <sub>1</sub>	0.870	0.855	10.299	b <sub>0</sub> =9.503
				b <sub>1</sub> =56.394
$X_1$ and $X_2$	0.939	0.924	7.471	$b_0=27.684$
				b <sub>1</sub> =38.576
				$b_2 = 12.875$
$X_1$ and $X_3$	0.877	0.847	10.609	$b_0 = 8.311$
				b <sub>1</sub> =44.313
				b <sub>3</sub> =6.743

- After adding X2 / X3, we see that R2 increased highest for X2.
- However, sometimes R2 might be inaccurate.
- Thus, if we look at adjusted R2, once again, for the X2 case, the increment is the highest.
- Thus, the best model currently is  $Y = 27.684 + 38.576X_1 + 12.875X_2$

#### c. STEP 3: CHECK OUT THE FINAL R<sup>2</sup> AND ADJUSTED R<sup>2</sup>

• Now we have 1 possible model left:

$$Y = a + b_1 X_1 + b_2 X_3 + b_3 X_2$$

• After regression analysis...

43 | P A G E

Independent Variable in the Model	R <sup>2</sup>	Adjusted R <sup>2</sup>	S <sub>e</sub>	Parameter Estimates
$X_1$ and $X_2$	0.939	0.924	7.471	$b_0=27.684$ $b_1=38.576$ $b_2=12.875$
$X_1$ , $X_2$ and $X_3$	0.943	0.918	7.762	$\begin{array}{c} b_0 = 26.440 \\ b_1 = 30.803 \\ b_2 = 12.567 \\ b_3 = 4.576 \end{array}$

- We see that after including X3, R2 increased but Adjusted R2 dropped.
- What should we do?
- Since Adjusted R2 is a better gauge than R2, we should NOT include X3 in the model.
- Thus, the best model that should be used (finally) is
- $Y = 27.684 + 38.576X_1 + 12.875X_2$

#### V. REFERENCES

- Ang, A. (2019a). How to Perform Simple Linear Regression Using Excel. Singapore.
- Ang, A. (2019b). Hypothesis Testing. Singapore.
- SUSS. (2014). BUS105e Study Guide Business Statistics. Singapore: Singapore University of Social Sciences (SUSS).

## 45 | PAGE

#### VI. ABOUT DR. ALVIN ANG



Dr. Alvin Ang earned his Ph.D., Masters and Bachelor degrees from NTU, Singapore. He is a scientist, entrepreneur, as well as a personal/business advisor. More about him at <u>www.AlvinAng.sg</u>.

## 46 | P A G E