D R .   A L V I N ' S   P U B L I C A T I O N S

MULTIPLE REGRESSION

DR. ALVIN ANG

CONTENTS

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 → 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.

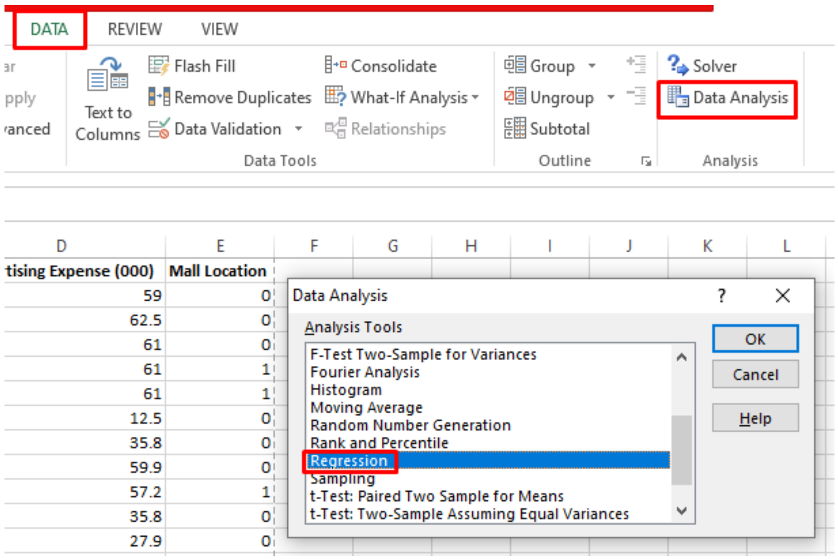
6. But we are still trying to "linearly regress" every Variable unto Y.

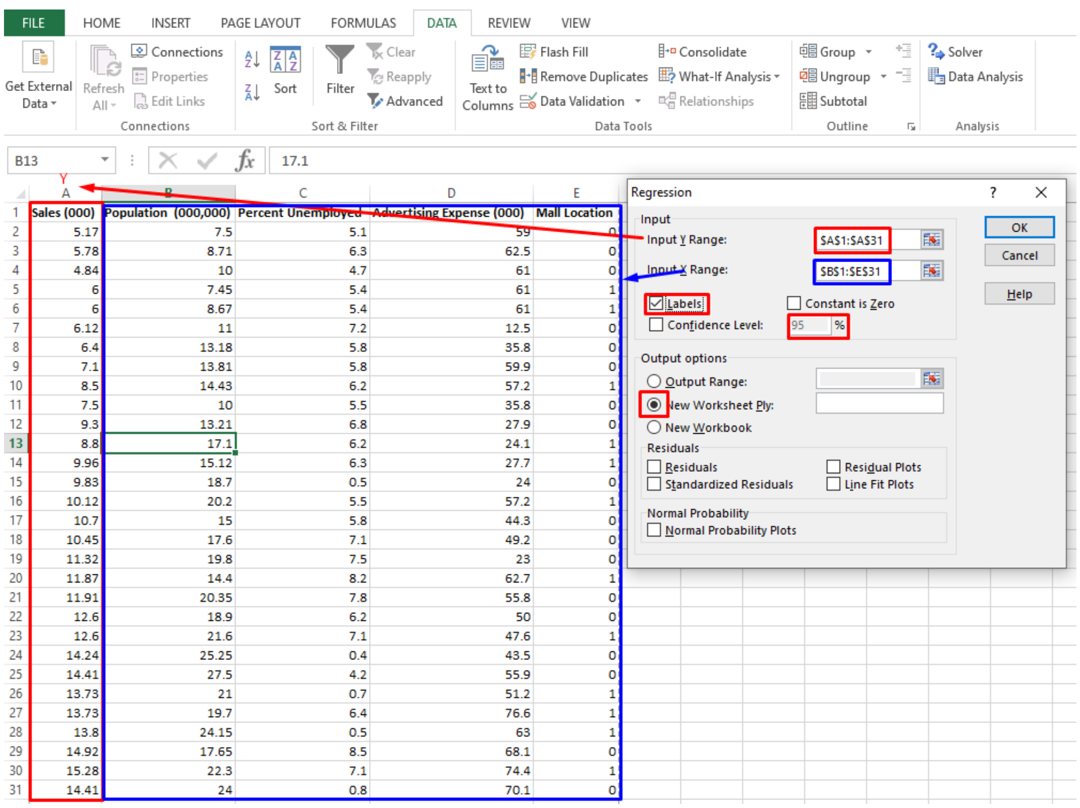
II. MR EXAMPLE

Given:

Sales (000)	Population (000,000)	Percent Unemployed	Advertising Expense (000)	Mall 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

A. USING EXCEL





	Sales (000)	Population (000,000)	Percent Unemployed	Advertising Expense (000)	Mall Location
2	5.17	7.5	5.1	59	0
3	5.78	8.71	6.3	62.5	0
4	4.84	10	4.7	61	0
5	6	7.45	5.4	61	1
6	6	8.67	5.4	61	1
7	6.12	11	7.2	12.5	0
8	6.4	13.18	5.8	35.8	0
9	7.1	13.81	5.8	59.9	0
10	8.5	14.43	6.2	57.2	1
11	7.5	10	5.5	35.8	0
12	9.3	13.21	6.8	27.9	0
13	8.8	17.1	6.2	24.1	1
14	9.96	15.12	6.3	27.7	1
15	9.83	18.7	0.5	24	0
16	10.12	20.2	5.5	57.2	1
17	10.7	15	5.8	44.3	0
18	10.45	17.6	7.1	49.2	0
19	11.32	19.8	7.5	23	0
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

B. QUESTION 1: FORMULATE THE MR EQUATION

SUMMARY OUTPUT

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	1.948183131		
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
Population (000,000) [X1]	0.55190731	0.050629062	10.90099815	5.468E-11	0.447634803	0.65617981	0.447634803	0.656179811
Percent Unemployed [X2]	0.20316264	0.117086169	1.735154932	0.09502688	-0.037980835	0.44430612	-0.037980835	0.444306121
Advertising Expense (000) [X3]	0.03135496	0.016062075	1.952111518	0.06221113	-0.001725501	0.06443543	-0.001725501	0.064435426
Mall Location [X4]	0.21979032	0.540028513	0.406997626	0.68747343	-0.89241922	1.33199987	-0.89241922	1.331999866

Multiple Regression Equation $\hat{Y} = a + b_1X_1 + b_2X_2 + \ldots + b_kX_k$

where

- X_1 is one of the independent variables.
- X_2 is the second independent variable.
- X_k is the k^{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)

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.

D. QUESTION 3: WHAT IS THE COEFFICIENT OF MULTIPLE DETERMINATION (MULITPLE R^2)?

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

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

Figure 2: R^2 (SUSS, 2014)

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

E. QUESTION 4: WHAT IS THE ADJUSTED COEFFICIENT OF MULTIPLE DETERMINATION (ADJUSTED MULITPLE R²)?

SUMMARY OUTPUT									
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	

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

Figure 3: R²adj (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^2 increases.
- But if the independent variable is not a good predictor, it still increases R^2.
- This makes R^2 inaccurate.
- R^2_{adj} will not necessarily increase when a new variable is added to the model.
- Here, $R^2_{adj} = 0.822$

F. QUESTION 4: WHAT IS THE MULTIPLE STANDARD ERROR OF ESTIMATE (S_{YX})?

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

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

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.

- $S_{y.x}$: Standard Error of Estimate = 1.396 x 10^3
- $S_{y.x}$ 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.

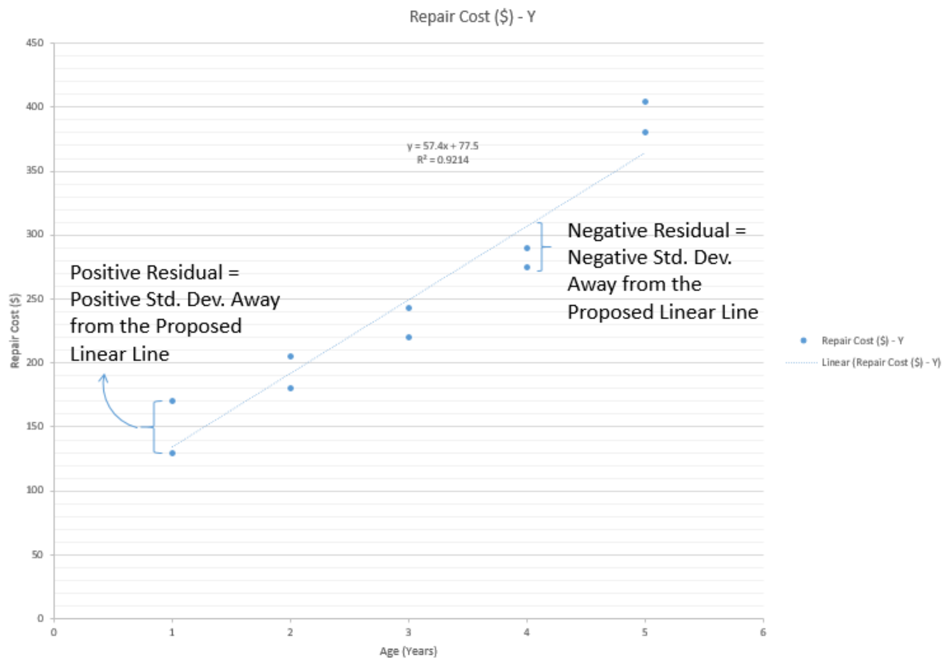


Figure 5: What are Residuals?

- 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 $S_{y.x}$ = 1.396, this shows that about 68% of the predictions should be within ±1.396 x 10^3 (±1σ) of the actual repair costs and about 95% should be within (1.396 x 10^3 x 2) = ±2.792 (±2σ) of actual repair costs.

G. QUESTION 5: WHAT IS THE ANOVA TABLE?

SUMMARY OUTPUT								
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

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	$n-1$	SS total		

Figure 6: ANOVA (SUSS, 2014)

$$\text{Total variation} = SS\ total = \sum\left(Y-\bar{Y}\right)^2$$

$$\text{Error variation} = SSE = \sum\left(Y-\hat{Y}\right)^2$$

$$\text{Regression variation} = SSR = \sum\left(\hat{Y}-\bar{Y}\right)^2 = (SS\ total - SSE)$$

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

$$\textbf{Global Test} \quad F = \frac{MSR}{MSE} = \frac{SSR/k}{SSE/[n-(k+1)]}$$

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 → Accept H1

H. QUESTION 6: WHAT IS THE SIGNIFICANCE F?

SUMMARY OUTPUT								
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	1.948183131					
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 = 7.22×10^{-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

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

- β1 refers to X1 (population)
- β2 refers to X2 (percent unemployed)
    - β3 refers to X3 (advertising expense)
    - β4 refers to X4 (mall location)
    - Thus, what H0 means is that β1 / β2 / β3 / β4 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 → 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.
    - We will not know which factor is important (β1 / β2 / β3 / β4), 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 → Global F Test
- Step 4: Formulate the Decision Rule

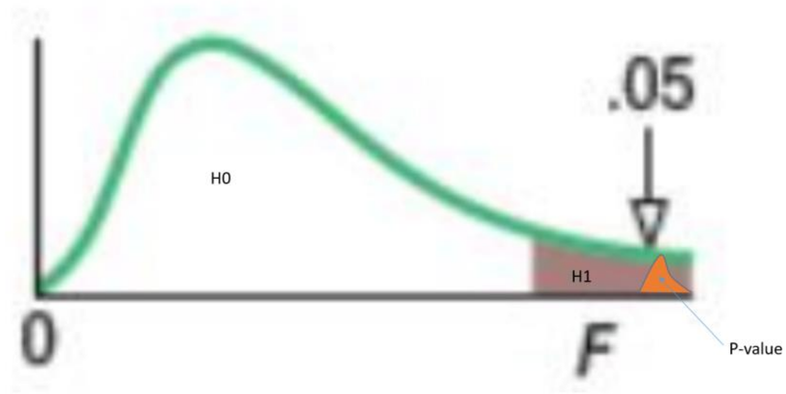


Figure 9: F Distribution 5%

- 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%.
- 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.
- But if the p-value is large > 5%, it will infiltrate the H0 area → Thus H0 is accepted.

- Since the p-value i.e. Significance F = $7.22 \times 10^{-10} < 5\%$
- Thus we accept H1 → The equation is important!

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

SUMMARY OUTPUT

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

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**
$H_0: \beta_1 = 0$ $H_1: \beta_1 \neq 0$	$H_0: \beta_2 = 0$ $H_1: \beta_2 \neq 0$	$H_0: \beta_3 = 0$ $H_1: \beta_3 \neq 0$	$H_0: \beta_4 = 0$ $H_1: \beta_4 \neq 0$

  - β1 refers to X1 (population)
  - β2 refers to X2 (percent unemployed)

- β3 refers to X3 (advertising expense)
- β4 refers to X4 (mall location)
- Thus, what H0 means is that β1 / β2 / β3 / β4 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, α = 5%
- Step 3: State the Test Statistic → Individual F Test
- Step 4: Formulate the Decision Rule

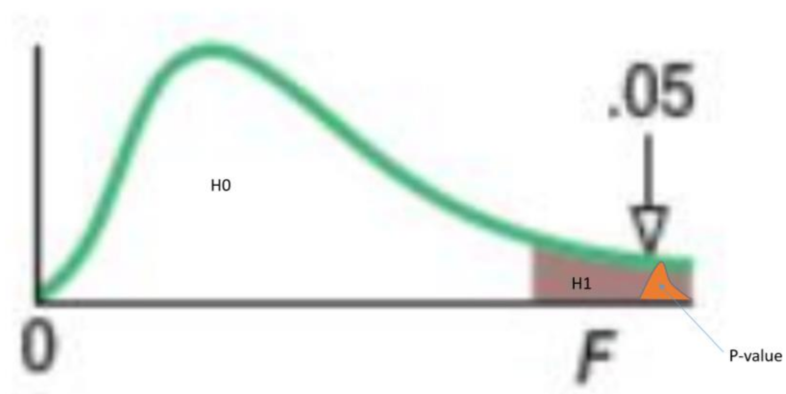


Figure 10: F Distribution 5%

- 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.

- Step 5: Make the Decision
    - P-value for population = 5.5×10^{-11} < 5% → Accept H1 → Retain
    - P-value for percent unemployed = 0.1 > 5% → Accept H0 → Drop off
    - P-value for advertising expense = 0.062 > 5% → Accept H0 → Drop off

- P-value for mall location = 0.687 > 5% → Accept H0 → 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.

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

SUMMARY OUTPUT	
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

- 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.

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

SUMMARY OUTPUT								
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

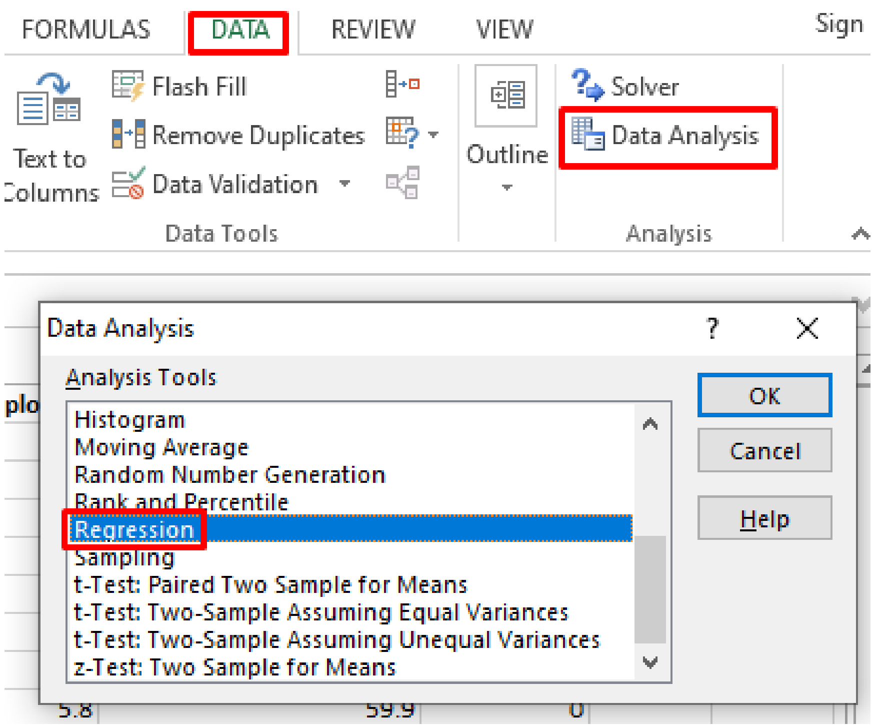
- The given coefficients are:
    - Population = 0.55
    - Percent Unemployed = 0.2
    - Advertising Expense = 0.03
    - Mall Location = 0.22
- However, this is not 100% accurate. Thus, a 95% confidence interval is attached to them.
- This means that the lower and upper limits (95%) are:
    - Population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

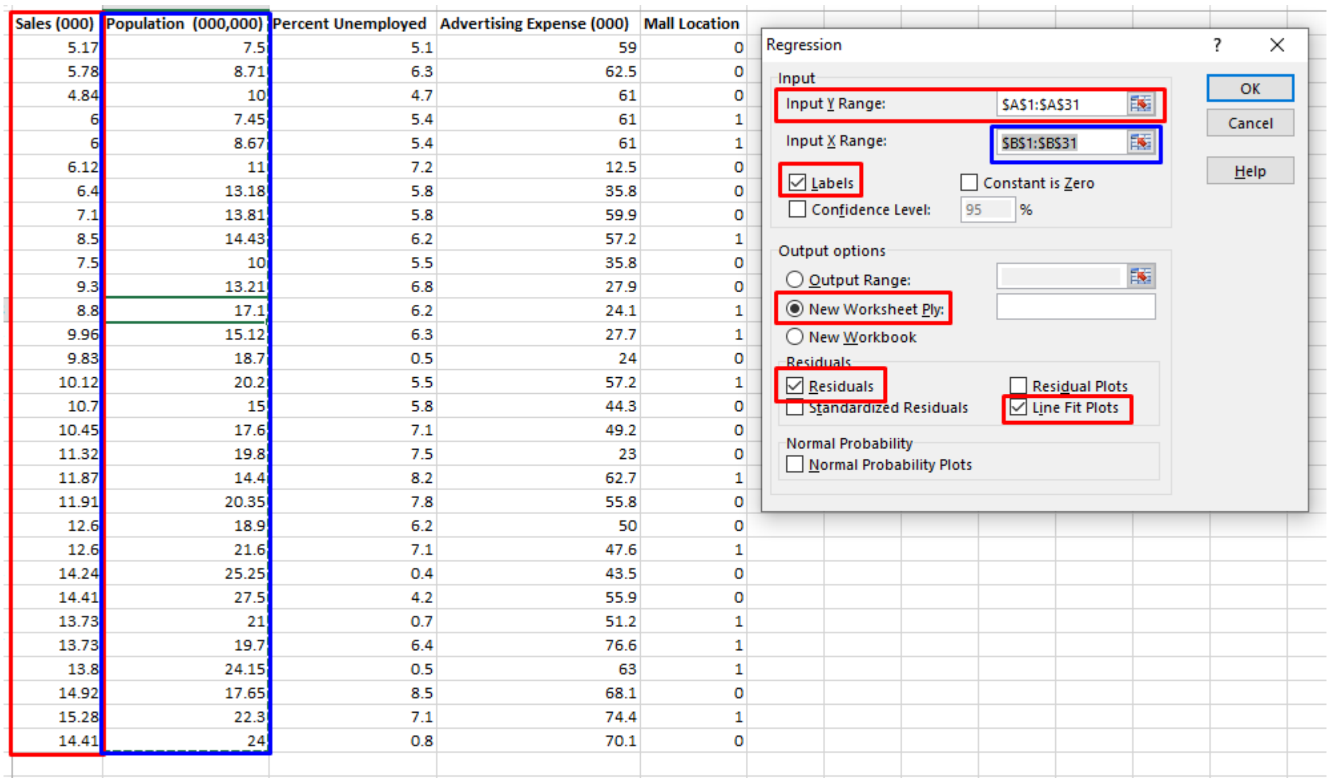
- Advertising Expense = between -0.002 and 0.064
- Mall Location = between -0.892 and 1.33

A. ASSUMPTION 1: LINEARITY

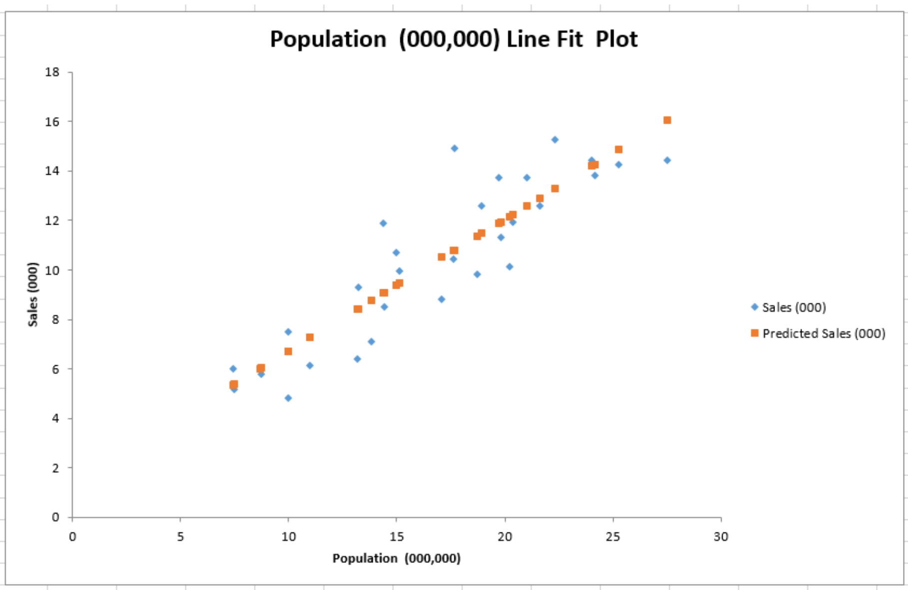
- LINEARITY: There must be a linear relationship between the dependent variable, Y, and each independent variable, X1, X2 etc..

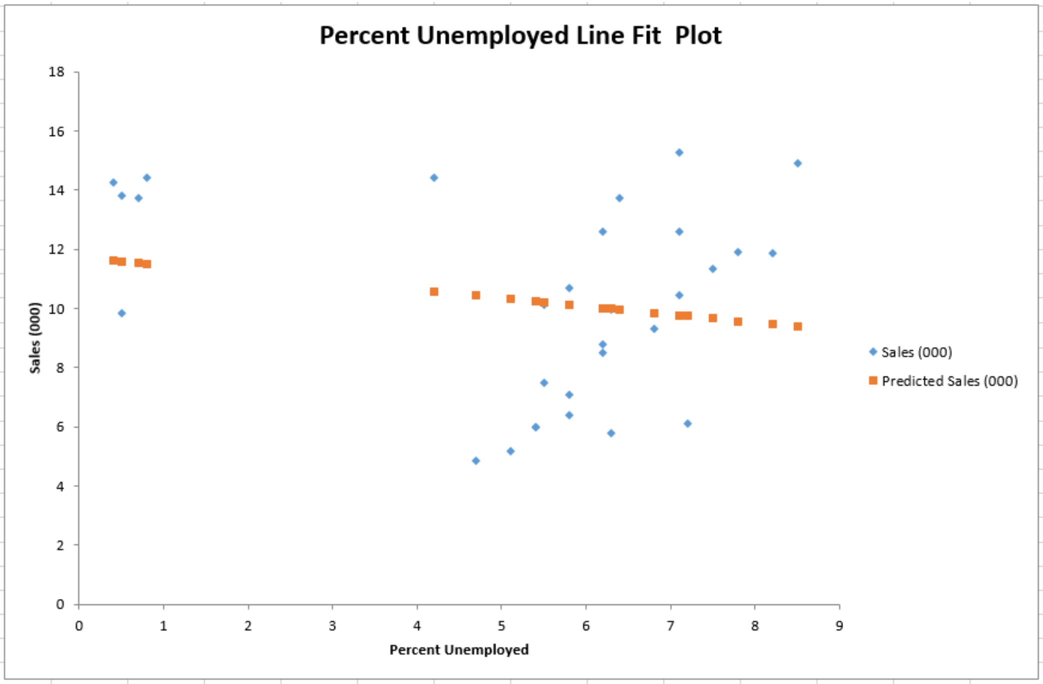
- We check this by using a ***Line Fit Plot:***

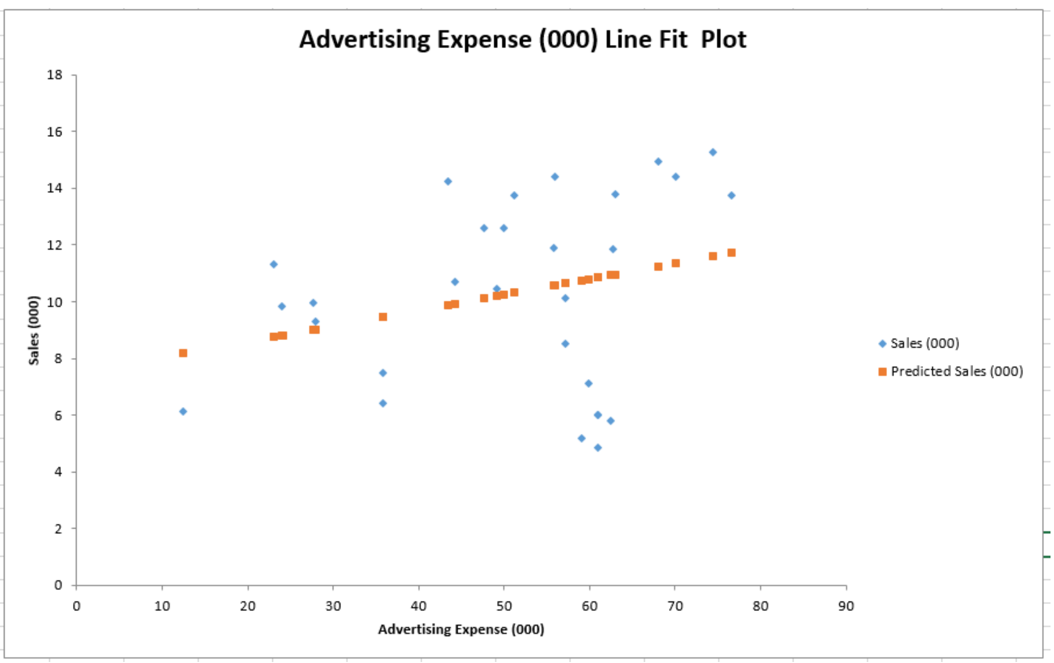


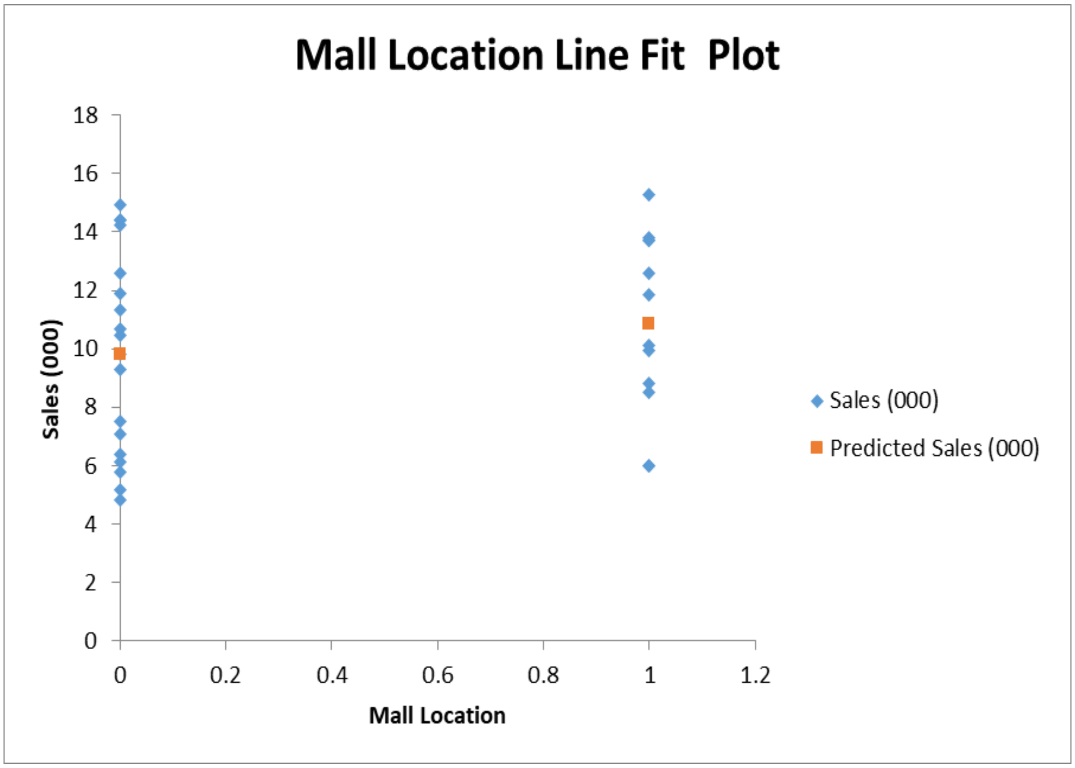


Sales (000)	Population (000,000)	Percent Unemployed	Advertising Expense (000)	Mall Location
5.17	7.5	5.1	59	0
5.78	8.71	6.3	62.5	0
4.84	10	4.7	61	0
6	7.45	5.4	61	1
6	8.67	5.4	61	1
6.12	11	7.2	12.5	0
6.4	13.18	5.8	35.8	0
7.1	13.81	5.8	59.9	0
8.5	14.43	6.2	57.2	1
7.5	10	5.5	35.8	0
9.3	13.21	6.8	27.9	0
8.8	17.1	6.2	24.1	1
9.96	15.12	6.3	27.7	1
9.83	18.7	0.5	24	0
10.12	20.2	5.5	57.2	1
10.7	15	5.8	44.3	0
10.45	17.6	7.1	49.2	0
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




Percent Unemployed Line Fit Plot
Sales (000)
Percent Unemployed
Sales (000)
Predicted Sales (000)


Advertising Expense (000) Line Fit Plot
Sales (000)
Advertising Expense (000)
Sales (000)
Predicted Sales (000)



- 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.
- This means that the variation in the residuals must be more or less the same.

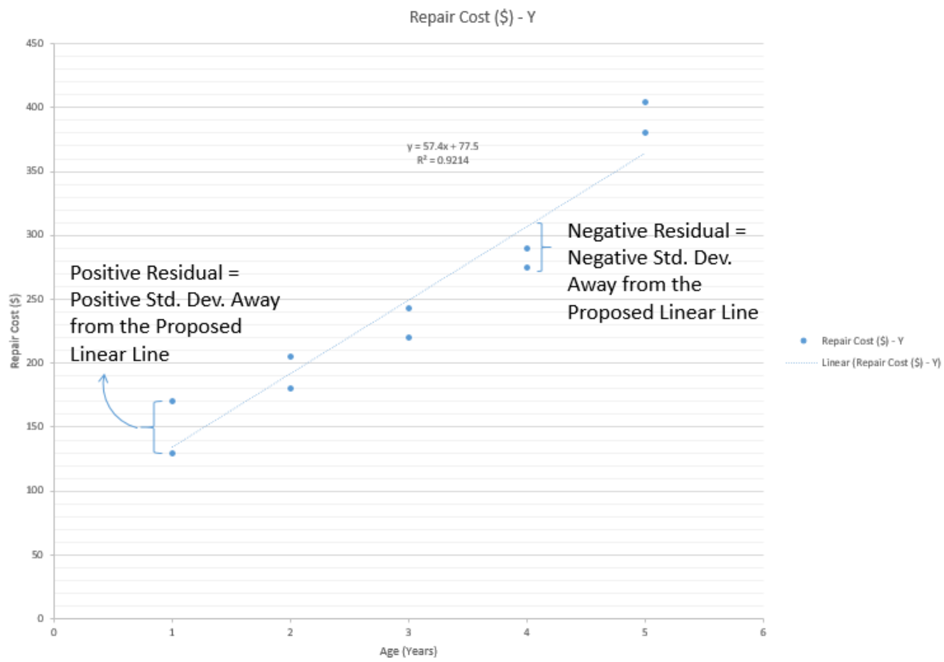
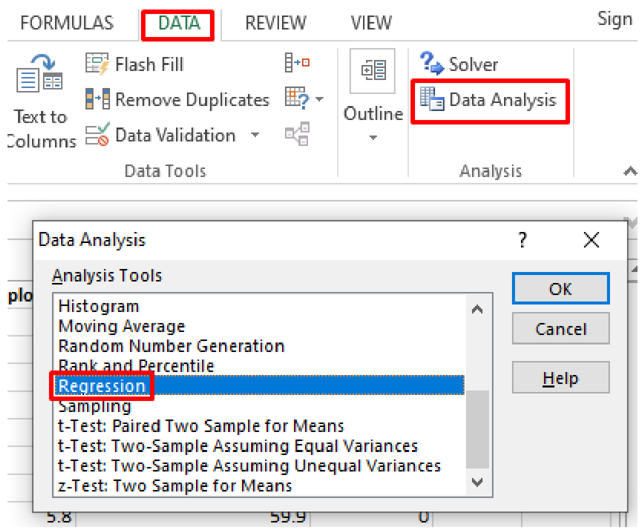
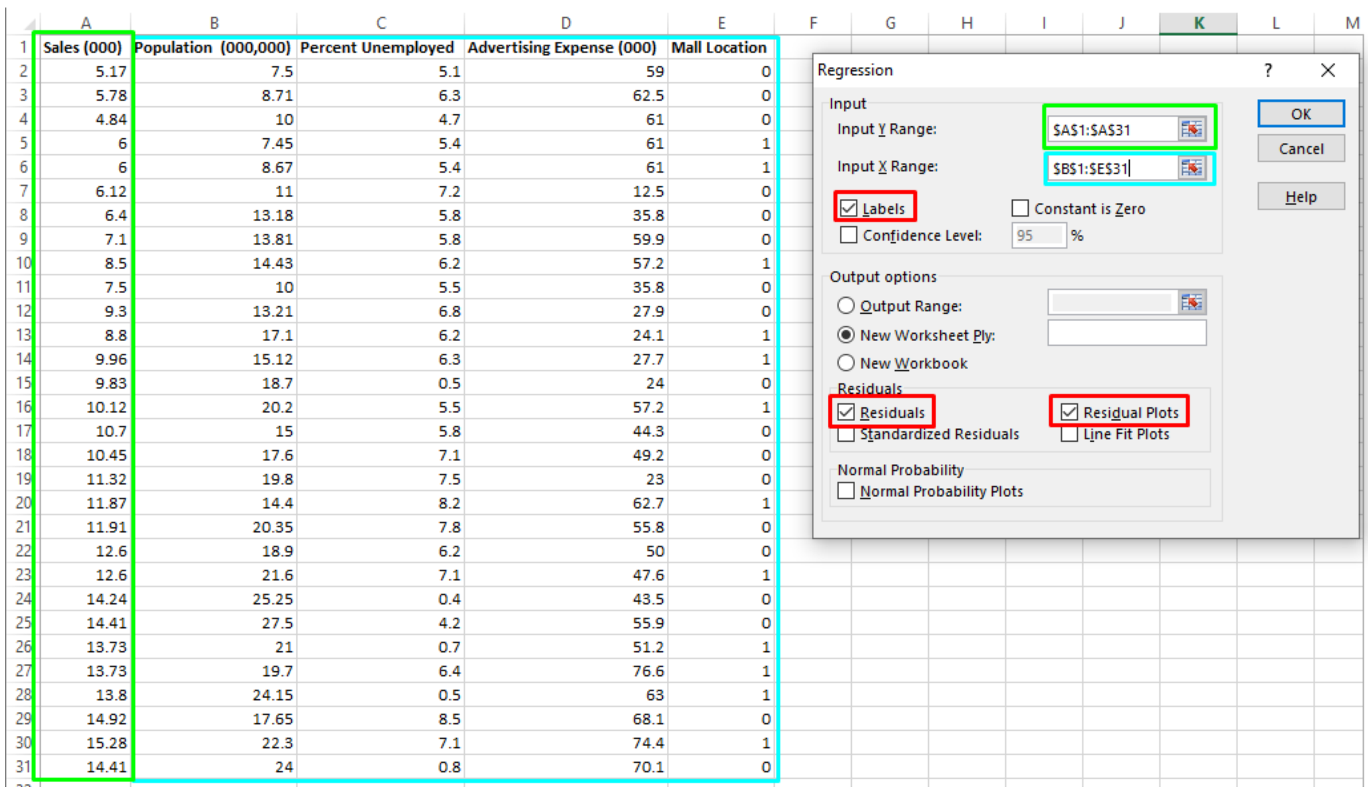


Figure 11: What are Residuals?

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





	A	B	C	D	E
1	Sales (000)	Population (000,000)	Percent Unemployed	Advertising Expense (000)	Mall Location
2	5.17	7.5	5.1	59	0
3	5.78	8.71	6.3	62.5	0
4	4.84	10	4.7	61	0
5	6	7.45	5.4	61	1
6	6	8.67	5.4	61	1
7	6.12	11	7.2	12.5	0
8	6.4	13.18	5.8	35.8	0
9	7.1	13.81	5.8	59.9	0
10	8.5	14.43	6.2	57.2	1
11	7.5	10	5.5	35.8	0
12	9.3	13.21	6.8	27.9	0
13	8.8	17.1	6.2	24.1	1
14	9.96	15.12	6.3	27.7	1
15	9.83	18.7	0.5	24	0
16	10.12	20.2	5.5	57.2	1
17	10.7	15	5.8	44.3	0
18	10.45	17.6	7.1	49.2	0
19	11.32	19.8	7.5	23	0
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

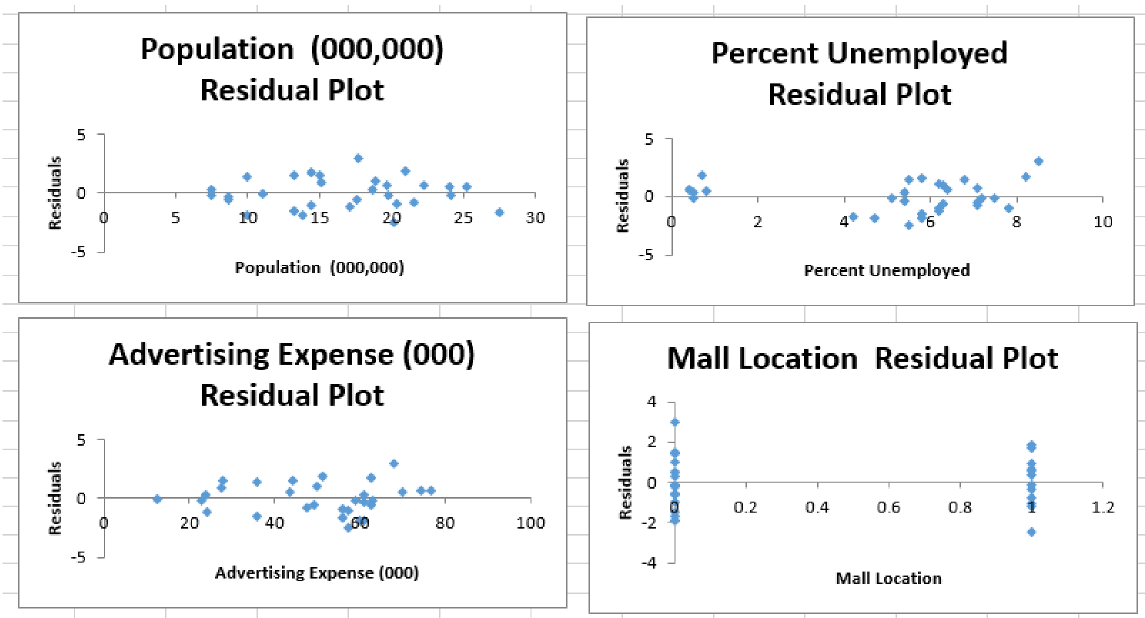
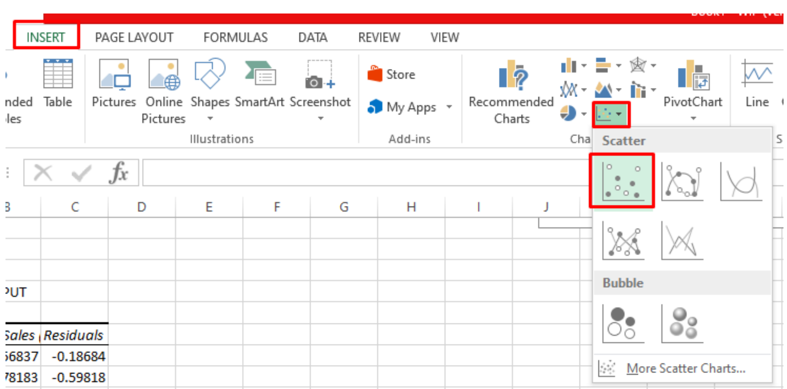


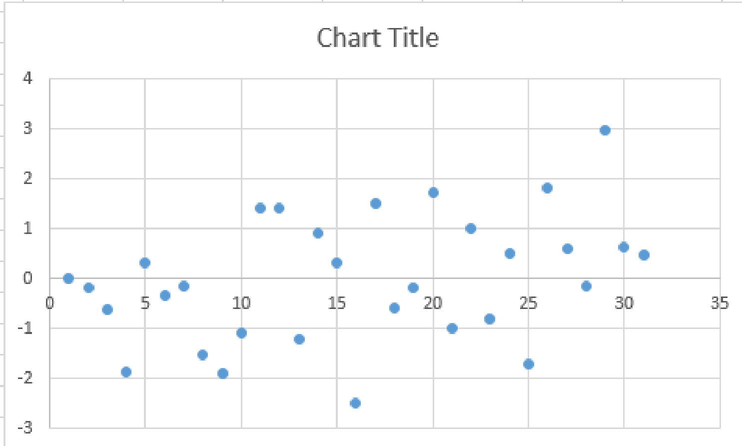
Figure 12: Individual Residual Plots for Each Variable

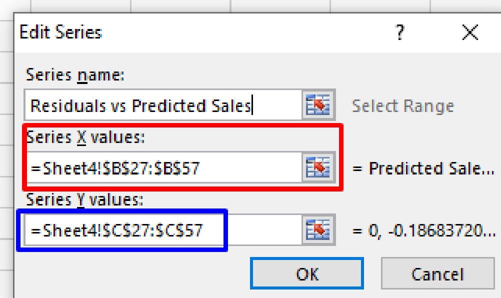
- Figure 12 shows Individual Residual Plots for each Variable.
- We see that they are scattered randomly and uniformly around the 0 line.
- This means that individually, they are Homoscedastictiy.
- But we are not done. We need to check Overall Homoscedasticity.



RESIDUAL OUTPUT

bservatio	cted Sales	Residuals
1	5.356837	-0.18684
2	6.378183	-0.59818
3	6.71805	-1.87805
4	5.672691	0.327309
5	6.346018	-0.34602
6	6.257149	-0.13715
7	7.906449	-1.50645
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
13	9.044546	0.915454
14	9.506227	0.323773
15	12.61068	-2.49068
16	9.177438	1.522562
17	11.03015	-0.58015
18	11.50411	-0.18411
19	10.13061	1.739395
20	12.89705	-0.98705
21	11.58986	1.010135
22	13.4074	-0.8074
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





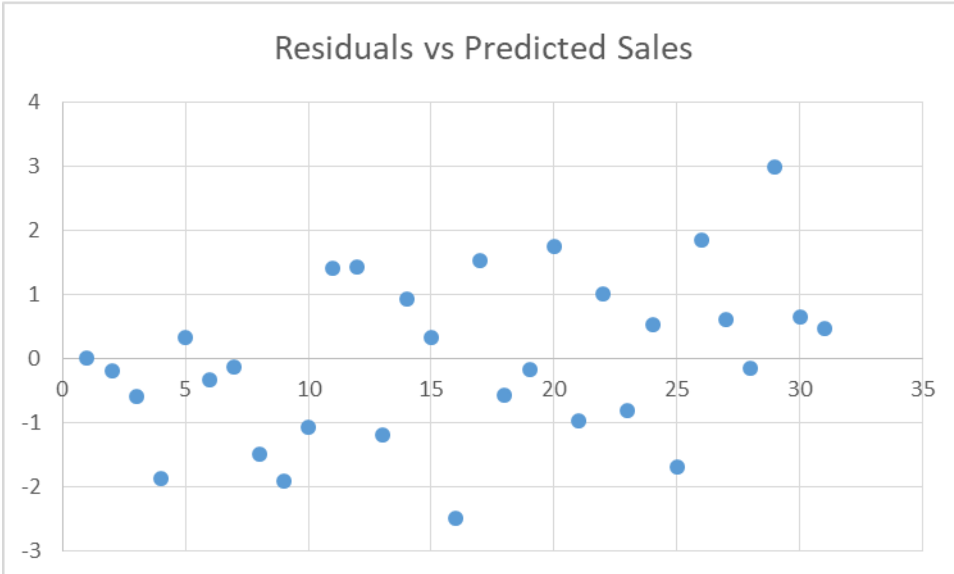
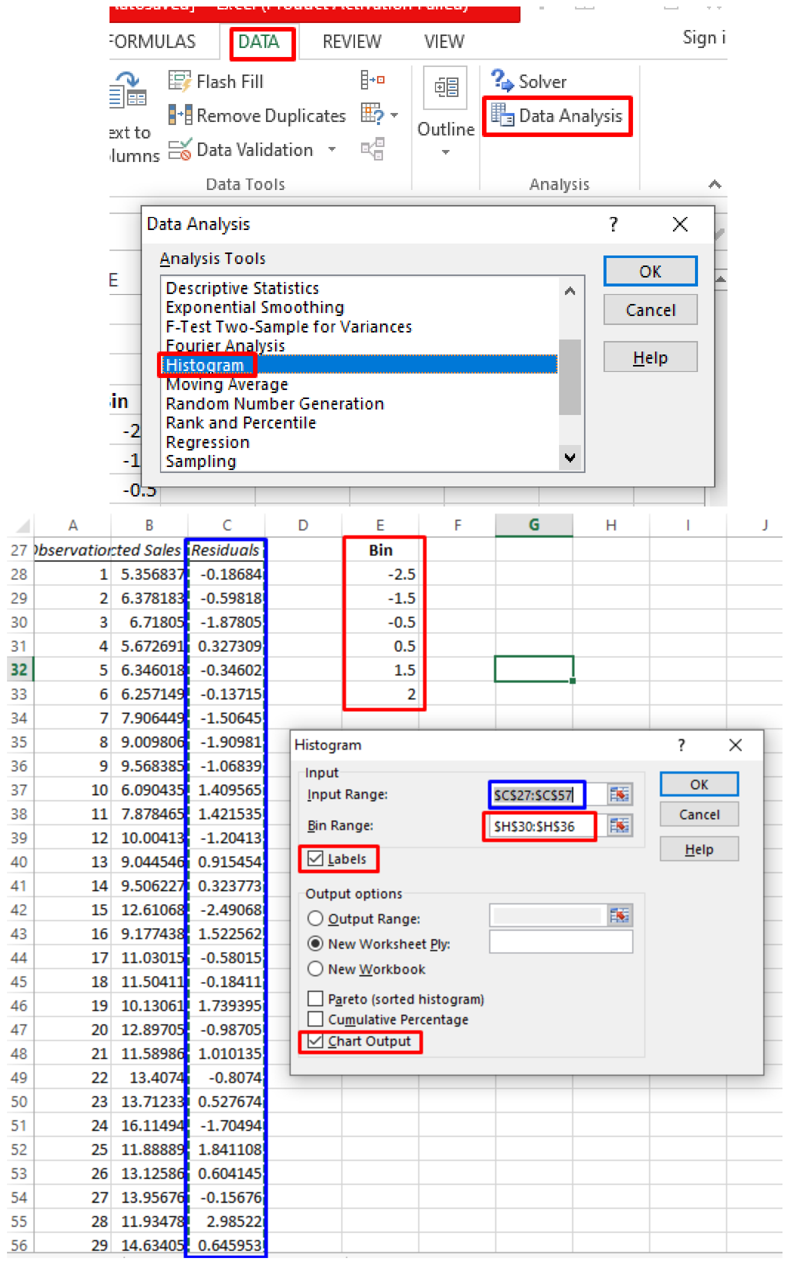


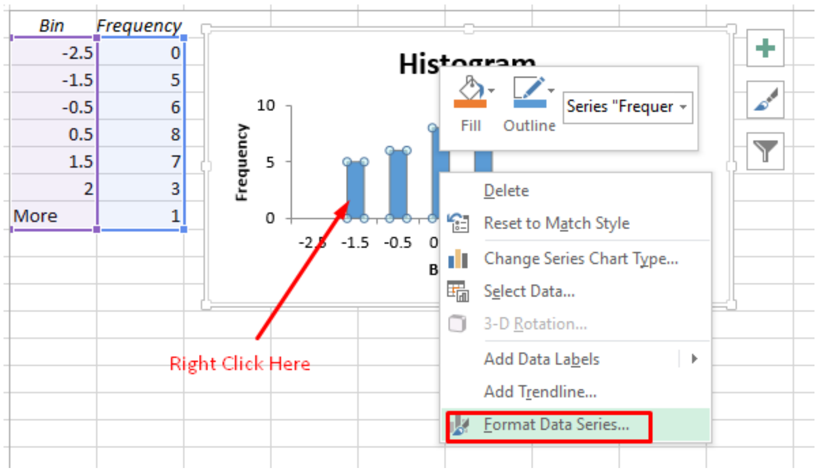
Figure 13: Overall Resdial Plot

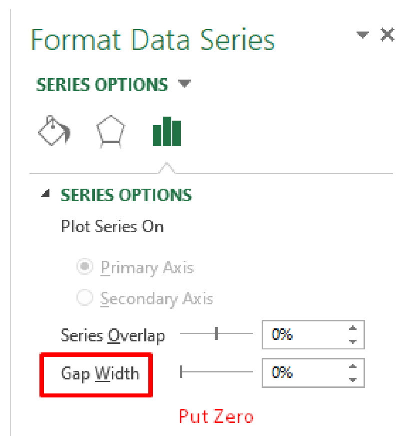
- Figure 13 shows the Overall Residual Plot. The Residuals vs Predicted Sales.
- Figure 13 was gotten from drawing a scatter plot from the ***Residual Output***.
- Since Figure 13 shows that the residuals are uniformly randomly scattered around 0, this means that we can assume Overall Homoscedasitity.

C. ASSUMPTION 3: NORMALLY DISTRIBUTED

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







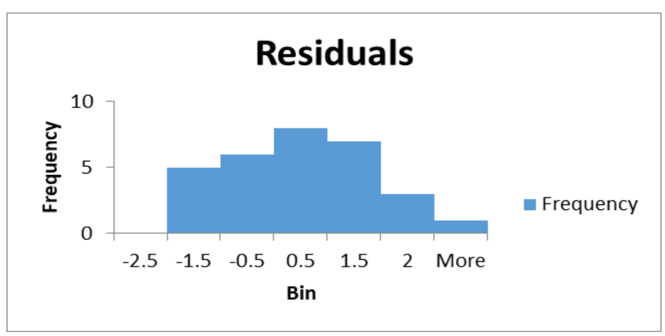


Figure 14: Histogram of the Residuals

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

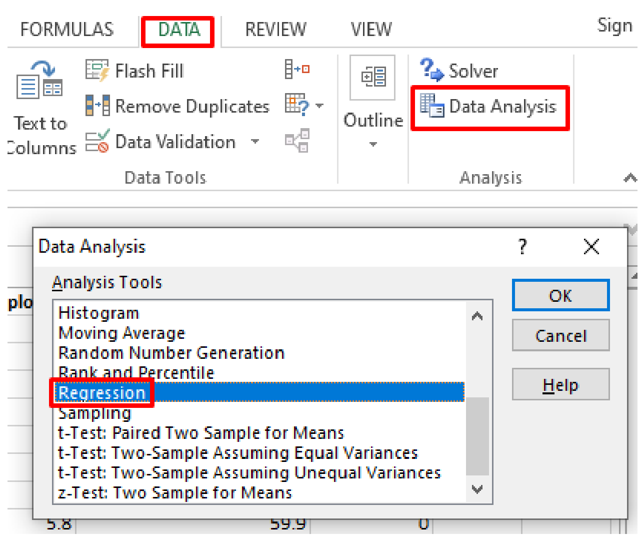
D. ASSUMPTION 4: MULTICOLLINEARITY

- Multicollinearity must NOT exist.
- Multicollinearity = Correlation between indenpendent 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.

Variance Inflation Factor $$VIF = \frac{1}{1-R_j^2}$$

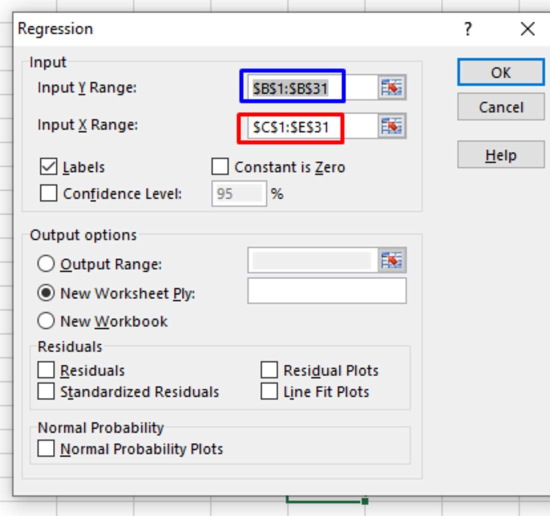
Figure 15: VIF (SUSS, 2014)

- There is no short cut to measure VIF in excel, thus we need to get the R^2 for each and every variable in order to get their VIF.



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
5.78	8.71	6.3	62.5	0
4.84	10	4.7	61	0
6	7.45	5.4	61	1
6	8.67	5.4	61	1
6.12	11	7.2	12.5	0
6.4	13.18	5.8	35.8	0
7.1	13.81	5.8	59.9	0
8.5	14.43	6.2	57.2	1
7.5	10	5.5	35.8	0
9.3	13.21	6.8	27.9	0
8.8	17.1	6.2	24.1	1
9.96	15.12	6.3	27.7	1
9.83	18.7	0.5	24	0
10.12	20.2	5.5	57.2	1
10.7	15	5.8	44.3	0
10.45	17.6	7.1	49.2	0
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

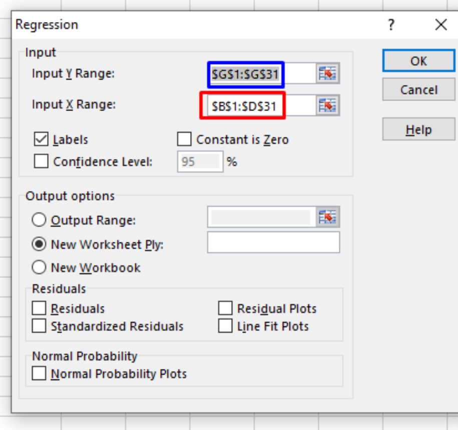


- Take note of something weird here!
- Input Y Range = Population
- Input X Range = Percent unemployed / Advertising / Mall Location!

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

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
5.78	8.71	62.5	0	6.3
4.84	10	61	0	4.7
6	7.45	61	1	5.4
6	8.67	61	1	5.4
6.12	11	12.5	0	7.2
6.4	13.18	35.8	0	5.8
7.1	13.81	59.9	0	5.8
8.5	14.43	57.2	1	6.2
7.5	10	35.8	0	5.5
9.3	13.21	27.9	0	6.8
8.8	17.1	24.1	1	6.2
9.96	15.12	27.7	1	6.3
9.83	18.7	24	0	0.5
10.12	20.2	57.2	1	5.5
10.7	15	44.3	0	5.8
10.45	17.6	49.2	0	7.1
11.32	19.8	23	0	7.5
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
12.6	21.6	47.6	1	7.1
14.24	25.25	43.5	0	0.4
14.41	27.5	55.9	0	4.2
13.73	21	51.2	1	0.7
13.73	19.7	76.6	1	6.4
13.8	24.15	63	1	0.5
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!

SUMMARY OUTPUT	
Regression Statistics	
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.

- We end up with this:

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

- We see that the VIF is less than 10 for each → No Multicollinearity!

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.

- 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:

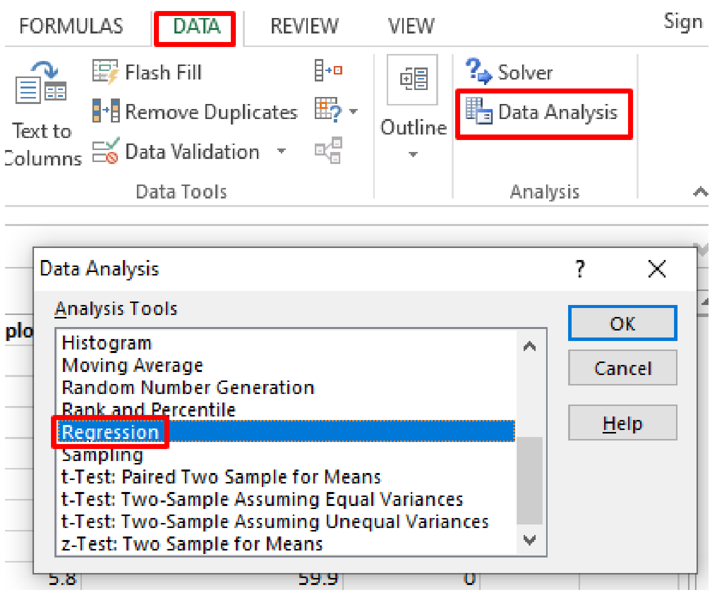
price	sq metres	size of garage	size of bedroom
65	1	0	2
73	1.1	0	2
85	1.15	1	2
87	1.4	0	3
98	1.7	1	3
105	1.8	1	4
95	1.9	0	3
125	1.9	1	4
125	2.1	2	4
137	2.1	2	4
150	2.3	2	4

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

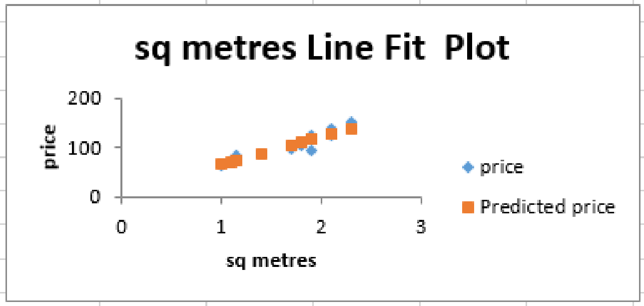
A. STEP 1: CHECK OUT INDIVIDUAL R^2 AND ADJUSTED R^2

- There are 3 possible models:

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



SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.932676854
R Square	0.869886115
Adjusted R Square	0.855429016
Standard Error	10.2990493
Observations	11



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 95.0%*	*pper 95.0%*
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^2	Adjusted R^2	S_e	Parameter Estimates
X_1	0.879	0.855	10.299	b_0=9.503 b_1=56.394
X_2	0.759	0.731	14.030	b_0=78.290 b_1=28.382
X_3	0.793	0.770	12.982	b_0=16.250 b_1=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)

$$Y = 9.503 + 56.394X_1$$

B. STEP 2: CHECK OUT ALTERNATIVE R^2 AND ADJUSTED R^2

- 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 Variable in the Model	R^2	Adjusted R^2	S_e	Parameter Estimates
X_1	0.870	0.855	10.299	b_0=9.503 b_1=56.394
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 and X_3	0.877	0.847	10.609	b_0=8.311 b_1=44.313 b_3=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^2 AND ADJUSTED R^2

- Now we have 1 possible model left:

$$Y = a + b_1X_1 + b_2X_3 + b_3X_2$$

- After regression analysis…

Independent Variable in the Model	R^2	Adjusted R^2	S_e	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	b_0=26.440 b_1=30.803 b_2=12.567 b_3=4.576

- 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).

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 www.AlvinAng.sg.