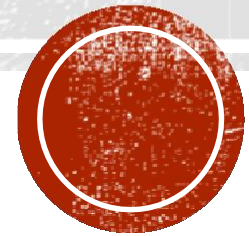


ISYS3401

Information Technology Evaluation

Tutorial Week 5



Helen Wang
Vincent Pang

Linear Regression

- Model the relationship between a **dependent variable** and one or more **independent / explanatory variables** by using **linear functions** where parameters are **estimated** from the data.
- A simple example:

$$y = \alpha + \beta x$$

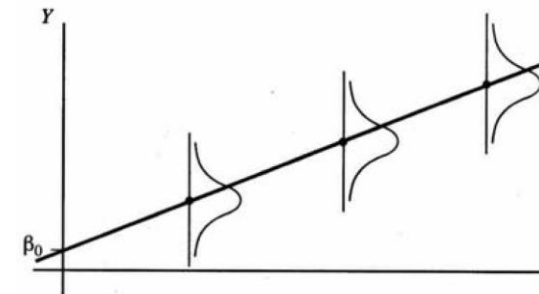
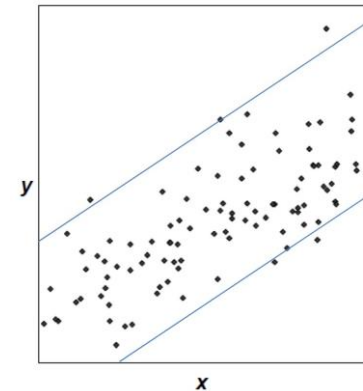
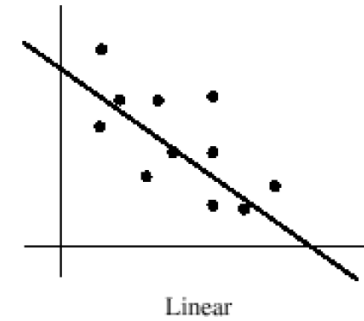
- α : intercept
- β : slope

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	31.91255948	2.803609497	11.3826692	3.92706E-08	25.85572941	37.96938955
X	-0.015073871	0.00298682	-5.046796127	0.000223707	-0.021526503	-0.008621239



Assumptions

1. Linearity: $E(y) = \alpha + \beta x$
 - Relationship between the independent and dependent variables to be linear
 - The expected value of y is best represented by a line as the scatter plot distributes around the line evenly
2. Constant variance
 - For all values of x , $SD(y)$ is the same
3. Normality
 - For a given x , y is Normally distributed



Residuals

- $E(y) = \alpha + \beta x$ the actual 'unobserved' association in real world
- $y = a + bx$ what we modelled / estimated
 - a and b are the least squares estimates of α and β

$$b = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2}$$

$$a = \bar{y} - b \bar{x}$$

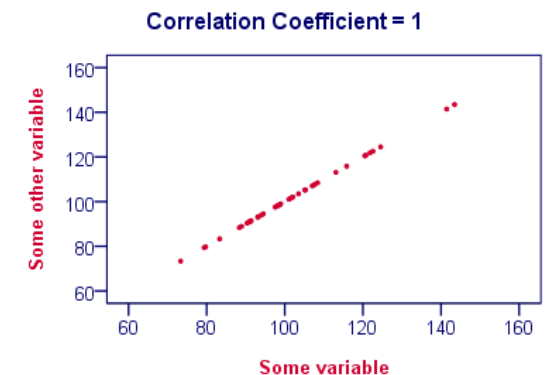
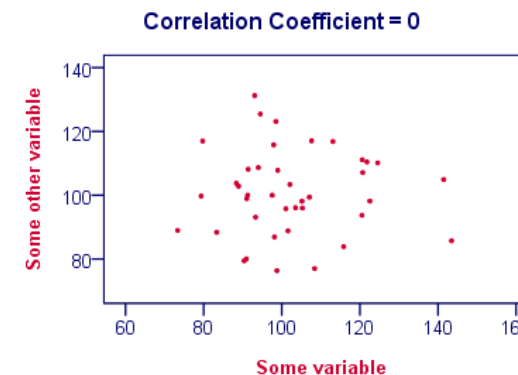
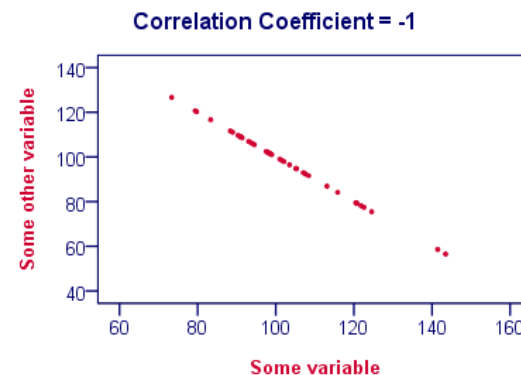
- Observed error / residuals:
 - difference between your observation and the actual association
 - Given pair (x_i, y_i)
 - $e_i = E(y) - y = y_i - (a + bx_i)$



ANOVA analysis

- Coefficient of Determination (R^2)
 - $R^2 = SS_R / SS_T = (SS_T - SS_E) / SS_T = 1 - SS_E / SS_T$
 - The percentage of the dependent variable (y) variation that is explained by a linear model
 - $[0, 1]$
- Sample correlation coefficient (r)
 - $[-1, 1]$
 - $r = 1 \Rightarrow$ perfect positive correlation
 - $r = -1 \Rightarrow$ perfect negative correlation
 - $r = 0 \Rightarrow$ no correlation

Regression Statistics	
Multiple R	0.517318229
R Square	0.26761815
Adjusted R Square	0.260576017
Standard Error	1.151153367
Observations	106



ANOVA (continued)

- Mean square: represent an estimate of population variance
 - $MS = SS/df$
- Error Mean Square: variance around the fitted regression line.
 - $MS_E = SS_E/(n-2)$
- F-Statistics: ratio of two variances; overall evaluation of the model

$$F = \frac{MS_R}{MS_E} \quad \text{with } 1 \text{ and } n-2 \text{ df}$$

- $F = 1$, no association between x and y
- $F > 1$, association
- Useful reading: <http://blog.minitab.com/blog/adventures-in-statistics-2/understanding-analysis-of-variance-anova-and-the-f-test>

Total SS			Residual SS or Error SS		Regression SS
SS_T	=		SS_E	+	SS_R
$n-1$	=		$n-2$	+	1

Source	SS	df
Regression	$\sum(\hat{y} - \bar{y})^2$	p
Residual	$\sum(y_i - \hat{y})^2$	$N - (p + 1)$
Total	$\sum(y_i - \bar{y})^2$	$N - 1$

$P = \# \text{variables}$



Question 1a

Univariate Regression

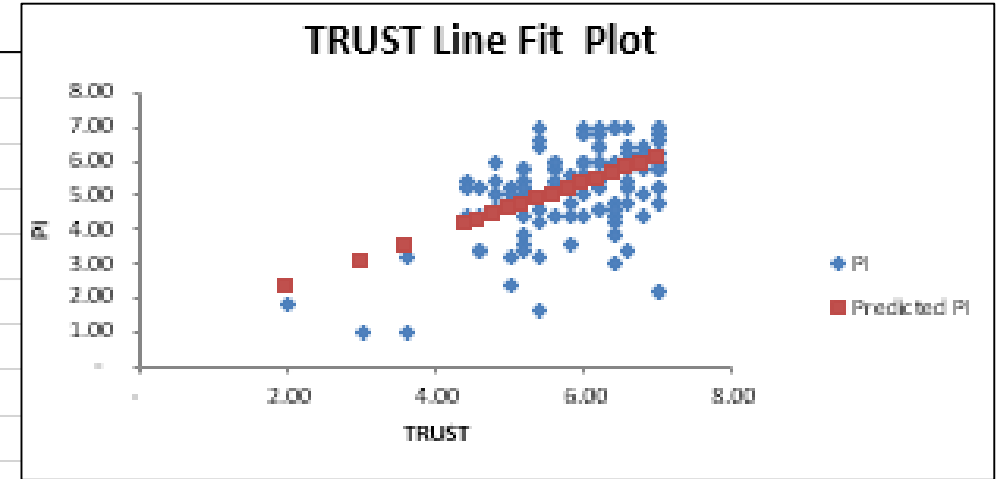
Two variables: study factor (independent variables) X and outcome factor (dependent variable) Y.

- a) Using EXCEL, plot these data PI as the outcome variable, and Trust as the predictor variable.
- b) Using EXCEL – Data Analysis Tool, carry out a simple linear regression using these data.
- c) Write a brief report summarising your results and conclusions.



Solution 1a

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.517318229					
R Square	0.26761815					
Adjusted R Square	0.260576017					
Standard Error	1.151153367					
Observations	106					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	1	50.35907058	50.35907058	38.00242669	1.36269E-08	
Residual	104	137.8160238	1.325154075			
Total	105	188.1750943				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	0.813538521	0.708799009	1.147770398	0.253696745	-0.592036512	2.219113553
TRUST	0.745622623	0.1209521	6.164610831	1.36269E-08	0.505770071	0.985475175



Summary report

- The plot of y against x shows an linear relationship with Y increases as X increases.
- The R-Square is 0.2676 or 26.76%
- The Correlation Coefficient r is 0.5173
- The linear relationship is estimated to be: $y = 0.8135 + 0.7453 x$.

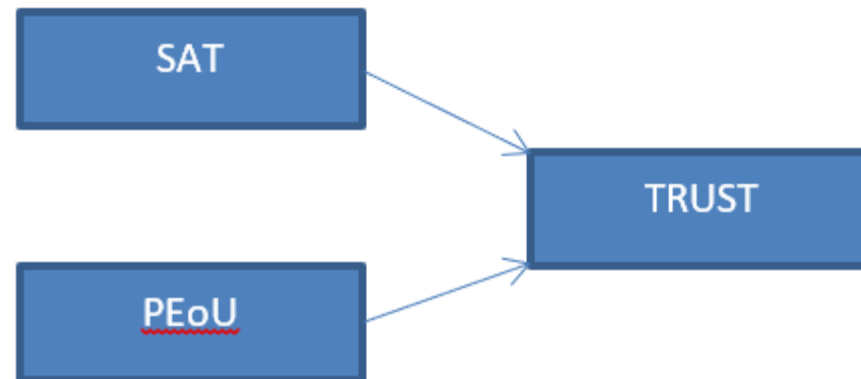


Question 1b

Multivariate Regression

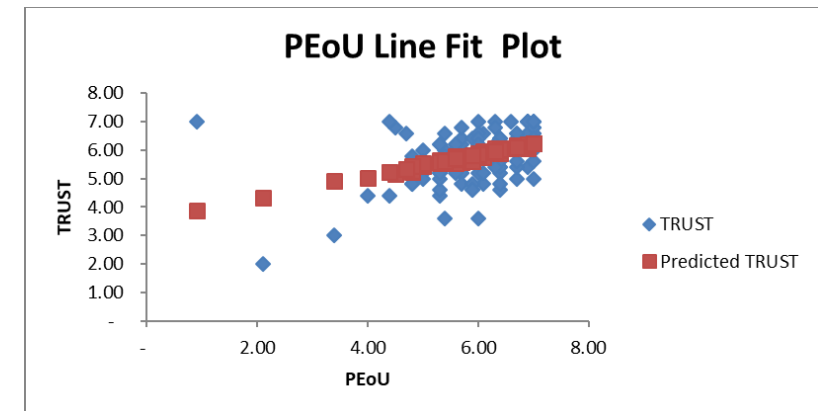
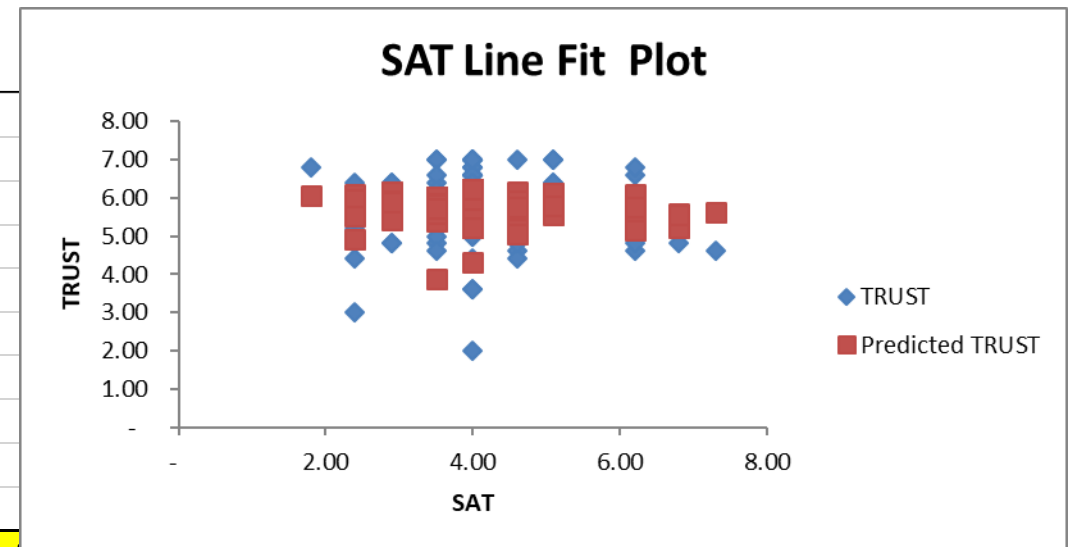
- *More than one independent variables*
- $y = a + b_1x_1 + b_2x_2 + \dots + b_kx_k + e$
 - $b_1 \dots b_k$: partial regression coefficients

- Using EXCEL, plot these data Trust as the outcome variable, and SAT and PEOU as the predictor variables
- Using EXCEL – Data Analysis Tool, carry out a simple linear regression using these data.
- Write a brief report summarising your results and conclusions.



Solution 1b

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.413388561					
R Square	0.170890103					
Adjusted R Square	0.154790881					
Standard Error	0.853900394					
Observations	106					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	2	15.47948344	7.739741721	10.61480548	6.43462E-05	
Residual	103	75.10202599	0.729145883			
Total	105	90.58150943				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	3.724898982	0.591322079	6.299272621	7.47297E-09	2.552151105	4.89764686
SAT	-0.052247778	0.083371926	-0.626683114	0.532253517	-0.217596331	0.113100774
PEoU	0.387425808	0.084489269	4.585503165	1.27807E-05	0.219861271	0.554990345



Summary report

- The R-Square is 0.1709 or 17.09%
- The Correlation Coefficient r is 0.4133
- The linear relationship is estimated to be: $y = 3.724 - 0.0052 x_1 + 0.3874 x_2$
- Looking at SAT plot chart, and the p-value, they do not look that great - the p-value for SAT is 0.5322.

Solution 1c

As we only have two dependent variables, you can choose any of the selection technique; say, we use Stepwise selection technique in our example.

SUMMARY OUTPUT		SAT			
<i>Regression Statistics</i>			As an individual construct, it seems SAT can only explain 0.16% (R-Square) of the relationship between Sat and Trust		
Multiple R	0.040400905				
R Square	0.001632233				
Adjusted R Square	-0.007967457				
Standard Error	0.932499017				
Observations	106				

SUMMARY OUTPUT			PEoU			
<i>Regression Statistics</i>			In constrast, it seems PEOU can explain 16.77% (R-Square) of the relationship between PEOU and Trust. Thus, it seems SAT does not explain much.			
Multiple R	0.409547024					
R Square	0.167728765					
Adjusted R Square	0.159726157					
Standard Error	0.851403732					
Observations	106					

As R-Square for SAT is only 0.16%, and PEOU is 16.77%, and if combined together, the R-Square is 17.09%. Thus, you can be brave and delete the SAT construct from the model because it does not really explain anything in the relationship.

Next, in part (d), you want to investigate further by breaking the data down to Store A and Store B, i.e. you analyse data from Store A and then Store B.

Tutorial Q2 - Multivariate Regression

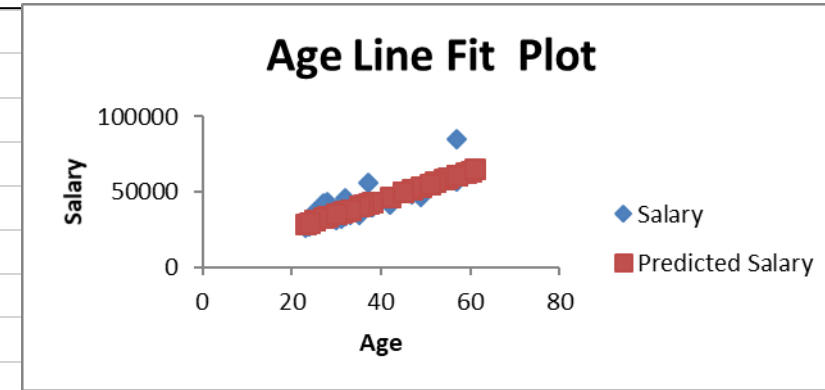
Salary Data tab provides salary and age data for 35 employees, along with an indicator of whether or not the employees have an MBA (Yes and No). The MBA indicator variable is categorical. Since regression analysis requires numerical data, we could include them by coding the variables. For example, coding “No” as 0 and “Yes” as 1.

- a) Convert variable MBA to binary values (No→0 and Yes→1)
- b) Perform two separate simple linear regressions for (i) Age and Salary and (ii) MBA and salary. Write a brief report summarising your results (including the regression models and hypotheses).
- c) Perform multivariate regression for Age and MBA on Salary. Write a brief report summarising your results (including regression model and the hypotheses).
- d) Compare and discuss the R-Square and coefficient estimate between Age and MBA.
- e) What will be the best regression model?



Solution 2b

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.881370267					
R Square	0.776813547					
Adjusted R Square	0.770050322					
Standard Error	6300.056543					
Observations	35					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	1	4558813188	4558813188	114.8584368	2.77815E-12	
Residual	33	1309793511	39690712.45			
Total	34	5868606699				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	7441.698957	3690.421971	2.016489988	0.051947154	-66.52099832	14949.91891
Age	944.0174762	88.08431545	10.71720285	2.77815E-12	764.8085889	1123.226363

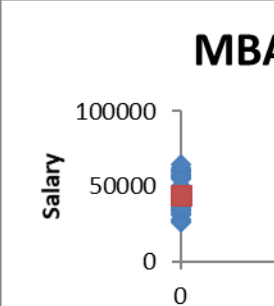


- The R-Square is 0.7768 or 77.68%
- The Correlation Coefficient r is 0.8814
- The linear relationship is estimated to be: $y = 7442 + 944\text{Age}$

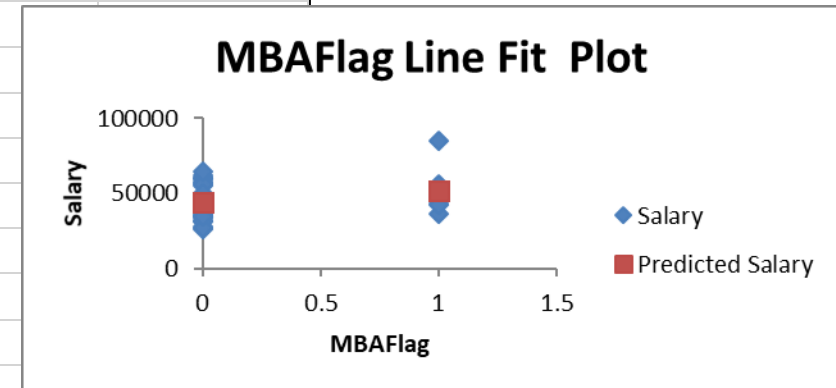
Solution 2b

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.217772555					
R Square	0.047424886					
Adjusted R Square	0.018558973					
Standard Error	13015.47878					
Observations	35					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	1	278318001.5	278318001.5	1.642937341	0.208855742	
Residual	33	5590288698	169402687.8			
Total	34	5868606699				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	44027.62069	2416.913733	18.21646345	8.70068E-19	39110.37273	48944.86865
MBAFlag	7482.212644	5837.401245	1.281771173	0.208855742	-4394.069487	19358.49477

MBA



A box plot showing the distribution of 'Salary'. The y-axis is labeled 'Salary' and has major tick marks at 0, 50000, and 100000. The plot shows a median salary around 40,000, with a box spanning from approximately 25,000 to 55,000. Whiskers extend from 0 to 100,000. There are several outliers represented by blue dots above the upper whisker, reaching up to 100,000. The title 'MBA' is positioned at the top right of the plot area.



- The R-Square is 0.04742 or 4.74%
- The Correlation Coefficient r is 0.2178
- The linear relationship is estimated to be: $y = 44028 + 7482\text{MBAFlag}$
- Note: P-value is greater than 0.05

Solution 2c

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.976118476					
R Square	0.952807278					
Adjusted R Square	0.949857733					
Standard Error	2941.914352					
Observations	35					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	2	5591651177	2795825589	323.0353318	6.05341E-22	
Residual	32	276955521.7	8654860.054			
Total	34	5868606699				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	893.5875971	1824.575283	0.489751015	0.627650922	-2822.950634	4610.125828
Age	1044.146043	42.14128238	24.77727265	1.8878E-22	958.3070599	1129.985026
MBAFlag	14767.23159	1351.801764	10.92411031	2.49752E-12	12013.7015	17520.76168

- The R-Square is 0.9528 or 95.28%
- The Correlation Coefficient r is 0.9761
- The linear relationship is estimated to be: $y = 893 + 1044\text{Age} + 14767\text{MBAFlag}$
- It is significance as both Age and MBAFlag are both < 0.05 when both variables are used together

Solution 2d:

- Coefficient of AGE changed from 944.017 to 1044.146
- Coefficient of MBA changed from 7482.213 to 14767.23
- These LARGE changes in coefficients indicate possible dependence between the two explanatory variables.
- A different regression may be needed. Example an interaction model may be MORE appropriate – by adding an extra term (Age*MBAFlag) in the regression model.

Solution 2d Model

$$\text{Salary} = 893.59 + 1044.15 \text{ Age} + 14767.23 \text{ MBA}$$

$$\text{No MBA: Salary} = 893.59 + 1044.15 \text{ Age}$$

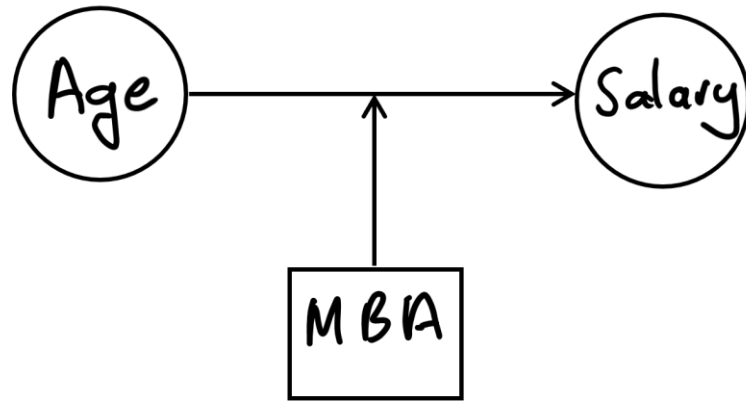
$$\text{MBA: Salary} = 15660.82 + 1044.15 \text{ Age}$$

The models suggest that the rate of salary increase for age is the same for both groups. However, individuals with MBAs might earn relatively higher salaries as they get older. In other words, the slope of *Age* depends on the value of *MBA*, or you can think of MBA is acting like a moderator. Such a dependence is called an **interaction**.

Employee	Salary	Age	MBAFlag	MBAFlag*Age
1	28260	25	0	0
2	43392	28	1	28
3	56322	37	1	37
4	26086	23	0	0
5	36807	32	0	0
6	57119	57	0	0
7	48907	45	0	0
8	34301	32	0	0
9	31104	25	0	0
10	60054	57	0	0
11	41420	42	0	0
12	36508	25	1	25

For Interaction, you need to multiply Columns Age and MBA together to create a new column called “Age * MBAFlag”. Run again with the new column, then you will have the best Regression Model.

Solution 2e Model



SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.989321416					
R Square	0.978756863					
Adjusted R Square	0.976701076					
Standard Error	2005.37675					
Observations	35					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	3	5743939086	1914646362	476.098288	5.31397E-26	
Residual	31	124667613.2	4021535.91			
Total	34	5868606699				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	3902.509386	1336.39766	2.920170772	0.006467654	1176.908389	6628.110383
Age	971.3090382	31.06887722	31.26308786	5.23658E-25	907.9436454	1034.674431
MBAFlag	-2971.080074	3026.24236	-0.98177202	0.333812767	-9143.142058	3200.981911
MBAFlag*Age	501.8483604	81.55221742	6.153705887	7.9295E-07	335.5215164	668.1752044

- The R-Square is 0.9798 or 97.98%
- The Correlation Coefficient r is 0.9893
- The linear relationship is estimated to be: $y = 3902 + 971\text{Age} - 2971\text{MBAFlag} + 502(\text{MBAFlag} * \text{Age})$