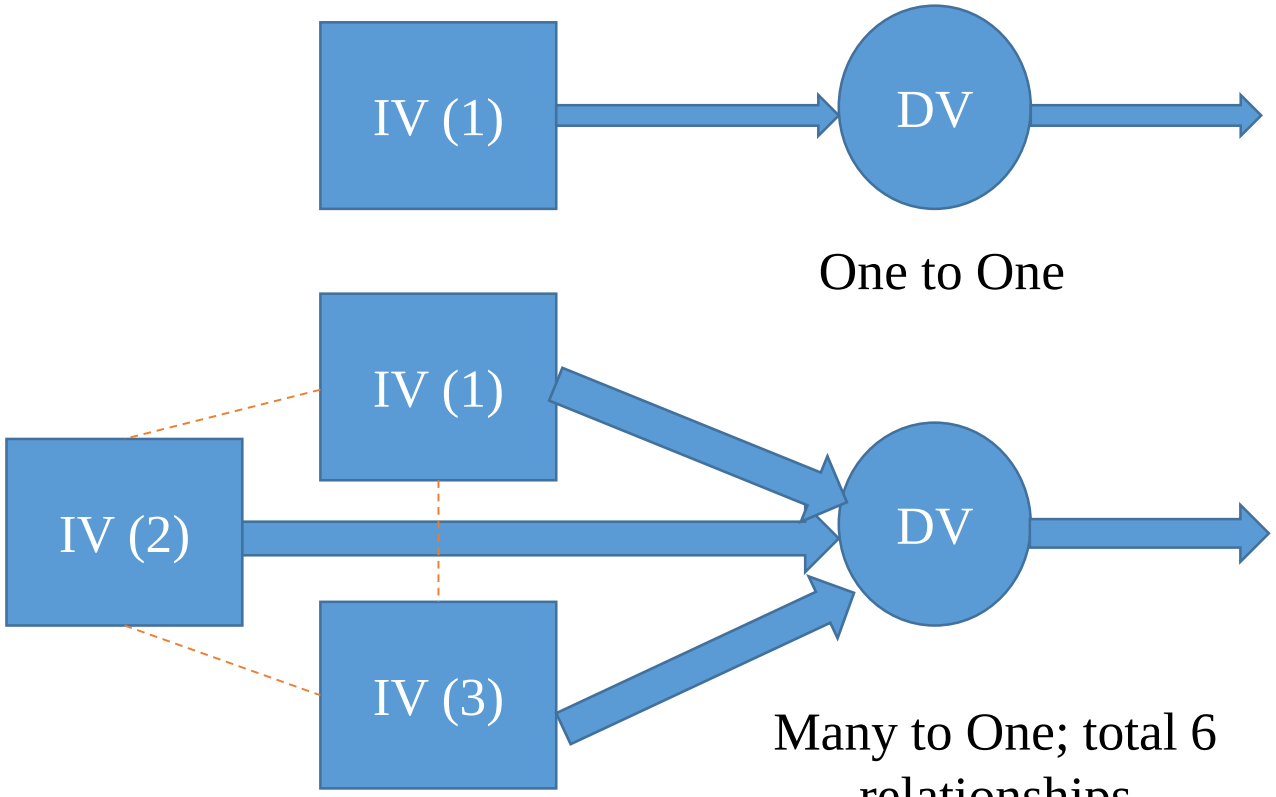
MULTIPLE REGRESSION

Relationship between DV and IVs



Predict Simple Linear Regression

One to One

Predict Multiple Regression

Many to One; total 6 relationships

Question 1:

Suppose you are an owner of a courier delivery service and interested to develop a model that will allow you to predict the total delivery time that will be taken up by the person to deliver the product to the customers. To develop the model, let's assume we have collected the data for last 10 trips and got four piece of information 1.) distance covered in miles, 2.) number of deliveries, 3.) gas price, and 4.) total delivery time taken (in hours). What will be the total delivery time for the 11^{th} trip?

- collected data for 10 trips

Data Collected for 10 Trips

X_1 (distance covered in miles)	X_2 (number of deliveries)	X_3 (gas price)	y (total delivery time in hours)
89	4	3.84	7
66	1	3.19	5.4
78	3	3.78	6.6
111	6	3.89	7.4
44	1	3.57	4.8
77	3	3.57	6.4
80	3	3.03	7
66	2	3.51	5.6
109	5	3.54	7.3
76	3	3.25	6.4

DV: total delivery time in hours (Y)

IV's: distance covered in miles ($\mathbf{X_1}$); number of deliveries ($\mathbf{X_2}$); and gas price ($\mathbf{X_3}$)

- Observe the relationship between DV and IV's and among IV's

Mulitiple Regression

Desirable: Strong linear relationship between IV & DV = Y & X1 and Y &X2

Mulitiple Regression

Desirable: Strong linear relationship between IV & DV = Y & X1 and Y &X2

No relationship Between Y & X3
Strong Linear relationship exist between X1 & X2: Not desirable
No relationship exist between X1 & X3 and X2 & X3: Desirable

Mulitiple Regression

Desirable: Strong linear relationship between IV & DV = Y & X1 and Y &X2

No relationship Between Y & X3
Strong Linear relationship exist between X1 & X2: Not desirable
No relationship exist between X1 & X3 and X2 & X3: Desirable

Simple Linear Regression: One to one

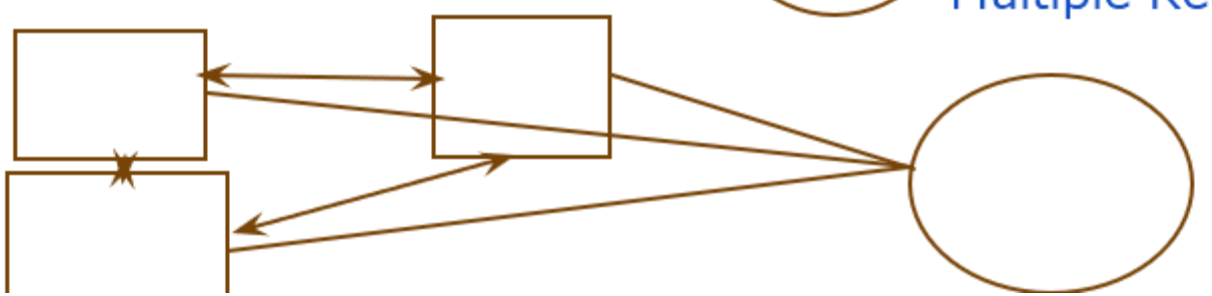
Mulitiple Regression

Desirable: Strong linear relationship between IV & DV = Y & X1 and Y &X2

No relationship Between Y & X3
Strong Linear relationship exist between X1 & X2: Not desirable
No relationship exist between X1 & X3 and X2 & X3: Desirable

Simple Linear Regression: One to one

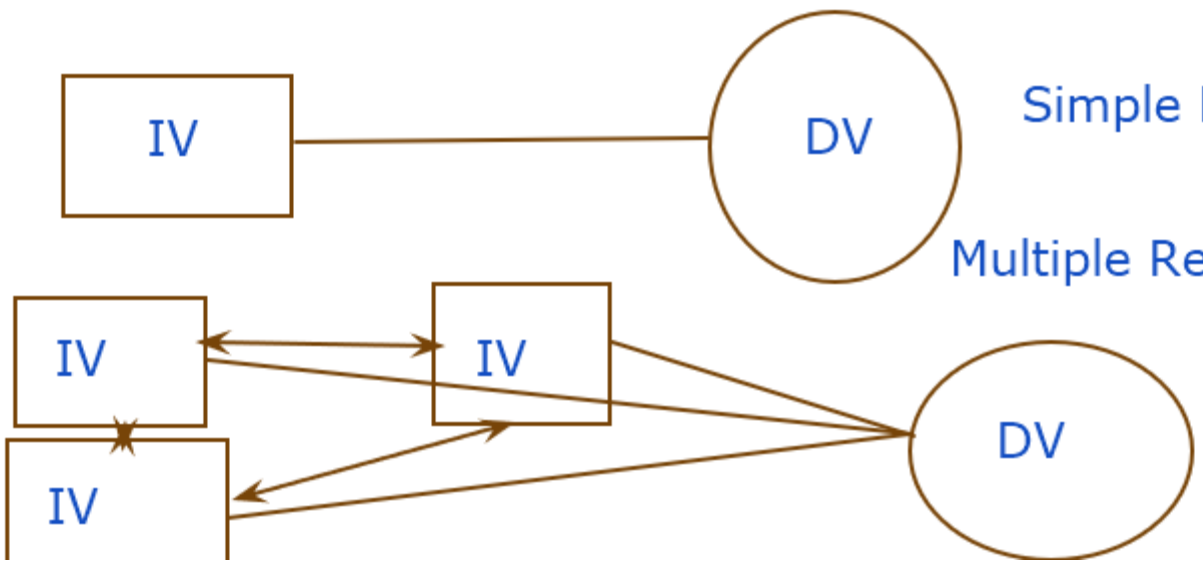
Multiple Regression model: Many to one



Mulitiple Regression

Desirable: Strong linear relationship between IV & DV = Y & X1 and Y &X2

No relationship Between Y & X3
Strong Linear relationship exist between X1 & X2: Not desirable
No relationship exist between X1 & X3 and X2 & X3: Desirable



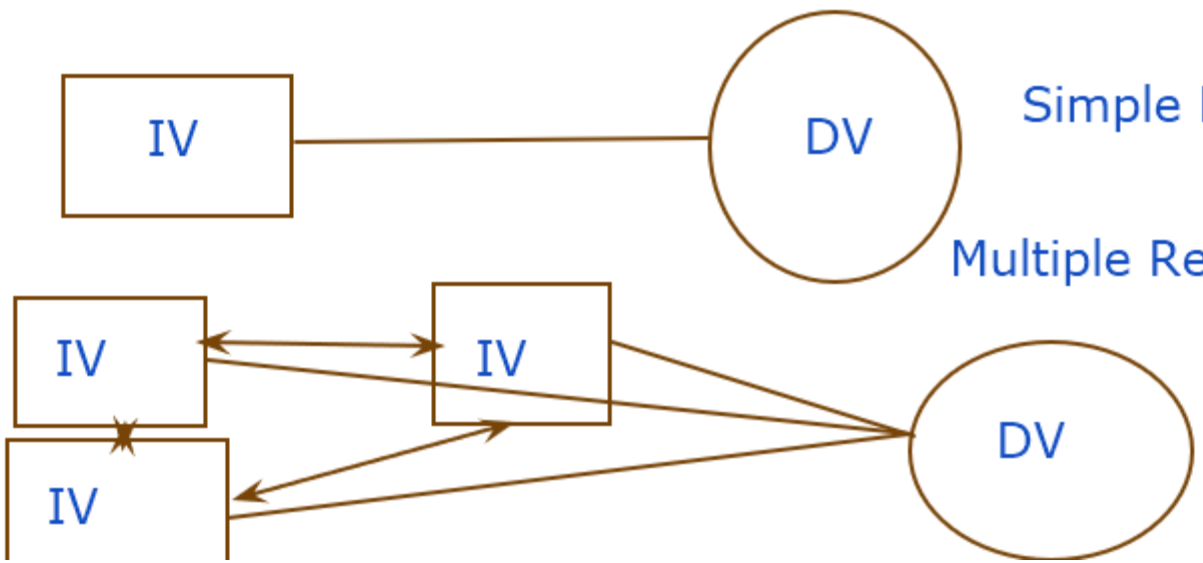
Simple Linear Regression: One to one

Multiple Regression model: Many to one

Mulitiple Regression

Desirable: Strong linear relationship between IV & DV = Y & X1 and Y &X2

No relationship Between Y & X3
Strong Linear relationship exist between X1 & X2: Not desirable
No relationship exist between X1 & X3 and X2 & X3: Desirable

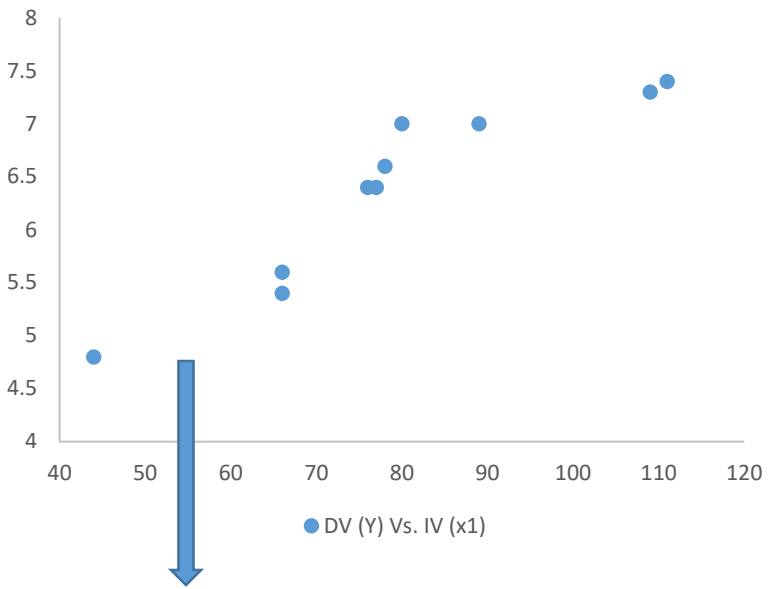


Simple Linear Regression: One to one

Multiple Regression model: Many to one

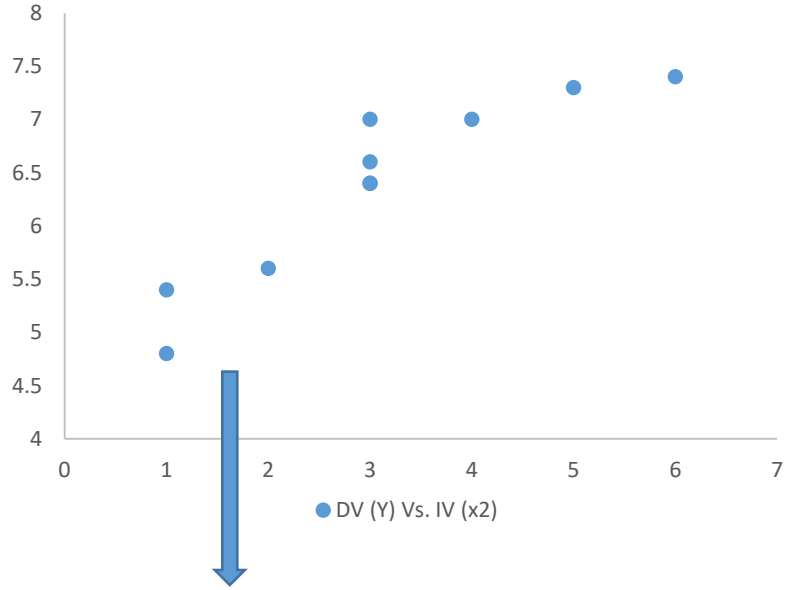
Scatter Plots (DV Vs. IVs)

DV (Y) Vs. IV ($\mathbf{X_1}$)



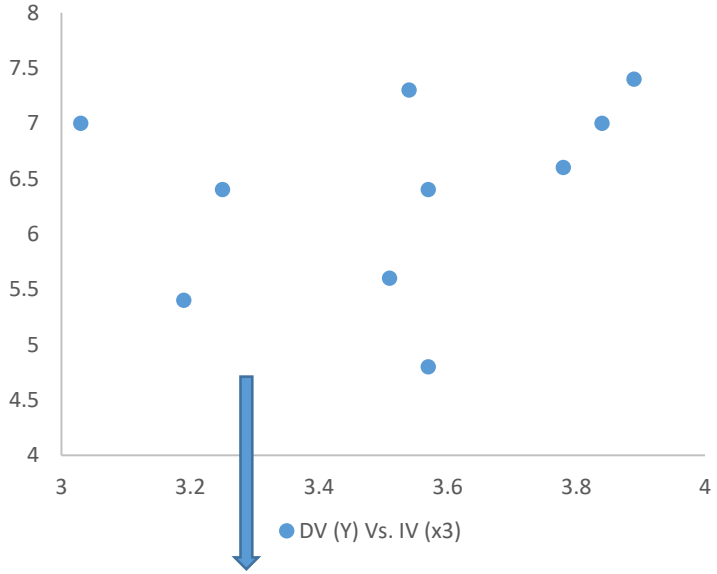
Shows Strong Linear Relationship

DV (Y) Vs. IV ($\mathbf{X_2}$)



Shows Strong Linear Relationship

DV (Y) Vs. IV ($\mathbf{X_3}$)

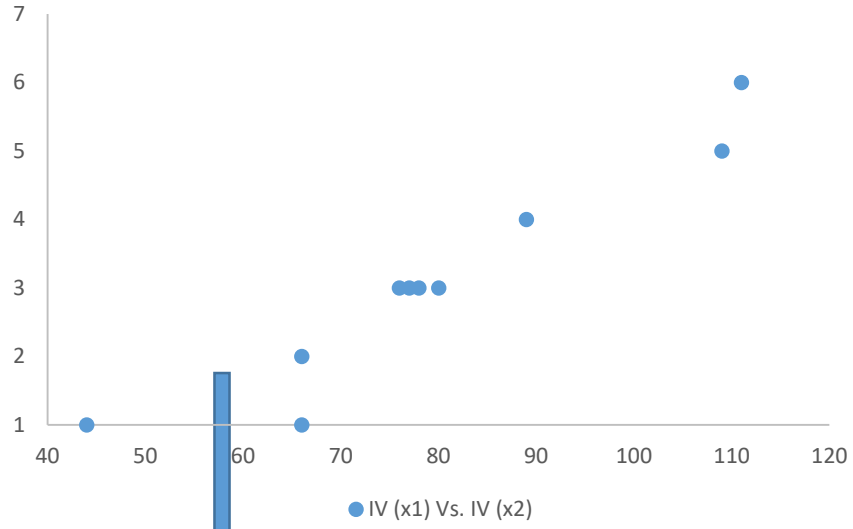


Shows No Linear Relationship; Data Points are Scattered all over the Space

- The IV ($\mathbf{X_1}$ & $\mathbf{X_2}$) shows linear relationship with the DV (Y) whereas IV ($\mathbf{X_3}$) does not show the linear relationship with the DV (Y)

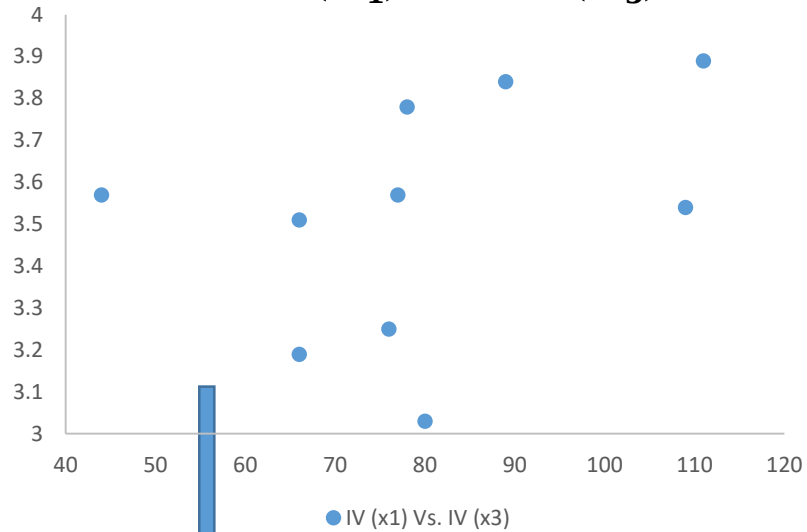
Scatter Plots (IV Vs. IVs)

IV ($\mathbf{X_1}$) Vs. IV ($\mathbf{X_2}$)



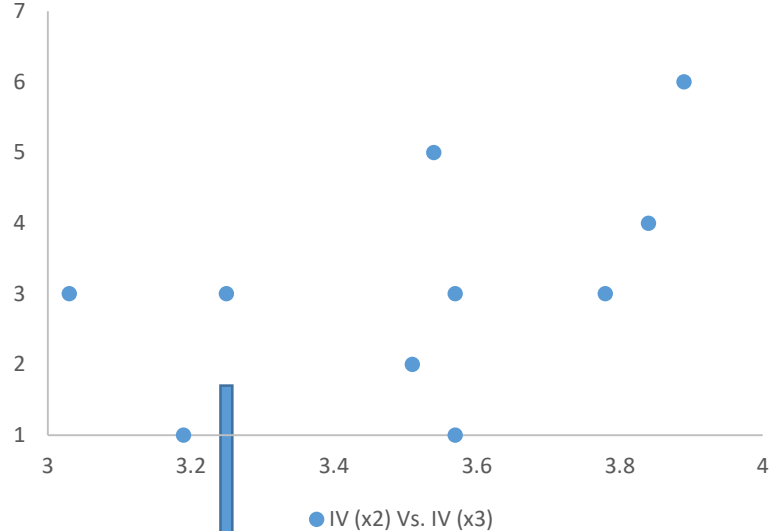
Shows Strong Linear Relationship
* **Multicollinearity Problem**

IV ($\mathbf{X_1}$) Vs. IV ($\mathbf{X_3}$)



Shows No Linear Relationship; Data Points are Scattered all over the Space

IV ($\mathbf{X_2}$) Vs. IV ($\mathbf{X_3}$)



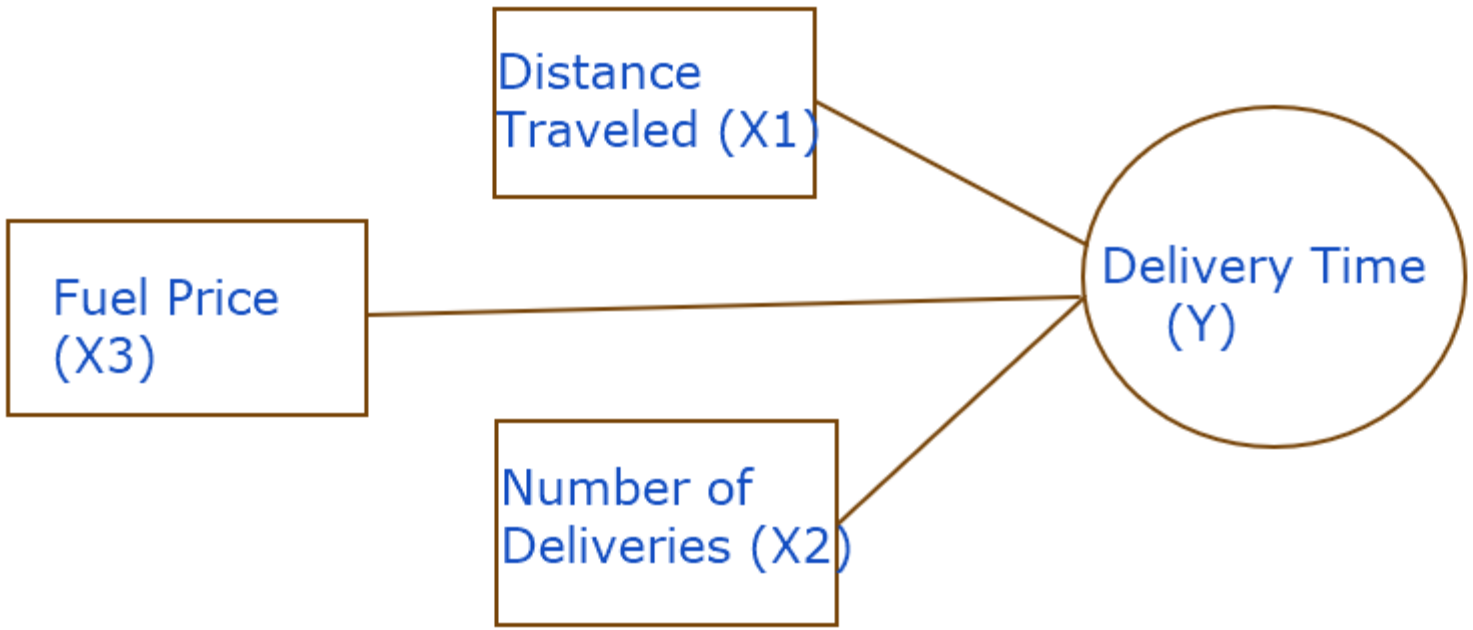
Shows No Linear Relationship; Data Points are Scattered all over the Space

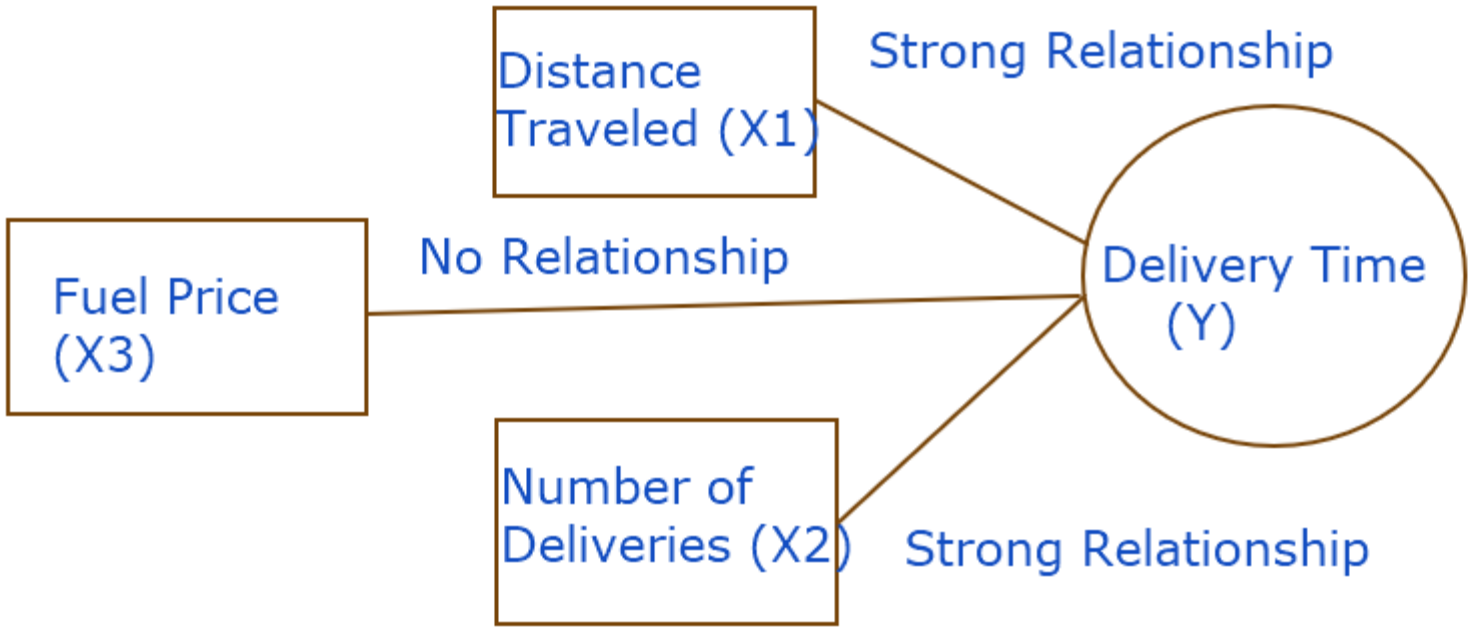
- The IV ($\mathbf{X_1}$) shows strong linear relationship with the IV ($\mathbf{X_2}$) whereas IV ($\mathbf{X_1}$) with IV ($\mathbf{X_3}$) and IV ($\mathbf{X_2}$) with IV ($\mathbf{X_3}$) does not show the linear relationship

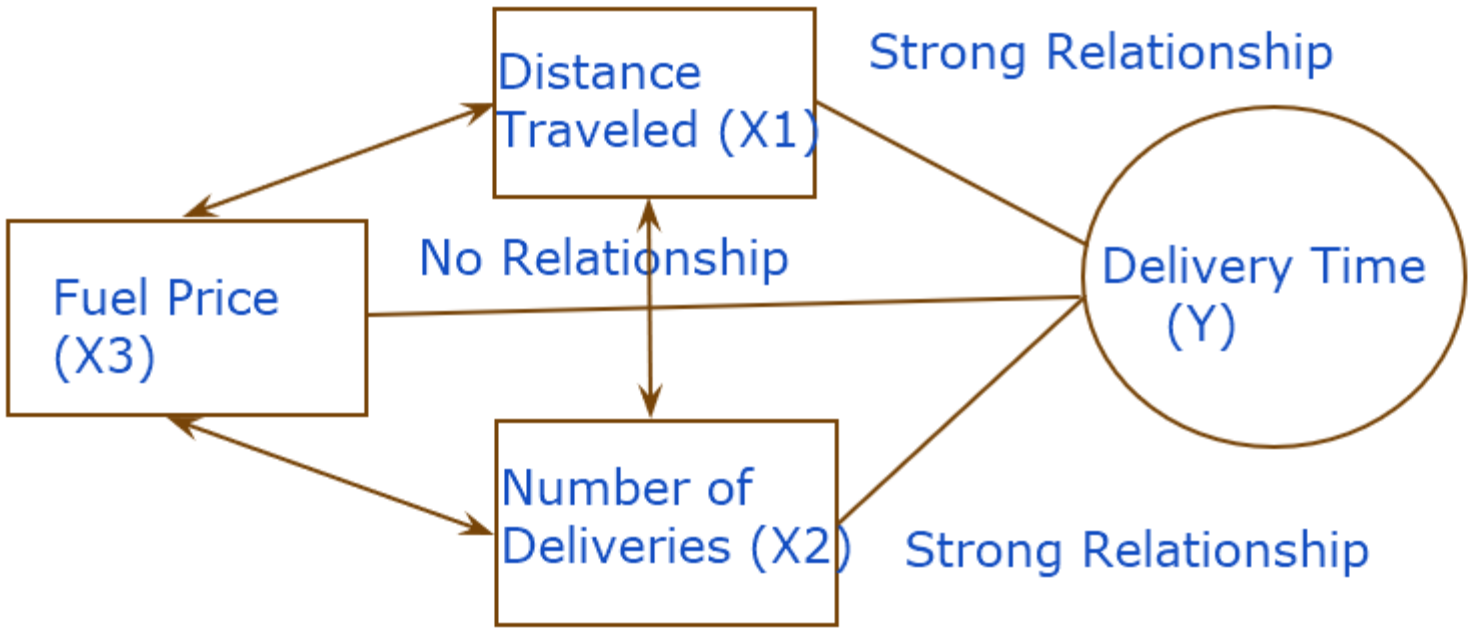
Correlation Coefficient for DV and IVs

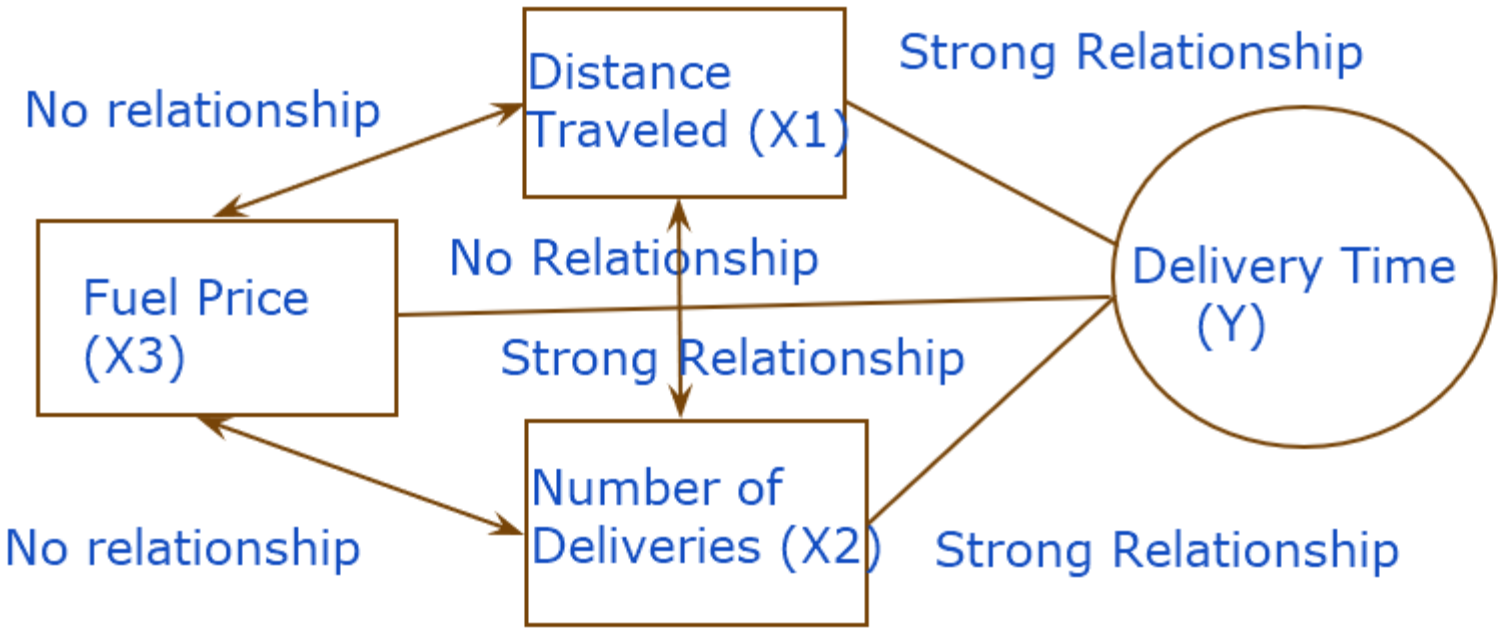
- "r" for DV (Y) Vs. IV ($\mathbf{X_1}$) = 0.928179 – Confirms Strong Linear Relationship
- "r" for DV (Y) Vs. IV ($\mathbf{X_2}$) = 0.916443 – Confirms Strong Linear Relationship
- "r" for DV (Y) Vs. IV ($\mathbf{X_3}$) = 0.267212 – does not have linear relationship
- "r" for IV ($\mathbf{X_1}$) Vs. IV ($\mathbf{X_2}$) = 0.955898 – Confirms Strong Linear Relationship
- "r" for IV ($\mathbf{X_1}$) Vs. IV ($\mathbf{X_3}$) = 0.355796 – does not have linear relationship
- "r" for IV ($\mathbf{X_2}$) Vs. IV ($\mathbf{X_3}$) = 0.498242 – does not have linear relationship

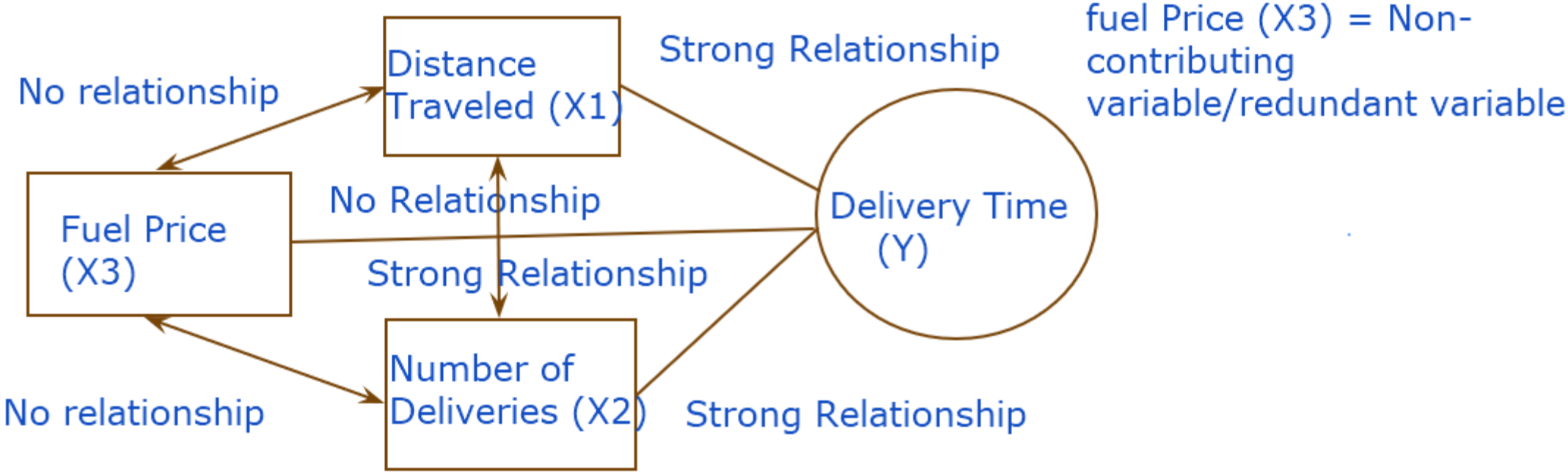
➢ Scatter Plot Results are confirmed with the correlation coefficients

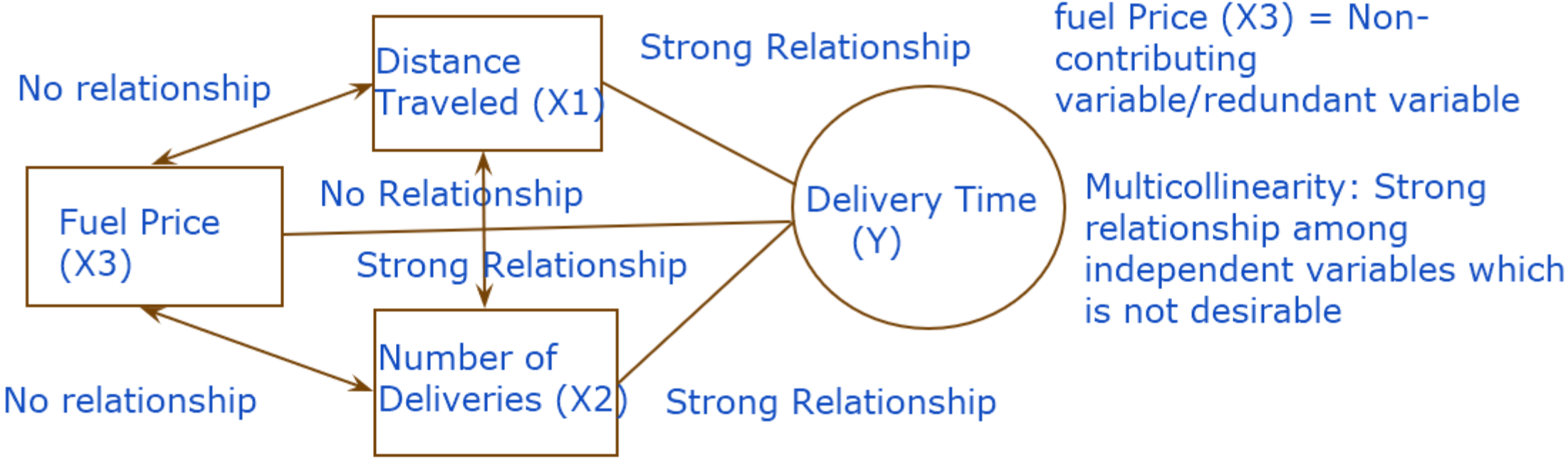

Distance Traveled (X1)
Fuel Price (X3)
Number of Deliveries (X2)
Delivery Time (Y)


Distance Traveled (X1)
Strong Relationship
Fuel Price (X3)
No Relationship
Delivery Time (Y)
Number of Deliveries (X2)
Strong Relationship


Distance Traveled (X1)
Strong Relationship
Fuel Price (X3)
No Relationship
Delivery Time (Y)
Number of Deliveries (X2)
Strong Relationship

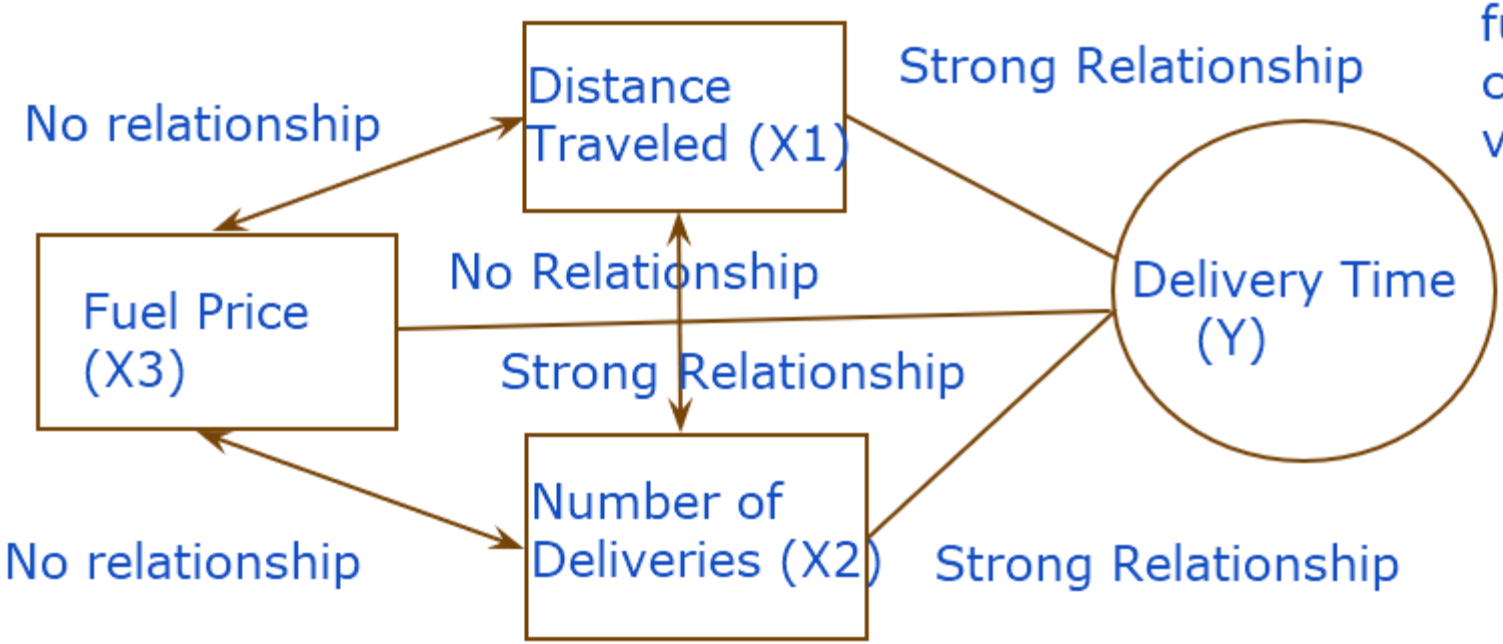

Distance Traveled (X1)
Strong Relationship
No relationship
Delivery Time (Y)
Fuel Price (X3)
No Relationship
Strong Relationship
Number of Deliveries (X2)
No relationship
Strong Relationship


Distance Traveled (X1)
Strong Relationship
fuel Price (X3) = Non-contributing variable/redundant variable
No relationship
Fuel Price (X3)
No Relationship
Delivery Time (Y)
Strong Relationship
Number of Deliveries (X2)
No relationship
Strong Relationship



fuel Price (X3) = Non-contributing variable/redundant variable

Multicollinearity: Strong relationship among independent variables which is not desirable



fuel Price (X3) = Non-contributing variable/redundant variable

Multicollinearity: Strong relationship among independent variables which is not desirable

Overfitting:included redundant variables in the model.

Univariate Regression Model using IV (X_1)

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.928179							
R Square	**0.861515**							
Adjusted R Square	0.844205							
Standard Error	**0.342309**							
Observations	10							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	1	5.831597	5.831597	**49.76813**	**0.000106676**			
Residual	8	0.937403	0.117175					
Total	9	6.769						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	**3.18556**	0.466951	6.822047	0.000135	2.108769788	4.262351	2.10877	4.262351
X Variable 1	**0.040257**	0.005706	7.054653	0.000107	0.027097763	0.053416	0.027098	0.053416

Same as **correlation coefficient "r"**

Coefficient of determination (R Square) indicates the amount of variability explained in the dependent variable that is accounted by independent variable (X_1)

Standard error shows the average distance that the observed values fall from the **regression** line

- **$Y = 3.18556 + 0.040257\ (X_1)$;** For one mile increase in X_1, the travels time is expected to increase by 0.040257 times

Univariate Regression Model using IV (X_2)

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.916443	Same as **correlation coefficient "r"**						
R Square	0.839868							
Adjusted R Square	0.819852							
Standard Error	**0.368091**							
Observations	10							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	1	5.68507	5.68507	41.95894	0.000193			
Residual	8	1.08393	0.135491					
Total	9	6.769						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	4.845415	0.265345	18.26079	8.32E-08	4.233528	5.457302	4.233528	5.457302
X Variable 1	0.498253	0.07692	6.477572	0.000193	0.320876	0.675631	0.320876	0.675631

- **$Y = 4.845415 + 0.498253(X_2)$;** For unit increase in the number of deliveries, the travels time is expected to increase by 0.498253

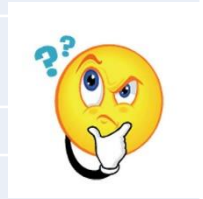
Univariate Regression Model using IV (X_3)

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.267211531							
R Square	**0.071402002**	**Very less**						
Adjusted R Square	-0.044672747							
Standard Error	**0.886402832**							
Observations	10							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	1	0.48332015	0.48332	**0.6151381**	**0.455453413**	**Not Significant**		
Residual	8	6.28567985	0.78571					
Total	9	6.769						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	**3.536488198**	3.64903876	0.969156	0.3608511	-4.878210281	11.95118668	-4.8782103	11.95118668
X Variable 1	**0.811348252**	1.03447735	0.784307	0.4554534	-1.574160801	3.196857305	-1.5741608	3.196857305

- **$Y = 3.536488198 + 0.811348252(X_3)$;** For unit increase in the gas price, the travels time is expected to increase by 0.811348252

Bivariate Regression Model using IV's (X_1 & X_2)

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.933487816							
R Square	0.871399503							
Adjusted R Square	0.834656504							
Standard Error	**0.352642426**							
Observations	10							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	2	5.8985032	2.94925	**23.716069**	**0.000762692**			
Residual	7	0.8704968	0.12436					
Total	9	6.769						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	3.732158132	0.8869736	4.20774	0.0039969	1.634798789	5.8295175	1.634798789	5.829517474
X Variable 1	0.026222566	0.0200161	1.31008	**0.2315209**	-0.02110789	0.073553	-0.021107895	0.073553027
X Variable 2	0.184040518	0.2509086	0.7335	**0.487089**	-0.40926407	0.7773451	-0.409264074	0.777345109



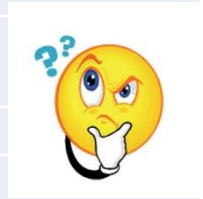
Shock!

The overall regression model is significant but the individual IV is not significant; why?

- **$Y = 3.732158132 + 0.026222566\ (X_1) + 0.1844040518\ (X_2)$;** For one mile increase in X_1, the travels time is expected to increase by **0.026222566** provided the value of variable 'X_2' is constant

Bivariate Regression Model using IV's (X_1 & X_2)

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.933487816							
R Square	0.871399503							
Adjusted R Square	0.834656504							
Standard Error	**0.352642426**							
Observations	10							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	2	5.8985032	2.94925	**23.716069**	**0.000762692**			
Residual	7	0.8704968	0.12436					
Total	9	6.769						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	3.732158132	0.8869736	4.20774	0.0039969	1.634798789	5.8295175	1.634798789	5.829517474
X Variable 1	0.026222566	0.0200161	1.31008	**0.2315209**	-0.02110789	0.073553	-0.021107895	0.073553027
X Variable 2	0.184040518	0.2509086	0.7335	**0.487089**	-0.40926407	0.7773451	-0.409264074	0.777345109



Shock!

The overall regression model is significant but the individual IV is not significant; why?

Multicollinearity Problem!

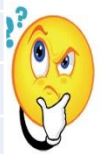
- **$Y = 3.732158132 + 0.026222566\ (X_1) + 0.1844040518\ (X_2)$;** For one mile increase in X_1, the travels time is expected to increase by **0.026222566** provided the value of variable 'X_2' is constant

Bivariate Regression Model using IV's (X_2 & X_3)

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.94211797							
R Square	0.88758627							
Adjusted R Square	0.855468062							
Standard Error	**0.329703013**							
Observations	10							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	2	6.008071	3.004036	**27.63499**	**0.00047629**			
Residual	7	0.760929	0.108704					
Total	9	6.769						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	7.324307017	1.457572	5.025005	0.001522	3.87769682	10.77092	3.877697	10.77092
X Variable 1	0.566500812	0.079463	7.129079	0.000189	0.37859975	0.754402	0.3786	0.754402
X Variable 2	-0.764987072	0.443787	-1.72377	**0.12841**	-1.81437682	0.284403	-1.81438	0.284403

Shock!

The overall regression model is significant but the individual IV is not significant; why?



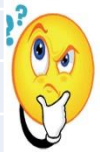
- **$Y = 7.324307017 + 0.566500812\ (X_1) - 0.764987072\ (X_2)$;** For one unit increase in X_1, the travels time is expected to increase by **0.566500812** provided the value of variable 'X_2' is constant

Bivariate Regression Model using IV's (X_2 & X_3)

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.94211797							
R Square	0.88758627							
Adjusted R Square	0.855468062							
Standard Error	**0.329703013**							
Observations	10							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	2	6.008071	3.004036	**27.63499**	**0.00047629**			
Residual	7	0.760929	0.108704					
Total	9	6.769						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	7.324307017	1.457572	5.025005	0.001522	3.87769682	10.77092	3.877697	10.77092
X Variable 1	0.566500812	0.079463	7.129079	0.000189	0.37859975	0.754402	0.3786	0.754402
X Variable 2	-0.764987072	0.443787	-1.72377	**0.12841**	-1.81437682	0.284403	-1.81438	0.284403

Shock!

The overall regression model is significant but the individual IV is not significant; why?



There is no relationship between DV and Gas Price

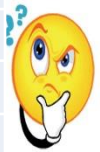
- **$Y = 7.324307017 + 0.566500812\ (X_1) - 0.764987072\ (X_2)$;** For one unit increase in X_1, the travels time is expected to increase by **0.566500812** provided the value of variable 'X_2' is constant

Bivariate Regression Model using IV's (X_3 & X_1)

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.93062562							
R Square	0.86606405							
Adjusted R Square	0.82779663							
Standard Error	**0.3598834**							
Observations	10							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	2	5.862388	2.9311938	**22.63189335**	**0.000879306**			
Residual	7	0.906612	0.1295161					
Total	9	6.769						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	3.86756989	1.482416	2.6089638	0.034966016	0.362212889	7.372927	0.362213	7.372927
X Variable 1	-0.21912293	0.44941	-0.487579	**0.640746929**	-1.28180896	0.843563	-1.28181	0.843563
X Variable 2	0.04137042	0.006419	6.4445363	0.000352083	0.026190818	0.05655	0.026191	0.05655

Shock!

The overall regression model is significant but the individual IV is not significant; why?



There is no relationship between DV and Gas Price

- **$Y = 3.86756989 - 0.21912293\ (X_1) + 0.04137042\ (X_2)$;** For one unit increase in X_1, the travels time is expected to decrease by **0.21912293** provided the value of variable 'X_2' is constant

Multiple Regression Model using IV (X_1, X_2 & X_3)

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.945877513							
R Square	0.894684269							
Adjusted R Square	0.842026404							
Standard Error	**0.344693628**							
Observations	10							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	3	6.056117819	2.01870594	**16.99051534**	**0.002452078**			
Residual	6	0.712882181	0.118813697					
Total	9	6.769						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	6.211377664	2.320572882	2.676657007	0.036699328	0.533140378	11.88961	0.53314	11.88961
X Variable 1	0.014121891	0.022207306	0.635911928	**0.548306757**	-0.040217429	0.068461	-0.04022	0.068461
X Variable 2	0.383150235	0.300056891	1.276925301	**0.248817455**	-0.351062526	1.117363	-0.35106	1.117363
X Variable 3	-0.606552713	0.526627587	-1.151767829	**0.293234725**	-1.895163998	0.682059	-1.89516	0.682059

Shock!

The overall regression model is significant but the individual IV is not significant; why?

- **$Y = 6.211377664 + 0.014121891(X_1) + 0.383150235(X_2) - 0.606552713(X_3)$;**

Multiple Regression Model using IV (X_1, X_2 & X_3)

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.945877513							
R Square	0.894684269							
Adjusted R Square	0.842026404							
Standard Error	**0.344693628**							
Observations	10							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	3	6.056117819	2.01870594	**16.99051534**	**0.002452078**			
Residual	6	0.712882181	0.118813697					
Total	9	6.769						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	6.211377664	2.320572882	2.676657007	0.036699328	0.533140378	11.88961	0.53314	11.88961
X Variable 1	0.014121891	0.022207306	0.635911928	**0.548306757**	-0.040217429	0.068461	-0.04022	0.068461
X Variable 2	0.383150235	0.300056891	1.276925301	**0.248817455**	-0.351062526	1.117363	-0.35106	1.117363
X Variable 3	-0.606552713	0.526627587	-1.151767829	**0.293234725**	-1.895163998	0.682059	-1.89516	0.682059

Shock!

The overall regression model is significant but the individual IV is not significant; why?

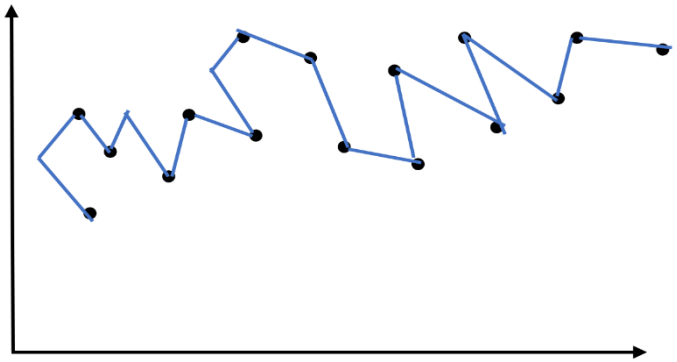
Multicollinearity & Overfitting Issue

- **$Y = 6.211377664 + 0.014121891(X_1) + 0.383150235(X_2) - 0.606552713(X_3)$;**

Important Issues in Regression Model

- **Multicollinearity:** It happens when independent variables in the regression model are highly correlated to each other.

- **Overfitting:** It occurs when the model is too complex and we have a independent variable in the regression model which is irrelevant and does not help in explaining any variability that is present in the dependent variable.



- Overfitting issue occurs when a model is too **closely fit the training set** and **getting a drastic difference of fitting in test set.**