Unit 3

Multiple Linear Regression

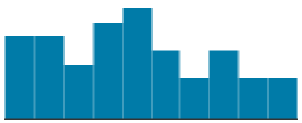
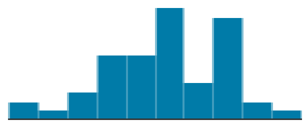
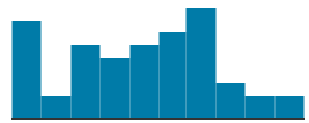
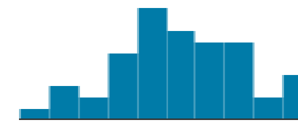
Introduction

- Multiple regression is an extension of simple linear regression.
- We consider the problem of regression when a study variable depends on more than one explanatory or independent variables, called as multiple linear regression model.
- Multiple linear regression is what you can use when you have a bunch of different independent variables!

Example 1: 50 start ups

- You have a dataset in front of you with information on 50 companies.

This dataset has data collected from New York, California and Florida about 50 business Startups "17 in each state". The variables used in the dataset are Profit, R&D spending, Administration Spending, and Marketing Spending.

# R&D Spend	# Administration	# Marketing Spend	State	# Profit
 0 165k	 51.3k 183k	 0 472k	New York 34% California 34% Other (16) 32%	 14.7k 192k
165349.2	136897.8	471784.1	New York	192261.83
162597.7	151377.59	443898.53	California	191792.06
153441.51	101145.55	407934.54	Florida	191050.39
144372.41	118671.85	383199.62	New York	182901.99
142107.34	91391.77	366168.42	Florida	166187.94
131876.9	99814.71	362861.36	New York	156991.12
134615.46	147198.87	127716.82	California	156122.51
130298.13	145530.06	323876.68	Florida	155752.6
120542.52	148718.95	311613.29	New York	152211.77
123334.88	108679.17	304981.62	California	149759.96
101913.08	110594.11	229160.95	Florida	146121.95

Example 1: 50 start ups

- You’ve been hired to analyze this information and create a model.
- You need to inform the guy who hired you what kind of companies will make the most sense in the future to invest in.
- To keep things simple, let’s say that your employer wants to make this decision based on last year’s profit.
- This means that the profits column is your dependent variable.
- The other columns are the independent variables.

Example 2: Region delivery Service

- You are a small business owner for regional delivery service who offers same day delivery for letters, packages and small cargo.
- You are able to use google maps to group individual deliveries into one group to reduce time and costs.

milesTraveled, (x_1)	numDeliveries, (x_2)	travelTime(hrs), (y)
89	4	7
66	1	5.4
78	3	6.6
111	6	7.4
44	1	4.8
77	3	6.4
80	3	7
66	2	5.6
109	5	7.3

As the owner you would like to estimate how long a delivery will take based on two factors:

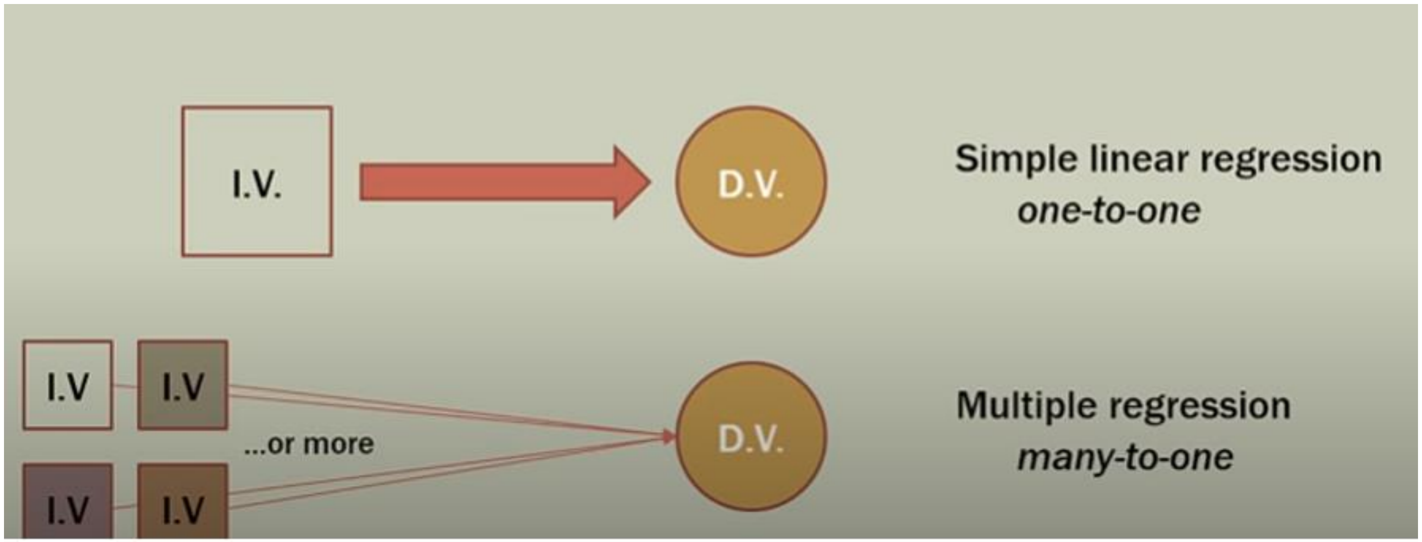
The total distance of the trip in miles

The number of deliveries that must be made during the trip.

Example 3: Examination Performance

- Let us assume we are having data of students like revision time, test anxiety, lecture attendance and gender.
- We want to use this data to predict examination performance.
- We will use multiple regression to understand whether this can be predicted.

Multiple Linear Regression



Simple Linear Regression

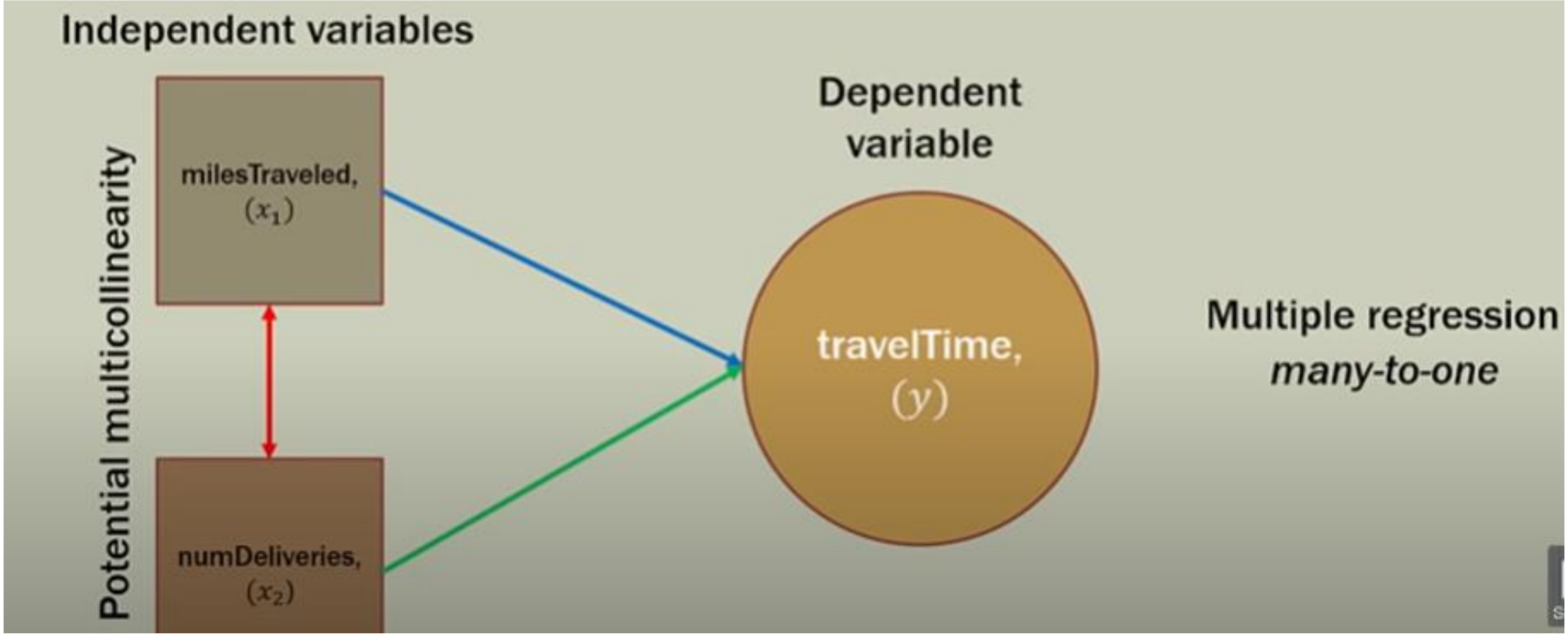
$$y = b_0 + b_1{*}x_1$$

Multiple Linear Regression

Dependent variable (DV) Independent variables (IVs)

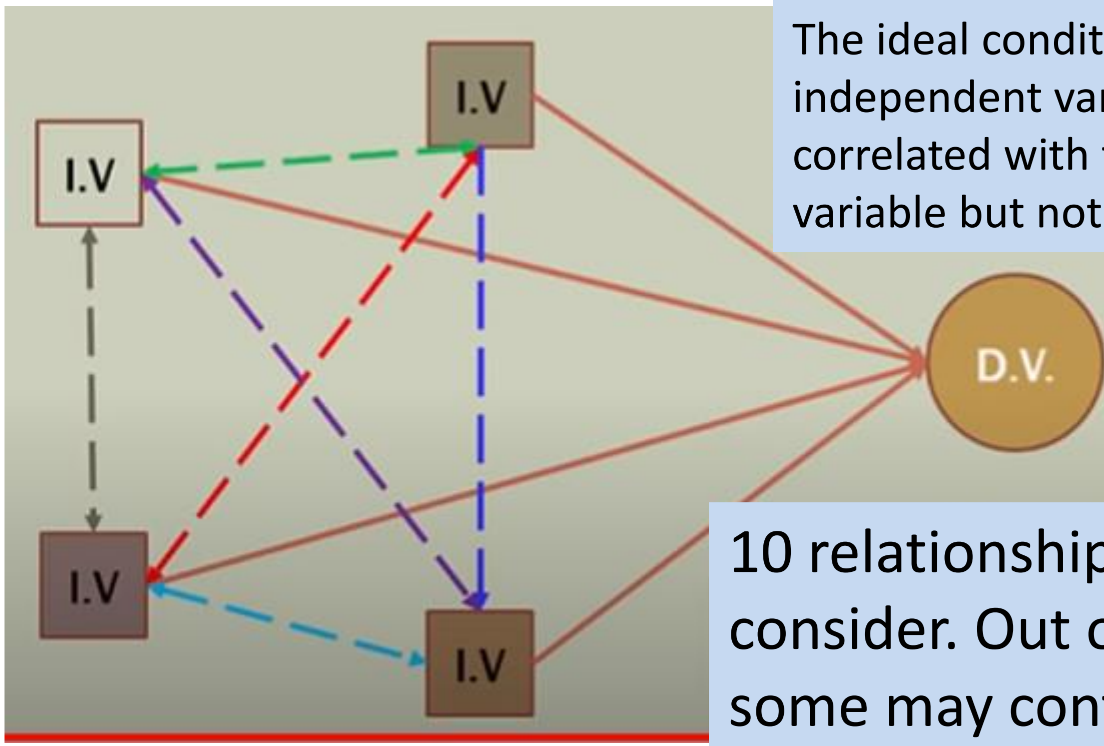
$$y = b_0 + b_1{*}x_1 + b_2{*}x_2 + \ldots + b_n{*}x_n$$

Considerations for MLR



- If independent variables are related to each other, then we will not know that the dependent variable is changing because of which variable

Considerations for MLR



The ideal condition would be for the independent variables to be correlated with the dependent variable but not with each other.

10 relationships to consider. Out of which some may contribute nothing.

Assumptions in Multiple Linear Regression

- There are some assumptions that absolutely have to be true:
  - There is a linear relationship between the dependent variable and the independent variables.
  - The independent variables aren't too highly correlated with each other.
  - Your observations for the dependent variable are selected independently and at random.
  - Regression residuals are normally distributed.

Multiple Regression model

Multiple Regression Model

$$y = \boxed{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p} + \boxed{\epsilon}$$

linear parameters error

Multiple Regression Equation

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p$$

error term assumed to be zero

Estimated Multiple Regression Equation

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots b_p x_p$$

$b_0, b_1, b_2, \ldots b_p$ are the estimates of $\beta_0, \beta_1, \beta_2, \ldots \beta_p$

$\hat{y}$ = predicted value of the dependent variable

Multiple Regression model

variables

Example $\hat{y} = 6.211 + 0.014x_1 + 0.383x_2 - 0.607x_3$

intercept

coefficients

Estimated Multiple Regression Equation $\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3$

$b_0, b_1, b_2, \ldots b_p$ are the estimates of $\beta_0, \beta_1, \beta_2, \ldots \beta_p$

$\hat{y}$ = predicted value of the dependent variable

Interpreting coefficients

$$\hat{y} = 27 + 9x_1 + 12x_2$$

x_1 = capital investment ($1000s)
x_2 = marketing expenditures ($1000s)
$\hat{y}$ = predicted sales ($1000s)

- In Multiple regression, each coefficient is interpreted as the estimate change in y corresponding to a one unit change in a variable, when all other variables are held constant.
- In this example, $9000 is an estimate of the expected increase in sales y, corresponding to a $1000 increase in capital investment(x1), when marketing expenditures(x2) are held constant

Interpreting coefficients

$$\hat{y} = 27 + 9x_1 + 12x_2$$

x_1 = capital investment (\$1000s)
x_2 = marketing expenditures (\$1000s)
$\hat{y}$ = predicted sales (\$1000s)

- In Multiple regression, each coefficient is interpreted as the estimate change in y corresponding to a one unit change in a variable, when all other variables are held constant.
- In this example, $9000 is an estimate of the expected increase in sales y, corresponding to a $1000 increase in capital investment(x1), when marketing expenditures(x2) are held constant

Example 4: Region delivery Service

- You are a small business owner for regional delivery service who offers same day delivery for letters, packages and small cargo.
- You are able to use google maps to group individual deliveries into one group to reduce time and costs.

As the owner you would like to estimate how long a delivery will take based on three factors:

  - The total distance of the trip in miles
  - The number of deliveries that must be made during the trip.
  - The daily price of gas/petrol

Preparation for Multiple linear regression

- Before conducting Multiple linear regression, a lot of pre-work is required:
  - Generate a list of potential variables: Independent and dependent
  - Collect data on variables
  - Check the relationships between independent variable and dependent variable using scatter plot and correlations
  - Check the relationships between independent variable using scatter plot and correlations
  - Conduct simple linear regression for each IV/DV pair (optional step)
  - Use the non-redundant independent variables in the analysis to find the best fitting plot.
  - Use the best fitting model to make predictions about the dependent variable

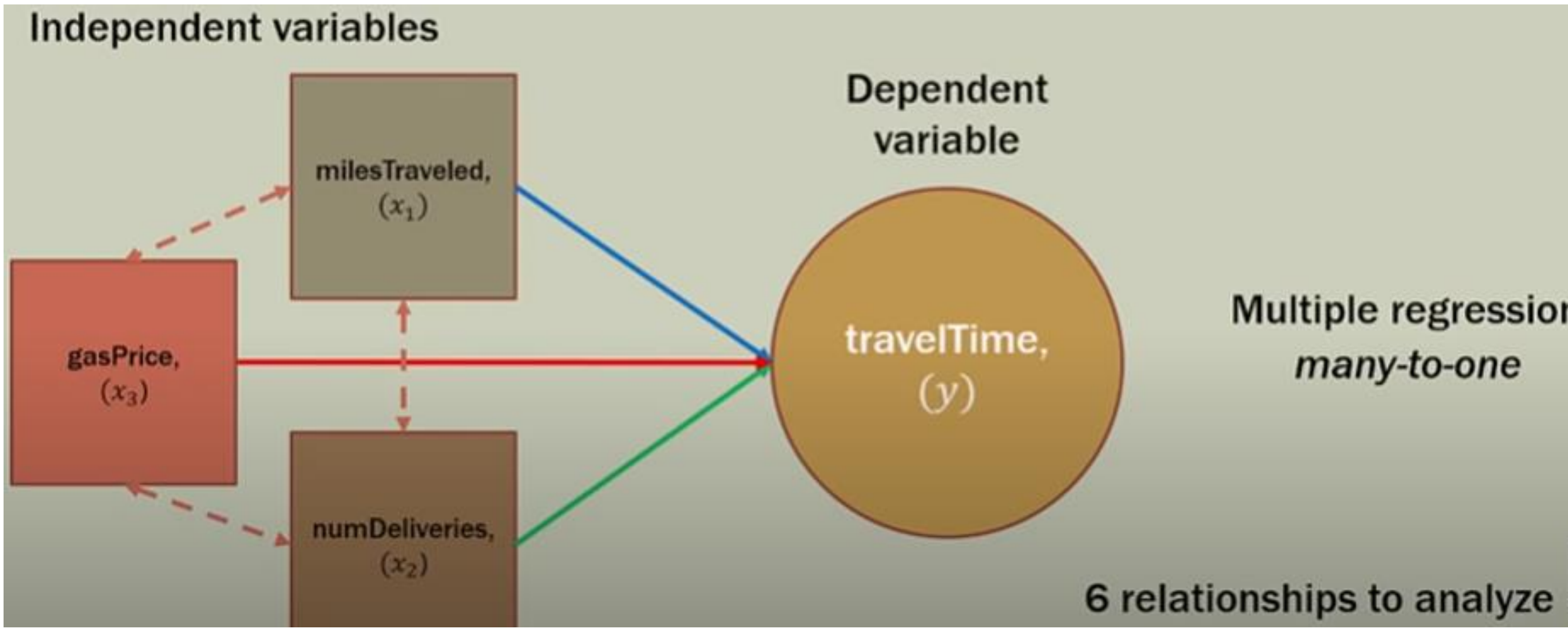
Example 4: Region delivery Service

- To conduct the analysis you take a random sample of past few trips and record four pieces of information for each trip.
  - The total distance of the trip in miles
  - The number of deliveries that must be made during the trip.
  - The daily price of gas/petrol
  - Total travel time in hours

milesTraveled,(x_1)	numDeliveries,(x_2)	gasPrice,(x3)	travelTime(hrs),(y)
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

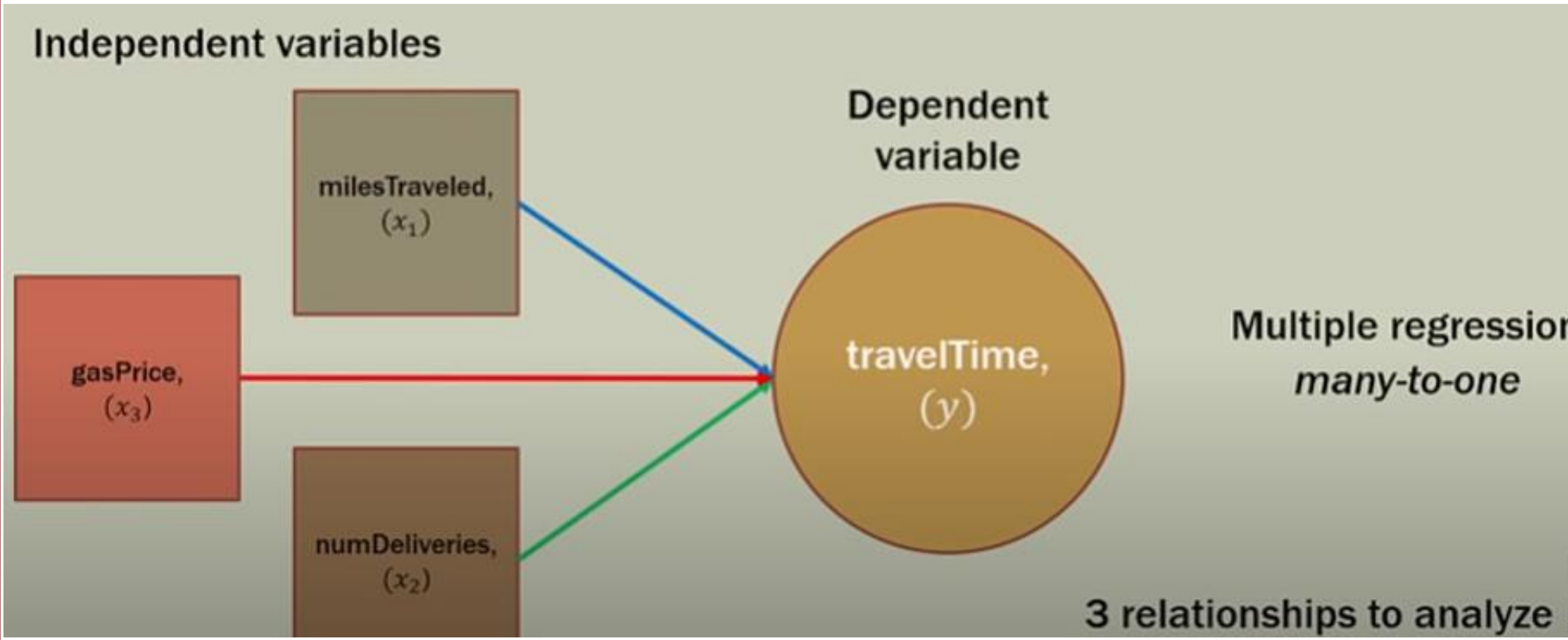
Example 4: Region delivery Service

- Sketching out the relationships



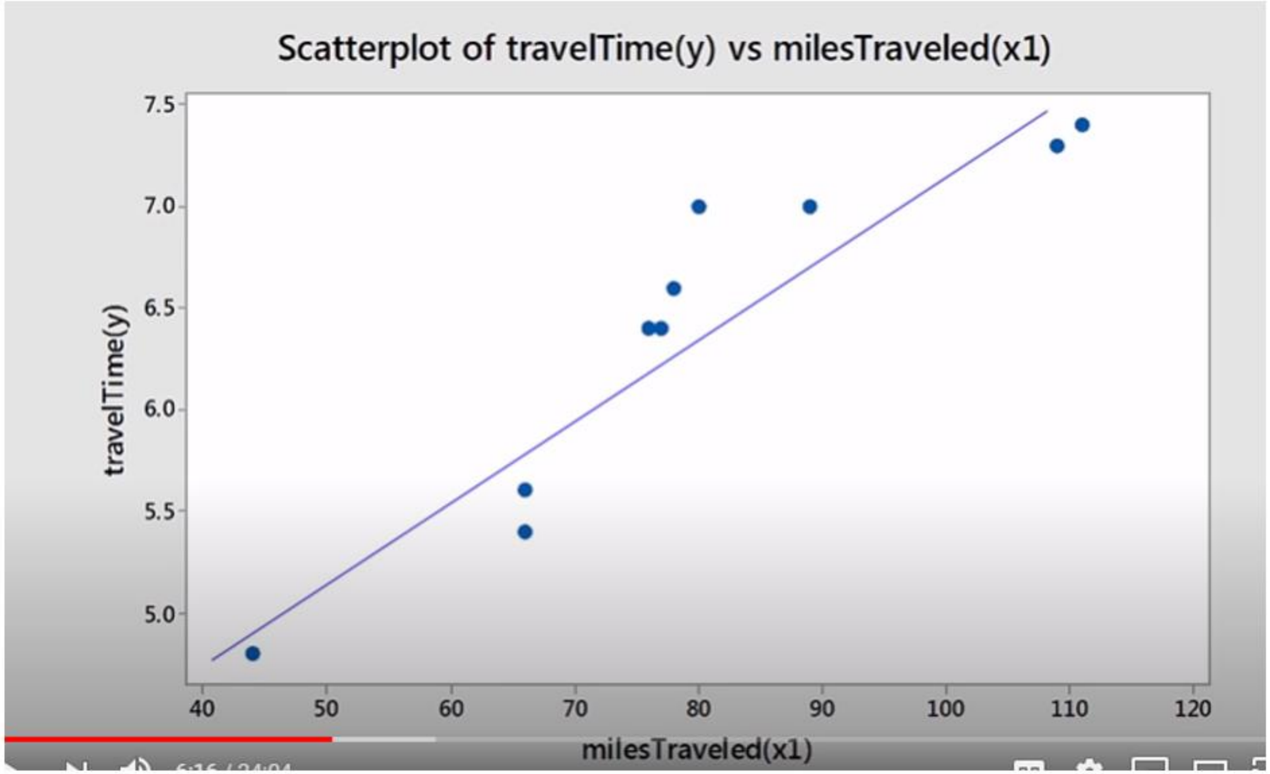
Example 4: Region delivery Service

- Relationships of IV to DV



Example 4: Region delivery Service

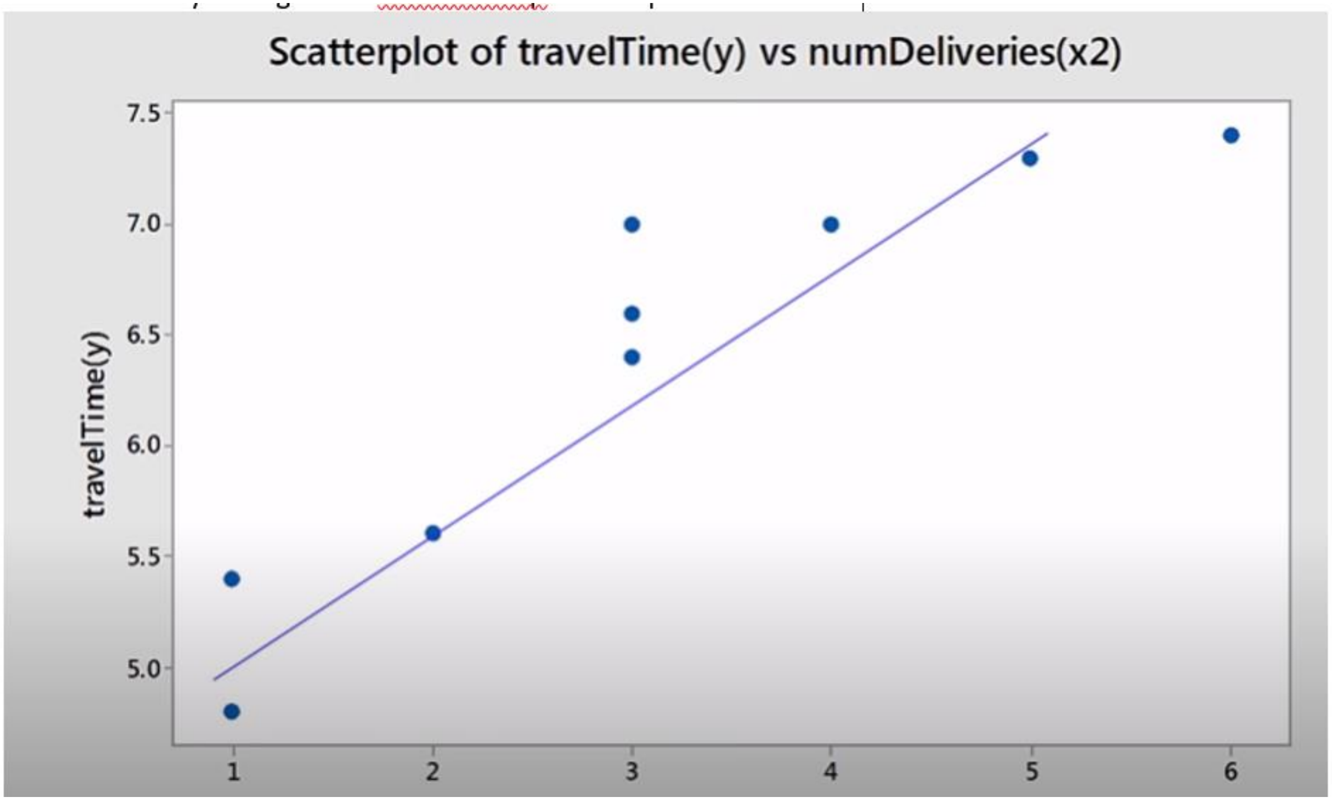
- Draw IV to DV scatter plots. Relevancy check



- x1 has relatively strong linear relationship with dependent variable

Example 4: Region delivery Service

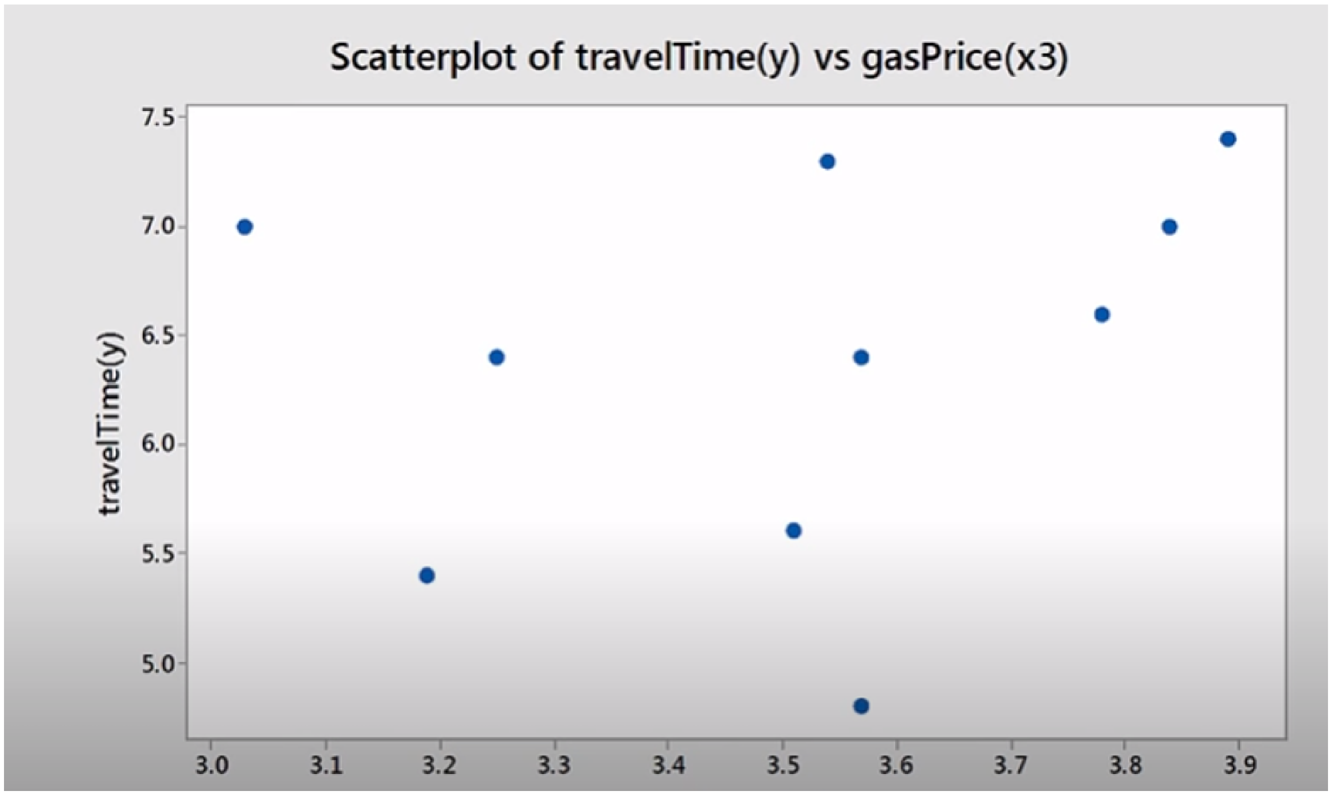
- Draw IV to DV scatter plots. Relevancy check



- x1 has relatively strong linear relationship with dependent variable

Example 4: Region delivery Service

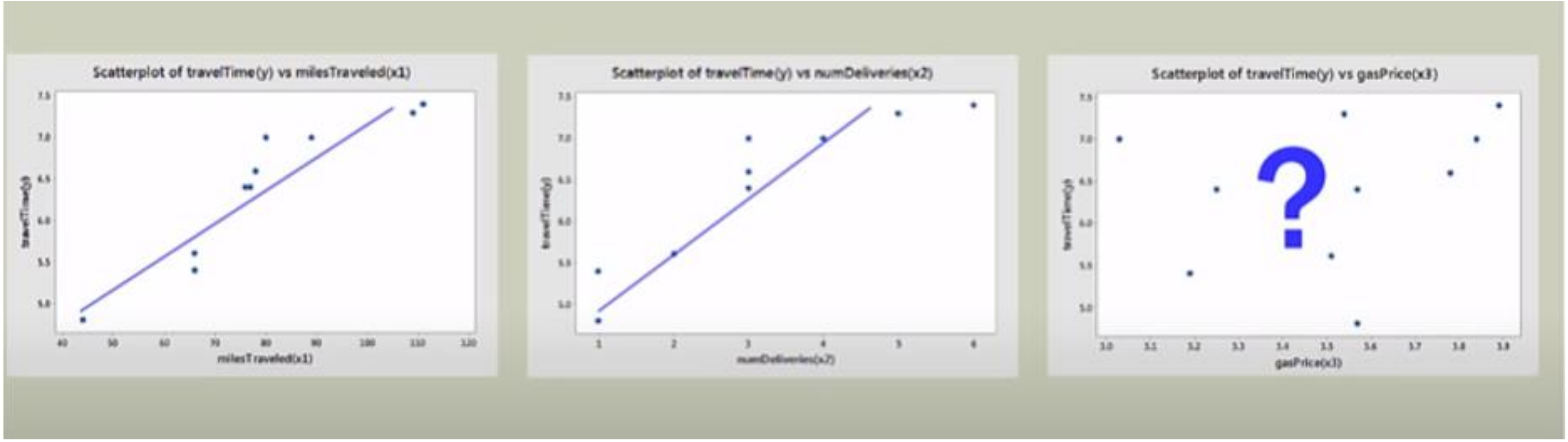
- Draw IV to DV scatter plots. Relevancy check



- Data points are scattered all over the place. there is no relationship

Example 4: Region delivery Service

- DV vs IV scatter plots.

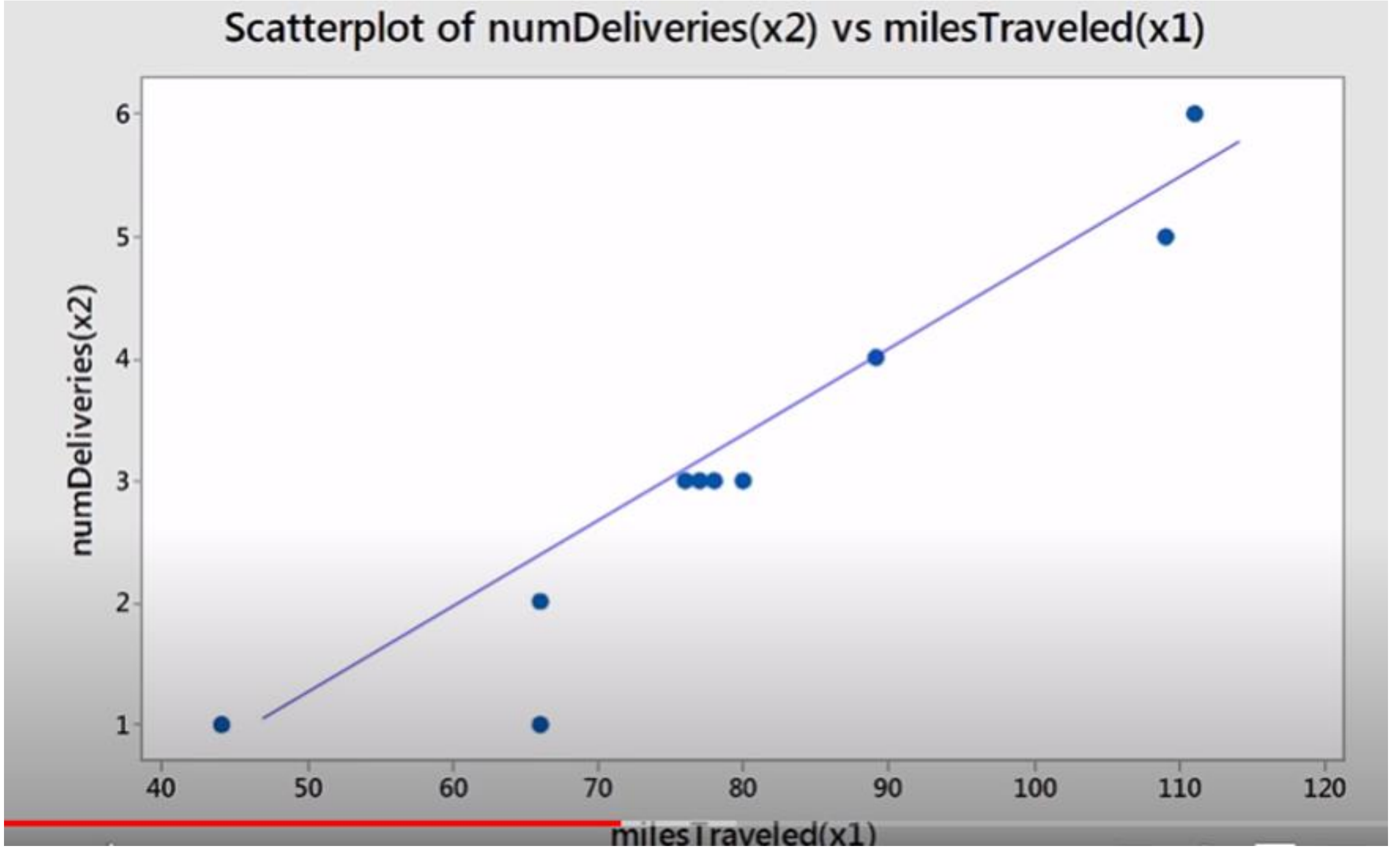


Example 4: Region delivery Service

- Scatter plot summary
- Dependent variable vs independent variable
- Travel time (y) appears highly correlated with milesTraveled (x1)
- Travel time (y) appears highly correlated with numDeliveries (x2)
- Travel time (y) does not appear highly correlated with gasPrices (x3)
- Since gas price (x3) does not appear correlated with the dependent variable, we would not use that variable in the multiple regression

Example 4: Region delivery Service

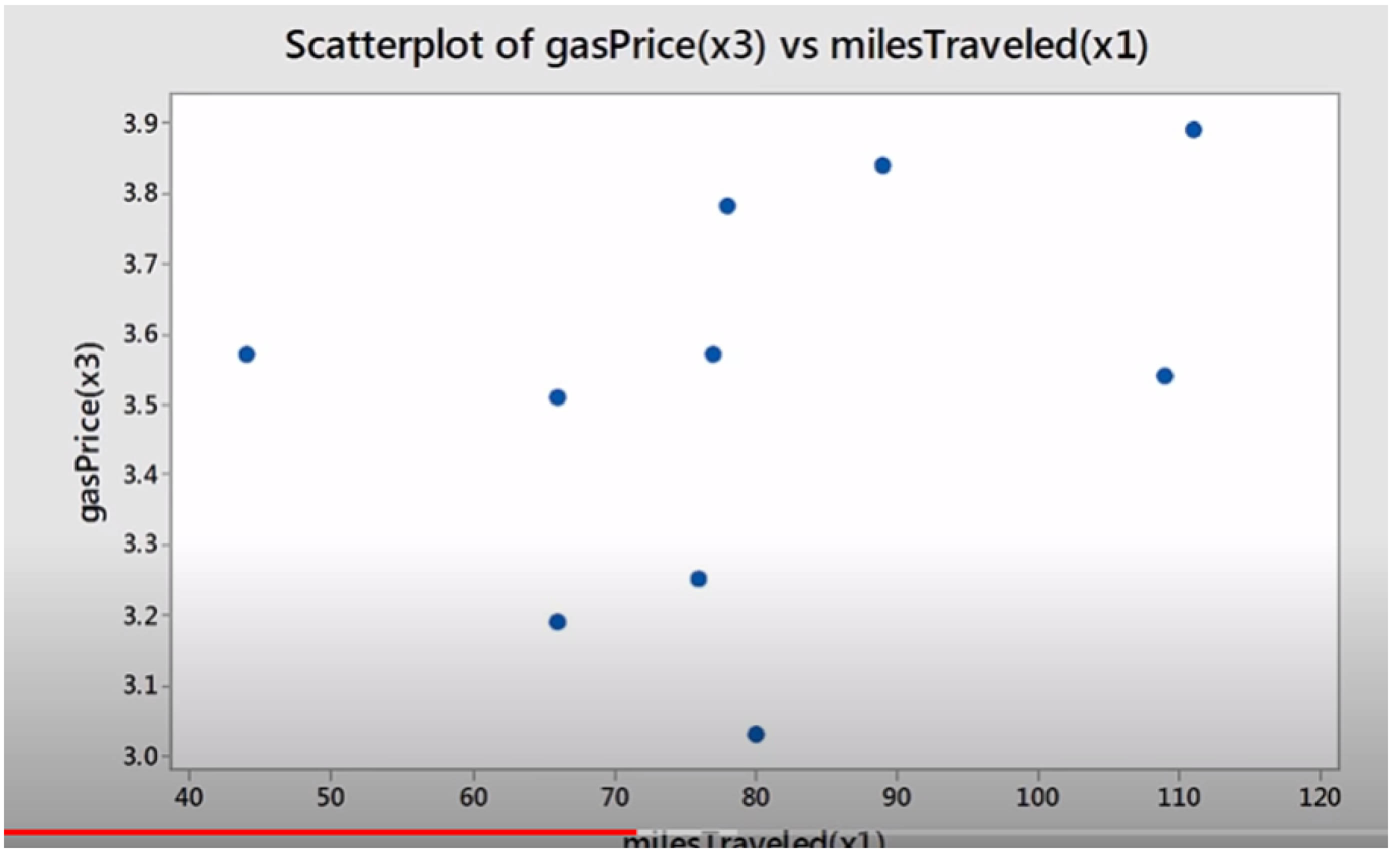
- Draw IV to IV scatter plots. that is multicollinearity check



- we have a potential problem. x2 and x1 are highly correlated

Example 4: Region delivery Service

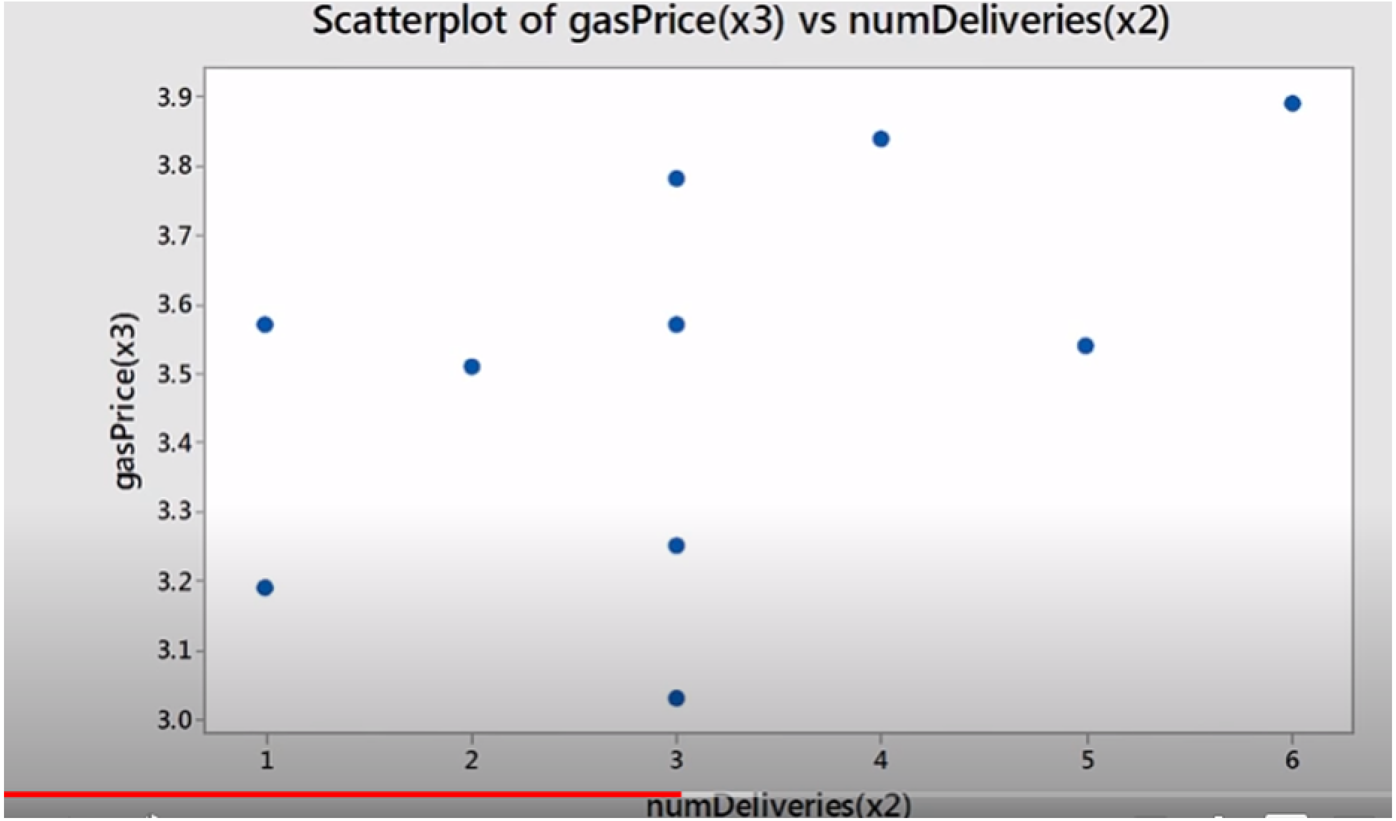
- Draw IV to IV scatter plots. that is multicollinearity check



- Non correlated

Example 4: Region delivery Service

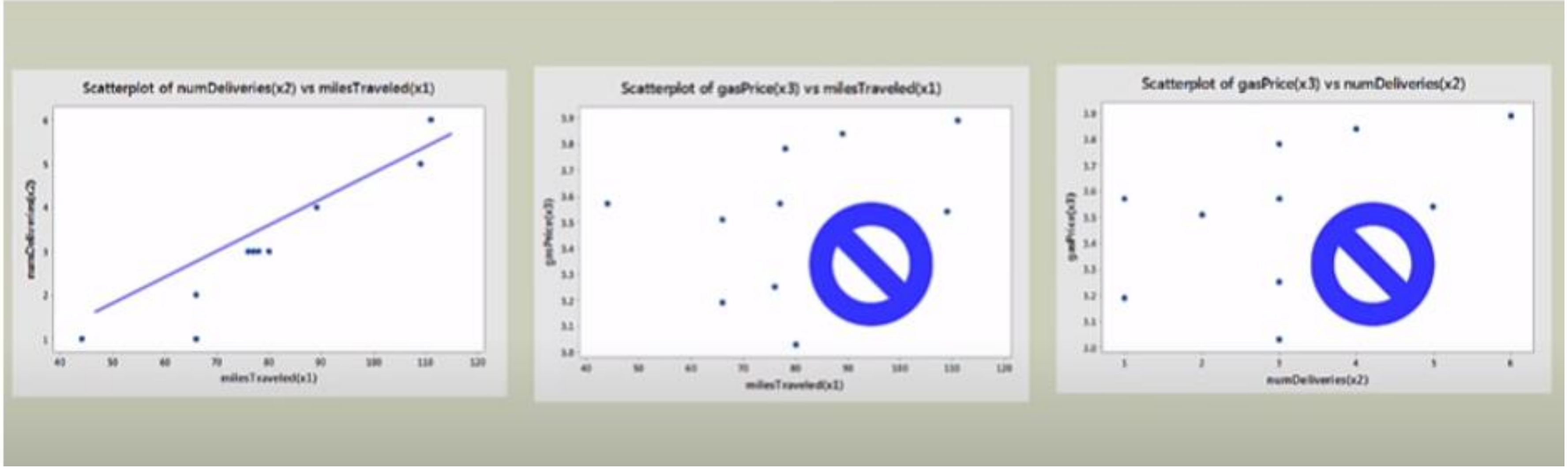
- Draw IV to IV scatter plots. that is multicollinearity check



- Non correlated

Example 4: Region delivery Service

- IV scatter plots.



Example 4: Region delivery Service

- Scatter plot summary for independent variables
  - NumDeliveries (x2) appears highly correlated with milesTraveled (x1), this is multicollinearity
  - milesTraveled (x1) does not appear highly correlated with gasPrice (x3)
  - gasPrice (x3) does not appear highly correlated with NumDeliveries (x2)
- Since NumDeliveries (x2) is highly correlated with miles Travelled, we would not use both in the multiple regression: they are redundant

Example 4: Region delivery Service

- Correlations

Correlation: milesTraveled(x1), numDeliveries(x2), gasPrice(x3), travelTime(y)

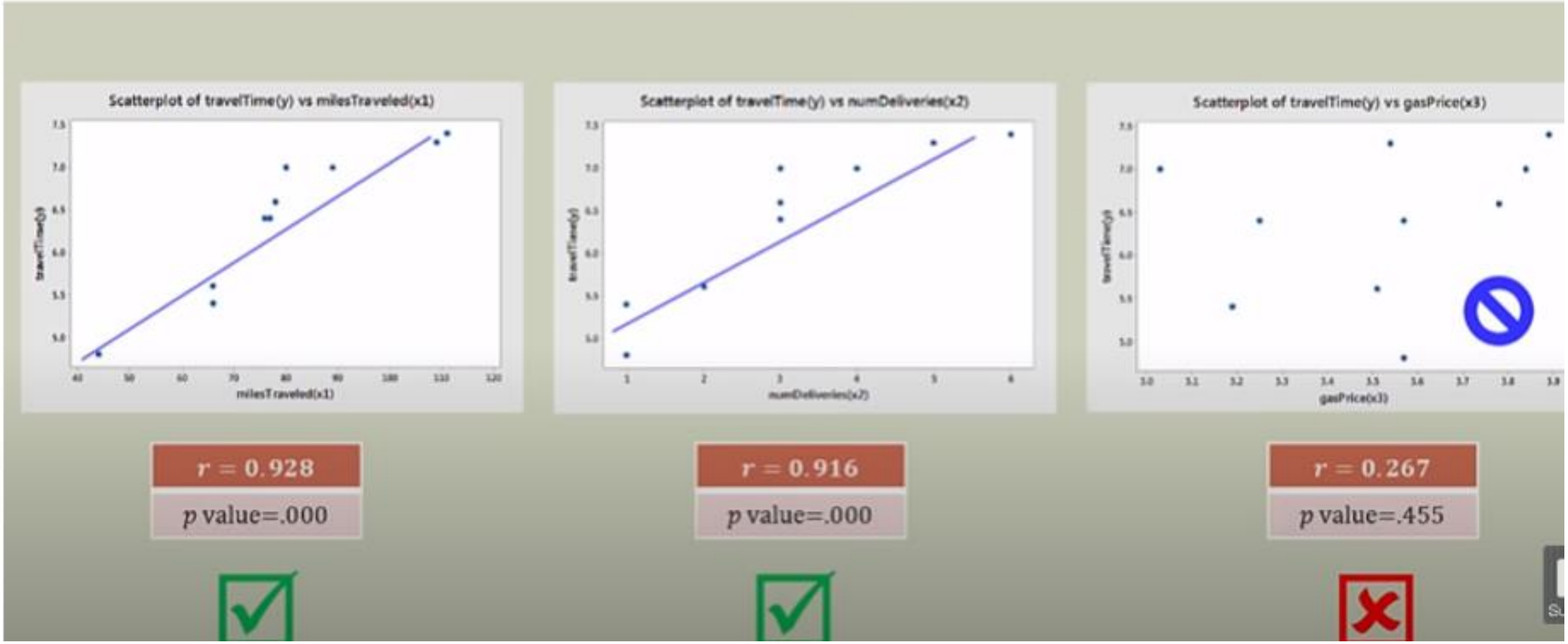
	milesTraveled(x1)	numDeliveries(x2)	gasPrice(x3)
numDeliveries(x2)	0.956		
	0.000		
gasPrice(x3)	0.356	0.498	
	0.313	*0.143*	
travelTime(y)	0.928	0.916	0.267
	0.000	*0.000*	*0.455*

Cell Contents: Pearson correlation
P-Value

- Correlation between y and x1 is o.928. p-value for this correlation is 0.000. it is < 0.001. threshold is 0.05. p value < 0.05 is considered to be considered to be significant. similarly for x2.
- For x3 correlation is very low and p-value is way above 0.05. therefore x3 is not significant.
- It verifies our scatter plots. Similarly observe values for all combinations.
- We can decide which variables to take for analysis.

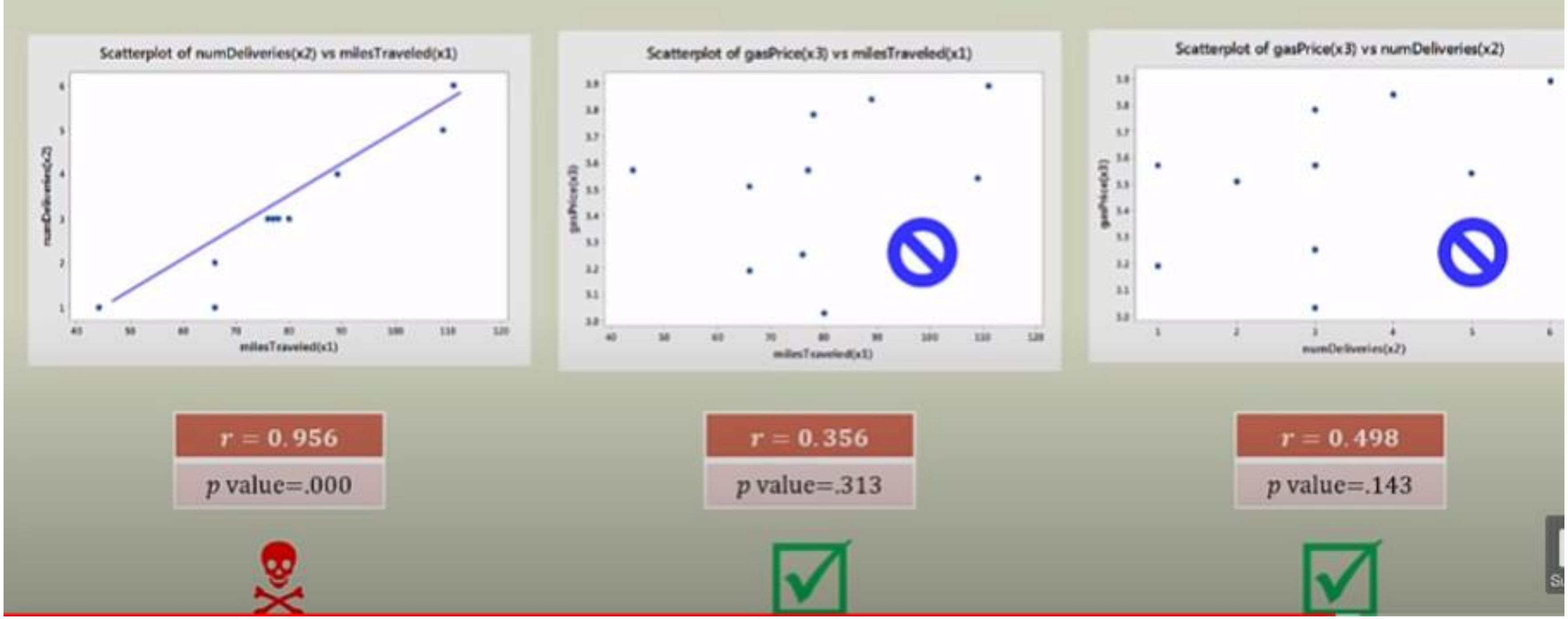
Example 4: Region delivery Service

- DV vs IV scatter plots.



Example 4: Region delivery Service

- IV scatter plots.



Example 4: Region delivery Service

- Correlation summary
- Correlation analysis confirms the conclusions reached by visual examination of the scatter plots

- Redundant multicollinear variables
  - MilesTravelled and NumDeliveries are both highly correlated with each other and therefore redundant, only one should be used in multiple regression analysis.
- Non contributing variables
  - gasPrice is not correlated with dependent variable and should be excluded

Example 4: Region delivery Service

- Evaluating basic models
- Regression analysis

TRAVELTIME(y) ON MILESTRAVLED(x_1)

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.928178501
R Square	0.86151533
Adjusted R Square	0.844204746
Standard Error	0.34230884
Observations	10

ANOVA

	df	SS	MS	F	Significance F
Regression	1	5.831597265	5.831597265	49.76812677	0.000106676
Residual	8	0.937402735	0.117175342		
Total	9	6.769			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	3.185560249	0.466950794	6.822046973	0.000134762	2.108769788	4.26235071	2.108769788	4.26235071
milesTraveled	0.040256781	0.005706416	7.054652845	0.000106676	0.027097763	0.053415799	0.027097763	0.053415799

Example 4: Region delivery Service

- Evaluating basic models
- Regression analysis

- R-square is the % of variation in dependent variable due to independent variable.
  - 86% of the variation is the value in above table. It is pretty high.
  - Adjusted R square is similar to R-square. It is adjusted for the number of independent variables. In this case it is 1. It always lower than R-square.
- Standard error of regression is average distance of the data points from regression line in dependent variable units.
  - Data points are on an average 0.342 hrs away from regression line.
  - This gives a measure of how tightly data points are around regression line.
  - It forms a channel around regression line: narrower the channel is, more tightly data points are around regression. Wider band means more scattered they are from the regression line.
  - SE shows how wide band is.
  - It is in units of dependent units. in this case it is hrs.

Example 4: Region delivery Service

- ANOVA table gives significance of overall model.
- next table:
- under coefficients : coefficients of miles travelled (in hrs). with increase in 1 hr mile travelled increases by 0.042 hrs.
- p-value is 0.001. it is significant. p-value is same as significance F in ANOVA. this is because we have only one indpendent variable.

Example 4: Region delivery Service

- Evaluating basic models
- Regression analysis

TRAVELTIME(y) ON MILESTRAVLED(x_1)

SUMMARY OUTPUT

$$\hat{y} = 3.1856 + 0.0403(\text{milesTraveled})$$
$$\hat{y} = 3.1856 + 0.0403(x_1)$$

Regression Statistics	
Multiple R	0.928178501
R Square	0.86151533
Adjusted R Square	0.844204746
Standard Error	0.34230884
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	5.831597265	5.831597265	49.76812677	0.000106676
Residual	8	0.937402735	0.117175342		
Total	9	6.769			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	3.185560249	0.466950794	6.822046973	0.000134762	2.108769788	4.26235071	2.108769788	4.26235071
milesTraveled	0.040256781	0.005706416	7.054652845	0.000106676	0.027097763	0.053415799	0.027097763	0.053415799

Example 4: Region delivery Service

- Evaluating basic models
- Regression analysis

TRAVELTIME(y) ON MILESTRAVLED(x_1)

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.928178501
R Square	0.86151533
Adjusted R Square	0.844204746
Standard Error	0.34230884
Observations	10

$$\hat{y} = 3.1856 + 0.0403(\text{milesTraveled})$$
$$\hat{y} = 3.1856 + 0.0403(x_1)$$

An increase in 1 mile will increase delivery time by .0403 hours.

84 mile trip estimate

$$\hat{y} = 3.1856 + 0.0403(84)$$
$$\hat{y} = 6.5708 \text{ hours } (6{:}34)$$
$$\hat{y} = 6.5708 \pm 2.31(0.3423)$$
$$\hat{y} = 5.7764 \; to \; 7.3615 \text{ hours}$$
$$\hat{y} = 5{:}47 \text{ to } 7{:}22 \; (\sim 95\% \text{ PI})$$

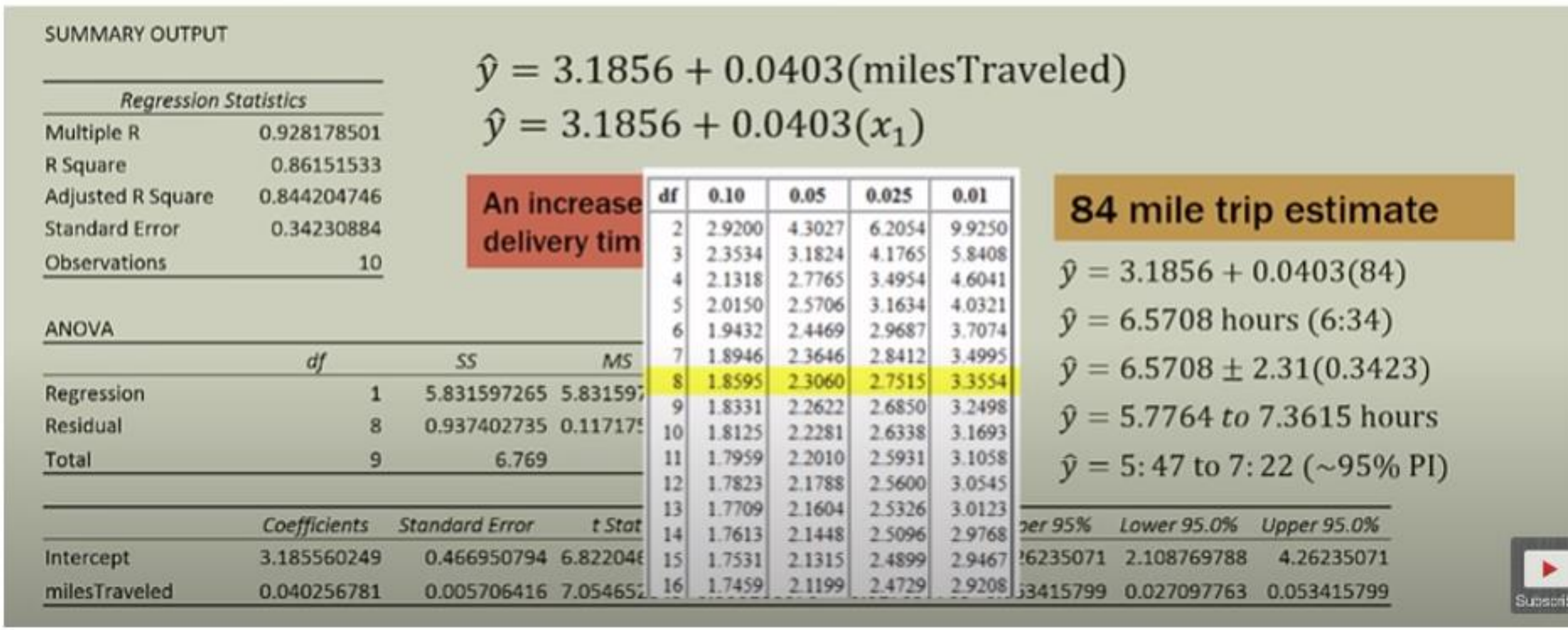
ANOVA

	df	SS	MS	F	Significance F
Regression	1	5.831597265	5.831597265	49.76812677	0.000106676
Residual	8	0.937402735	0.117175342		
Total	9	6.769			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	3.185560249	0.466950794	6.822046973	0.000134762	2.108769788	4.26235071	2.108769788	4.26235071
milesTraveled	0.040256781	0.005706416	7.054652845	0.000106676	0.027097763	0.053415799	0.027097763	0.053415799

Example 4: Region delivery Service

- Evaluating basic models
- Regression analysis



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.928178501
R Square	0.86151533
Adjusted R Square	0.844204746
Standard Error	0.34230884
Observations	10

ANOVA

	df	SS	MS
Regression	1	5.831597265	5.831597
Residual	8	0.937402735	0.117175
Total	9	6.769	

	Coefficients	Standard Error	t Stat	per 95%	Lower 95.0%	Upper 95.0%
Intercept	3.185560249	0.466950794	6.822048	26235071	2.108769788	4.26235071
milesTraveled	0.040256781	0.005706416	7.05465	53415799	0.027097763	0.053415799

$$\hat{y} = 3.1856 + 0.0403(\text{milesTraveled})$$

$$\hat{y} = 3.1856 + 0.0403(x_1)$$

An increase
delivery tim

df	0.10	0.05	0.025	0.01
2	2.9200	4.3027	6.2054	9.9250
3	2.3534	3.1824	4.1765	5.8408
4	2.1318	2.7765	3.4954	4.6041
5	2.0150	2.5706	3.1634	4.0321
6	1.9432	2.4469	2.9687	3.7074
7	1.8946	2.3646	2.8412	3.4995
8	1.8595	2.3060	2.7515	3.3554
9	1.8331	2.2622	2.6850	3.2498
10	1.8125	2.2281	2.6338	3.1693
11	1.7959	2.2010	2.5931	3.1058
12	1.7823	2.1788	2.5600	3.0545
13	1.7709	2.1604	2.5326	3.0123
14	1.7613	2.1448	2.5096	2.9768
15	1.7531	2.1315	2.4899	2.9467
16	1.7459	2.1199	2.4729	2.9208

84 mile trip estimate

$$\hat{y} = 3.1856 + 0.0403(84)$$

$$\hat{y} = 6.5708 \text{ hours } (6{:}34)$$

$$\hat{y} = 6.5708 \pm 2.31(0.3423)$$

$$\hat{y} = 5.7764 \text{ to } 7.3615 \text{ hours}$$

$$\hat{y} = 5{:}47 \text{ to } 7{:}22 \ (\sim 95\% \text{ PI})$$

using t-table for 95% interval, df =8 we get critical t of 2.31.
2.31 is multiplied by 0.3423 in 84 mile trip estimate. 0.3423 is Standard Error.

Example 4: Region delivery Service

Regression Analysis: travelTime(y) versus numDeliveries(x2)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	5.6851	5.6851	41.96	0.000
numDeliveries(x2)	1	5.6851	5.6851	41.96	0.000
Error	8	1.0839	0.1355		
Lack-of-Fit	4	0.6639	0.1660	1.58	0.334
Pure Error	4	0.4200	0.1050		
Total	9	6.7690			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.368091	83.99%	81.99%	70.27%

Mini Tab software is used to generate the above.

Regression Analysis: travelTime(y) versus numDeliveries(x2)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	5.6851	5.6851	41.96	0.000
numDeliveries(x2)	1	5.6851	5.6851	41.96	0.000
Error	8	1.0839	0.1355		
Lack-of-Fit	4	0.6639	0.1660	1.58	0.334
Pure Error	4	0.4200	0.1050		
Total	9	6.7690			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.368091	83.99%	81.99%	70.27%

S is standard error of regression. R-sq(pred) is used for how well our model is predicting if additional data points are added.

Example 4: Region delivery Service

Regression Analysis: travelTime(y) versus numDeliveries(x2)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	5.6851	5.6851	41.96	0.000
numDeliveries(x2)	1	5.6851	5.6851	41.96	0.000
Error	8	1.0839	0.1355		
Lack-of-Fit	4	0.6639	0.1660	1.58	0.334
Pure Error	4	0.4200	0.1050		
Total	9	6.7690			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.368091	83.99%	81.99%	70.27%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	4.845	0.265	18.26	0.000	
numDeliveries(x2)	0.4983	0.0769	6.48	0.000	1.00

Regression Equation

travelTime(y) = 4.845 + 0.4983 numDeliveries(x2)

Example 4: Region delivery Service

Regression Analysis: travelTime(y) versus numDeliveries(x2)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	5.6851	5.6851	41.96	0.000
numDeliveries(x2)	1	5.6851	5.6851	41.96	0.000
Error	8	1.0839	0.1355		
Lack-of-Fit	4	0.6639	0.1660	1.58	0.334
Pure Error	4	0.4200	0.1050		
Total	9	6.7690			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.368091	83.99%	81.99%	70.27%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	4.845	0.265	18.26	0.000	
numDeliveries(x2)	0.4983	0.0769	6.48	0.000	1.00

Regression Equation

travelTime(y) = 4.845 + 0.4983 numDeliveries(x2)

Example 4: Region delivery Service

Regression Analysis: travelTime(y) versus numDeliveries(x2)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	5.6851	5.6851	41.96	0.000
numDeliveries(x2)	1	5.6851	5.6851	41.96	0.000
Error	8	1.0839	0.1355		
Lack-of-Fit	4	0.6639	0.1660	1.58	0.334
Pure Error	4	0.4200	0.1050		
Total	9	6.7690			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.368091	83.99%	81.99%	70.27%

An increase in 1 delivery will increase delivery time by .4983 hours.

4 delivery estimate

$$\hat{y} = 4.845 + 0.4983(4)$$
$$\hat{y} = 6.838 \text{ hours } (6{:}50)$$

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	4.845	0.265	18.26	0.000	
numDeliveries(x2)	0.4983	0.0769	6.48	0.000	1.00

Regression Equation

travelTime(y) = 4.845 + 0.4983 numDeliveries(x2)

Example 4: Region delivery Service

Regression Analysis: travelTime(y) versus gasPrice(x3)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	0.4833	0.4833	0.62	0.455
gasPrice(x3)	1	0.4833	0.4833	0.62	0.455
Error	8	6.2857	0.7857		
Lack-of-Fit	7	5.0057	0.7151	0.56	0.777
Pure Error	1	1.2800	1.2800		
Total	9	6.7690			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.886403	7.14%	0.00%	0.00%

- huge standard error of regression. R-sq value is too low.

Example 4: Region delivery Service

Regression Analysis: travelTime(y) versus gasPrice(x3)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	0.4833	0.4833	0.62	0.455
gasPrice(x3)	1	0.4833	0.4833	0.62	0.455
Error	8	6.2857	0.7857		
Lack-of-Fit	7	5.0057	0.7151	0.56	0.777
Pure Error	1	1.2800	1.2800		
Total	9	6.7690			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.886403	7.14%	0.00%	0.00%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	3.54	3.65	0.97	0.361	
gasPrice(x3)	0.81	1.03	0.78	0.455	1.00

Regression Equation

travelTime(y) = 3.54 + 0.81 gasPrice(x3)

Example 4: Region delivery Service

Regression Analysis: travelTime(y) versus gasPrice(x3)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	0.4833	0.4833	0.62	0.455
gasPrice(x3)	1	0.4833	0.4833	0.62	0.455
Error	8	6.2857	0.7857		
Lack-of-Fit	7	5.0057	0.7151	0.56	0.777
Pure Error	1	1.2800	1.2800		
Total	9	6.7690			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.886403	7.14%	0.00%	0.00%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	3.54	3.65	0.97	0.361	
gasPrice(x3)	0.81	1.03	0.78	0.455	1.00

Regression Equation

travelTime(y) = 3.54 + 0.81 gasPrice(x3)

- Therefore we will not consider gas price.

Example 4: Region delivery Service

- Model options summary

F	p−value	S	$R^2(adj)$	$R^2(pred)$	x_1	x_2	x_3
49.77	< 0.001	0.34230	84.42%	79.07%	X		
41.96	< 0.001	0.36809	81.99%	70.27%		X	
0.62	0.455	0.88640	0.00%	0.00%			X

- For first model, on an average data points are 0.3423 hrs away from regression line which is least among all models (i.e. second and third)
- which model has most data points clustered around regression line?

  first model

- If we are only using one independent variable then we will choose x1. we have highest F-statistics and lowest standard error, highest R-square adjusted and highest r-square predicted.

Example 4: Region delivery Service

- Model options summary

F	p–value	S	$R^2(adj)$	$R^2(pred)$	x_1	x_2	x_3
49.77	< 0.001	0.34230	84.42%	79.07%	X		
41.96	< 0.001	0.36809	81.99%	70.27%		X	
~~0.62~~	~~0.455~~	~~0.88640~~	~~0.00%~~	~~0.00%~~			~~X~~

F	p–value	S	$R^2(adj)$	$R^2(pred)$	x_1	x_2	x_3
49.77	< 0.001	0.34230	84.42%	79.07%	X		
41.96	< 0.001	0.36809	81.99%	70.27%		X	
0.62	0.455	0.88640	0.00%	0.00%			X

Example 4: Region delivery Service

- Two variable regression

Regression Analysis: travelTime(y) versus milesTraveled(x1), numDeliveries(x2)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	5.89850	2.94925	23.72	0.001
milesTraveled(x1)	1	0.21343	0.21343	1.72	0.232
numDeliveries(x2)	1	0.06691	0.06691	0.54	0.487
Error	7	0.87050	0.12436		
Total	9	6.76900			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.352642	87.14%	83.47%	59.95%

Example 4: Region delivery Service

- Two variable regression

Regression Analysis: travelTime(y) versus milesTraveled(x1), numDeliveries(x2)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	5.89850	2.94925	23.72	0.001
milesTraveled(x1)	1	0.21343	0.21343	1.72	0.232
numDeliveries(x2)	1	0.06691	0.06691	0.54	0.487
Error	7	0.87050	0.12436		
Total	9	6.76900			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.352642	87.14%	83.47%	59.95%

Ex 4:Region delivery Service (two variable regression)

Regression Analysis: travelTime(y) versus milesTraveled(x1), numDeliveries(x2)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	5.89850	2.94925	23.72	0.001
milesTraveled(x1)	1	0.21343	0.21343	1.72	0.232
numDeliveries(x2)	1	0.06691	0.06691	0.54	0.487
Error	7	0.87050	0.12436		
Total	9	6.76900			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.352642	87.14%	83.47%	59.95%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	3.732	0.887	4.21	0.004	
milesTraveled(x1)	0.0262	0.0200	1.31	0.232	11.59
numDeliveries(x2)	0.184	0.251	0.73	0.487	11.59

Regression Equation

Regression equation:

Travel time (y) = 3.732 + 0.0262 milesTravelled(x1) + 0.184 numDeliveries(x2)

Ex 4:Region delivery Service (two variable regression)

Regression Analysis: travelTime(y) versus milesTraveled(x1), numDeliveries(x2)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	5.89850	2.94925	23.72	0.001
milesTraveled(x1)	1	0.21343	0.21343	1.72	0.232
numDeliveries(x2)	1	0.06691	0.06691	0.54	0.487
Error	7	0.87050	0.12436		
Total	9	6.76900			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.352642	87.14%	83.47%	59.95%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	3.732	0.887	4.21	0.004	
milesTraveled(x1)	0.0262	0.0200	1.31	0.232	11.59
numDeliveries(x2)	0.184	0.251	0.73	0.487	11.59

Regression Equation

- P-value for x1 is 0.232 and for x2 it is 0.487. In both cases p value is greater than 0.001
- Therefore both x1 and x2 are insignificant
- F-value is 23.72 and p value is 0.001. That shows that overall model is significant
- However, neither of coefficients (x1, and x2) are significant
- such strange relationship is because of strong correlation between independent variables.

Ex 4:Region delivery Service (two variable regression)

Regression Analysis: travelTime(y) versus milesTraveled(x1), numDeliveries(x2)

Analysis of Variance

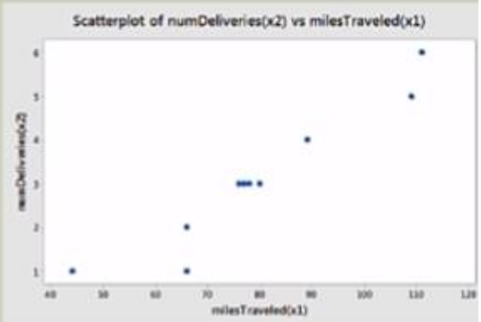
Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	5.89850	2.94925	23.72	0.001
milesTraveled(x1)	1	0.21343	0.21343	1.72	0.232
numDeliveries(x2)	1	0.06691	0.06691	0.54	0.487
Error	7	0.87050	0.12436		
Total	9	6.76900			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.352642	87.14%	83.47%	59.95%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	3.732	0.887	4.21	0.004	
milesTraveled(x1)	0.0262	0.0200	1.31	0.232	11.59
numDeliveries(x2)	0.184	0.251	0.73	0.487	11.59



$r = 0.956$

p value=.000

- scatter plot shows that x1 and x2 are strongly correlated. they have almost linear relationship. r=0.956 shows that they are strongly correlated.

Ex 4:Region delivery Service (two variable regression)

Regression Analysis: travelTime(y) versus milesTraveled(x1), gasPrice(x3)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	5.86239	2.93119	22.63	0.001
milesTraveled(x1)	1	5.37907	5.37907	41.53	0.000
gasPrice(x3)	1	0.03079	0.03079	0.24	0.641
Error	7	0.90661	0.12952		
Total	9	6.76900			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.359883	86.61%	82.78%	68.11%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	3.87	1.48	2.61	0.035	
milesTraveled(x1)	0.04137	0.00642	6.44	0.000	1.14
gasPrice(x3)	-0.219	0.449	-0.49	0.641	1.14

Regression Equation

travelTime(y) = 3.87 + 0.04137 milesTraveled(x1) - 0.219 gasPrice(x3)

- x3 has negative intercept.
- If we hold x1 constant and increase the price of gas then travel time would decrease by 0.219 hrs.
- That is, if gas price goes up and travel time goes down
- Model can not be accepted

Ex 4:Region delivery Service (two variable regression)

Regression Analysis: travelTime(y) versus milesTraveled(x1), gasPrice(x3)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	5.86239	2.93119	22.63	0.001
milesTraveled(x1)	1	5.37907	5.37907	41.53	0.000
gasPrice(x3)	1	0.03079	0.03079	0.24	0.641
Error	7	0.90661	0.12952		
Total	9	6.76900			

If gasPrice is held constant, then travelTime is expected to increase by 0.04137 hours for each additional mile traveled.

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.359883	86.61%	82.78%	68.11%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	3.87	1.48	2.61	0.035	
milesTraveled(x1)	0.04137	0.00642	6.44	0.000	1.14
gasPrice(x3)	-0.219	0.449	-0.49	0.641	1.14

Regression Equation

travelTime(y) = 3.87 + 0.04137 milesTraveled(x1) - 0.219 gasPrice(x3)

Ex 4:Region delivery Service (two variable regression)

Regression Analysis: travelTime(y) versus milesTraveled(x1), gasPrice(x3)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	5.86239	2.93119	22.63	0.001
milesTraveled(x1)	1	5.37907	5.37907	41.53	0.000
gasPrice(x3)	1	0.03079	0.03079	0.24	0.641
Error	7	0.90661	0.12952		
Total	9	6.76900			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.359883	86.61%	82.78%	68.11%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	3.87	1.48	2.61	0.035	
milesTraveled(x1)	0.04137	0.00642	6.44	0.000	1.14
gasPrice(x3)	-0.219	0.449	-0.49	0.641	1.14

Regression Equation

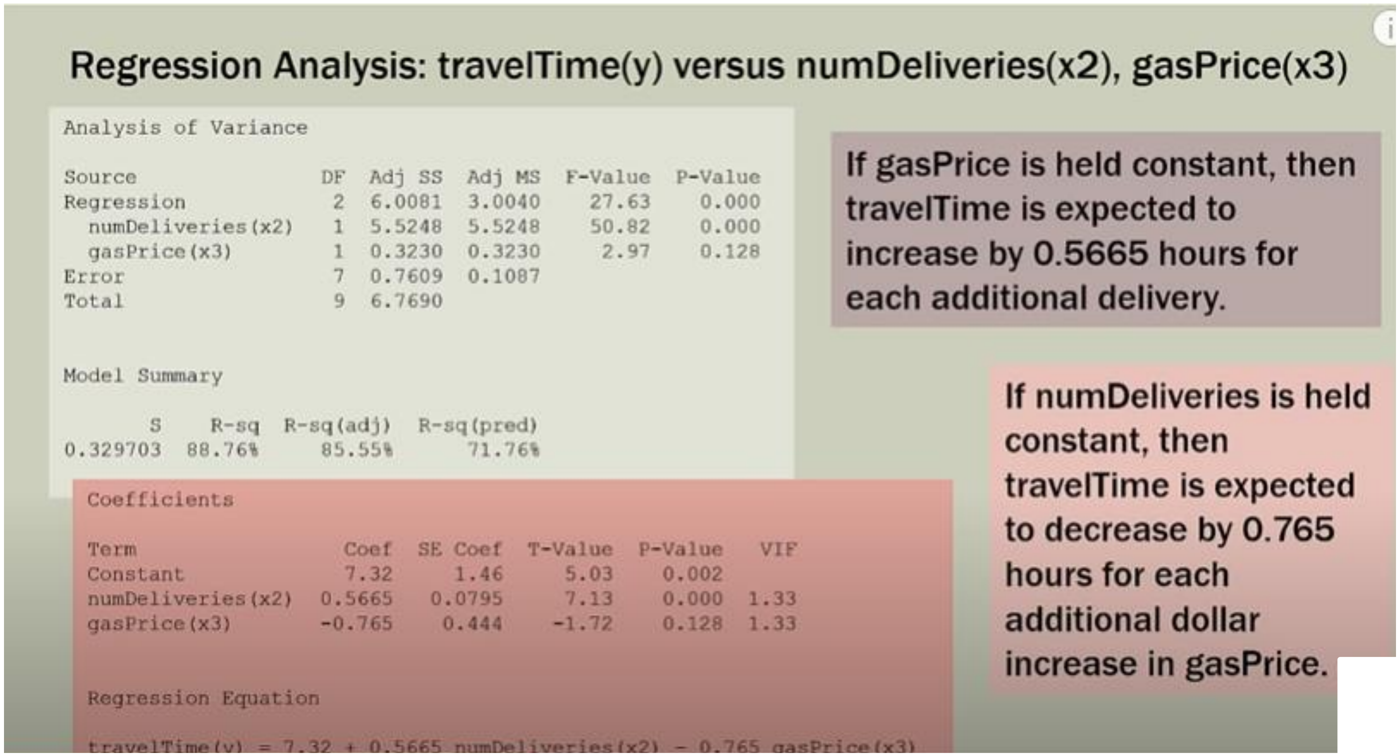
travelTime(y) = 3.87 + 0.04137 milesTraveled(x1) - 0.219 gasPrice(x3)

If gasPrice is held constant, then travelTime is expected to increase by 0.04137 hours for each additional mile traveled.

If milesTraveled is held constant, then travelTime is expected to decrease by 0.219 hours for each additional dollar increase in gasPrice.

- If gas price is held constant then conclusion makes sense.
- It does not make sense if miletravelled is held constant.
- this is because correlation shows that gasPrice(x3) has no relation with y.

Ex 4:Region delivery Service (two variable regression)



Regression Analysis: travelTime(y) versus numDeliveries(x2), gasPrice(x3)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	6.0081	3.0040	27.63	0.000
numDeliveries(x2)	1	5.5248	5.5248	50.82	0.000
gasPrice(x3)	1	0.3230	0.3230	2.97	0.128
Error	7	0.7609	0.1087		
Total	9	6.7690			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.329703	88.76%	85.55%	71.76%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	7.32	1.46	5.03	0.002	
numDeliveries(x2)	0.5665	0.0795	7.13	0.000	1.33
gasPrice(x3)	-0.765	0.444	-1.72	0.128	1.33

Regression Equation

travelTime(y) = 7.32 + 0.5665 numDeliveries(x2) - 0.765 gasPrice(x3)

If gasPrice is held constant, then travelTime is expected to increase by 0.5665 hours for each additional delivery.

If numDeliveries is held constant, then travelTime is expected to decrease by 0.765 hours for each additional dollar increase in gasPrice.

- If gasprice held constant makes sense then relation between x2 and y makes sense
- If numdeliveries is held constant then conclusion does not make sense

Ex 4:Region delivery Service (two variable regression)

MODEL OPTIONS SUMMARY

F	p–value	S	$R^2(adj)$	$R^2(pred)$	x_1	x_2	x_3	VIF
49.77	< 0.001	0.34230	84.42%	79.07%	X			1.00
41.96	< 0.001	0.36809	81.99%	70.27%		X		1.00
0.62	0.455	0.88640	0.00%	0.00%			X	1.00
23.72	0.001	0.35264	83.47%	59.95%	X	X		11.59
22.63	0.001	0.35988	82.78%	68.11%	X		X	1.14
27.63	< 0.001	0.32970	85.55%	71.76%		X	X	1.33

Example 4: Region delivery Service

- Full regression model

Example 4: Region delivery Service

- Full regression model: statistical values

Regression Analysis: travelTime(y) versus milesTraveled(x1), numDeliveries(x2), gasPrice(x3)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	6.05612	2.01871	16.99	0.002
milesTraveled(x1)	1	0.04805	0.04805	0.40	0.548
numDeliveries(x2)	1	0.19373	0.19373	1.63	0.249
gasPrice(x3)	1	0.15761	0.15761	1.33	0.293
Error	6	0.71288	0.11881		
Total	9	6.76900			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.344694	89.47%	84.20%	57.49%

Example 4: Region delivery Service

- Full regression model (coefficients)

Regression Analysis: travelTime(y) versus milesTraveled(x1), numDeliveries(x2), gasPrice(x3)

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	6.21	2.32	2.68	0.037	
milesTraveled(x1)	0.0141	0.0222	0.64	0.548	14.94
numDeliveries(x2)	0.383	0.300	1.28	0.249	17.35
gasPrice(x3)	-0.607	0.527	-1.15	0.293	1.71

Regression Equation

travelTime(y) = 6.21 + 0.0141 milesTraveled(x1) + 0.383 numDeliveries(x2)
- 0.607 gasPrice(x3)

Example 4: Region delivery Service

- Full regression model

MODEL OPTIONS SUMMARY

F	p–value	S	$R^2(adj)$	$R^2(pred)$	x_1	x_2	x_3	VIF
49.77	< 0.001	0.34230	84.42%	79.07%	X			1.00
41.96	< 0.001	0.36809	81.99%	70.27%		X		1.00
0.62	0.455	0.88640	0.00%	0.00%			X	1.00
23.72	0.001	0.35264	83.47%	59.95%	X	X		11.59
22.63	0.001	0.35988	82.78%	68.11%	X		X	1.14
27.63	< 0.001	0.32970	85.55%	71.76%		X	X	1.33
16.99	0.002	0.34469	84.20%	57.49%	X	X	X	below
					14.94	17.35	1.71	

Example 4: Region delivery Service (full regression model)

MODEL OPTIONS SUMMARY

F	p–value	S	$R^2(adj)$	$R^2(pred)$	x_1	x_2	x_3	VIF
49.77	< 0.001	0.34230	84.42%	79.07%	X			1.00
41.96	< 0.001	0.36809	81.99%	70.27%		X		1.00
0.62	0.455	0.88640	0.00%	0.00%			X	1.00
23.72	0.001	0.35264	83.47%	59.95%	X	X		11.59
22.63	0.001	0.35988	82.78%	68.11%	X		X	1.14
27.63	< 0.001	0.32970	85.55%	71.76%		X	X	1.33
16.99	0.002	0.34469	84.20%	57.49%	X	X	X	below
					14.94	17.35	1.71	

- Single IV shows x1

Multiple Linear Regression Model

- A linear regression model that contains more than one predictor variable is called a *multiple linear regression model*.
- Multiple linear regression model with two predictor variables,

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

- The model is linear because it varies linearly with the change in the parameters β_0, β_1 , β_2
  - The model describes a plane in the three-dimensional space of Y, x_1 and x_2
- The parameter β_0 is the intercept of this plane, parameters β_1 and β_2 are referred to as partial regression coefficients.

Multiple Linear Regression Model

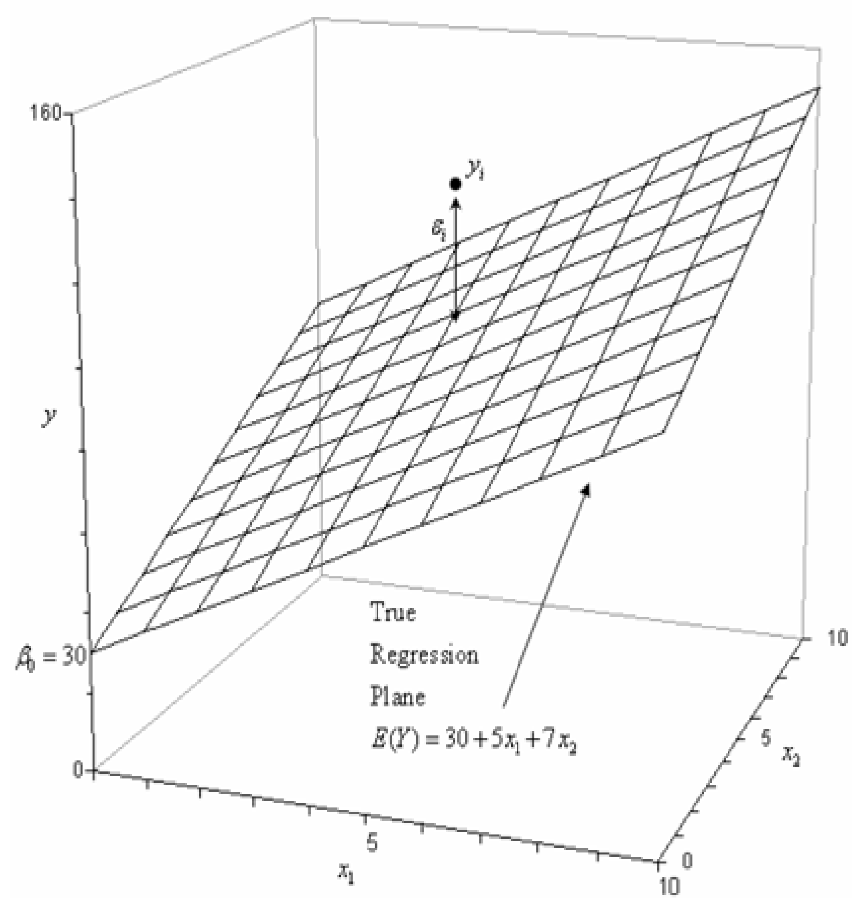
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

- Parameter β_1 represents the change in the mean response corresponding to a unit change x_1 when x_2 is held constant.
- Parameter β_2 represents the change in the mean response corresponding to a unit change x_2 when x_1 is held constant.

Multiple Linear Regression Model

- Example of a multiple linear regression model with two predictor variables,

- This regression model is a first order multiple linear regression model.
- This is because the maximum power of the variables in the model is 1.
- Notice observed data point and the corresponding random error,



Multiple Linear Regression Model

- The true regression model is usually never known
- However, the regression model can be estimated by calculating the parameters of the model for an observed data set.
- This is explained in Estimating Regression Models Using Least Squares.

Multiple Linear Regression Model

- A linear regression model may also take the following form:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon$$

- All multiple linear regression models can be expressed in the following general form:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon$$

$$Y = 500 + 5x_1 + 7x_2 - 3x_3 - 5x_4 + 3x_5 + \epsilon$$

Estimating Regression Models Using Least Squares

- Consider a multiple linear regression model with k predictor variables: (level represent sample number)

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon$$

Let each of the k predictor variables, x_1, x_2 ... x_k, have n levels. Then x_{ij} represents the i th level of the j th predictor variable x_j. For example, x_{51} represents the fifth level of the first predictor variable x_1, while x_{19} represents the first level of the ninth predictor variable, x_9. Observations, y_1, y_2 ... y_n, recorded for each of these n levels can be expressed in the following way:

$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + ... + \beta_k x_{1k} + \epsilon_1$$

$$y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + ... + \beta_k x_{2k} + \epsilon_2$$

..

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \epsilon_i$$

..

$$y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + ... + \beta_k x_{nk} + \epsilon_n$$

Estimating Regression Models Using Least Squares

$$
\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + ... + \beta_k x_{1k} + \epsilon_1 \\
y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + ... + \beta_k x_{2k} + \epsilon_2 \\
&.. \\
y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \epsilon_i \\
&.. \\
y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + ... + \beta_k x_{nk} + \epsilon_n
\end{aligned}
$$

or

$$y = X\beta + \epsilon$$

$$
y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix}
\quad
X = \begin{bmatrix} 1 & x_{11} & x_{12} & . & . & . & x_{1n} \\ 1 & x_{21} & x_{22} & . & . & . & x_{2n} \\ . & . & . & & & & . \\ . & . & . & & & & . \\ . & . & . & & & & . \\ 1 & x_{n1} & x_{n2} & . & . & . & x_{nn} \end{bmatrix}
\quad
\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_n \end{bmatrix}
\text{ and } \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ . \\ \epsilon_n \end{bmatrix}
$$

Estimating Regression Models Using Least Squares

- The matrix X is referred to as the *design matrix*.
- It contains information about the levels of the predictor variables at which the observations are obtained.
- The vector β contains all the regression coefficients.
- To obtain the regression model, β should be known.
- β is estimated using least square estimates.

$$\hat{\beta} = (X'X)^{-1}X'y$$

- Knowing the estimates, the multiple linear regression model can now be estimated as:

$$\hat{y} = X\hat{\beta}$$

Estimating Regression Models Using Least Squares

- Estimated regression model is also referred to as the *fitted model*.
- The observations, y_i may be different from the fitted values $y_i^\wedge$ obtained from this model
- The difference between these two values is the residual, e_i
- The vector of residuals, e is obtained as:

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

Estimating Regression Models Using Least Squares

- The fitted model can also be written as follows,

$$\begin{aligned} \hat{y} &= X\hat{\beta} \\ &= X(X'X)^{-1}X'y \\ &= Hy \end{aligned}$$

where $H = X(X'X)^{-1}X'$

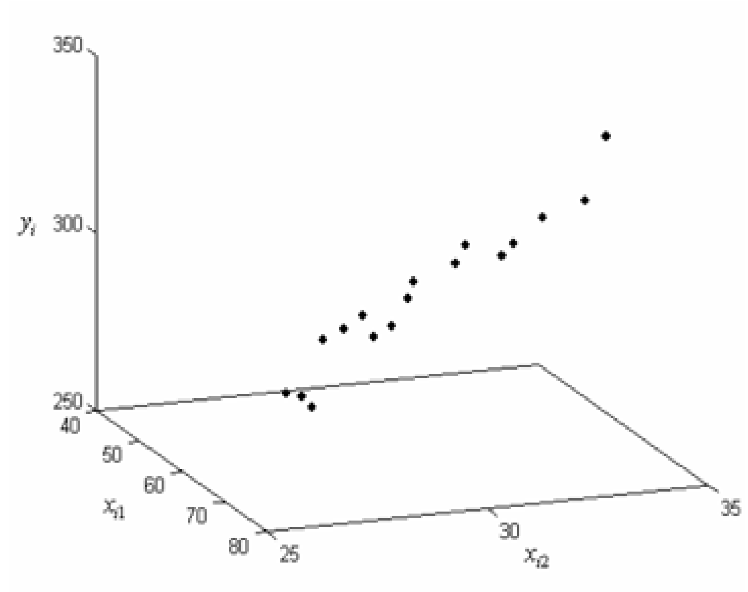
- The matrix H is referred to as the hat matrix.
- It transforms the vector of the observed response values, y to the vector of fitted values, y^

Example Multiple Regression Model Fitting

- An analyst studying a chemical process expects the yield (y) to be affected by the levels of two factors, x_1 and x_2.
- The analyst wants to fit a first order regression model to the data.
- Factors, x_1 and x_2 are not dependent on each other

Example

Observation Number	Factor 1 (x_{i1})	Factor 2 (x_{i2})	Yield (y_i)
1	41.9	29.1	251.3
2	43.4	29.3	251.3
3	43.9	29.5	248.3
4	44.5	29.7	267.5
5	47.3	29.9	273.0
6	47.5	30.3	276.5
7	47.9	30.5	270.3
8	50.2	30.7	274.9
9	52.8	30.8	285.0
10	53.2	30.9	290.0
11	56.7	31.5	297.0
12	57.0	31.7	302.5
13	63.5	31.9	304.5
14	65.3	32.0	309.3
15	71.1	32.1	321.7
16	77.0	32.5	330.7
17	77.8	32.9	349.0



- The first order regression model applicable this data set having two predictor (independent) variables is:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

Example

Observation Number	Factor 1 (x_{i1})	Factor 2 (x_{i2})	Yield (y_i)
1	41.9	29.1	251.3
2	43.4	29.3	251.3
3	43.9	29.5	248.3
4	44.5	29.7	267.5
5	47.3	29.9	273.0
6	47.5	30.3	276.5
7	47.9	30.5	270.3
8	50.2	30.7	274.9
9	52.8	30.8	285.0
10	53.2	30.9	290.0
11	56.7	31.5	297.0
12	57.0	31.7	302.5
13	63.5	31.9	304.5
14	65.3	32.0	309.3
15	71.1	32.1	321.7
16	77.0	32.5	330.7
17	77.8	32.9	349.0

$$X = \begin{bmatrix} 1 & 41.9 & 29.1 \\ 1 & 43.4 & 29.3 \\ . & . & . \\ . & . & . \\ . & . & . \\ 1 & 77.8 & 32.9 \end{bmatrix} \quad y = \begin{bmatrix} 251.3 \\ 251.3 \\ . \\ . \\ . \\ 349.0 \end{bmatrix}$$

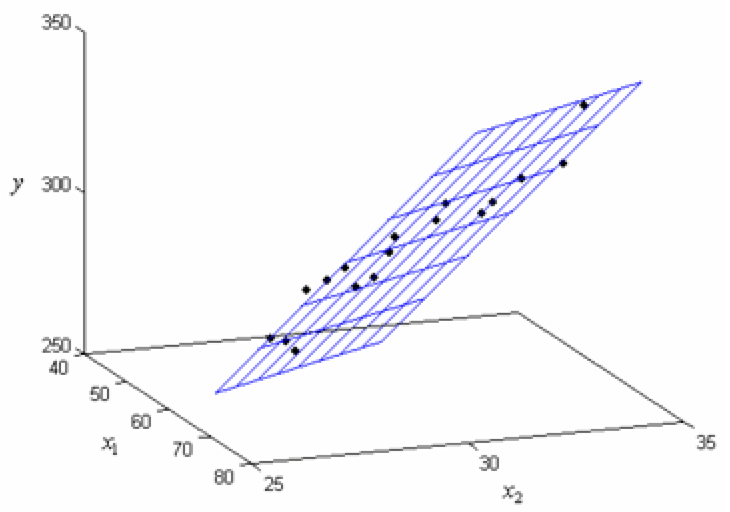
Example

$$X = \begin{bmatrix} 1 & 41.9 & 29.1 \\ 1 & 43.4 & 29.3 \\ . & . & . \\ . & . & . \\ . & . & . \\ 1 & 77.8 & 32.9 \end{bmatrix} \quad y = \begin{bmatrix} 251.3 \\ 251.3 \\ . \\ . \\ . \\ 349.0 \end{bmatrix}$$

- The least square estimates $\hat{\beta}$ can now be obtained:

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'y \\ &= \begin{bmatrix} 17 & 941 & 525.3 \\ 941 & 54270 & 29286 \\ 525.3 & 29286 & 16254 \end{bmatrix}^{-1} \begin{bmatrix} 4902.8 \\ 276610 \\ 152020 \end{bmatrix} \\ &= \begin{bmatrix} -153.51 \\ 1.24 \\ 12.08 \end{bmatrix} \end{aligned}$$

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} -153.51 \\ 1.24 \\ 12.08 \end{bmatrix}$$



$$\begin{aligned} \hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \\ &= -153.5 + 1.24x_1 + 12.08x_2 \end{aligned}$$

Hypothesis Tests in Multiple Linear Regression

- As in the case of simple linear regression, these tests can only be carried out if it can be assumed that the random error terms, ϵ_i are normally and independently distributed with a mean of zero and variance of σ^2
- Three types of hypothesis tests can be carried out for multiple linear regression models:
- **Test for significance of regression**: checks the significance of the whole regression model.
- **t test**: checks the significance of individual regression coefficients.
- **F test**: can be used to simultaneously check the significance of a number of regression coefficients

  It can also be used to test individual coefficients.

Test for Significance of Regression

- For multiple linear regression analysis, ANOVA is used
- The test is used to check if a linear statistical relationship exists between the response variable and at least one of the predictor variables.
- The statements for the hypotheses are:

$$H_0: \quad \beta_1 = \beta_2 = ... = \beta_k = 0$$
$$H_1: \quad \beta_j \neq 0 \text{ for at least one } j$$

- The test for H_0 is carried out using the following statistic:

$$F_0 = \frac{MS_R}{MS_E}$$

Where MS_R is the regression mean square and MS_E is the error mean square

Test for Significance of Regression

- If the null hypothesis, H_0 is true then the statistic F_0 follows the F distribution with k degrees of freedom in the numerator and n-(k+1) degrees of freedom in the denominator.
- The null hypothesis H_0 is rejected if the calculated statistic, F_0 is such that

$$F_0 > f_{\alpha,k,n-(k+1)}$$

Test for Significance of Regression for example 5

Observation Number	Factor 1 (x_{i1})	Factor 2 (x_{i2})	Yield (y_i)
1	41.9	29.1	251.3
2	43.4	29.3	251.3
3	43.9	29.5	248.3
4	44.5	29.7	267.5
5	47.3	29.9	273.0
6	47.5	30.3	276.5
7	47.9	30.5	270.3
8	50.2	30.7	274.9
9	52.8	30.8	285.0
10	53.2	30.9	290.0
11	56.7	31.5	297.0
12	57.0	31.7	302.5
13	63.5	31.9	304.5
14	65.3	32.0	309.3
15	71.1	32.1	321.7
16	77.0	32.5	330.7
17	77.8	32.9	349.0

$$X = \begin{bmatrix} 1 & 41.9 & 29.1 \\ 1 & 43.4 & 29.3 \\ . & . & . \\ . & . & . \\ . & . & . \\ 1 & 77.8 & 32.9 \end{bmatrix} \quad y = \begin{bmatrix} 251.3 \\ 251.3 \\ . \\ . \\ . \\ 349.0 \end{bmatrix}$$

- The null hypothesis for the model is:

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

Test for Significance of Regression for example

- The statistic to test H_0

$$F_0 = \frac{MS_R}{MS_E}$$

- First calculate MS_R

$$SS_R = y'\left[H - (\frac{1}{n})J\right]y$$

$$\begin{aligned} H &= X(X'X)^{-1}X' \\ &= \begin{bmatrix} 0.27552 & 0.25154 & . & . & -0.04030 \\ 0.25154 & 0.23021 & . & . & -0.029120 \\ . & . & . & . & . \\ . & . & . & . & . \\ -0.04030 & -0.02920 & . & . & 0.30115 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} SS_R &= y'\left[H - (\frac{1}{n})J\right]y \\ &= 12816.35 \end{aligned}$$

$$\begin{aligned} MS_R &= \frac{SS_R}{dof(SS_R)} \\ &= \frac{12816.35}{2} \\ &= 6408.17 \end{aligned}$$

Test for Significance of Regression for example 5

- Calculate MS_E

$$\begin{aligned} SS_E &= y'\left[I - H\right]y \\ &= 423.37 \end{aligned}$$

$$\begin{aligned} MS_E &= \frac{SS_E}{dof(SS_E)} \\ &= \frac{SS_E}{(n-(k+1))} \\ &= \frac{423.37}{(17-(2+1))} \\ &= 30.24 \end{aligned}$$

- The statistic to test the significance of regression

$$\begin{aligned} f_0 &= \frac{MS_R}{MS_E} \\ &= \frac{6408.17}{423.37/(17-3)} \\ &= 211.9 \end{aligned}$$

Test for Significance of Regression for example

- F_0 =211.9
- Given significance level, α = 0.1, k = 2 and number of samples(n) = 17
- The critical value for this test, corresponding to a significance level of 0.1, is:

$$\begin{aligned} f_{\alpha,k,n-(k+1)} &= f_{0.1,2,14} \\ &= 2.726 \end{aligned}$$

$$f_0 > f_{0.1,2,14}$$

- Hypothesis: β1 = β2 = 0 is rejected
- It is concluded that at least one coefficient out of β1 and β2 is significant

Test for Significance of Regression for example

- It can be concluded that a regression model exists between yield and either one or both of the factors in the table
- The analysis of variance is summarized in the following table.

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares	F Statistic	P Value
Regression	2	12816.35	6408.17	211.9	0.00
Error	14	423.37	30.24		
Total	16	13239.72			

Test on Individual Regression Coefficients (*t* Test)

- The t test is used to check the significance of individual regression coefficients in the multiple linear regression model.
- Adding a significant variable to a regression model makes the model more effective, while adding an unimportant variable may make the model worse.
- The hypothesis statements to test the significance of a particular regression coefficient, β_j are

$$H_0: \quad \beta_j = 0$$
$$H_1: \quad \beta_j \neq 0$$

- The test statistic for this test is based on the t distribution (and is similar to the one used in the case of simple linear regression models

Test on Individual Regression Coefficients (*t* Test)

$$T_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

where the standard error, $se(\hat{\beta}_j)$

- The analyst would accept the null hypothesis if the test statistic lies in the acceptance region:

$$-t_{\alpha/2,\, n-(k+1)} < t < t_{\alpha/2,\, n-(k+1)}$$

This test measures the contribution of a variable while the remaining variables are included in the model.

For the model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$

if the test is carried out for β_1

then the test will check the significance of including the variable x1 in the model that contains x2 and x3

Hence the test is also referred to as partial or marginal test

Test on Individual Regression Coefficients (*t* Test) Example

$$(t_0)_{\hat{\beta}_1} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{1.24}{0.3946} = 3.1393$$

$$(t_0)_{\hat{\beta}_2} = \frac{\hat{\beta}_2}{se(\hat{\beta}_2)} = \frac{12.08}{3.93} = 3.0726$$

- The critical value of t test at a significance of 0.1 are:

$$t_{\alpha/2,n-(k+1)} = t_{0.05,14} = 1.761$$
$$-t_{\alpha/2,n-(k+1)} = -t_{0.05,14} = -1.761$$

- Since t_0 for β_1 and β_2 are do not fall within confidence interval,

  β_1 and β_2 are non zero and are significant