Regression analysis in Minitab

Regression analysis:

a. **Simple linear regression**
b. **Multiple linear regression**

Simple Linear Regression:

Example:

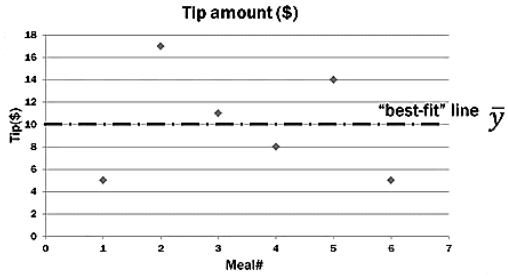
Let's assume that you are a small restaurant owner or a very business minded server / waiter at a nice restaurant. Here in the U.S. "tips" are a very important part of a waiter's pay. Most of the time the dollar amount of the tip is related to the dollar amount of the total bill.

As the waiter or owner, you would like to develop a model that will allow you to make a prediction about what amount of tip to expect for any given bill amount. Therefore one evening, you collect data for six meals.

Unfortunately when you begin to look at your data, you realize you only collected data for the tip amount and not the meal amount also! So this is the best data you have.

Q: How might you predict the tip amount for future meals using only this data? How would you logically model the data?

Meal(#)	Tips Amount($)
1	5
2	17
3	11
4	8
5	14
6	5



With only one variable, and no other information, the best prediction for the next measurement is the mean of the sample itself. The variability in the tip amounts can only be explained by the tips themselves.

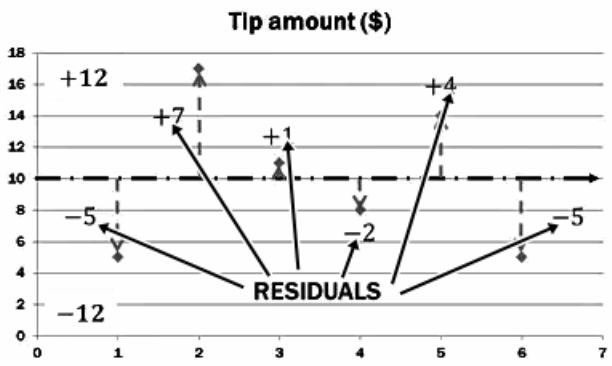
Ans:

The mean of the data set will be the best fit line in this case when one does not have any idea about the dependent variable. That is the best or only possible logical estimate of the dependent variable can be constructed with the given dataset.

Goodness of fit of the Tips:

Meal#	Tip amount ($)
1	5.00
2	17.00
3	11.00
4	8.00
5	14.00
6	5.00

$\bar{y} = \$10$



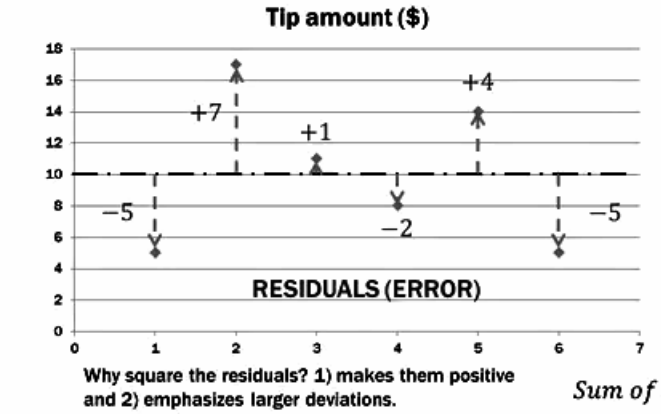
$\hat{y} - y_i = Residuals\ or\ Erros, (e_i)$

$\sum e_i = 0\ or\ sum\ of\ residuals\ is\ zero.$

Here, the sum of residuals above the line and the sum of residuals below the line gives zero. This suggests that the error is evenly distributed around the predicted value of the estimated value.

Regression analysis in Minitab

Squaring the residuals:

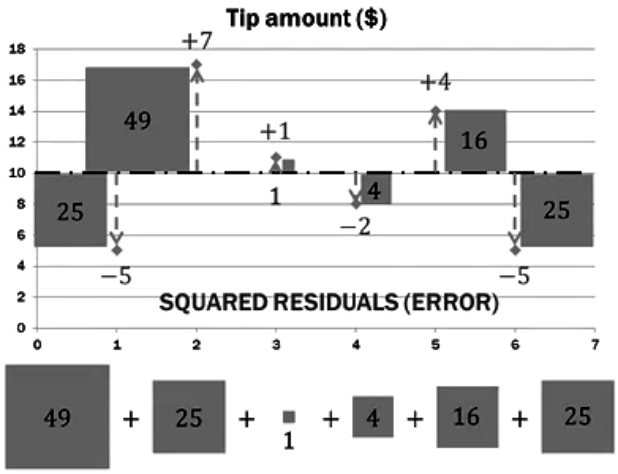


Why square the residuals? 1) makes them positive and 2) emphasizes larger deviations.

Meal#	Residual	Residual²
1	−5	25
2	+7	49
3	+1	1
4	−2	4
5	+4	16
6	−5	25

$Sum\ of\ squared\ errors\ (SSE) = 120$

The goal of simple linear regression is to create a linear model that minimizes the sum of squares of the residuals / error (SSE).



Meal#	Residual	Residual²
1	−5	25
2	+7	49
3	+1	1
4	−2	4
5	+4	16
6	−5	25

$= 120$ $Sum\ of\ squared\ errors\ (SSE) = 120$

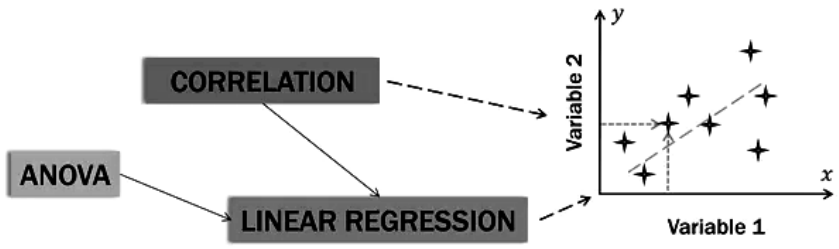
If our regression model is significant, it will "eat up" much of the raw SSE we had when we assumed (like this problem) that the independent variable did not even exist. The regression line will/should literally "fit" the data better.It will minimize the residuals.

When conducting simple linear regression with **TWO** variables, we will determine how good that line 'fits" the data by **comparing It to THIS TYPE; where we pretend the second variable does not even exist.**

Q. If a two-variable regression model looks like this example, what does the other variable do to help explain the dependent variable?

Ans: <u>NOTHING</u>.

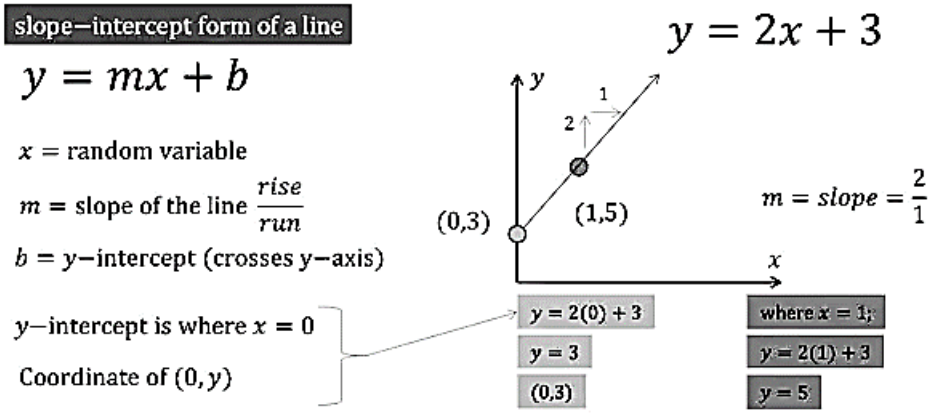
- Simple linear regression is really a **comparison of two models**
  a. One is where the independent variable does not even exist
  b. And the other uses the best fit regression line
- If there is only one variable, the best prediction for other values is the **mean** of the "dependent" variable
- The difference between the best-fit line and the observed value is called the residual (or error)
- The residuals are squared and then added together to generate **sum of squares (LITERALLY) residuals / error, SSE**.
- Simple linear regression is designed to find the best fitting line through the data that minimizes the SSE.



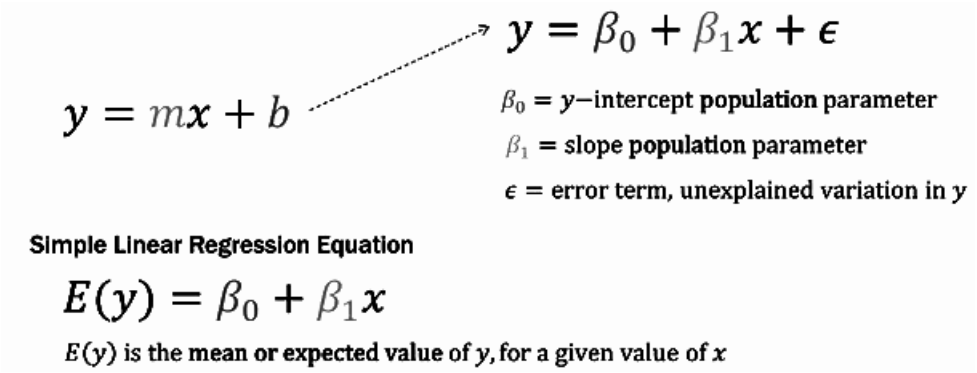
The value of one variable**, is a function of** the other variable.
The value of y**, is a function of** x; $y = f(x)$.
The value of the dependent variable**, is a function of the** independent variable.

Regression analysis in Minitab

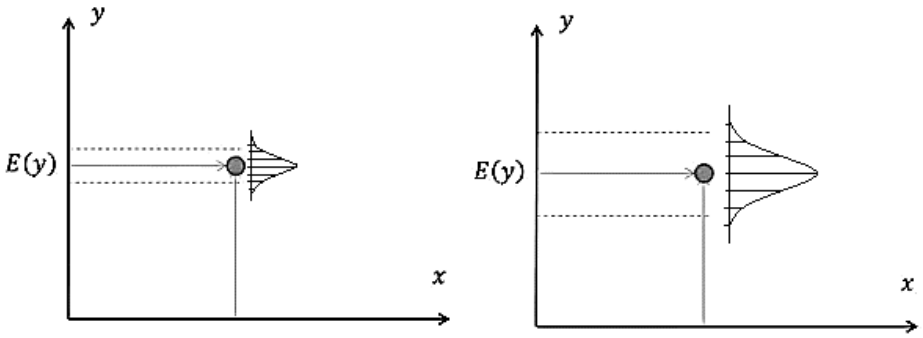
Algebra of lines:



Simple linear regression model:

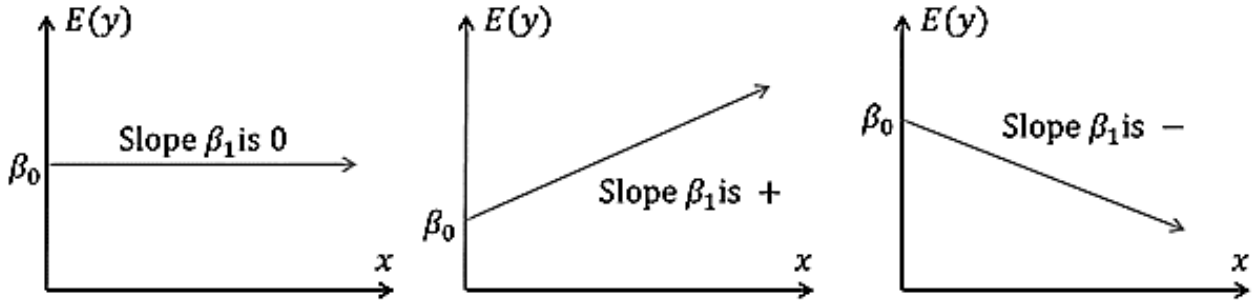


Distribution of y values:



General Regression Line:

$E(y) = \beta_0 + \beta_1 x$



$E(y) = \beta_0 + 0(x)$ $E(y) = \beta_0 + \beta_1 x$ $E(y) = \beta_0 - \beta_1 x$

REGRESSION EQUATION WITH ESTIMATES:

If we actually knew **the population parameters**, β_0 *and* , β_1 , we could use the Simple Linear Regression Equation.

Regression analysis in Minitab

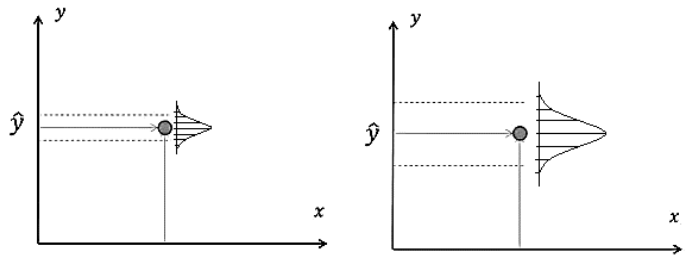
$$E(y) = \beta_0 + \beta_1 x$$

In reality we almost never have the population parameters. Therefore we will estimate them using **sample data**. When using sample data, we have to change our equation a little bit.

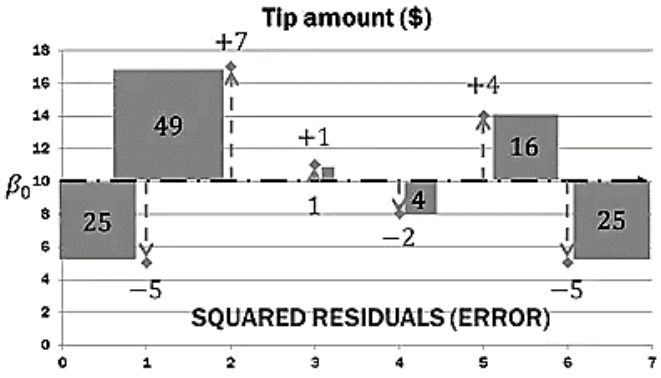
$\hat{y}$ *is the point estimator of the* $E(y)$ *and* $\hat{y}$ *is the mean value of* y *for a given value of* x.

$$\hat{y} = b_0 + b_1 x$$

Distribution of sample values:



When the slope, $\beta_1 = 0$:



When concluding simple linear regression with TWO variables, we will determine how good the regression line "fits" the data by comparising it to THIS TYPE, where we pretend the second variable does not even exists, therefore the slop is , $\beta_1 = 0$.

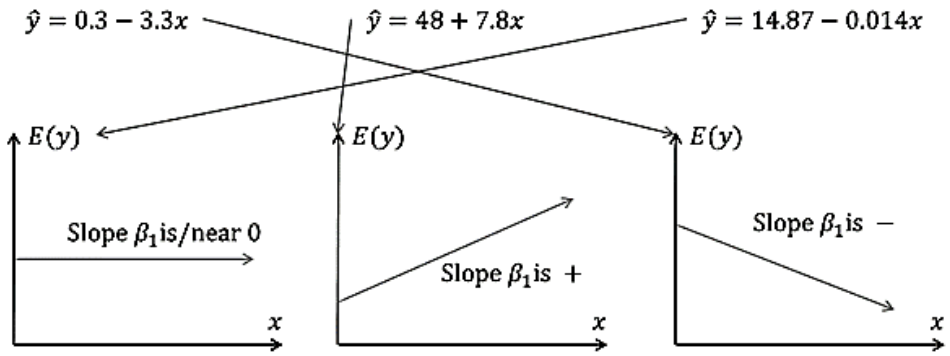
In this situation,the value of $\hat{y}$ is 10 for every value of x.

$\hat{y} = b_0 + b_1 x$ $\qquad$ $b_0 = 10$

Since x=0, $\hat{y} = b_0 = 10$

Sum of squared errors (SSE) = 120

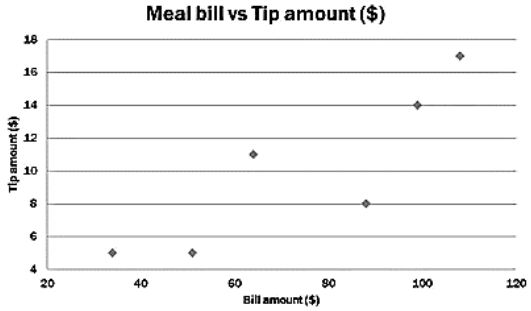
Pettern matching to general linear model:



Getting ready for least square:

Obs	Bill($)	Tips($)
1	34	5
2	108	17
3	64	11
4	88	8
5	99	14
6	51	5

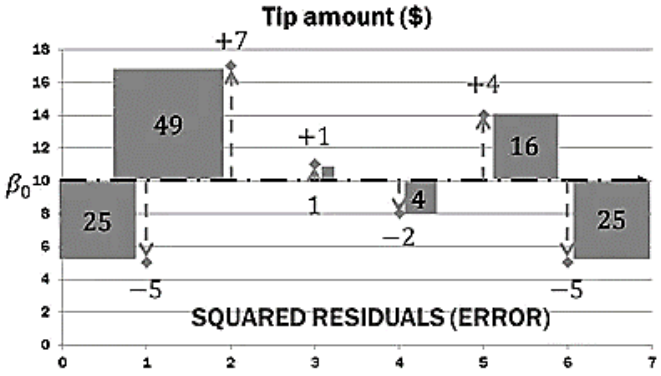
Scatter Plot:



Regression analysis in Minitab

You want to know to what degree the tip amount can be predicted by the bill. So the tip is the DEPENDENT variable, bill is the INDEPENDENT variable.

When the slope, $\beta_1 = 0$:



When concluding simple linear regression with TWO variables, we will determine how good the regression line "fits" the data by comparising it to THIS TYPE, where we pretend the second variable does not even exists, therefore the slop is , $\beta_1 = 0$.

In this situation,the value of $\hat{y}$ is 10 for every value of x.

$\hat{y} = b_0 + b_1 x$ $\qquad$ $b_0 = 10$

Since x=0, $\hat{y} = b_0 = 10$

Sum of squared errors (SSE) = 120

When we conduct regression on two or more variable then we will compare the two variable model with the above mentioned model, with the expectation that the additional information about the bill amount will improvise the predictive capacity of the new model. If the new model is not good in predicting inspite of the avalibality of additional information then that model is not significantly different from this model and the additional information is of no use.

Least Square Method:

The goal of this method is to $min \sum (y_i - \hat{y}_i)^2$.

y_i: observed value of dependent variable (tip amount)

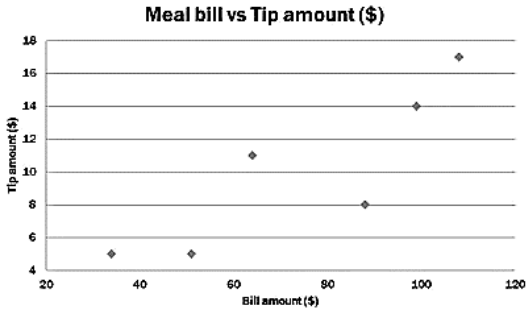
$\hat{y}_i$: estimated(predicted)value of the dependent variable (predicted tip amount)

Plain English. The goal Is to minimize the sum of the squared differences between the observed values (y_i) of the dependent variable and the estimated value $(\hat{y}_i)$ of the dependent variable , that is provided by the regression line. sum of squared residuals.

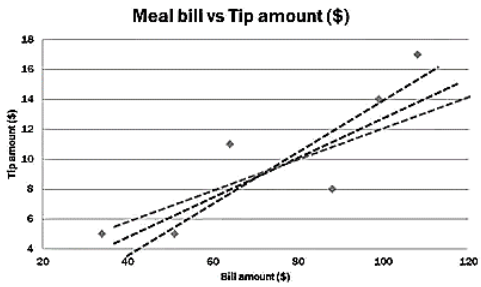
Not only that, but the sum of squared residuals should be much smaller than when we just used the dependent variable alone; $\beta_1 = 0$, $\hat{y} = 10$ for all values of x. That sum opf the squared residual was 120.

Step 1: (Sctter Plot)

Obs	Bill($)	Tips($)
1	34	5
2	108	17
3	64	11
4	88	8
5	99	14
6	51	5



Step 2: (Look a visual line)



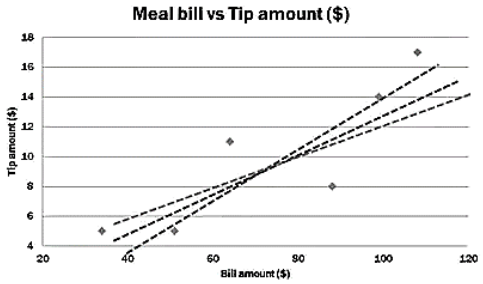
Q. Does the data seem to fall along a line?

Ans: In this case, **YES**: Proceed.

If not...if it's a BLOB with no linear pattern, **then stop**.

Regression analysis in Minitab

Step 3: (Crrelation)



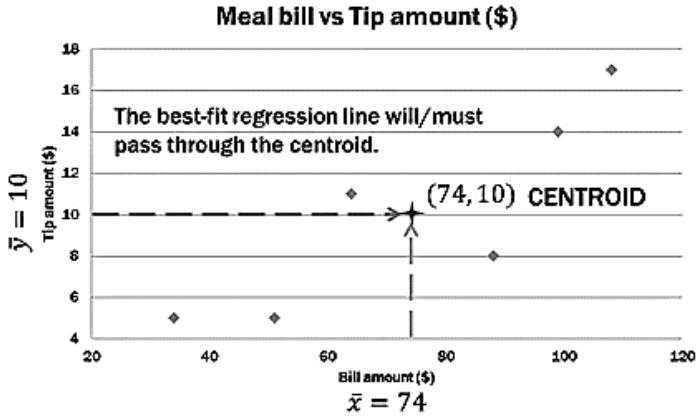
Q: What is the correlation coefficeint among the dependent and independent variable?

Ans: in this case, r=.866 .

Proceed further if the correlation coefficent suggests significantly strong association among the depent and independent variables.

Step 4(Descriptive statistics/ Centroid) :

Obs	Bill($)	Tips($)
1	34	5
2	108	17
3	64	11
4	88	8
5	99	14
6	51	5
Mean	$\bar{x} = 74$	$\bar{y} = 10$



The best fit line must pass through $(\bar{x}, \bar{y})$.

Step 5 (Calculation):

$$\hat{y} = b_0 + b_1 x$$

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

$\bar{x}$: $Mean\ of\ the\ dependent\ variable$.

$\bar{y}$: $Mean\ of\ the\ independent\ variable$.

x_i: $Value\ of\ the\ dependent\ variable$.

y_i: $Value\ of\ the\ independent\ variable$.

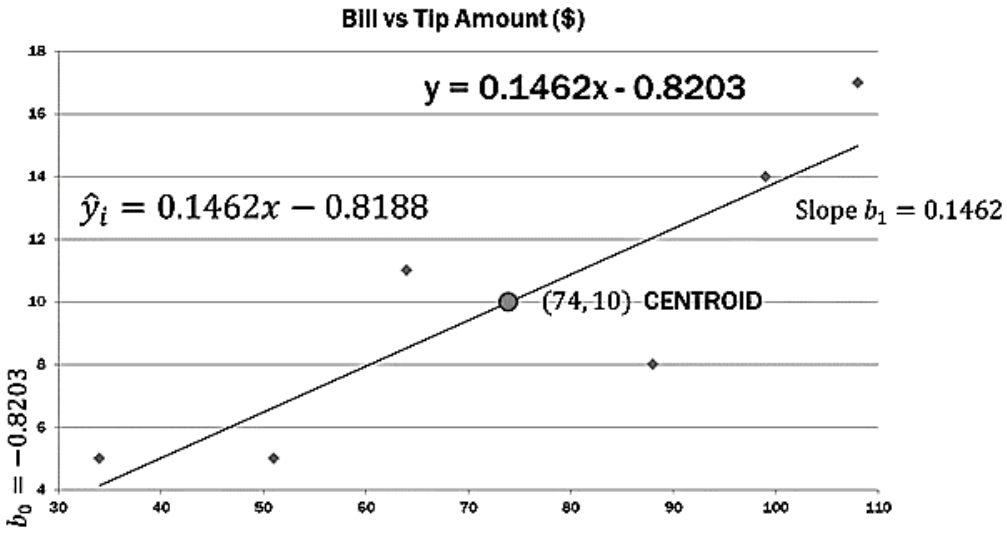
Meal	Total Bill	Tips Amount	Bill deviation	Tip deviation	Product deviation	Bill deviation squared
	x_i	y_i	$(y_i - \bar{y})$	$(x_i - \bar{x})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	34	5				
2	108	17				
3	64	11				
4	88	8				
5	99	14				
6	51	5				

Regression analysis in Minitab

After all the calculation done, we will land up an regression equation of this form in this case,

$\hat{y} = .1462x - .8188$

Step 6 (Plot on graph) :



Step 7 (Interpretation):

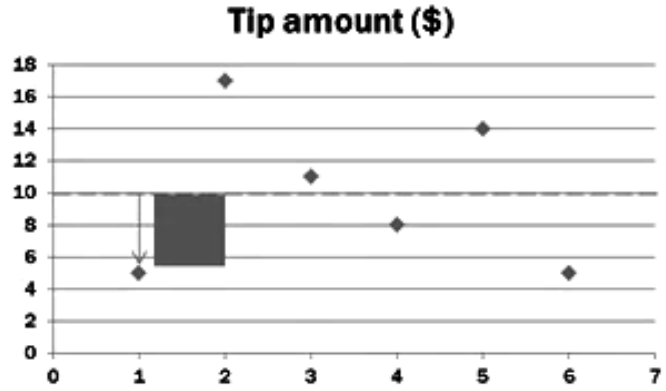
$\hat{y} = .1462x - .8188$ → It the bill amount (x) Is zero, then the expected/predicted tip amount Is $-0.6166 or negative 82.cents1 Does this make sense? NO. The Intercept may or may not make sense in the °real world."

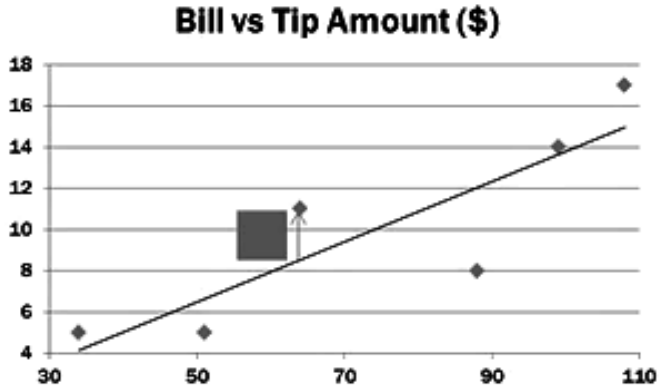
For every $1 the bill amount (x) Increases, we would expect the tip amount to increase by $0.1462 or about 15-cents.

Step 8 (fit and coefficient of determination of the model):

Q: Is this model good enough to work with?

Here we are comparing two models. The first model is made with the mean line of the independent variable and the second model consists the simple linear equation constructed with the dependent and independent dataset.





SSE = 120

SST = 120

SST=120

With the only dependent variable, the only sum of squares Is due to error. Therefor it is also the total, and the maximum sum of squares for the data under analysis.

SST=120

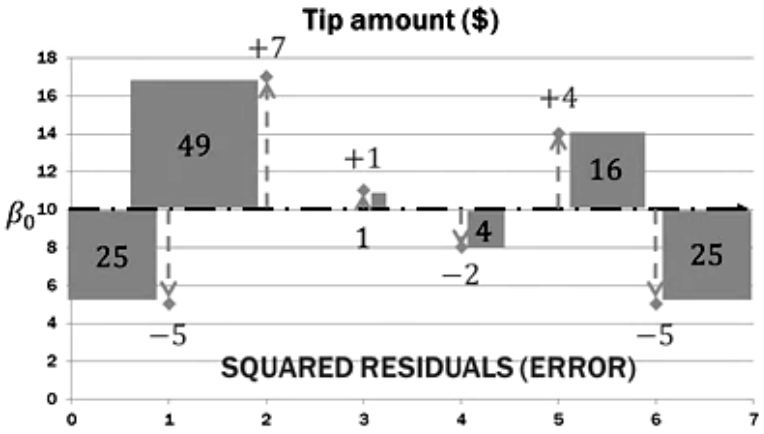
SSE=?

SST-SSE=SSR

With the both IV and DV the total sum of squares remains the same. But (idealy) the error sum of Squares will be reduced significantly. The difference between SST and SSE is due to regression or SSR.

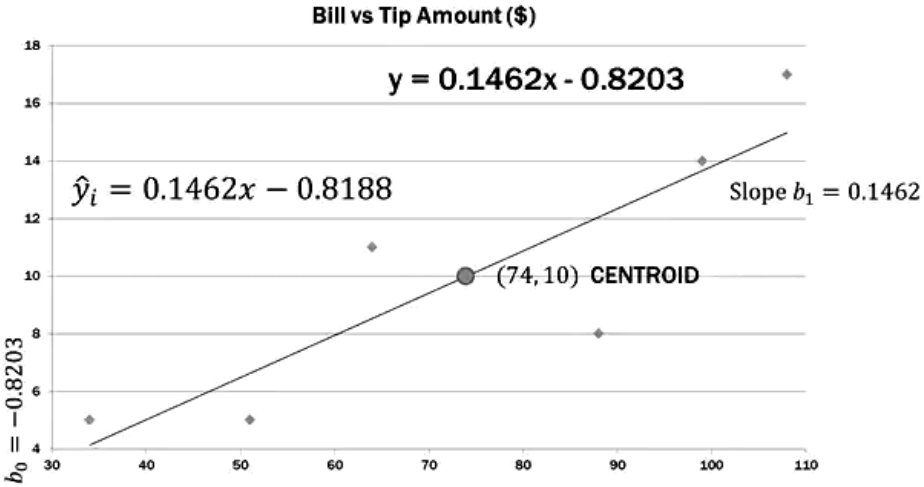
Regression analysis in Minitab

When the slope, $\beta_1 = 0$:

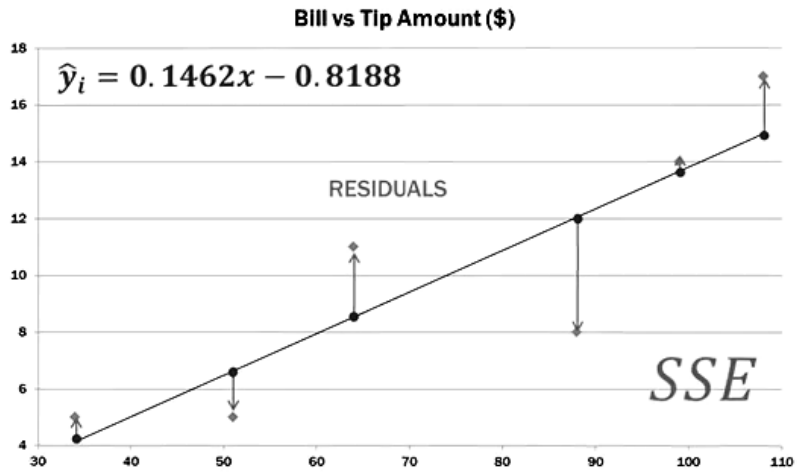


Having only the DV, the best prediction for the tip of the next meal is the mean of the tips
$\hat{y} = \mathbf{10}$

Since the mean line is flat. Its slope is zero, $\beta_1 = 0$.

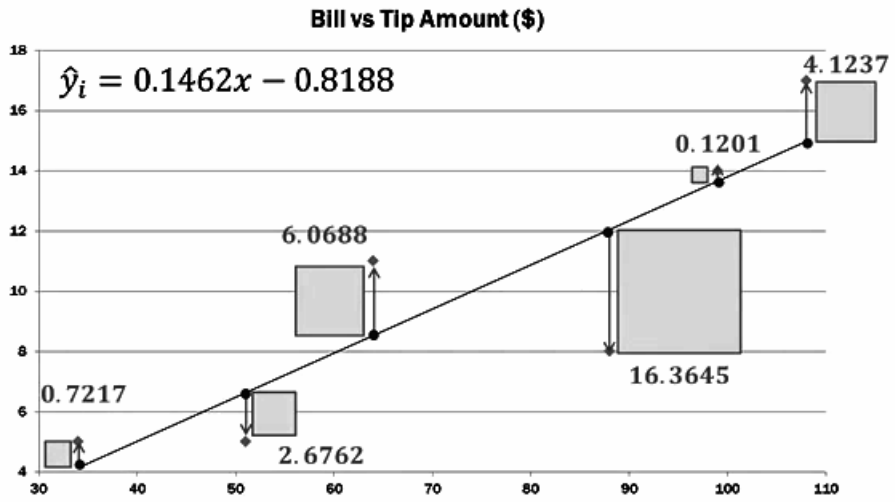


Meal	Total Bill	Tips Amount	Regression Eq	Predicated Tip	Error	Sum sq errors
	x_i	y_i	$.1462x - .8188$	$\hat{y}$	$y_i - \hat{y}$	$\sum(y_i - \hat{y})^2$
1	34	5				
2	108	17				
3	64	11				
4	88	8				
5	99	14				
6	51	5				
	$\overline{x}$	$\overline{y}$				



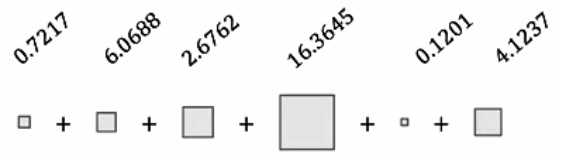
The values on the regression line denotes the estimated values of the variable y and the diamond above and below the line denotes the actual value of the dependent variable. The distance between the predicated value and the observed value is the error. The sum of error terms is zero or approximately zero. To get sum of squared errors the above mention table is constructed and the required calculation is done using Minitab or excel.

Regression analysis in Minitab



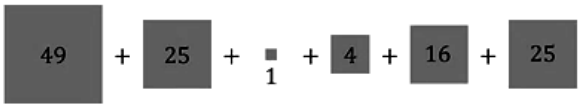
Sum of squared errors comparison:

DV and IV (Tips amount as the function of bill amount) :

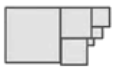


SSE=30.075

DV (Tips amount only):



SSE=120



= 30.075

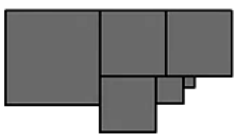
So when we conducted the regression, the SSE decreased to 30.075 from 120. That is, 30.075 of the sum of squares was explained or allocated to errors.

Where did the 89.925 go?

The 89.925 is the sum of squares due to regression.

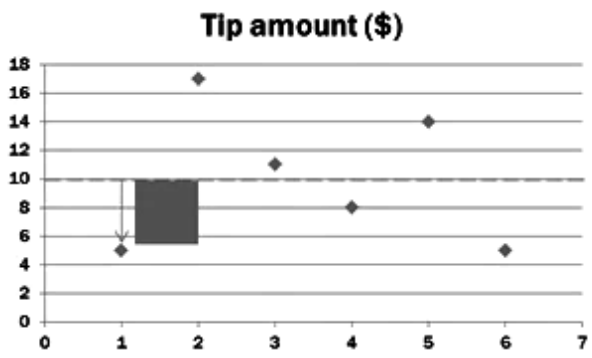
So,

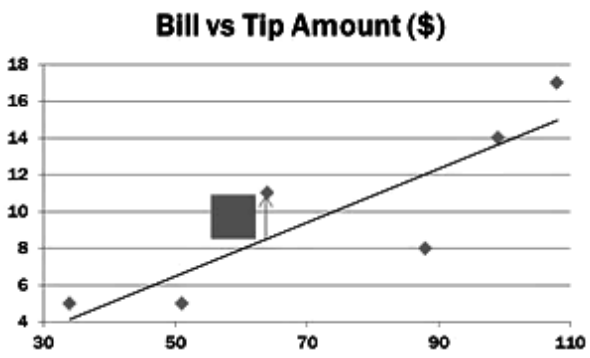
SST=SSE+SSR



Sum of Squared Errors Comparison

= 120





SSE = 120
SST = 120
SST=120

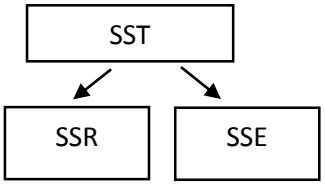
SST=120
SSE=30.075
SST-SSE=SSR=120-30.075=89.925

Regression analysis in Minitab

Coefficient of determination:

Q: How well does the estimated regression equation fit our data?

This is where regression begins to look a lot like ANOVA; the total sum of squares is partitioned or allocated to SSE and SSR.



If SSR is large, it uses up more of SST and therefor SSE is smaller relative to SST. The coefficient of determination quantifies this ratio as a percentage.

Coefficient of determination = $r^2 = \frac{SSR}{SST}$

This is where the regression looks like ANOVA. The total sum of squares is partitioned or allocated to in SSR and SSE.

Analysis of Variance

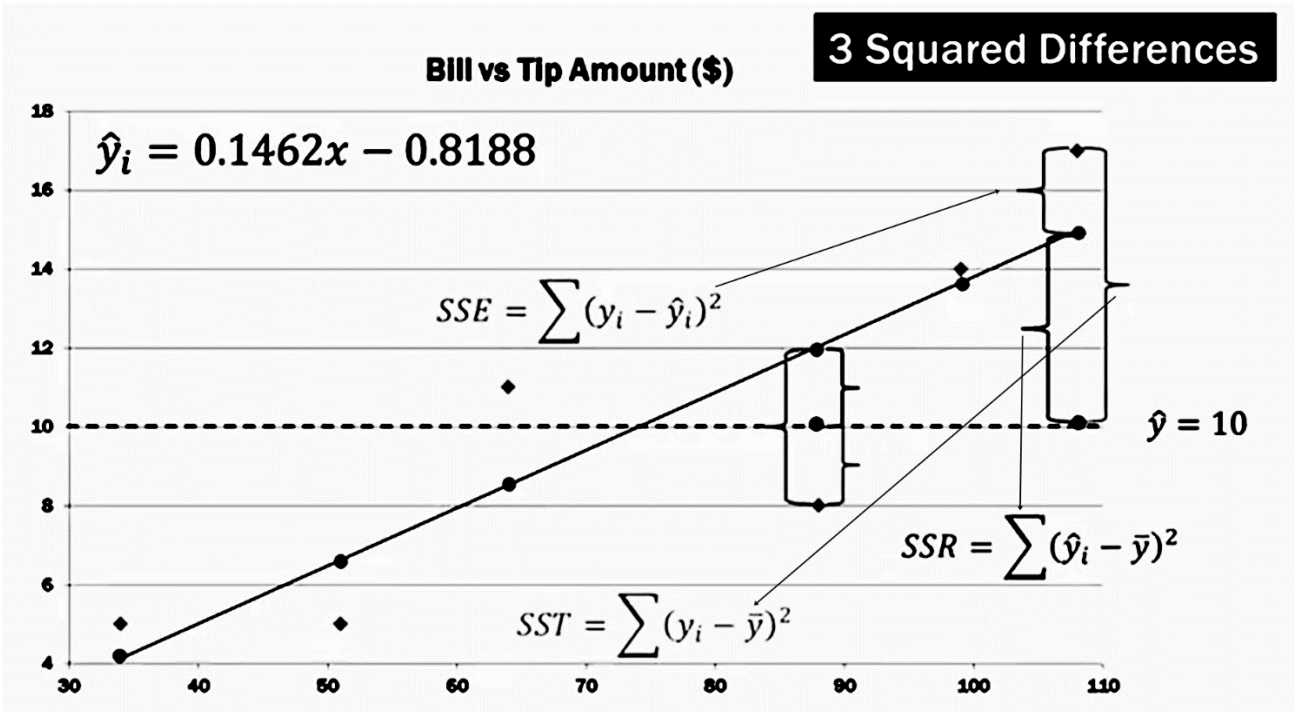
Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	89.93	89.925	11.96	0.026
C1	1	89.93	89.925	11.96	0.026
Error	4	30.07	7.519		
Total	5	120.00			

Coefficient of determination = $r^2 = \frac{SSR}{SST}$

Coefficient of determination = $r^2 = \frac{89.925}{120}$

Coefficient of determination = $r^2 = .7493 \; or \; 74.93\%$

We can conclude that 74.93% of the total sum of squares can be explained by using the estimated regression equation to predict the tip amount. The remainder is error.



This show how the mean values, observed values, predicated value and SST, SSE and SSR are related. Also this shows why the relation **SST=SSR+SSE.**

Regression analysis in Minitab

Multiple Linear Regression

Example:

REGIONAL DELIVERY SERVICE

Let's assume that you are a small business owner for Regional Delivery Service, Inc. (RDS) who offers same-day delivery for letters, packages, and other small cargo. You are able to use Google Maps to group individual deliveries into one trip to reduce time and fuel costs. Therefore some trips will have more than one delivery. As the owner, you would like to be able to estimate how long a delivery will take based on two factors:

1) The total distance of the trip in miles and

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

To conduct your analysis you take a random sample of 10 past trips and record three pieces of information for each trip:

1) Total miles traveled,

2) Number of deliveries,

3) Total travel time in hours.

Miles traveled (x_1)	**Number of deliveries** (x_2)	**Traveled time (in hours)** (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
106	5	7.3
79	3	6.4

Remember that in this case you would like to be able to predict the total travel time using both the miles traveled and the number of deliveries in each trip.

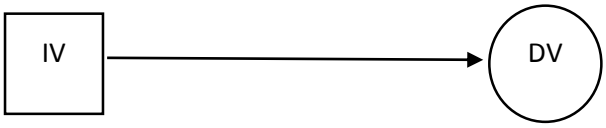
In what ways the traveled time DEPEND on the first two measure?

Travel time is the dependent variable, miles traveled and number of deliveries are the independent variables.

Note: Some prefer predictor variable(s) and response variable instead of independent and dependent variable respectively.

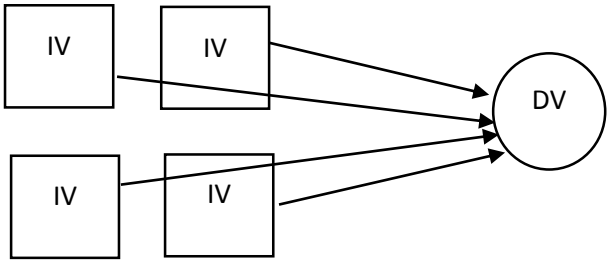
Multiple linear regression is an extension of simple linear regression.

Simple Linear Regression



One to One

Multiple Linear Regression:



Many to one

Regression analysis in Minitab

Addition concepts related to multiple linear regression:

Adding more independent variables to a multiple regression procedure does not mean the regression will be "better" or offer better predictions; in fact it can make things worse. This is called **OVERFITTING**.

The addition of more independent variables creates more relationships among them. So not only are the independent variables potentially related to the dependent variable, they are also potentially related to each other. When this happens, it is called **MULTICOLLINEARITY**.

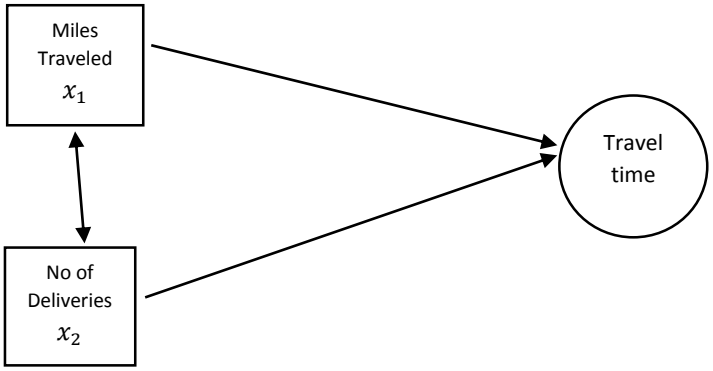
The ideal is for all of the independent variables to be correlated with the dependent variable but **NOT** with each other.

Because of **Multicollinearity** and **overfitting**, there is a fair amount of prep-work to do **BEFORE** conducting multiple regression analysis if one is to do it properly.

- **Correlation**
- **Scatter plot**
- **Simple regression**

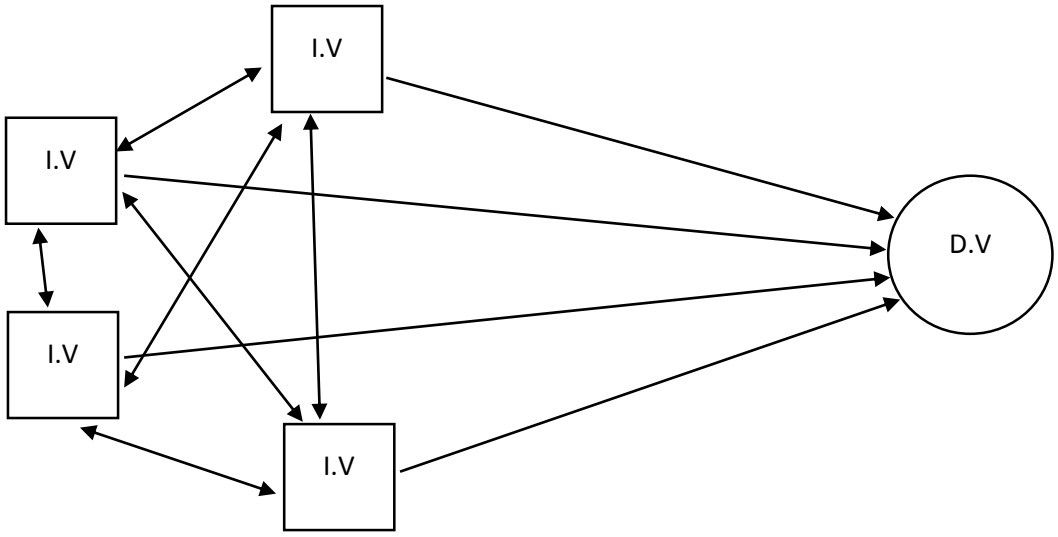
Additional relationship among the dependent and independent variables:

Independent variables **Dependent variables**



Multiple linear regression

Many to one



Multiple Regression

Many to one

The ideal is for all of the independent variables to be correlated with the dependent variable but **NOT** with each other.

Here, **total 10 relationships** are under consideration.

Some independent variables, or sets of independent variables, are better at predicting the DV than others.

Also there may be some of the independent variables which contributes nothing to the model.

Multiple linear regression model:

Multiple Regression Model:	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$
Multiple Regression Equation:	$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$
Estimated Multiple Regression Equation:	$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$

The predictor variable y is the original value of the data which can be replaced by the equation of the model. (Here this is assumed that the original value of the predictor variable can be written with a linear equation, otherwise this will not be appropriate to fit a linear model to a non-linear dataset.) This model is linear in a sense that the coefficient of the equation is of linear form.

Another important assumption of this model is that, the error is distributed normally. In previous case the sum of error zero, is an indication in the similar direction. So, in the above mentioned case $E(\varepsilon) = 0$.

Now, $\hat{y}$ *is the predicted value of the dependent variable and* $b_0, b_1, b_2, \ldots, b_p$ *is the linear estimate of* $\beta_0, \beta_1, \ldots, \beta_p$.

Example:	$\hat{y} = 6.211 + .014x_1 + .383x_2 - .607x_3$
Estimated Multiple Regression Equation:	$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$

Intercept

How to interpret the coefficients of multiple linear regression:

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

x_1: *capital investment* ($1000s)
x_2: *marketing expenditure* ($1000s)
$\hat{y}$: *predicted sales amount* ($1000s)

In multiple regression each coefficient is interpreted as the estimated change in y corresponding to a one unit change in a variable, **when all the other variables are held constant**.

So in this example. $9000 is an estimate of the expected Increase In sales y, corresponding to a $1000 Increase In capital Investment(x_1) when marketing expenditures(x_2) are held constant.

Review:

- Multiple regression is an extension of simple linear regression.
- Two or more Independent variables are used to predict / explain the variance in one dependent variable.
- Two problems may arise:
  a. **Over fitting**
  b. **Multicollinearity**
- Over fitting is caused by adding too many Independent variables; they account for more variance but add nothing to the model.
- Multicollinearity happens when some/all of the independent variables are correlated with each other.
- In multiple regression, each coefficient is interpreted as the estimated change in y corresponding to a one unit change in a variable, when all other variables are held constant.

Regression analysis in Minitab

REGIONAL DELIVERY SERVICE:

Let's assume that you are a small business owner for Regional Delivery Service, Inc. (RDS) who offers same-day delivery for letters, packages, and other small cargo. You are able to use Google Maps to group individual deliveries into one trip to reduce time and fuel costs. Therefore some trips will have more than one delivery. As the owner, you would like to be able to estimate how long a delivery will take based on three factors:

1) The total distance of the trip in miles,
2) The number of deliveries that must be made during the trip, and
3) The daily price of gas/petrol in U.S. dollars.

Steps to be followed before conducting a multiple linear regression:

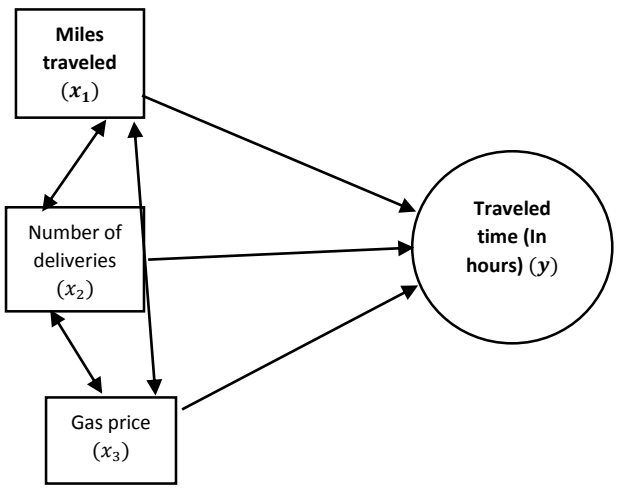
1. Generate a list of potential variables: independent(s) and dependent.
2. Collect the data on the variable
3. **Check the relationships between each independent variable and the dependent variable using the scatter plot and correlations.**
4. **Check the relationships among the independent variables using scatterplots and correlations.**
5. (Optional) Conduct simple linear regressions for each IV/DV pair.
6. **Use non redundant independent variables in the analysis to find the best fit model.**
7. Use the best fitting model to maim predictions about the dependent variable.

To conduct your analysis you take a random sample of 10 past trips and record four pieces of information for each trip:
1) total miles traveled,
2) Number of deliveries,
3) The daily gas price, and
4) total travel time (in hours)

Miles traveled (x_1)	**Number of deliveries** (x_2)	**Gas price** (x_3)	**Traveled time (in hours)** (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
106	5	3.54	7.3
79	3	3.25	6.4

Relationship among the variables:

Independent variables **Dependent variables**



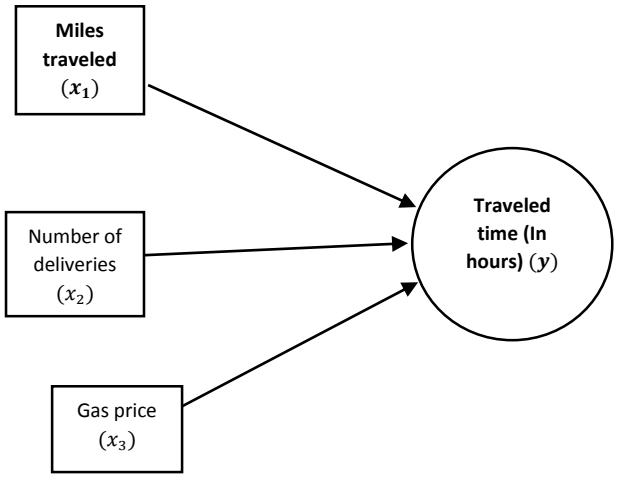
Multiple linear regression
Many to one

Total 6 relationships to be analyzed.

Regression analysis in Minitab

Relationship among the dependent variable to independent variables:

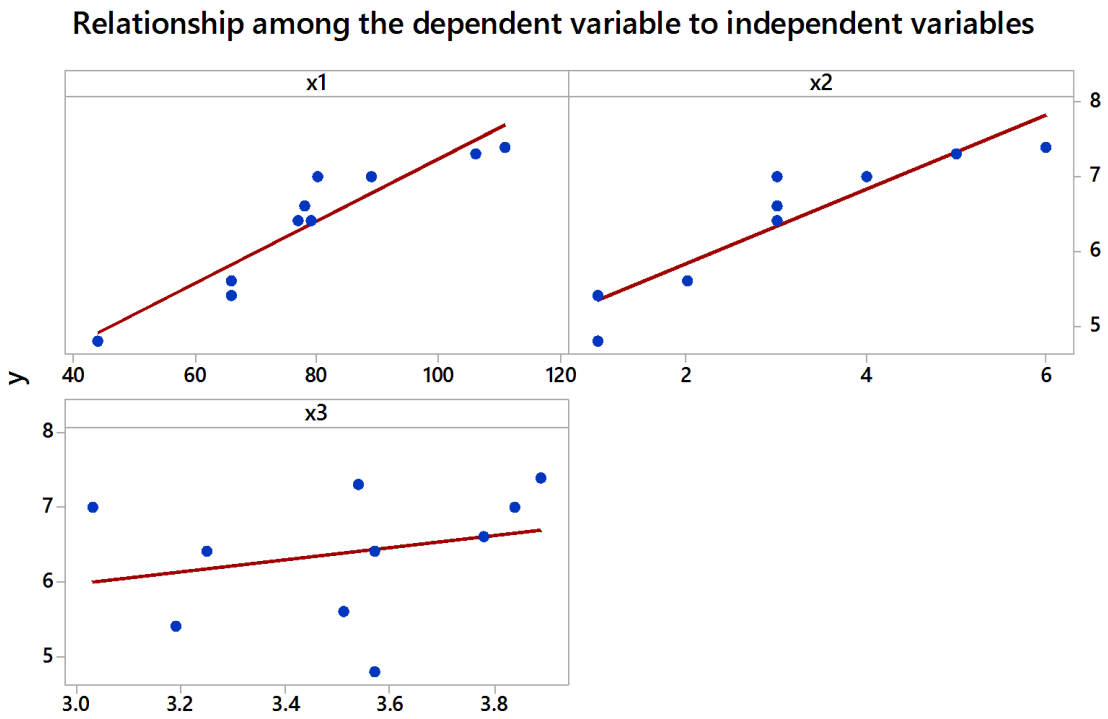
Independent variables **Dependent variables**



Multiple linear regression
Many to one

Total 3 relationships to be analyzed.

Independent to dependent variables scatter plot (Relevancy Check):



Interpreting the scatter plots:

This is a visual tool. Some times this can be misleading but in general cases this provides a rough idea about the data. In this case the in first two scatter plots, shows a linear pattern but the third scatter plot looks quite non-linear in nature. Again, this is difficult to conclude anything but this shows an indication that the third variables or gas price in this case may not be linearly related with the dependent variable in this case travel time.

Scatter plot summary:

- **Dependent variable vs independent variables**
  a. Travel time(y) appears highly correlated with miles Traveled (x_1)
  b. Travel Time(y) appears highly correlated with number Deliveries(x_2)
  c. Travel Time(y) **DOES NOT** appear highly correlated with gas Price (x_3)

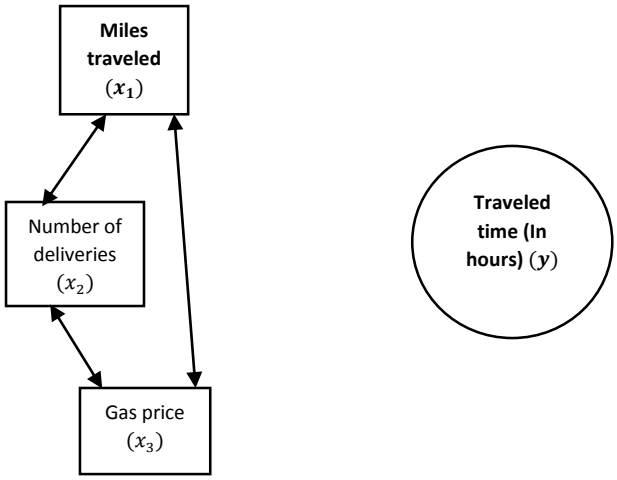
Since gas Price (x_3)does **NOT APPEAR CORRELATED** with the dependent variable we would **NOT** use that variable in the multiple regression.

Note: *For now, we will keep gas Price in and then take it out later for learning purposes.*

Regression analysis in Minitab

Relationship among the independent variables:

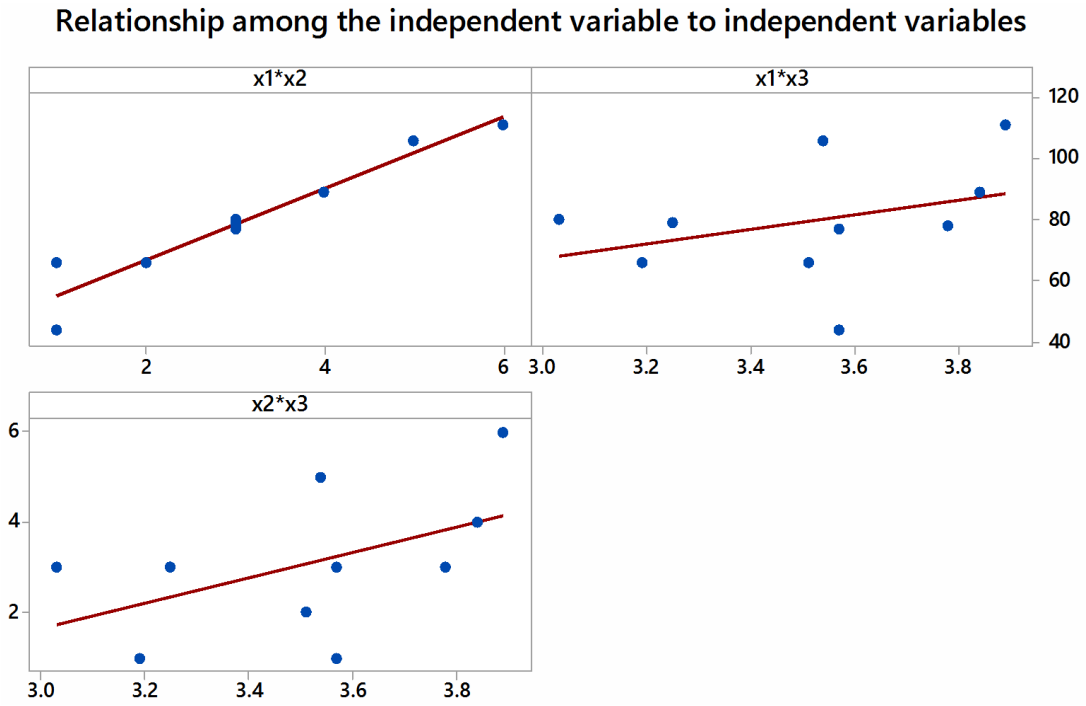
Independent variables **Dependent variables**



Multiple linear regression
Many to one

Total 3 relationships to be analyzed.

Independent to independent variables scatter plot (Check for Multicollinearity):



In the check for **Multicollinearity**, the linear pattern is not desired. So, this is quite evident from the scatter plot that the linear relationship among miles traveled and the number of deliveries are quite strong. Other two scatter plot does not suggests any strong linear pattern. This scatter plots suggests that the number of delivers and the miles traveled may be related in a linear relationship but the price of gas and the miles traveled and price of gas and the number of deliveries may not be related in a linear fashion.

Real life example of Multicollinearity:

In tea, the result of sugar free and sugar cube are same. Both of them makes a cup of tea sweet. In this, case this is difficult to understand the sweetness caused by two different components and measure the variability caused by each component.

Scatter plot summary:

- Number deliveries (x_2) APPEARS highly correlated with miles Traveled(x_1); this is Multicollinearity.
- Miles Traveled (x_1) does not appear highly correlated with gas Price(x_3).
- Gas Prices (x_3) does not appear correlated with number Deliveries(x_2).

Regression analysis in Minitab

Since number deliveries is **HIGHLY CORRELATED** with miles traveled, we would **NOT** use **BOTH** in the multiple regression; they are redundant.

Note: For now, we will keep both in and then take one out later for learning purposes.

Correlation analysis:

Correlation: x1, x2, x3, y

	x1	x2	x3
x2	0.959		
	0.000		
x3	0.348	0.498	
	0.325	0.143	
y	0.935	0.916	0.267
	0.000	0.000	0.455

Cell Contents: Pearson correlation
P-Value

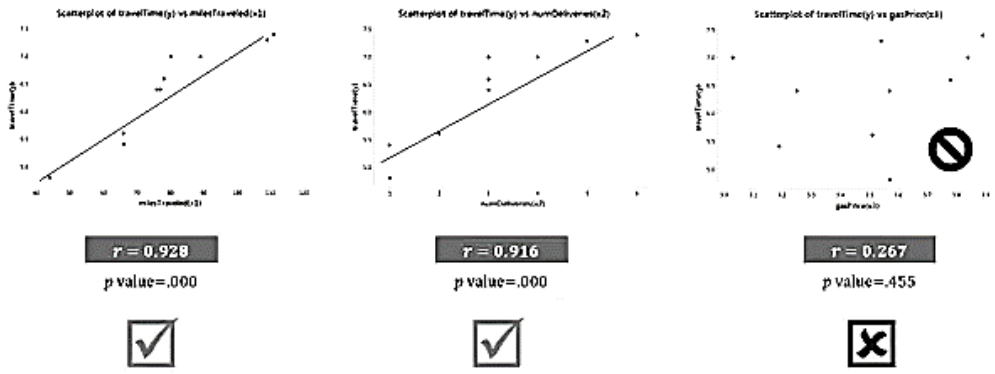
Relation among dependent variables and independent variables:

This correlation matrix shows that the correlation among dependent variable (y) and independent variables (x_1) ***and*** (x_2) *are high and they are statistically significant. This is a statistical evidence towards the visualization of the data in scatter plot.*

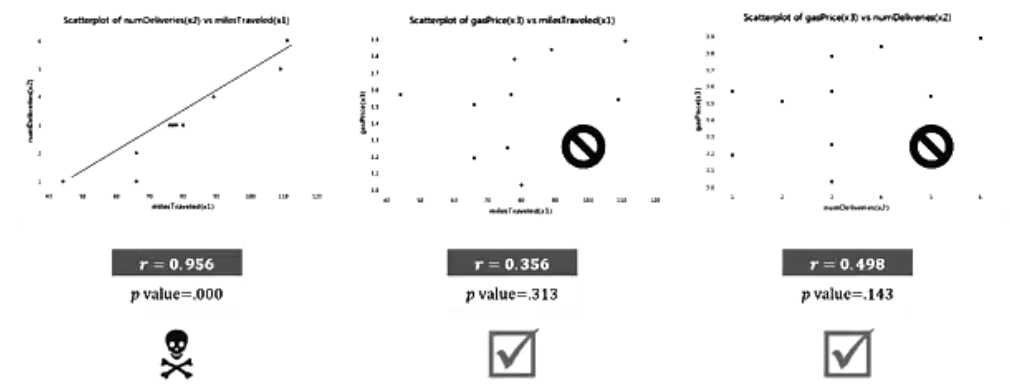
Relation among dependent variables and dependent variables:

From the correlation matrix this can be shown that the dependent variable (x_1) *and* (x_2) *are highly correlated among them. This is may cause Multicollinearity. This risk is called potential Multicollinearity risk.*

Independent variable vs dependent variable scatter plot:



Independent variables scatter plot:



Regression analysis in Minitab

Correlation summary:

- Correlation analysis confirms the conclusions reached by visual examination of the scatterplots.
- **Redundant multicollinear variables:**
  - Miles traveled and number deliveries are both highly correlated with each other and therefore are redundant; only one should be used in the multiple regression analysis.
- **Non-contributing variables:**
  - Gas Price is NOT correlated with the depended variable and should be excluded

Conduct simple linear regressions for each IV/DV pair:

In this part we will conduct regression analysis using Minitab and interpret the result using the non-contributing and redundant variable under consideration and will observe how the affect the analysis.

In this context, we will consider the following statistics:

Coefficients
- Values
- T-statistic
- P-value

Analysis of variance
- F-value
- P-value

R-square, R-squared (adjusted), R-squared (predicted)
VIF (variance inflation factor)
Mallows C_p

Here,
R-square is the statistic which represent the percentage of variation explained by the model. R-square is also known as the coefficient of determination. The formula of R-square is

$$R^2 = 1 - \frac{SSE}{SST}$$

SST=SSR+SSE,
SST= total sum of squares,
SSE= sum of squares due to error and
SSR= sum of squares due to regression.
Since in this case only two variables under consideration R-square is nothing but the correlation coefficient between the dependent and independent variable:

$$r^2{}_{xy} = \frac{\sqrt{\sum(x-\overline{x})(y-\overline{y})}}{\sqrt{\sum(x-\overline{x})}\sqrt{\sum(y-\overline{y})}} = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$$

Adjusted R square is the measure of same thing. This is just adjusted with respect to number of independent variable present and the size of the sample considered for the model. Here is the formula of adjusted R square;

$$Adjusted\ R^2 = 1 - \left|\left(1-R^2\right)\frac{(n-1)}{(n-1-k)}\right|$$

In this case number of independent variables are one. This is taken into account and the new measure has been calculated.
Sum of squared error due to regression is represented by

$$\mathbf{SSE} = \sum(y-\hat{y})^2$$

This is also a measure of goodness of the regression. How small the SSE is, that much better is the model. The advantage of SSE is that, SSE is in the units of the predicted variable.
To obtain the estimate of average deviation from the regression line, we use the square root of the SSE divided by n−k. k is the number of independent variable used in the regression equation.

The **standard error of the estimate (SEE)** therefore becomes

Regression analysis in Minitab

$$\mathbf{SEE} = \sqrt{\frac{\boldsymbol{SSE}}{\boldsymbol{n-k}}}$$

Since the regression equation is based on the sample data, the validity of the regression equation must be validated for the population data.
If the sample regression model is same as the population regression model and the population regression model has a slope of zero, the sample regression model must have the same.
So, to test the validity of the regression line, a test of, $\beta_1$or the slope of the regression line has been conducted. Here, we are interested in testing the population slop using the slope of the regression equation.

$$\boldsymbol{H_0}: \boldsymbol{\beta_1 = 0} \text{ Against } \boldsymbol{H_1}: \boldsymbol{\beta_1 \neq 0}$$

Here the formula for t-stat will be

$$\boldsymbol{t = \frac{b_1 - \beta_1}{S_{b_1}}}$$

Since, the alternative hypothesis is two side and there are k independent variables are in the equation there for the calculated value of the t statistics is expected to follow t distribution with α/2 level of significance and (n-k) degree of freedom.

The formula of $S_{b_1}$is $S_{b_1} = \frac{SSE}{\sqrt{SS_x}}$. Since, the exact value of the SSE is not known for the population parameter this is going to be replaced by the estimate of SSE, which is

$$\mathbf{SEE} = \sqrt{\frac{\boldsymbol{SSE}}{\boldsymbol{n-k}}}$$

If the p-value in the test is lower than the value selected for the threshold in the test and this can be concluded that the null hypothesis is significant. In the above mentioned case this has been conducted for a coefficient of x, this test can be also applied for the intercept.

Confidence interval and prediction interval:

For a given value of $x_i$the confidence interval to estimate $\hat{y}$ will be:

$$\boldsymbol{\hat{y} \pm t_{\alpha/2,(n-k)} S_e \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{SS_{xx}}}}$$

$$\boldsymbol{SS_{xx} = \frac{\sum (x_i - \bar{x})^2}{n}}$$

Not only for the regression but the confidence interval and prediction interval can be constructed for the regression coefficients using to respective standard error and other estimates.

For more than two variable, the slight modification in this tests has been introduced.

$$\boldsymbol{H_0}: \boldsymbol{\beta_1 = \beta_2 = \cdots = \beta_p = 0} \text{ Against } \boldsymbol{H_1}: \boldsymbol{\beta_i \neq 0}, \mathbf{i = 1(1)p}$$

This is easy to conclude that the test reduces to ANOVA testing and the Minitab also shows output for this tests and further post-hoc tests related to the pairwise problem in a format of ANOVA table.

$$\boldsymbol{VIF = (1 - R_i^2)^{-1}}$$

If the VIF is equal to 1 there is no Multicollinearity among factors, but if the VIF is greater than 1, the predictors may be moderately correlated. The output above shows that the VIF for the Publication and Years factors are about 1.5, which indicates some correlation, but not enough to be overly concerned about. **A VIF between 5 and 10 indicates high correlation that may be problematic. And if the VIF goes above 10, you can assume that the regression coefficients are poorly estimated due to Multicollinearity.**

Regression analysis in Minitab

Regression Analysis: y versus x1 [Travel times (y) on Miles traveled(x_1)]

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	5.91197	5.91197	55.19	0.000
x1	1	5.91197	5.91197	55.19	0.000
Error	8	0.85703	0.10713		
Lack-of-Fit	7	0.83703	0.11958	5.98	0.305
Pure Error	1	0.02000	0.02000		
Total	9	6.76900			

Here in the ANOVA table, **p-value** related to source **Regression** suggests the significance of the **total regression model**. If the, p-value in the ANOVA table corresponding to regression is not significant then the model is not valid. The p-value corresponding to the variable x1, suggests that the how much appropriate is variable x1 in the model.

If the p-value related to Regression and p-values related to variables are not significant (>.05) and p-value associated with the error is significant then the model is not relevant in most of the cases.

Model Summary → *Standard Error of the Estimate*

S	R-sq	R-sq(adj)	R-sq(pred)
0.327306	87.34%	85.76%	81.05%

In model summary S is nothing but the estimated value of the standard error of the fitted value and the observed value of the dependent variable. If S is a high value, then on an average the observed points are quite far away from the regression model, which is not desired for predictability. This is an important statistic. As much low the S is, the fit of the model is better. R-sq is the coefficient of determination. The formula of the R-square has been discussed. This is a measure of variability of the dependent variable explained by the model. R-sq(adj) is the measure of same things just the sample size and the number of independent variables include in the regression model are under consideration. R-sq predicted is a measure of expected values of data point can be explained by the model when prediction is made based on the regression model using the sample data for the population data.

Coefficients

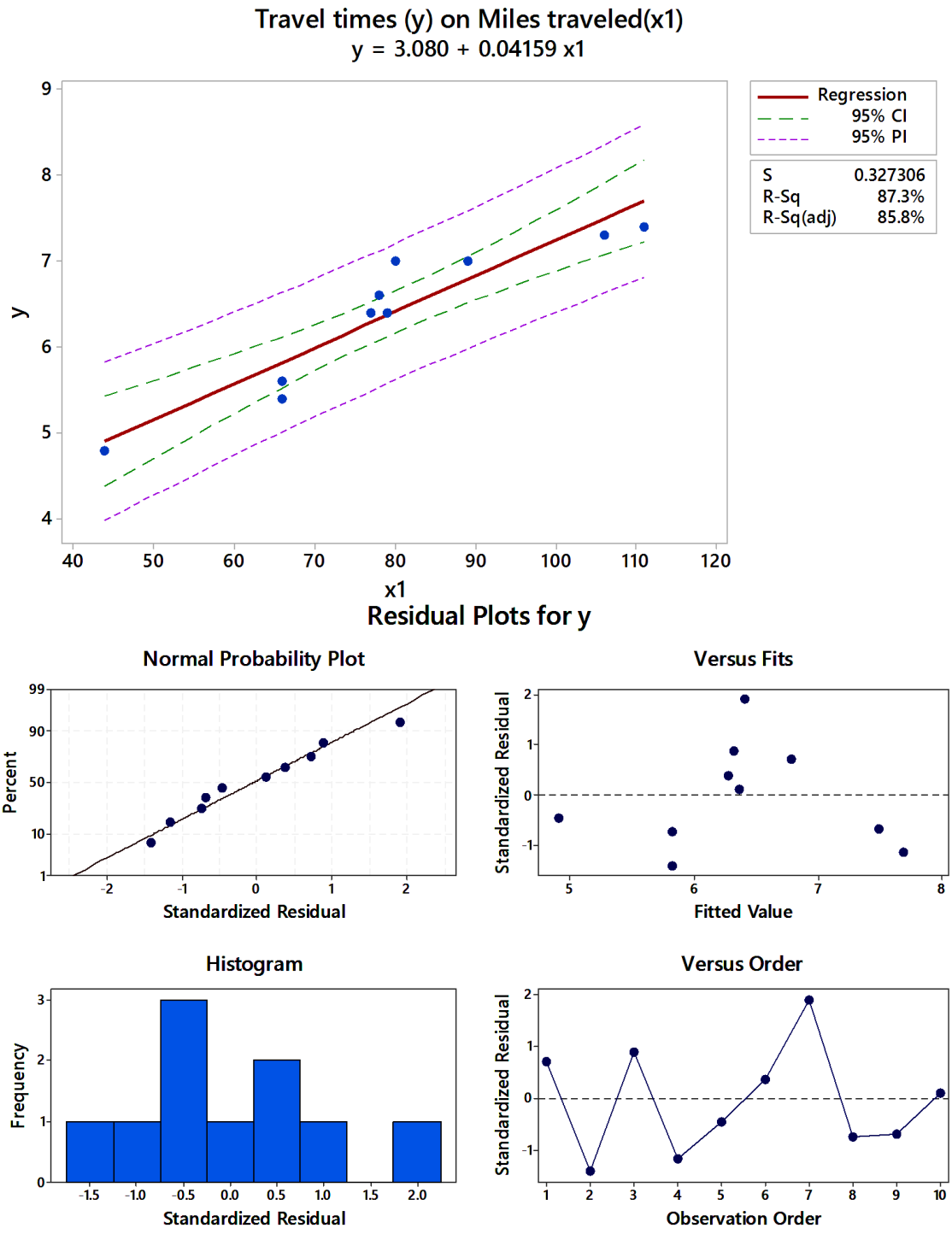
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	3.080	0.457	6.73	0.000	
x1	0.04159	0.00560	7.43	0.000	1.00

Interpretation of the coefficient of regression can be made in the following fashion that, on an average extra one mile travel with increase the delivery time by .4159 hour (this is important to remember that this is in the unit of dependent variable) or 25 minutes approximately.

Regression Equation

y = 3.080 + 0.04159 x1

Regression analysis in Minitab



Regression analysis in Minitab

Regression Analysis: y versus x2 [Travel times (y) on number of deliveries(x_2)]:

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	5.6851	5.6851	41.96	0.000
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

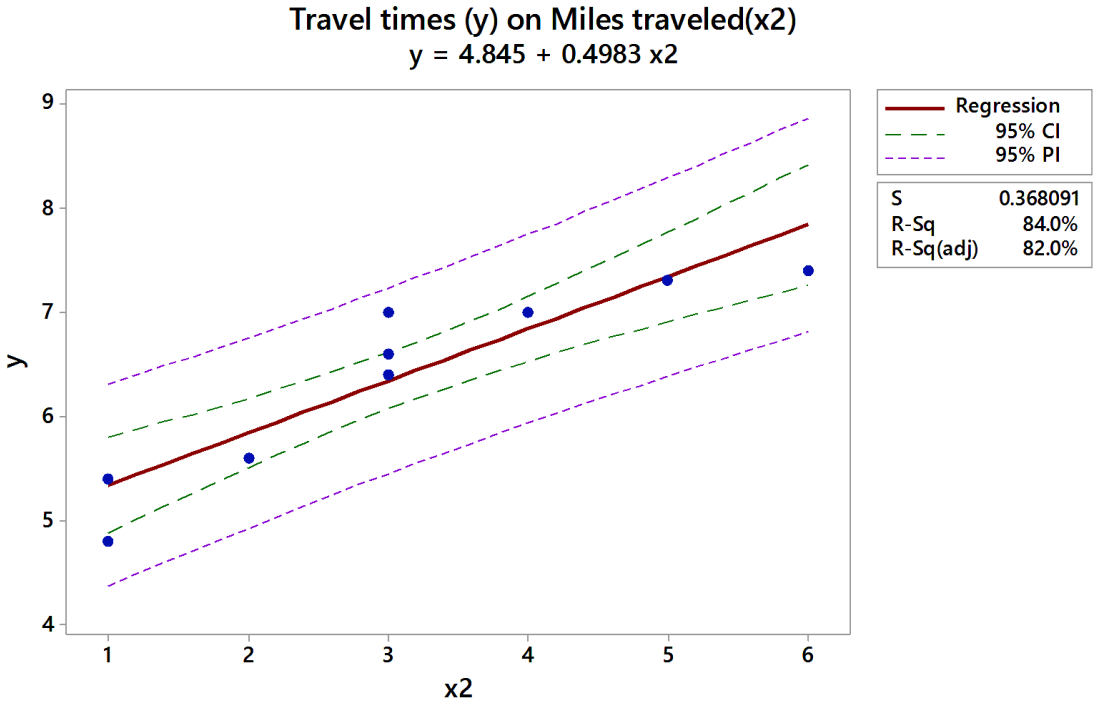
S	R-sq	R-sq(adj)	R-sq(pred)
0.368091	83.99%	81.99%	70.27%

Coefficients

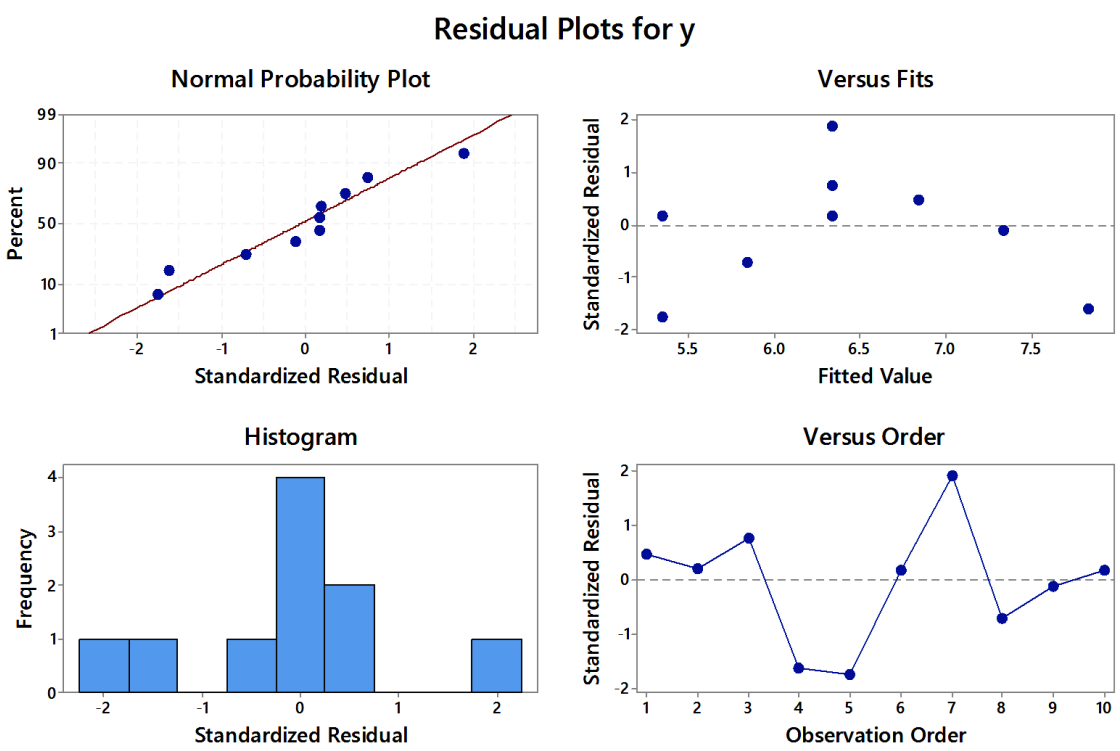
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	4.845	0.265	18.26	0.000	
x2	0.4983	0.0769	6.48	0.000	1.00

Regression Equation

y = 4.845 + 0.4983 x2



Regression analysis in Minitab



Regression Analysis: y versus x3

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	0.4833	0.4833	0.62	0.455
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%

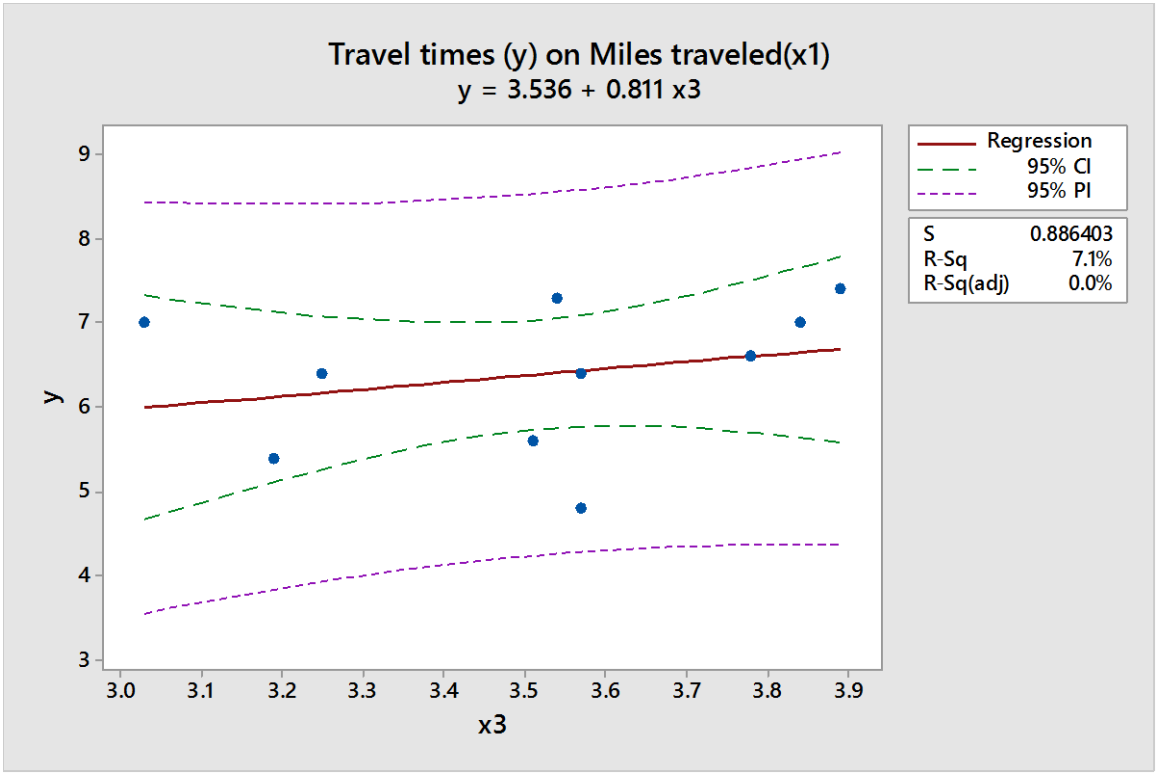
Coefficients

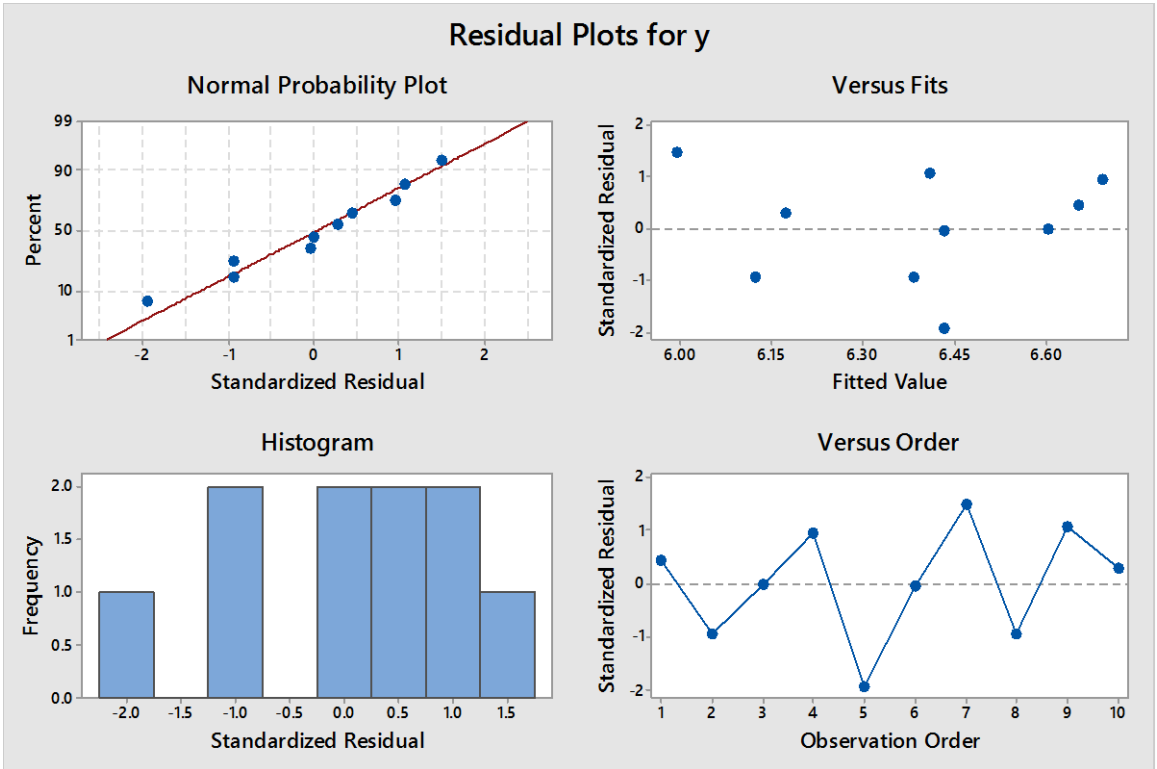
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	3.54	3.65	0.97	0.361	
x3	0.81	1.03	0.78	0.455	1.00

Regression Equation

y = 3.54 + 0.81 x3

Regression analysis in Minitab





Regression analysis in Minitab

Partial regression:

Regression Analysis: y versus x1, x2

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	5.94568	2.97284	25.28	0.001
x1	1	0.26061	0.26061	2.22	0.180
x2	1	0.03371	0.03371	0.29	0.609
Error	7	0.82332	0.11762		
Total	9	6.76900			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.342954	87.84%	84.36%	55.51%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	3.508	0.932	3.76	0.007	
x1	0.0309	0.0208	1.49	0.180	12.54
x2	0.136	0.254	0.54	0.609	12.54

Regression Equation

y = 3.508 + 0.0309 x1 + 0.136 x2

In this case, the p-value associated with the Regression seems to be significant but the value corresponding to the variable x1 and x2 are not significant. This calls for further investigation. In the end we will see that, this is not a suitable model, which calls for an important observation that although, the overall regression model may be significant in an analysis but this does not guarantees that the model will be valid without further investigation.

The model summary suggests, that the model is capable of explaining the 87.84% variation with coefficient of determination and 84.36% variation by considering the sample size and the number of independent variables in the model. The model also, is expected to estimate 55.1% percent predicted values for population data. S, the standard error of the estimate is quite high in this case in comparison with the simple linear regression model used in the first case. This is not and good indication. Another important observation can be made that the model is explaining the variation quite good for the sample data points but there is a sharp decrease when it comes to prediction making for the population data.

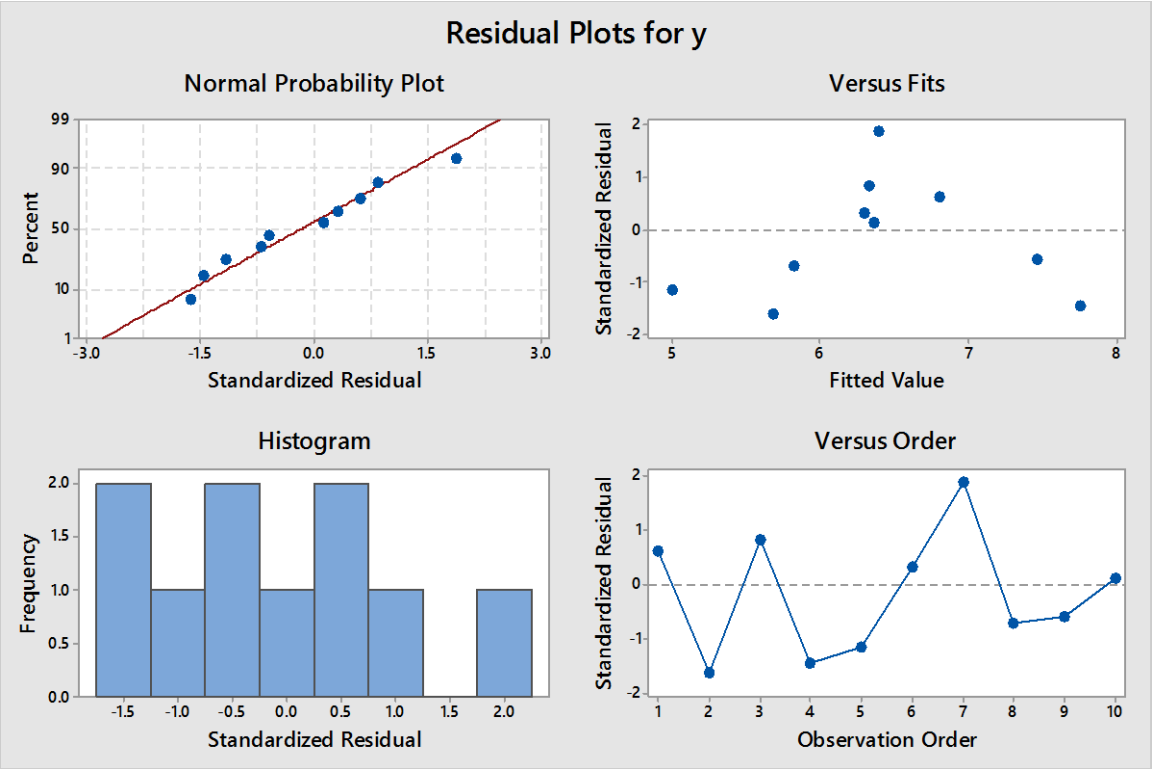
If we look at the coefficients of the model then it suggests that none of the variables used in the model is statistically significant. This has already been shown in the ANOVA table. The p-values corresponding to the x1 and x2 are in the coefficient section are as same as the p-values corresponding to them in the ANOVA table. Here one additional important factor is the VIF or variance inflation factor.

These two variables are statistically not significant because of co-linearity. This has been found earlier in the scatter plot and the correlation matrix. That has indicated this problem of potential co-linearity in this pair of variables. So, the VIF is way high then it is expected to be.

So, we should suspect the model because:

1. Scatter plots shows highly dependence between these independent variables
2. In correlation matrix the value of regression coefficient of variable x1 and x2 is high and the p-value is significant, which indicates existence of strong linear relation among them.
3. P-values corresponding to the variables are not significant in the ANOVA table and coefficient section.
4. The VIF corresponding to the variables are way higher than the expected values.
5. R-squared predicted is way lower than the R-squared predicted.

Regression analysis in Minitab



Regression Analysis: y versus x1, x3

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	5.93761	2.96881	25.00	0.001
x1	1	5.45429	5.45429	45.92	0.000
x3	1	0.02564	0.02564	0.22	0.656
Error	7	0.83139	0.11877		
Total	9	6.76900			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.344630	87.72%	84.21%	70.20%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	3.70	1.42	2.61	0.035	
x1	0.04260	0.00629	6.78	0.000	1.14
x3	-0.199	0.429	-0.46	0.656	1.14

Regression Equation

y = 3.70 + 0.04260 x1 - 0.199 x3

Here, for the ANOVA table this can be stated that the although the model and the independent variable x1 is significant but the variable x3 or the gas price is not significant in the model. So, the model must be analyzed further.

Regression analysis in Minitab

In model summary, this can be observed that the although the model explains and predicts the fair amount of variation in the dataset and the prediction for population does not drops very much but the estimated value of standard error of the regression is high.

In the multiple linear regression, the coefficient associated with a variable is interpreted in the following fashion:
The change in the independent variable which is under consideration causes the value of coefficient in dependent variable when the other independent variables in the model are held constant.

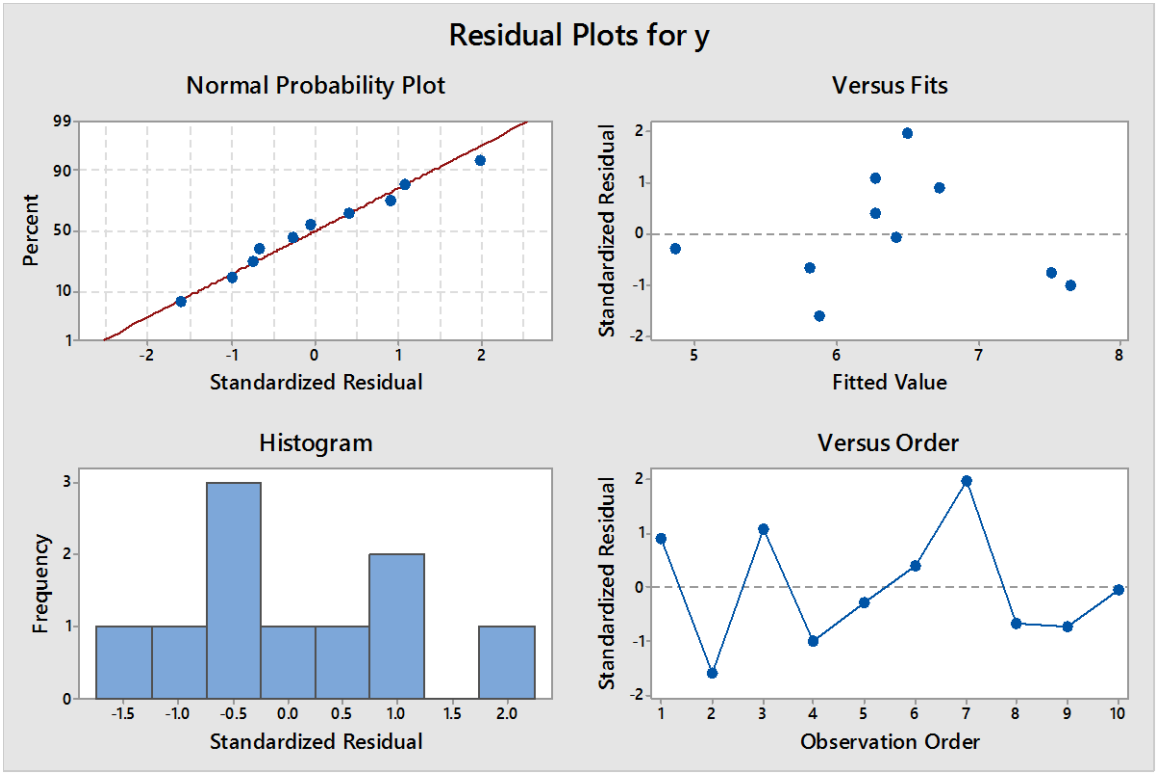
In this case, in the gas price (x3) holds constant then for each extra miles traveled the delivery time will increase by .042 hours or two and a half minutes.

For example, here in the regression equation given by the Minitab suggests that if the variable x1 held constant then one unit positive change in x3 (the price of gas) will reduce the delivery time by .199 hours or 12 minutes approximately. This is not fairly logical in real life, scenario.

In real life scenario, people drive slow to save fuel when the price of the fuel is higher than the normal rate. But in this case, the opposite relationship is been reflected in the regression equation. This regression suggests, when the price of the gas is higher by one unit the delivery reduces by 12 minutes which defies the basic logic. So, this model may not be adequate for predicting.

Reason for suspecting the model:

1. One of the independent variable is statistically insignificant.
2. The estimate of the standard error due to regression is high.
3. The interpretation of coefficient associated with the variable not statistically significant not makes sense in the real world scenario.



Regression analysis in Minitab

Regression Analysis: y versus x2, x3

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	6.0081	3.0040	27.63	0.000
x2	1	5.5248	5.5248	50.82	0.000
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%

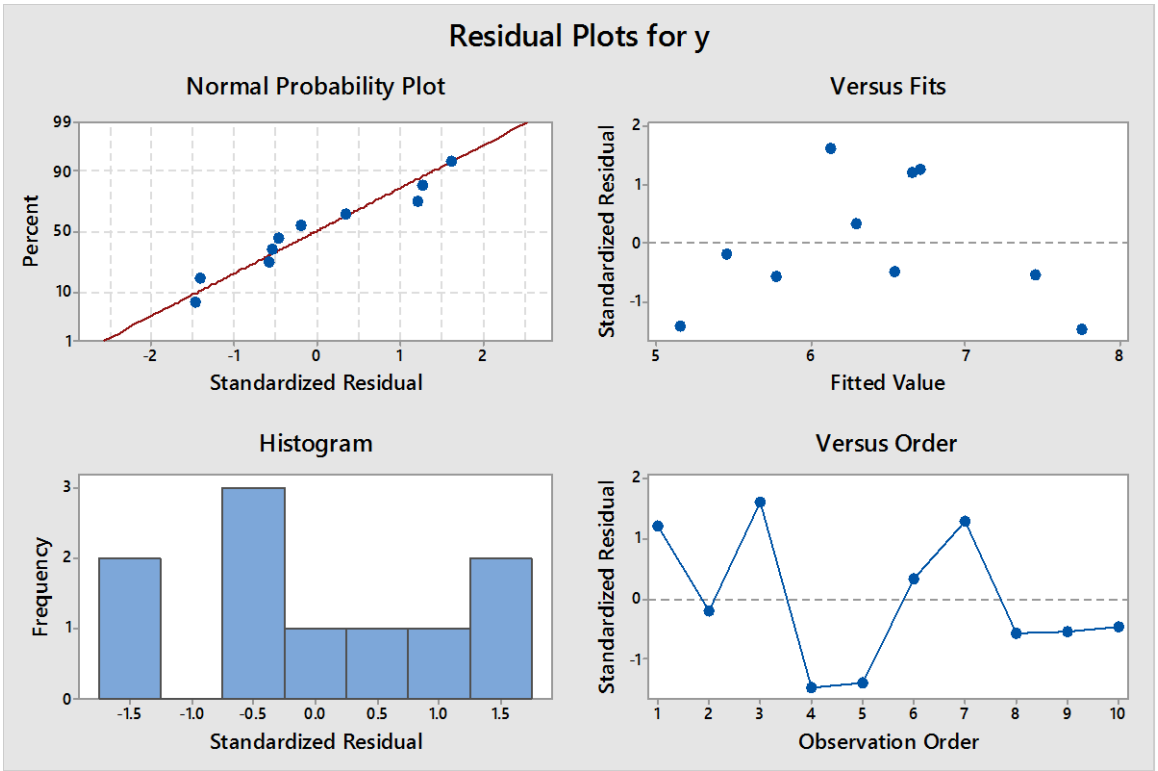
Coefficients

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

Regression Equation

y = 7.32 + 0.5665 x2 - 0.765 x3

In this case, the same problem occurs as the previous model and the result could be interpreted in the similar manner.



Regression analysis in Minitab

Regression Analysis: y versus x1, x2, x3

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	6.06847	2.02282	17.33	0.002
x1	1	0.06040	0.06040	0.52	0.499
x2	1	0.13086	0.13086	1.12	0.330
x3	1	0.12279	0.12279	1.05	0.345
Error	6	0.70053	0.11675		
Total	9	6.76900			

Model Summary

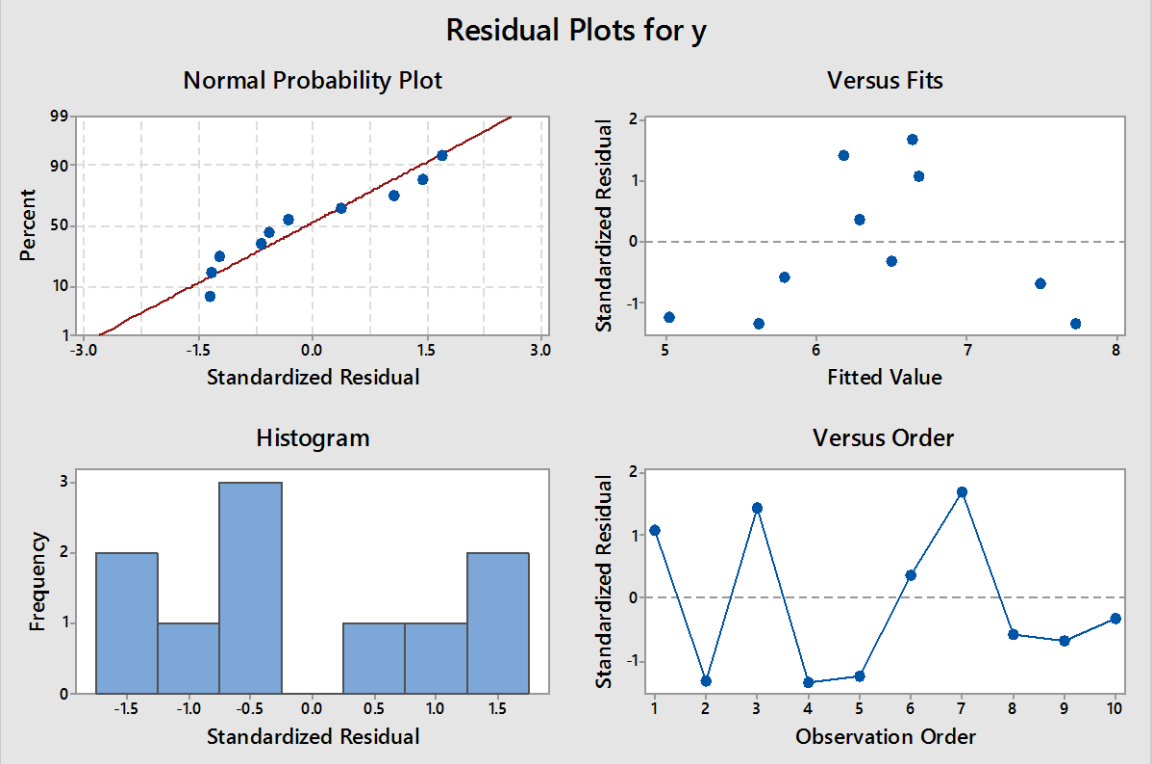
S	R-sq	R-sq(adj)	R-sq(pred)
0.341693	89.65%	84.48%	47.78%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	5.89	2.50	2.35	0.057	
x1	0.0176	0.0244	0.72	0.499	17.49
x2	0.342	0.323	1.06	0.330	20.46
x3	-0.557	0.543	-1.03	0.345	1.86

Regression Equation

y = 5.89 + 0.0176 x1 + 0.342 x2 - 0.557 x3



Summary of all models:

F	p-value	S	$R^2(adj)$	$R^2(pre)$	X1	X2	X3	VIF	Remark
49.77	<.001	.34320	84.42%	79.07%	X			1.00	Looks ok
41.96	<.001	.36809	81.99%	70.25%		X		1.00	Looks ok
0.62	.455	.88640	0.00%	0.00%			X	1.00	Not a significant variable
23.72	.001	.35624	83.47%	59.95%	X	X		11.59	Multicollinearity
22.63	.001	.38988	82.78%	68.11%		X	X	1.14	Negative coefficient
27.63	<.001	.32970	85.55%	71.76	X		X	1.33	Negative coefficient
16.99	.002	.34469	84.20%	57.49%	X	X	X	Below	Multicollinearity
					14.94	17.35	1.71		

Regression analysis in Minitab

Comparison among the models under study:

After, the analysis one of the major problem is to select the best model. So, one of the easy way to determine them is to eliminate them one by one.

1. From the results this can be easily stated that the 3rd model or model with only x3 will be eliminated due to higher p-value.
2. The model contenting x1 and x2 will be eliminated due to the higher value of VIF.
3. The model containing x1, x2, and x3 will also be eliminated because of the same reason and at the same time the p-values associated with the variables are not significant. R square predicted has dropped significantly from the r squared adjusted.
4. Among the left four model, model containing x1 and x3 are seem to be most preferred if we consider S and $R^2(adj)$. After this model, the model with the only variable x1 is the preferred choice.
5. But the whole scenario changes when we take $R^2(pre)$ and VIF under consideration. The $R^2(pre)$ of the model with the single variable x1 is way more than the $R^2(pre)$ of the other model. In fact the model has a problematic coefficient which does not have any real life significance and the VIF of the model is slightly higher than the simple linear regression with x1.
6. If all the measures are same then this is convention, to use the regression equation with less number of variable because this reduces the inherent chance of higher variability due to higher number of variables in the model.

Best Subsets Regression: y versus x1, x2, x3

Response is y

vars	R-sq	R-Sq (adj)	R-Sq (pred)	Mallows CP	S	X1	X2	X3
1	87.3	85.8	81.1	1.9	.32731	X		
1	84.0	82.0	70.3	3.3	.36809		X	
2	88.8	85.5	71.8	2.5	.32970			X
2	87.8	84.4	55.5	3.1	.34295	X	X	
3	89.7	84.5	47.8	4.0	.34169	X	X	X

1. Look at R-sq (adj). Which are the highest values?
2. Look at R-Sq (pred). Which are the highest values?
3. Examine the difference between R-Sq (adj) and R-Sq (pred). A large drop-off indicates over fitting; too many variables in the model.
4. Look at Mallows Cp. Look for one that is low and approximately equals the number of predictors plus the constant (1).
5. Using all the information above choose the best regression model.

Stepwise regression:

Stepwise regression is another tool of Minitab which behaves in a similar manner as the best subset regression. In this method Minitab adds a variable at a time and provides the best regression model.

Stepwise Selection of Terms

Candidate terms: x1, x2, x3

	------Step 1-----	
	Coef	P
Constant	3.080	
x1	0.04159	0.000
S	0.327306	
R-sq	87.34%	
R-sq(adj)	85.76%	
R-sq(pred)	81.05%	
Mallows' Cp	1.9	

α to enter = 0.05,
α to remove = 0.05

Regression analysis in Minitab

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	5.9120	5.9120	55.19	0.000
x1	1	5.9120	5.9120	55.19	0.000
Error	8	0.8570	0.1071		
Total	9	6.7690			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.327306	87.34%	85.76%	81.05%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	3.080	0.457	6.73	0.000	
x1	0.04159	0.00560	7.43	0.000	1.00

Regression Equation

y = 3.080 + 0.04159 x1

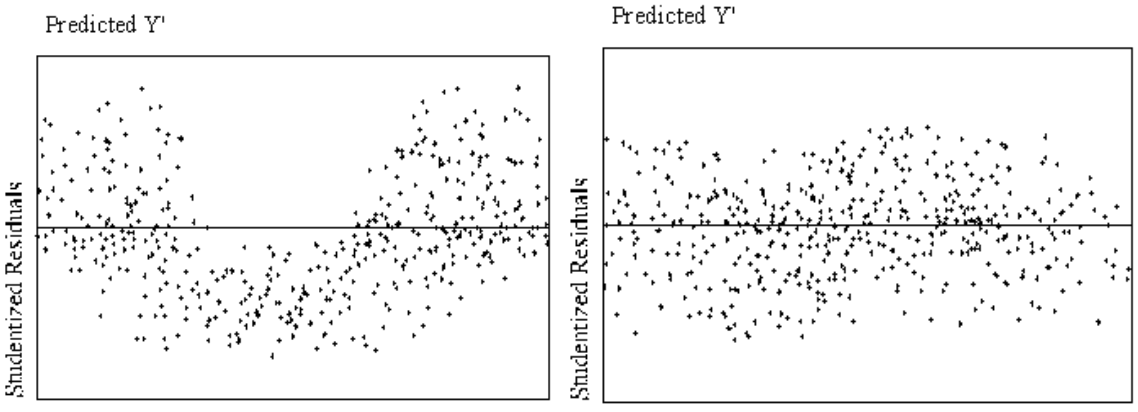
Basic residual analysis and validation of assumption of regression:

Assumptions of Regression model:

- The sample is representative of the population for the inference prediction.
- The error is a random variable with a mean of zero conditional on the explanatory variables.
- The independent variables are measured with no error. (Note: If this is not so, modeling may be done instead using errors-in-variables model techniques).
- The predictors are linearly independent, i.e. it is not possible to express any predictor as a linear combination of the others.
- The errors are uncorrelated, that is, the variance–covariance matrix of the errors is diagonal, and each non-zero element is the variance of the error.
- The variance of the error is constant across observations (homoscedasticity). If not, weighted least squares or other methods might instead be used.

The residual plot provides a visual tool to check the weather the linear relationship among the dependent variable and independent variable holds true in the regression model or not. If the relationship of the dependent variable is not linear then the estimator will under estimate the values of parameters of the regression model. This under-estimation carries two risks: increased chance of a Type II error for that IV, and in the case of multiple regression, an increased risk of Type I errors (over-estimation) for other IVs that share variance with that IV. A preferable method of detection is examination of residual plots (Plots of the standardized residuals as a function of standardized predicted values).

Example of curvilinear and linear relationships with standardized residuals by standardized predicted values.



Curvilinear Relationship Linear Relationship

Regression analysis in Minitab