Example (1)

A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)

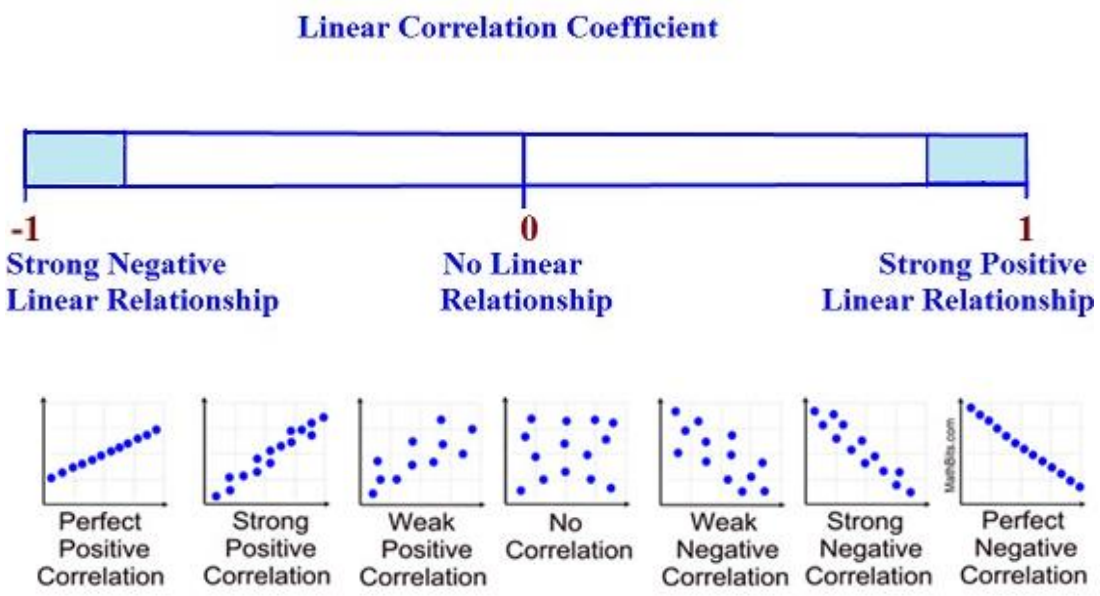
A random sample of 10 houses is selected

- Dependent variable (Y) = house price in $1000s
- Independent variable (X) = square feet

House Price in $1000s (Y)	Square Feet (X)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Coefficient of correlation:

A measure of the strength of the linear relationship between two variables



Value of Correlation Coefficient

Value of Correlation Coefficient	Degree of Correlation
+1	Perfect Positive Correlation
-1	Perfect Negative Correlation
$\pm 0.75 \le r < \pm 1$	High degree Positive (Negative) Correlation
$\pm 0.5 \le r < \pm 0.75$	Moderate degree Positive (Negative) Correlation
$0 < r < \pm 0.5$	Weak degree of Positive (Negative) Correlation
Zero	No Correlation (No relationship between X&Y)

$$r_{xy} = \frac{n\sum XY - \sum X \sum Y}{\sqrt{n\sum X^2 - (\sum X)^2}\sqrt{n\sum Y^2 - (\sum Y)^2}}$$

$$r_{xy} = \frac{10(5085975) - (2865 * 17150)}{\sqrt{10(30983750) - 17150^2}\sqrt{10(853423) - 2865^2}} = 0.76$$

There is a (positive & strong) relationship between y and x

Computations needed for the r_{xy}, $\hat{Y}$, b_0 , b_1

House Price in $1000s (Y)	Square Feet (X)	XY	X^2	Y^2
245	1400	343000	1960000	60025
312	1600	499200	2560000	97344
279	1700	474300	2890000	77841
308	1875	577500	3515625	94864
199	1100	218900	1210000	39601
219	1550	339450	2402500	47961
405	2350	951750	5522500	164025
324	2450	793800	6002500	104976
319	1425	454575	2030625	101761
255	1700	433500	2890000	65025
2865	**17150**	**5085975**	**30983750**	**853423**

$$\boldsymbol{n = 10}, \boldsymbol{\sum Y = 2865}, \boldsymbol{\sum X = 17150}, \boldsymbol{\sum XY = 5085975},$$

$$\boldsymbol{\sum X^2 = 30983750}, \quad \boldsymbol{\sum Y^2 = 853423}$$

Interpretation of $\mathbf{b_o}$ and b1

$$b_1 = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} = \frac{10(\mathbf{5085975}) - (\mathbf{2865} * \mathbf{17150})}{10(\mathbf{30983750}) - (\mathbf{17150})^2}$$

$$= 0.10977$$

b_1 estimates the change in the mean value of Y as a result of a one-unit increase in X.

Here, b_1 = 0.10977 tells us that the mean value of a house increases by .10977($1000) = $109.77, on average, for each additional one square foot of size

$$\boxed{b_0 = \bar{Y} - b_1\bar{X} = 286.5 - 0.10977(1715) = 98.24}$$

b_0 is the estimated mean value of Y when the value of X is zero (if X = 0 is in the range of observed X values)

Because a house cannot have a square footage of 0, b_0 has no practical application

Simple Linear Regression Model

$$\hat{Y} = 0.98.24 + 0.10977X$$

Making Predictions

Predict the price for a house with 2000 square feet:

$$\hat{Y} = 0.98.24 + 0.10977(2000) = 317.85$$

Measures of Variation

The most useful way for the test the significance of the regression is use the "Analysis of Variance" which separates the total variation of the dependent variable into two independent parts:

- Variation attributable to the relationship between X and Y
  (Explained variation).

- Variation in Y attributable to factors other than X
  "The variation error" (Unexplained variation).
  Total variation =

  Explained variation (Regression) + unexplained variation (Error)

$$(Y_i - \bar{Y}) = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$$

Total sum of squares = Regression sum of squares + Error sum of squares

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2$$

SST = SSR + SSE

$\bar{Y}$= Mean value of the dependent variable

Y_i =Observed value of the dependent variable

$\hat{Y}_i$= Predicted value of Y for the given X_i value

Computations needed for SSE " e_i^2 " /SST/SSR / SE

Y	X	$\hat{Y} = 0.98.24 + 0.10977X$	$e_i = Y_i - \hat{Y}_i$	e_i^2	$(Y - \bar{Y})^2$	$(X - \bar{X})^2$
1400	245	251.9232	-6.92316	47.9302	1722.25	99225
1600	312	273.8767	38.12329	1453.39	650.25	13225
1700	279	284.8535	-5.85348	34.2633	56.25	225
1875	308	304.0628	3.937162	15.5012	462.25	25600
1100	199	218.9928	-19.9928	399.714	7656.25	378225
1550	219	268.3883	-49.3883	2439.21	4556.25	27225
2350	405	356.2025	48.79749	2381.19	14042.25	403225
2450	324	367.1793	-43.1793	1864.45	1406.25	540225
1425	319	254.6674	64.33264	4138.69	1056.25	84100
1700	255	284.8535	-29.8535	891.231	992.25	225
17150	2865			13666	32600.5	1571500

$$SST = (Y - \bar{Y})^2 = 32600.5 \quad , \qquad SSE = e^2 = (Y - \hat{Y})^2 = 13666$$

$$SSR = SST - SSE = 32600.5 - 13666 = 18934.5$$

The Coefficient of Determination (r^2)

The coefficient of determination (r^2) is used to measure the quality of the regression model. It is considered as a measure for the percentage of variation in the values of the dependent variable (Y) that can be explained by the variation in the independent variable(X).
r^2: value varies from 0 to 1. $(0 \leq r^2 \leq 1)$

$r^2 = 1$ all of the variability in Y is explained when X is known (The sample data points all lie on the fitted regression line.

$r^2 = 0$ none of variability in Y is explained by X (The scatter diagram suggests no linear)

$$r^2 = \frac{SSR}{SST} = \frac{1893.5}{32600.5} = 0.58$$

Or

$$r^2 = 1 - \left[\frac{SSE}{SST}\right] = 1 - \left[\frac{13665.6}{32600.5}\right] = 0.58$$

The Standard error of estimate:

The standard error of estimate a measure of amount by the actual Y values differs from the estimated, or $\hat{Y}$, value.

(Measures the spread of the data points about the fitted line in the Y direction)

The standard error of estimate is denoted by $S_{Y.X}$ *or* $\hat{\sigma}$and is given by**:**

$$S_{Y.X} = \hat{\sigma} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum\left(Y_i - \hat{Y}_i\right)^2}{n-2}} = \sqrt{\frac{13665.565}{10-2}} = 41.33$$

Inferences about the Slope

The variance & Standard error of estimators

The standard error of the regression slope coefficient (b1) is estimated by:

$$se\ (b_1) = \sqrt{Var\ (b_1)} = \frac{S_{Y.X}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}} = \frac{41.33}{\sqrt{1571500}} = 0.03297$$

Tests for Significance:

Tests for statistical significance tell us what the probability that the relationship which found is only due to random chance. These tests tell us about the probability of making an error if we assume that a relationship exists.

Similarly, the tests for significance of regression are a test to determine whether there is a linear relationship between the dependent variable and a subset of explanatory variables.

There are many ways to test the significance of the regression coefficient: (t- test , Analyses of Variance)

t- Test:

T-test can be applied as a significance test through the following steps:-

1- State the null and alternative hypothesis

- **Null hypothesis**: $H_0 : \beta_1 = 0$ no linear relationship (the slope is zero)

Those mean the independent variable has no statistically significant effect on the dependent variable.

- **Alternative hypothesis**: $H_1 : \beta_1 \neq 0$ linear relationship does exist (the slope is not zero)

2- Calculate the test statistic " t " According to the formula :

$$t_c = \frac{b_1 \_ \beta_1}{Se(b_1)} = \frac{b_1}{Se(b_1)} = \frac{0.10977}{0.03297} = 3.329$$

3- Selected the level of significance (α). (0.05)

4- **Assign the tabulated value (Critical value) :**

$$t_{(\alpha/2, n-2)} = t_{(0.05/2, 10-2)} = 2.306$$

5- Determine the rejection region (Decision rule) and make a decision :

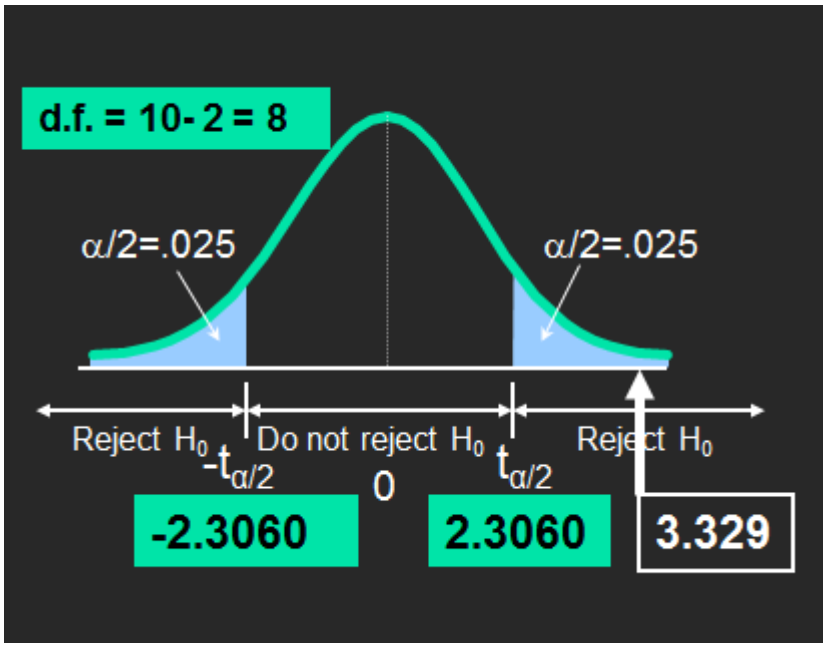
We will compare the computed value with critical value.

Reject the null hypothesis at the α significance level if

$$t_c > t_{(\alpha/2,n-2)} \quad or \quad -t_c < -t_{(\alpha/2,n-2)}$$

By other mean: $|t_c| > t_{(\alpha/2,n-2)}$

Decision: Reject H_0 There is sufficient evidence that square footage affects house price



F Test for Significance

Testing a hypothesis for a population slope, β_1 , using the F test
Step (1): State the null and alternate hypotheses:

$H_0 : \beta_1 = 0$ No linear relationship (the slope is zero)

$H_1 : \beta_1 \neq 0$ Linear relationship does exist (the slope is not zero)

Step (2): Select the level of significance (α).

Step (3): The critical value:

The critical value is read from the F - table and determined by:

- Level of significance (α). (0.05)
- Degrees of freedom

$$df_{SSR} = K$$

$$df_{SSE} = n - k - 1$$

The critical value

$$\mathbf{F_{(\alpha\ ,\ K,n-k-1)} = F_{(0.05\ ,1,10-1-1)} = F_{(0.05\ ,1,18)} = 5.32}$$

Step (4): The test statistic

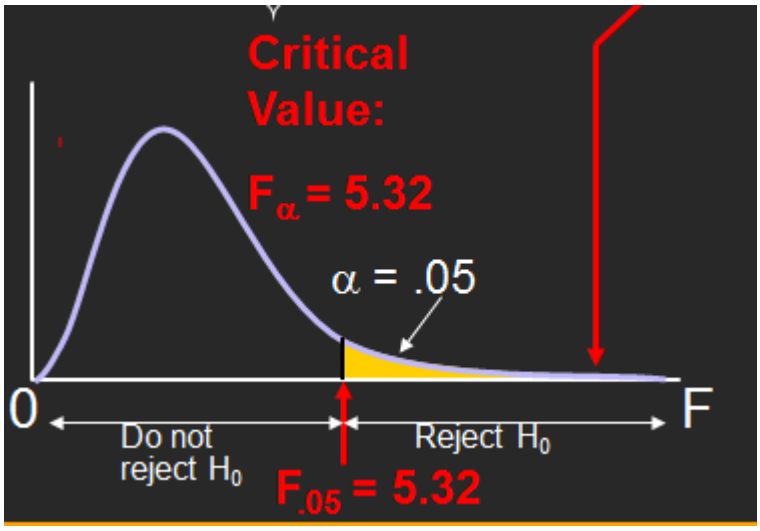
The F test statistic is found by dividing the mean of explained sum of squares (MSR) by the mean unexplained sum of squares (MSE)

$$F_{Stat} = \frac{MSR}{MSE} = \frac{\frac{SSR}{K}}{\frac{SSE}{n-k-1}} =$$

$$\frac{1834.5/1}{13665.6/\ (10-1-1)} = 11.08$$

Step (5) : Formulate the decision Rule and make a decision

Reject H_o If $\mathbf{F_c > F_{(\alpha,1,n-k-1)}}$



Decision:

Reject H0 at $\alpha = 0.05$

There is sufficient evidence that house size affects selling price

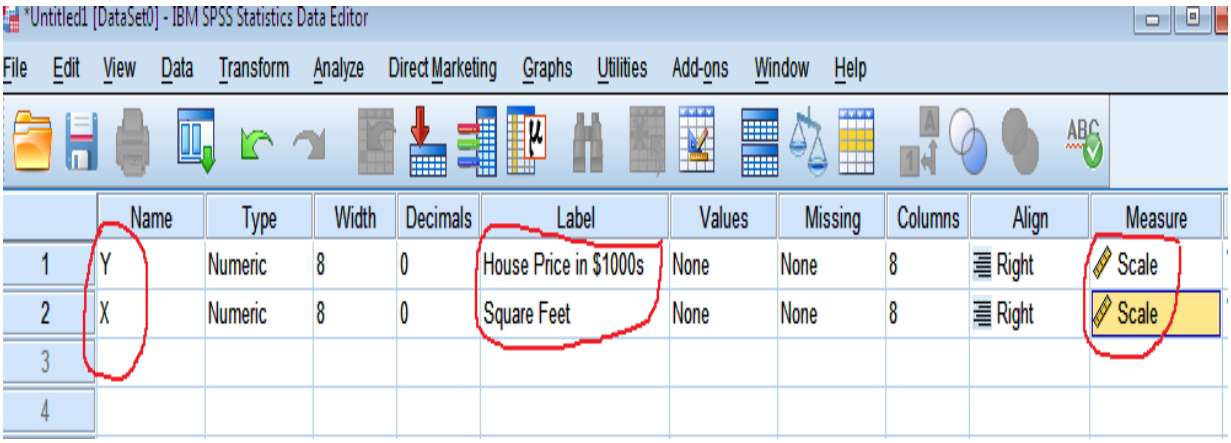
ANOVA TABLE

It is convenient to summarize the calculation of the F statistic in (ANOVA Table)

Source of variation (S.V)	**Sum of Squares (S.S)**	**Degrees of freedom (df)**	**Mean Squares (MS)**	**F- ratio**
Regression	SSR = 18934.5	K = 1	MSR =SSR/k = 18934.5	F = MSR / MSE = 11.08
Error	SSE = 13665.6	n-k-1= 10-1-1=8	MSE = SSE/n-k-1 = 13665.6/8	
Total variation	SST= 32600.5	n-1= 10-1=9		

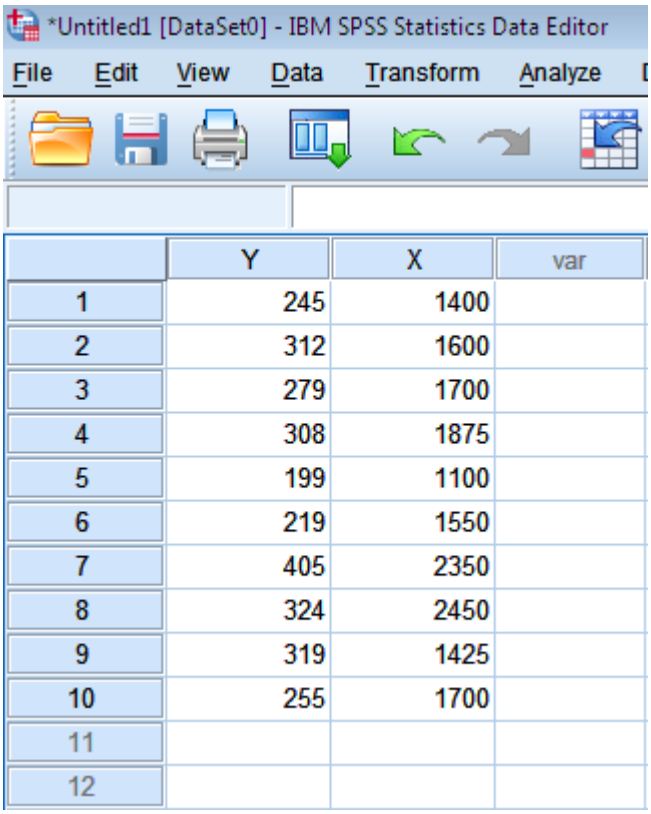
Using SPSS Data Analysis

Step (1) : Definition of variables



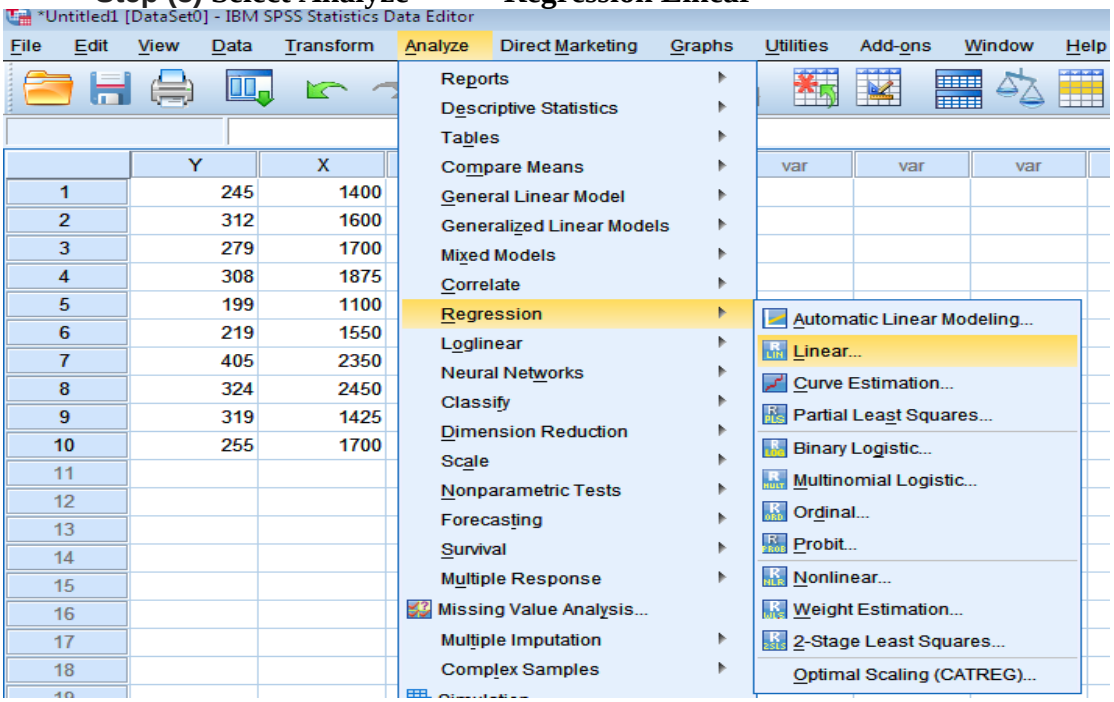
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	Y	Numeric	8	0	House Price in $1000s	None	None	8	Right	Scale
2	X	Numeric	8	0	Square Feet	None	None	8	Right	Scale
3										
4										

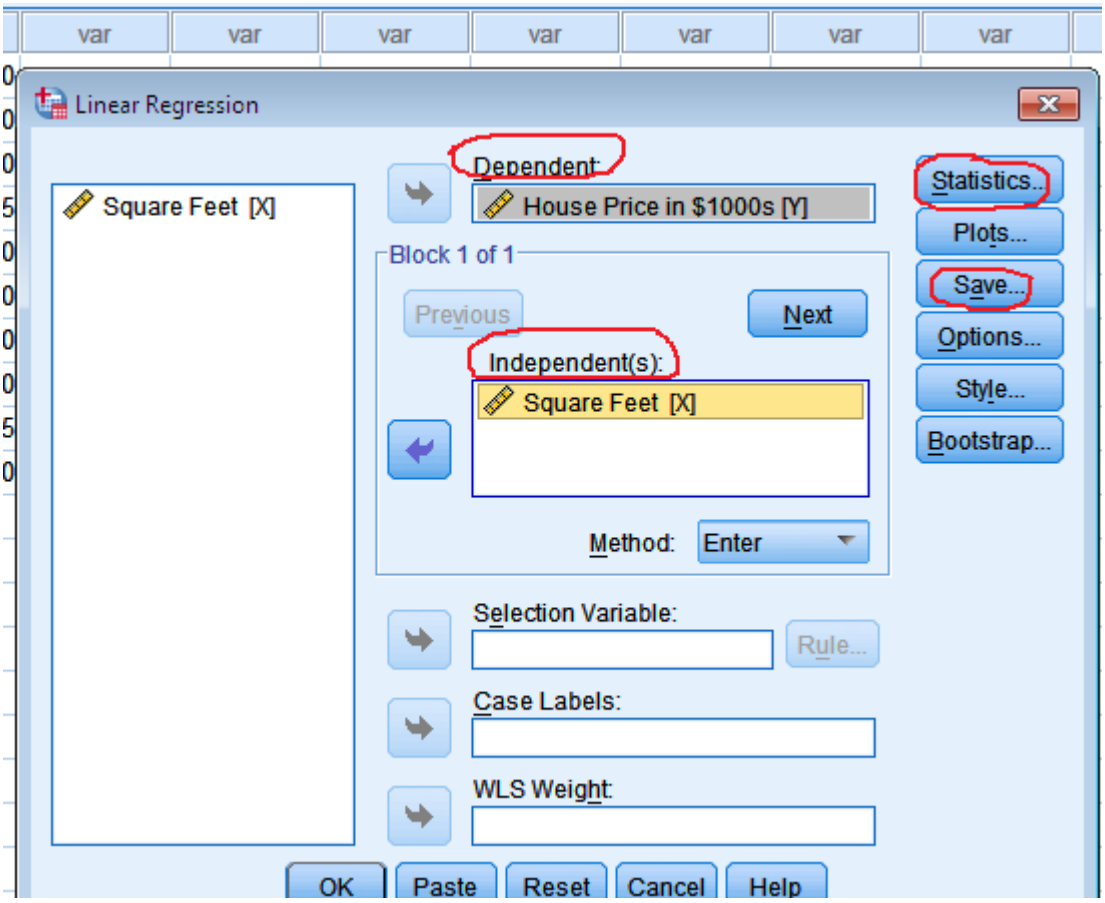
Step (2) : Data entry

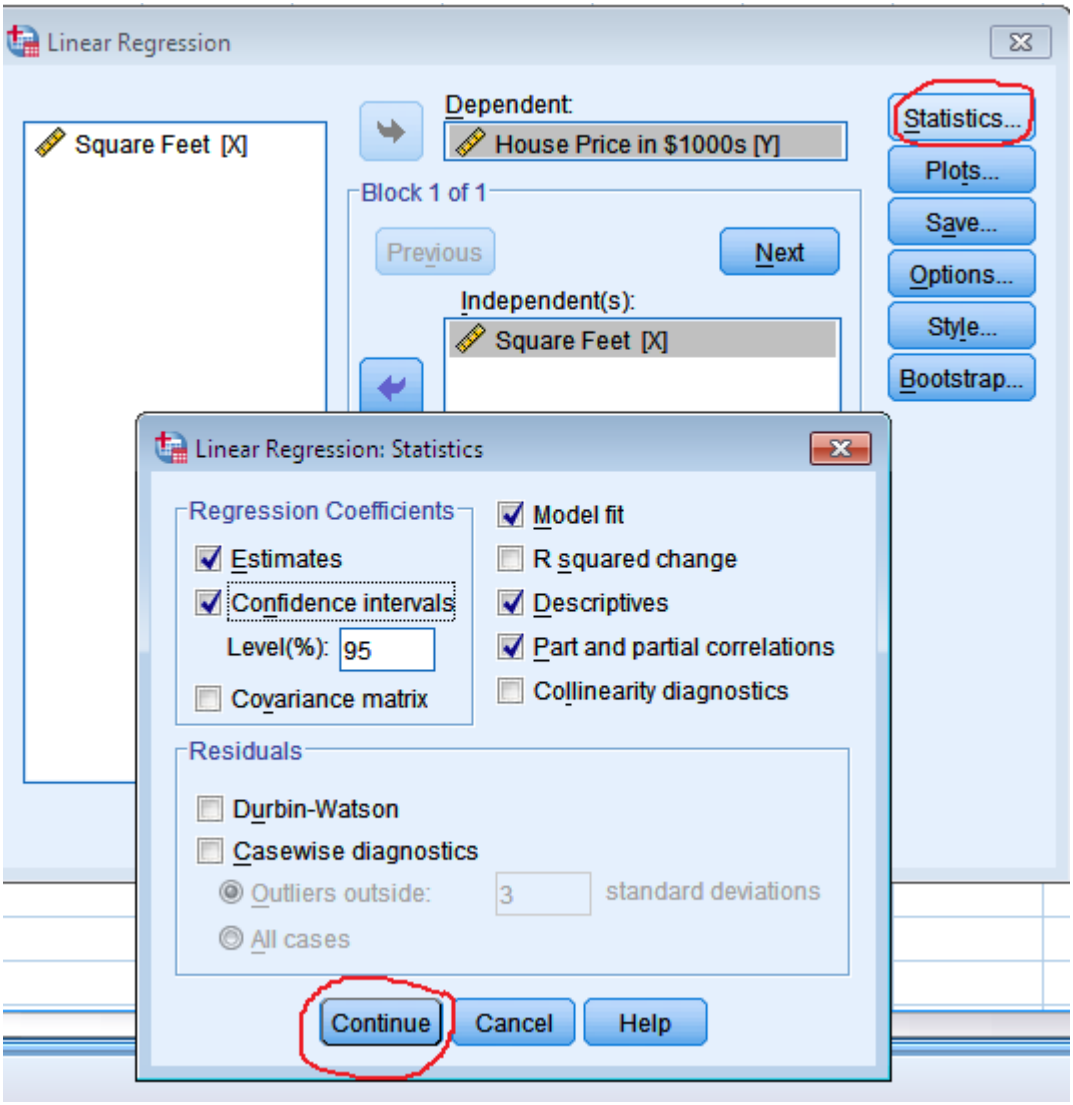


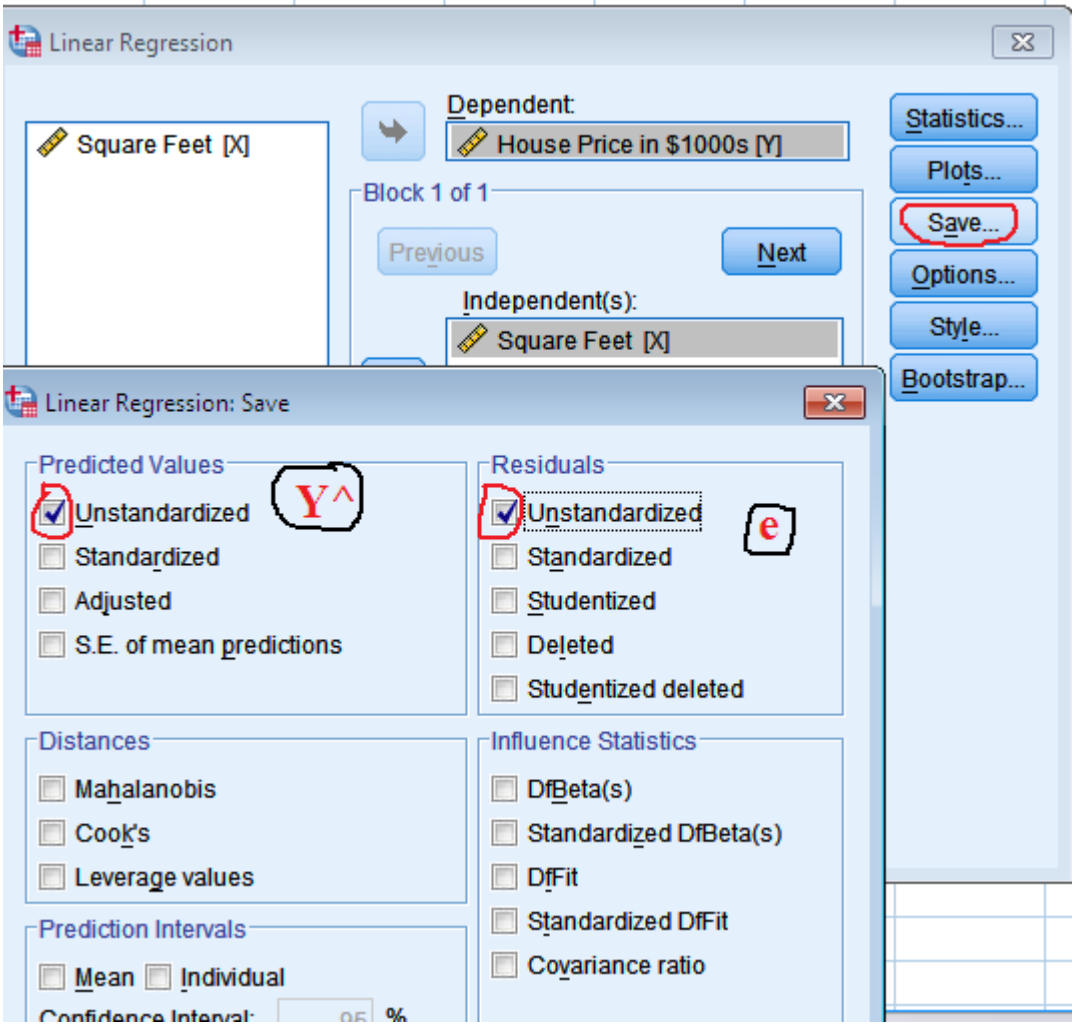
	Y	X	var
1	245	1400	
2	312	1600	
3	279	1700	
4	308	1875	
5	199	1100	
6	219	1550	
7	405	2350	
8	324	2450	
9	319	1425	
10	255	1700	
11			
12			

Step (3) Select Analyze >>> Regression Linear






Linear Regression
Square Feet [X]
Dependent:
House Price in $1000s [Y]
Block 1 of 1
Previous
Next
Independent(s):
Square Feet [X]
Statistics...
Plots...
Save...
Options...
Style...
Bootstrap...
Linear Regression: Statistics
Regression Coefficients
Estimates
Confidence intervals
Level(%): 95
Covariance matrix
Model fit
R squared change
Descriptives
Part and partial correlations
Collinearity diagnostics
Residuals
Durbin-Watson
Casewise diagnostics
Outliers outside: 3 standard deviations
All cases
Continue
Cancel
Help


Linear Regression
Square Feet [X]
Dependent:
House Price in $1000s [Y]
Block 1 of 1
Previous
Next
Independent(s):
Square Feet [X]
Statistics...
Plots...
Save...
Options...
Style...
Bootstrap...
Linear Regression: Save
Predicted Values
Unstandardized
Y^
Standardized
Adjusted
S.E. of mean predictions
Residuals
Unstandardized
e
Standardized
Studentized
Deleted
Studentized deleted
Distances
Mahalanobis
Cook's
Leverage values
Influence Statistics
DfBeta(s)
Standardized DfBeta(s)
DfFit
Standardized DfFit
Covariance ratio
Prediction Intervals
Mean
Individual
%

Output

Regression

[DataSet0]

Descriptive Statistics

	Mean	Std. Deviation	N
House Price in $1000s	286.50	60.185	10
Square Feet	1715.00	417.865	10

Correlations

		House Price in $1000s	Square Feet
Pearson Correlation	House Price in $1000s	1.000	.762
	Square Feet	.762	1.000
Sig. (1-tailed)	House Price in $1000s	.	.005
	Square Feet	.005	.
N	House Price in $1000s	10	10
	Square Feet	10	10

r

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Square Feet[b]	.	Enter

a. Dependent Variable: House Price in $1000s

b. All requested variables entered.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.762[a]	.581	.528	41.330

a. Predictors: (Constant), Square Feet

b. Dependent Variable: House Price in $1000s

r2

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	18934.935	1	18934.935	11.085	.010[b]
	Residual	13665.565	8	1708.196		
	Total	32600.500	9			

a. Dependent Variable: House Price in $1000s

b. Predictors: (Constant), Square Feet

F

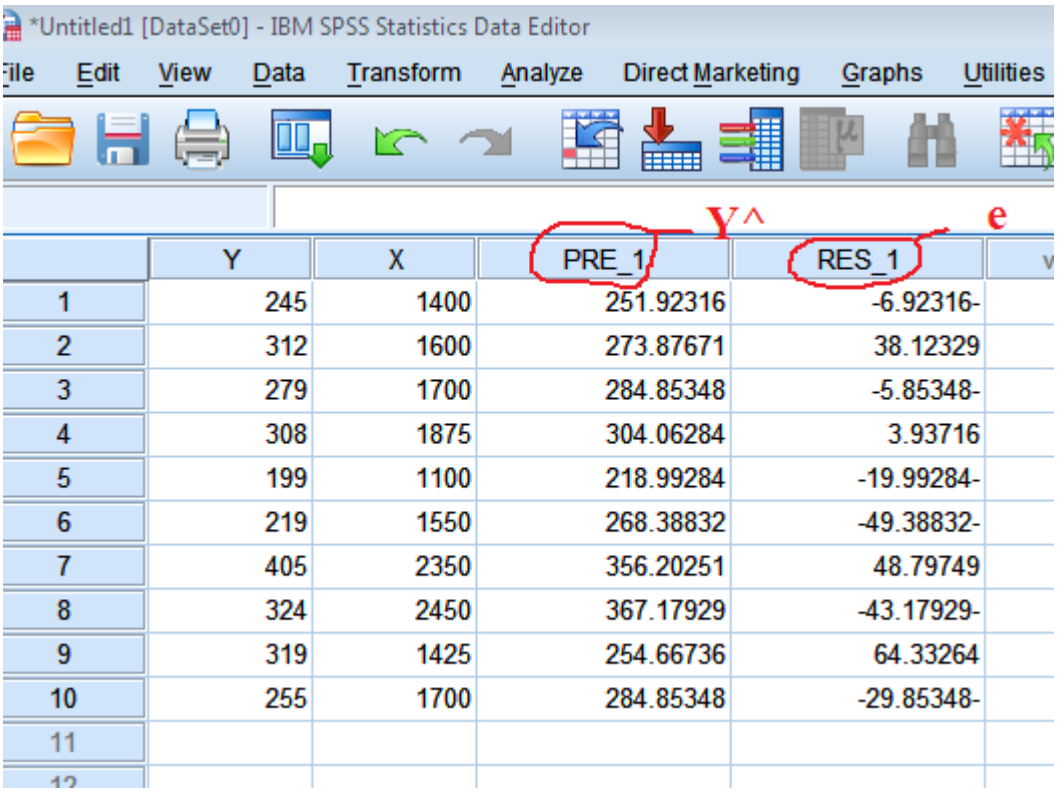
b0 , b1

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	98.248	58.033		1.693	.129	-35.577-	232.074
	Square Feet	.110	.033	.762	3.329	.010	.034	.186

a. Dependent Variable: House Price in $1000s

r



	Y	X	PRE_1	RES_1	v
1	245	1400	251.92316	-6.92316-	
2	312	1600	273.87671	38.12329	
3	279	1700	284.85348	-5.85348-	
4	308	1875	304.06284	3.93716	
5	199	1100	218.99284	-19.99284-	
6	219	1550	268.38832	-49.38832-	
7	405	2350	356.20251	48.79749	
8	324	2450	367.17929	-43.17929-	
9	319	1425	254.66736	64.33264	
10	255	1700	284.85348	-29.85348-	
11					