Statistics for Management and Economics

OPRE 6301 (Statistics and Data Analysis)
by
Ayşegül Toptal Bilhan

Chapter 16
(Simple Linear Regression and Correlation)

Simple Linear Regression

- Managerial decisions often are based on the relationship between two or more variables.
- Regression analysis can be used to develop an equation showing how the variables are related.
- The variable being predicted is called the dependent variable and is denoted by y.
- The variables being used to predict the value of the dependent variable are called the independent variables and are denoted by x.

Simple Linear Regression (Cont'd)

- Simple linear regression involves one independent variable and one dependent variable.
- The relationship between the two variables is approximated by a straight line.
- Regression analysis involving two or more independent variables is called multiple regression.

Simple Linear Regression Model

- The equation that describes how y is related to x and an error term is called the regression model.
- The simple linear regression model is:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where:

β_0 and β_1 are called parameters of the model,

ε is a random variable called the error term.

Simple Linear Regression Equation

The simple linear regression equation is:

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

- Graph of the regression equation is a straight line.
- β_0 is the *y* intercept of the regression line.
- β_1 is the slope of the regression line.
- $E(y)$ is the expected value of *y* for a given *x* value.

Estimated Single Linear Regression Equation

The estimated simple linear regression equation

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

- The graph is called the estimated regression line.
- b_0 is the *y* intercept of the line.
- b_1 is the slope of the line.
- $\hat{y}$ is the estimated value of *y* for a given *x* value.

Least Squares Method

- Least Squares Criterion

$$\min \sum (y_i - \hat{y}_i)^2$$

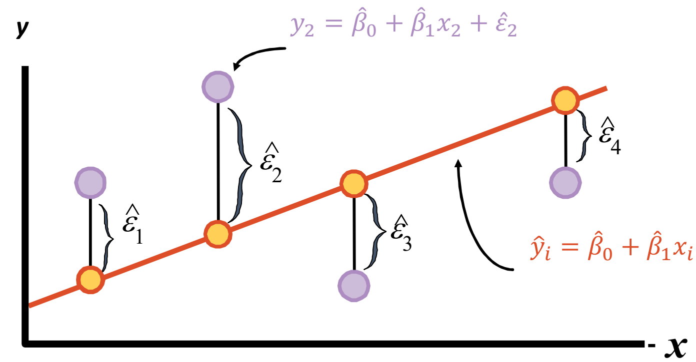
where:

y_i = <u>observed</u> value of the dependent variable for the *i* th observation

$\hat{y}_i$ = <u>estimated</u> value of the dependent variable for the *i* th observation

Least Squares Method (Graphically)

LS minimizes $\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \hat{\varepsilon}_1^2 + \hat{\varepsilon}_2^2 + \hat{\varepsilon}_3^2 + \hat{\varepsilon}_4^2$



Least Squares Method (Cont'd)

- Slope for the Estimated Regression Equation

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

- *y*-Intercept for the Estimated Regression Equation

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

where:

x_i = value of independent variable for *i*th observation

y_i = value of dependent variable for *i*th observation

$\bar{x}$ = mean value for independent variable

$\bar{y}$ = mean value for dependent variable

Least Squares Method (Cont'd)

Example: Reed Auto Sales

Reed Auto periodically has a special week-long sale. As part of the advertising campaign, Reed runs one or more television commercials during the weekend preceding the sale. Data from a sample of 5 previous sales are shown below.

Number of TV Ads (x)	Number of Cars Sold (y)
1	14
3	24
2	18
1	17
3	27
$\sum x = 10$	$\sum 100$
$\bar{x} = 2$	$\bar{y} = 20$

Least Squares Method (Cont'd)

Example (Cont'd):

Slope for the estimated regression equation:

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{20}{4} = 5$$

y-Intercept for the estimated regression equation:

$$b_0 = \bar{y} - b_1\bar{x} = 20 - 5\times 2 = 10$$

Estimated regression equation:

$$\hat{y} = 10 + 5x$$

Least Squares Method (Cont'd)

Example (Cont'd):

Excel Worksheet showing data:

	A	B	C	D
1	Week	TV Ads	Cars Sold	
2	1	1	14	
3	2	3	24	
4	3	2	18	
5	4	1	17	
6	5	3	27	
7				

Least Squares Method (Cont'd)

Example (Cont'd): Excel's Chart Tools for Scatter Diagram & Estimated Regression Equation

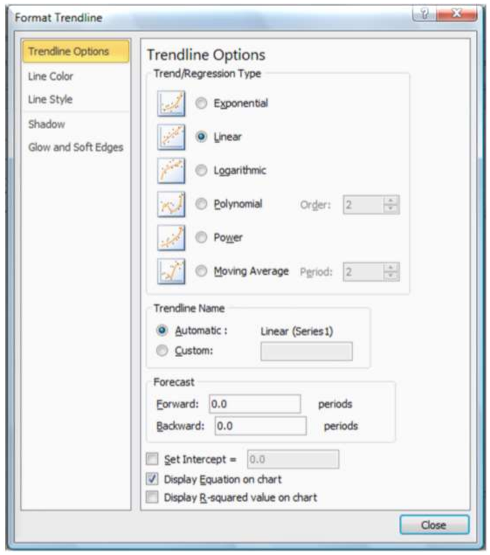
1) Select cells B2:C6
2) Click the **Insert** tab on the Ribbon
3) In the **Charts** group, click **Scatter**
4) When the list of scatter diagram subtypes appears, Click **Scatter with only Markers**
5) In the **Chart Layouts** group, click **Layout 1**
6) Right-click on the **Chart Title** to display a list of options; choose **Delete**
7) Select the **Horizontal (Value) Axis Title** and replace it with **TV Ads**
8) Select the **Vertical (Value) Axis Title** and replace it with **Cars Sold**
9) Right-click on the **Series 1 Legend Entry** to display a list of options; choose **Delete**
10) Position the mouse pointer over any **Vertical (Value) Axis Major Gridline** in the scatter diagram and right-click to display a list of options; choose **Delete**

Least Squares Method (Cont'd)

Example (Cont'd): Excel's Chart Tools for Scatter Diagram & Estimated Regression Equation

11) Position the mouse pointer over any data point in the scatter diagram and right-click to display a list of options; choose **Add Trendline**
12) When the **Format Trendline** dialog box appears, Select **Trendline Options** and then
    Choose **Linear** from the **Trend/Regression Type** list
    Choose **Display Equation on Chart**
    Click **Close**

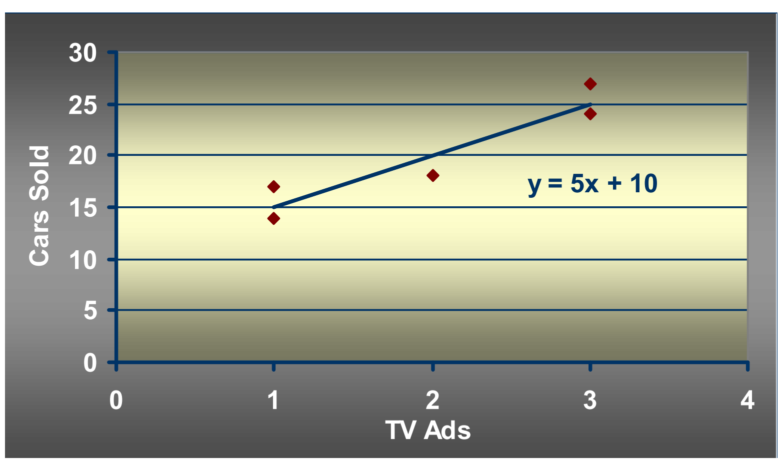
Least Squares Method (Cont‘d)



UT DALLAS

Least Squares Method (Cont‘d)

Example (Cont’d):



UT DALLAS

Coefficient of Determination

- Relationship Among SST, SSR, SSE

SST = SSR + SSE

$$\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i - \hat{y}_i)^2$$

where:

SST = total sum of squares

SSR = sum of squares due to regression

(measures the amount of variation in y explained by variation in the ***independent variable*** x.)

SSE = sum of squares due to error

(measures the amount of variation in y that remains unexplained, i.e. due to ***error***)

Coefficient of Determination (Cont'd)

- The coefficient of determination (a.k.a. R-squared) is:

R^2 = SSR/SST

- R^2 can vary from 0 to 1, since SST is fixed and 0<SSR<SST.
- A larger R^2 implies better regression, everything else being equal.

- Sample correlation coefficient r_{xy} is given by

$$r_{xy} = (\text{sign of } b_1)\sqrt{\text{Coefficient of Determination}}$$

$$r_{xy} = (\text{sign of } b_1)\sqrt{R^2}$$

Assumptions About the Error Term ε

Model assumptions

1. The error ε is a random variable with mean of zero.
2. The variance of ε, denoted by σ_{ε}^{2}, is the same for all values of the independent variable.
3. The values of ε are independent.
4. The error ε is a normally distributed random variable for all values of *x.*

Testing for Significance

- To test for a significant regression relationship, we must conduct a hypothesis test to determine whether the value of β_1 is zero.
- Two tests are commonly used:

***t* Test** and ***F* test**

- Both the *t* test and *F* test require an estimate of σ_{ε}^{2}, the variance of ε in the regression model.

Testing for Significance (Cont'd)

- An Estimate of σ_ε^2

$$s_\varepsilon^2 = \text{SSE}/(n - 2)$$

where:

$$\text{SSE}=\sum(y_i - \hat{y}_i)^2 = \sum(y_i - b_0 - b_1 x_i)^2$$

- The resulting s_ε is called the standard error of the estimate.

$$s_\varepsilon = \sqrt{\frac{\text{SSE}}{n-2}}$$

Testing for Significance: t Test

- Hypotheses

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

- Test Statistic

$$t = \frac{b_1}{s_{b_1}} \quad \text{where} \quad s_{b_1} = \frac{s_\varepsilon}{\sqrt{\sum(x_i - \bar{x})^2}}$$

(s_{b_1} is an estimate for the standard deviation of the sampling distribution of b_1)

- Rejection Region

$$t \leq -t_{n-2,\alpha/2} \text{ or } t \geq t_{n-2,\alpha/2}$$

Predicting the Particular Value of y for a Given x

The confidence interval to predict a one-time occurrence for a particular value of the dependent variable when the independent variable x_g is given

$$\hat{y} \pm t_{\frac{\alpha}{2},n-2} s_{\varepsilon} \sqrt{1 + \frac{1}{n} + \frac{(x_g - \bar{x})^2}{(n-1)s_x^2}}$$

where x_g is the given value of x and $\hat{y} = b_0 + b_1 x_g$

Using Excel's Regression Tool

Example (Revisited):

Excel Worksheet showing data:

	A	B	C	D
1	Week	TV Ads	Cars Sold	
2	1	1	14	
3	2	3	24	
4	3	2	18	
5	4	1	17	
6	5	3	27	
7				

Using Excel’s Regression Tool (Cont’d)

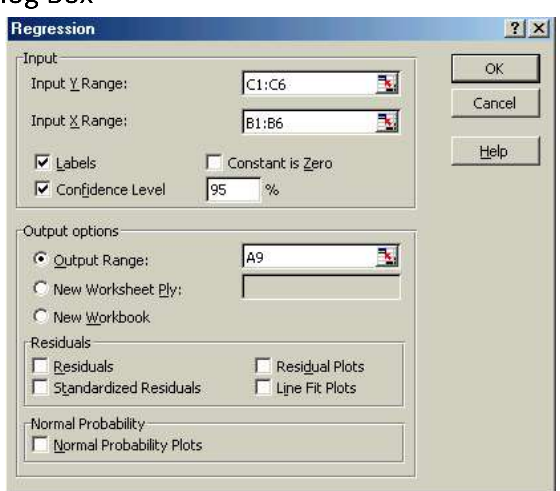
1) Select the **Tools** menu

2) Choose the **Data Analysis** option

3) Choose **Regression** from the list of Analysis Tools

Using Excel’s Regression Tool (Cont’d)

Excel Regression Dialog Box



Using Excel’s Regression Tool (Cont’d)

	A	B	C	D	E	F	G	H	I
1	Week	TV Ads	Cars Sold						
2	1	1	14						
3	2	3	24						
4	3	2	18						
5	4	1	17						
6	5	3	27						
7									
8	SUMMARY OUTPUT								
9									
10	*Regression Statistics*								
11	Multiple R	0.936585812							
12	R Square	0.877192982							
13	Adjusted R Square	0.83625731							
14	Standard Error	2.160246899							
15	Observations	5							
16									
17	ANOVA								
18		*df*	*SS*	*MS*	*F*	*Significance F*			
19	Regression	1	100	100	21.42857	0.018986231			
20	Residual	3	14	4.666667					
21	Total	4	114						
22									
23		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
24	Intercept	10	2.366431913	4.225771	0.024236	2.468950436	17.53104956	2.468950436	17.53104956
25	TV Ads	5	1.08012345	4.6291	0.018986	1.562561893	8.437438107	1.562561893	8.437438107
26									

Regression Statistics Output