Spring 2021

Section 3 (DSSA)

Regression – Part 1

- Regression analysis is a valuable tool that can be used to:
  - Understand the relationship between variables.
  - Predict the value of one variable based on another.

- Types of regression models:
  - Simple linear regression models contain *only two* variables.
  - Multiple regression models have *more* variables.

- Important terminologies:
  - The variable to be predicted is called the: *dependent or response* variable.
  - The other variable is called the: *independent, explanatory or predictable* variable.

- Scatter diagrams (plots) are a graphical way to investigate the relationship between variables:
  - The independent variable is normally plotted on the X-axis.
  - The dependent variable is normally plotted on the Y-axis.

Simple Linear Regression Model:

- Regression models are used to test, if there is a relationship between variables.
- There is some random error that cannot be predicted:

$$\boldsymbol{Y = \beta_0 + \beta_1 X + \varepsilon}$$

- $\boldsymbol{Y}$ = dependent variable (response) → **Actual value**
- $\boldsymbol{X}$ = independent variable (explanatory or predictable)
- $\boldsymbol{\beta_0}$ = intercept "value of Y when X is zero"
- $\boldsymbol{\beta_1}$ = slope of the regression line "unit increase in Y given unit increase in X"
- $\boldsymbol{\varepsilon}$ = random error

- True values of the slope and the intercept are not known so they are estimated using sample data.

$$\hat{Y} = \beta_0 + \beta_1 X$$

- $\hat{Y}$ = dependent variable (response) → **Predicted value**
- Error = Actual value – Predicted value
-

The following formulas can be used to compute the intercept and the slope:

$$\bar{X} = \frac{\Sigma X}{n} = \text{average (mean) of } X \text{ values}$$

$$\bar{Y} = \frac{\Sigma Y}{n} = \text{average (mean) of } Y \text{ values}$$

$$b_1 = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\Sigma(X - \bar{X})^2}$$

$$b_0 = \bar{Y} - b_1\bar{X}$$

Example 1:

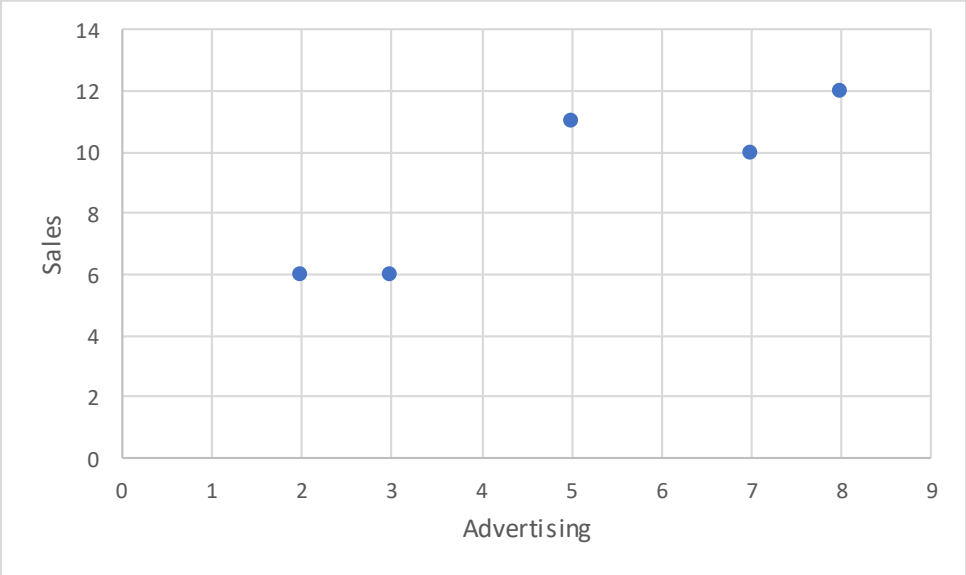
Judith Thompson runs a florist shop on the Gulf Coast of Texas, specializing in floral arrangements for weddings and other special events. She advertises weekly in the local newspapers and is considering increasing her advertising budget. Before doing so, she decides to evaluate the past effectiveness of these ads. Five weeks are sampled, and the advertising dollars and sales volume for each of these is shown in the following table.

- Draw a scatter diagram
- Develop a regression equation that would help Judith evaluate her advertising, and use the model to predict sales if the advertising budget is increased to 30.
- Measure how the model is fit by calculating r.
- Finally utilize excel QM to build the regression Model

SALES ($1,000)	ADVERTISING ($100)
11	5
6	3
10	7
6	2
12	8

Answer:

- **Scatter Diagram**



- **Regression Model:**

SALES Y	ADVERTISING X	$(X-\bar{X})^2$	$(X-\bar{X})(Y-\bar{Y})$
11	5	$(5-5)^2=0$	$(5-5)(11-9)=0$
6	3	$(3-5)^2=4$	$(3-5)(6-9)=6$
10	7	$(7-5)^2=4$	$(7-5)(10-9)=2$
6	2	$(2-5)^2=9$	$(2-5)(6-9)=9$
12	8	$(8-5)^2=9$	$(8-5)(12-9)=9$
$\sum Y = 45$	$\sum X = 25$	$\sum(X-\bar{X})^2 = 26$	$\sum(X-\bar{X})(Y-\bar{Y}) = 26$
$\bar{Y} = 45/5$	$\bar{X} = 25/5$		
$= 9$	$= 5$		

$$b_1 = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2} = \frac{26}{26} = 1$$

$$b_0 = \bar{Y} - b_1\bar{X} = 9 - (1)(5) = 4$$

The regression equation is

$$\hat{Y} = 4 + 1X$$

- **The sales when the advertising budget is increased to 30 is:**

$$Y = 4 + (1)\ (30) = 34$$

Measuring the fit of the regression model:

- How do we know the model is actually helpful in predicting Y based on X?
- We could just take the average error, but the positive and negative errors will cancel each other.

3 Measures of Variability:

- **SST : Sum of the squares total** [Total variability about the mean (Model Variability)]

$$\text{SST} = \sum(Y - \bar{Y})^2$$

- **SSE : Sum of the squared error [**Variability about the regression line (Unexplained Variability)]

$$\text{SSE} = \sum e^2 = \sum(Y - \hat{Y})^2$$

- **SSR : Sum of squares due to regression [**Total variability that is explained by the regression model (Explained Variability)]

$$\text{SSR} = \sum(\hat{Y} - \bar{Y})^2$$

Important relationship (check slide 20 in Lecture 5):

SST = SSE + SSR

Back to our example:

Y	X	$\hat{Y} = 4 + 1X$	$(Y - \hat{Y})^2$	$(Y - \bar{Y})^2$
11	5	9	$(11-9)^2 = 4$	$(11-9)^2 = 4$
6	3	7	$(6-7)^2 = 1$	$(6-9)^2 = 9$
10	7	11	$(10-11)^2 = 1$	$(10-9)^2 = 1$
6	2	6	$(6-6)^2 = 0$	$(6-9)^2 = 9$
12	8	12	$(12-12)^2 = 0$	$(12-9)^2 = 9$
$\sum Y = 45$	$\sum X = 25$		$\sum(Y - \hat{Y})^2 = 6$	$\sum(Y - \bar{Y})^2 = 32$
$\bar{Y} = 9$	$\bar{X} = 5$		SSE	SST

To get SSR we need to add another column or utilize the SST = SSR + SSE

So SSR = SST - SSE = 32 – 6 = 26

Coefficient of Determination (r^2):

- The proportion of the variability in Y explained by regression equation.

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

Back to our example:

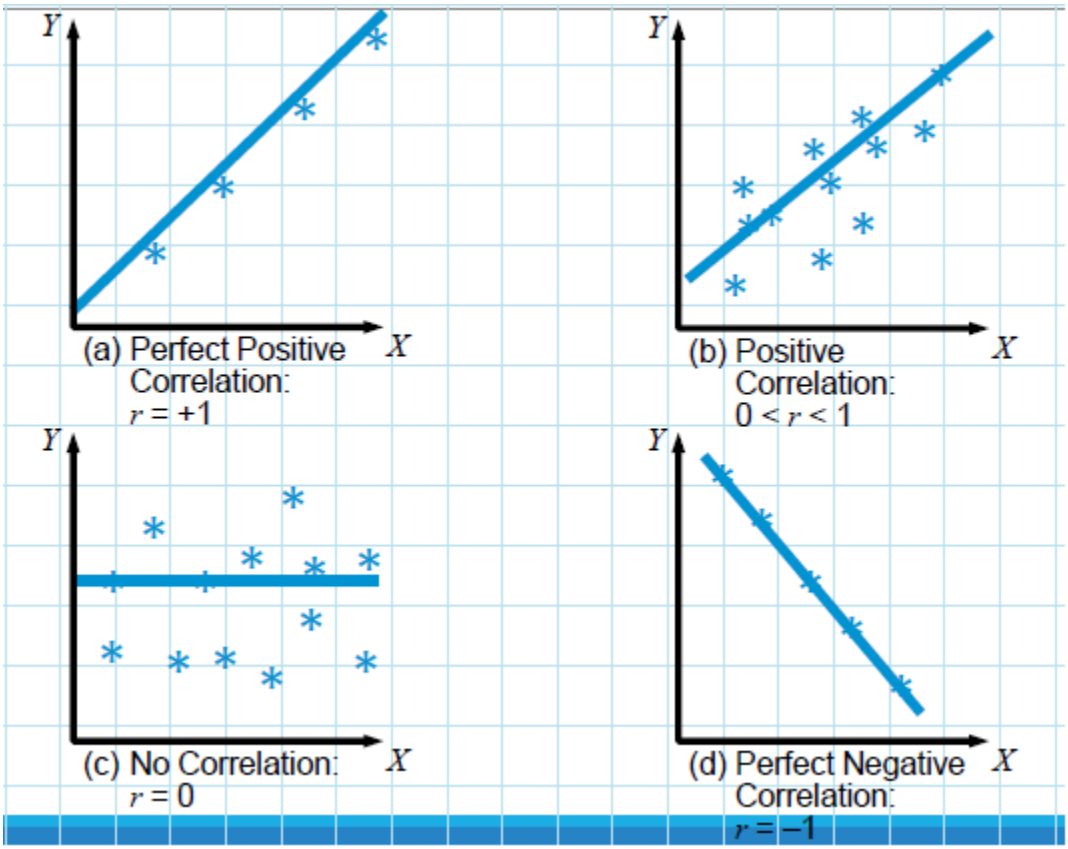
$$r^2 = \frac{SSR}{SST} = \frac{26}{32} = 0.8125$$

Indicating that about 81% of the variability in sales can be explained by the regression model with advertising as the independent variable

Correlation Coefficient (r):

- An expression of the strength of the linear relationship between the variables.
- It will always be between +1 and -1

$$r = \pm\sqrt{r^2}$$



Back to our example:

$$r = \sqrt{0.8125} = 0.901$$

Indicating that there is a very strong positive (directly proportional) linear relationship between the sales and the advertising

Use the excel QM add-in to build the regression Model:

1. Open the Excel QM.

2. Click on the "By chapter tab" and choose (chapter 4: regression models), then choose multiple regression for both simple or multiple regression examples.

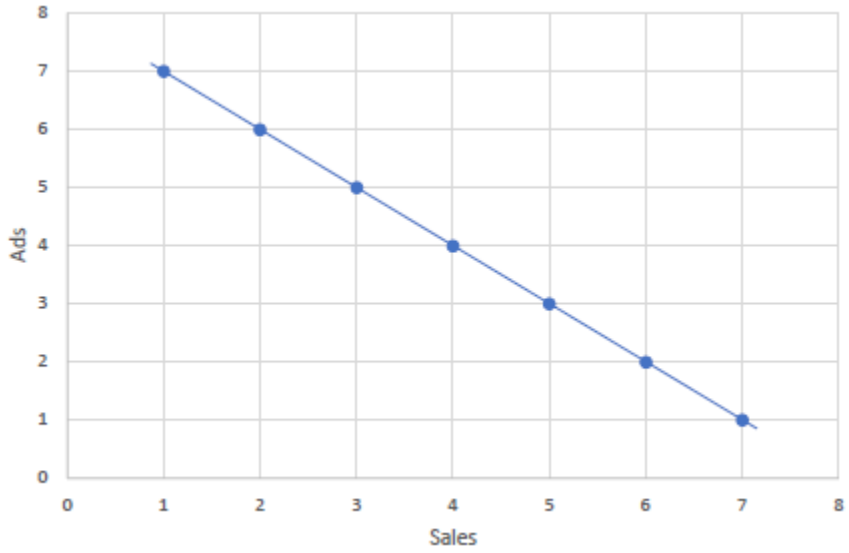
3. Enter the number of past observations = 5 and the number of independent (X) variables = 1. You can also enter a name or title for the problem. This will initialize the size of the spreadsheet.

4. Enter the data in the shaded part under Y and X1 and the calculations will be automatically added.

5. The intercept is 4 (the coefficient in the Y column) and the slope is 1 (the coefficient in the x1 column), resulting in the regression equation which is the equation found earlier.

6. Try changing values in the forecasting row, and check the effect on Y.

Example 2:



Given the following Scatter diagram and linear regression line

Which of the following statements are true?

1. r = 0
2. r = 1
3. r = -1
4. r^2 = 100%
5. r^2 = 0
6. SSR = SST
7. SSR > SST
8. SSR < SST

Answer:

1. r = 0 (false)
2. r = 1 (false)
3. r = -1 (true)
4. r^2 = 100% (true)
5. r^2 = 0 (false)
6. SSR = SST (true)
7. SSR > SST (false)
8. SSR < SST (false)