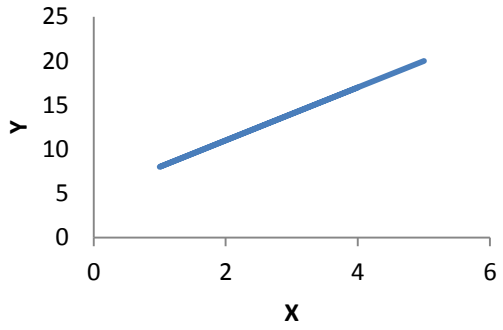
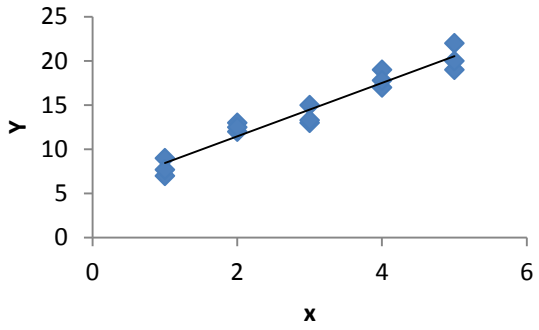
Chapter 12:

Regression Analysis:

Deterministic Relationship	**Non Deterministic Relationship**
For fixed x there is fixed y	For fixed x the y will be random
$y = \beta_0 + \beta_1 x$	$Y = \beta_0 + \beta_1 x + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$ $\boldsymbol{\mu_{Y.x} = \beta_0 + \beta_1 x}$
$\boldsymbol{\beta_1}$: The slope of a line is the change in *y* for a 1-unit increase in *x*. $\boldsymbol{\beta_0}$: The *y*-intercept is the height at which the line crosses the vertical axis and is obtained by setting $x = 0$ in the equation.	$\boldsymbol{\beta_1}$: The slope of a line is the ***expected*** change in *Y* associated with a 1-unit increase in the value of *x*. $\boldsymbol{\beta_0}$: The *y*-intercept is the height at which the line ***expected*** to crosses the vertical axis.
	

The estimated regression equation:

$$Y = b_0 + b_1 x = \hat{\beta}_0 + \hat{\beta}_1 x$$

The least square estimates:

$$b_1 = \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

$b_0 = \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

b_1 called the regression coefficient .

Purposes of the Estimated Regression Line:

for a fixed x value x^*: $(x = x^*)$ the equation gives:

1. A point estimate of the expected value of Y when $x = x^*$.

2. A point prediction of the Y value of that will result from a single new observation made at $x = x^*$ and it is then called the predicted value or the fitted value (under the condition that x^*should be in the actual range of x).

Estimate the variance:

$$\hat{\sigma}^2 = s^2 = \frac{Error\ Sum\ of\ Squares}{n-2} = \frac{SSE}{n-2} = \frac{\sum(y_i - \hat{y}_i)^2}{n-2}$$

Coefficient of Determination:

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

Where the regression some of squares is given by

$$SSR = SST - SSE$$

The sample correlation coefficient (Person's correlation coefficient):

The Pearson correlation evaluates the linear relationship between two continuous variables. A relationship is linear when a change in one variable is associated with a proportional change in the other variable.

$$\hat{\rho} = r = \frac{S_{xy}}{\sqrt{\sum(x_i - \bar{x})^2\ \sum(y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}\ S_{yy}}}$$

Spearman's Rank Correlation Coefficient:

The Spearman correlation evaluates the monotonic relationship between two continuous or ordinal variables. The Spearman correlation coefficient is based on the ranked values for each variable rather than the raw data.

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}\ ,$$

where $d_i = Rank(X_i) - Rank(Y_i)$.

To calculate the Spearman's correlation coefficient for a given set of data, we follow the steps:

1. Rank the values of X from low to high.
2. Rank the values of Y from low to high.
3. Compute $d_i = Rank(X_i) - Rank(Y_i)$ for each pairs of observations.
4. Square each d_i and compute $\sum d_i^2$, the sum of the square values.
5. Compute the Spearman's correlation coefficient

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)},$$

Where n is the number of pairs of X's and Y's.

❖ **Note that the value of r_s is between 1 and -1.**

Example: the table below shows the number of hrs studied by 10 students and the grades they obtained:

Number of hrs studied (x)	Grade in exam (y)	Rank(X)	Rank(y)	d_i	d_i^2
9	56	5	4	1	1
4	44	2	2	0	0
11	79	7	8	-1	1
13	72	8	7	1	1
10	70	6	6	0	0
5	54	3	3	0	0
18	94	10	10	0	0
15	85	9	9	0	0
2	33	1	1	0	0
8	65	4	5	-1	1
				$\sum d_i^2 =$	4

Spearman's correlation coefficient is: $r_s = \frac{1-6(4)}{10(100-1)} = 0.97.$

Notes:

1. The correlation can range in value from -1 to 1.
2. The absolute value of the coefficient reflects the strength of the correlation. So, the correlation of -0.7 is stronger than a correlation of +0.5. A common mistake among students occurs when they assume that a positive correlation is stronger than a negative correlation.

Correlation and causation:

Correlation is a measure of the degree to which two variables are related. A strong correlation between two variables, whether positive or negative, does not mean that one of the variables caused the other. For example, if the correlation between math score and high blood pressure is 0.93. This does not mean high blood pressure causes poor mathematics performance or that poor mathematics performance causes high blood pressure. The correlation of -0.93 shows that there is a very strong relationship between math test scores and systolic blood pressure, but that correlation tells us nothing about what causes that relationship. Correlation coefficient only indicates there is a relationship between two variables, but does not explain why the relationship occurs.