Chapter 2

We study the relationship between two variables or more. These variables could be either quantitative or qualitative.

Response Variable (Dependent): It measures an outcome of a study.
Explanatory Variable (Independent): Explains or causes changes in the response variables.

Example 2.5 page 82.

Goal: To show that changes in one or more Independent variables actually causes changes in the Dependent variable.

Section 2.1 Scatterplots

The scatterplot allows us to study the association between two variables. The association could be positive or negative.

Example 2.10 –Positive Association (Figure 2.2 page 86)
Example 2.12 –Negative Association(Figure 2.5 page 89)

Homework: 2.13, 2.27, 2.28, 2.29, and 2.31 pages 95-99.

Section 2.2 Linear Correlation(r)

The linear correlation coefficient, **r**, measures the strength of the linear association between two quantitative variables, positive or negative.

Rules for interpreting r:

a. The value of r always falls between –1 and 1. A positive value of r indicates positive correlation and a negative value of r indicates negative correlation.
b. The closer r is to 1, the stronger the positive correlation and the closer r is to –1, the stronger the negative correlation. Values of r closer to zero indicate no linear association.
c. The larger the absolute value of r, the stronger the relationship between the two variables.
d. r measures only the strength of linear relationship between two variables.
e. Changing the unit of measurement on the variables the value of r remains the same.

Formula for computing r: $r = \frac{ss_{xy}}{\sqrt{ss_x ss_y}}$ where

$$ss_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n} = \sum_{i=1}^{n} x_i y_i - n\overline{X}\overline{Y} \quad ;$$

$$ss_x = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = \sum_{i=1}^{n} x_i^2 - n\left(\overline{X}\right)^2 \quad ; \quad ss_y = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = \sum_{i=1}^{n} y_i^2 - n\left(\overline{Y}\right)^2$$

Note: The book is using different notation

HW: 2.53 page 106

Section 2.3 Least – Squares Regression

The **least square method** is a procedure that is used to find the line that provides the best approximation for the relationship between x and y. We refer to this equation of the line developed using the least square method as the **regression line**.

Regression Line: $\hat{Y} = a + bx$ **where**

a = y-intercept of the line
b = slope of the line
$\hat{Y}$ = estimated value of the dependent variable

Least Square Method : The values of a and b can be computed using the following equations.

$$b = \frac{SS_{xy}}{SS_x} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{X}\overline{Y}}{\sum_{i=1}^{n} x_i^2 - n\left(\overline{X}\right)^2} \quad \text{and} \quad a = \overline{Y} - b\overline{X}$$

where n = total number of observations.

r^2 – Square $= 100r^2$

The square of the correlation, r^2, is percentage of the variation explained in the y variable (dependent) using one or more x variables (independent). The prediction of y using one or more x variables is good, if the r^2 – Square is 75% or higher.

Note: Talk about

(a) Restriction on the range of the x values on the prediction
(b) The interpretation of the y-intercept and slope

Homework: 2.74 (Use Minitab), 2.78 pages 120-121.

2.4 Cautions about Regression and Correlation

Residuals: A residual is the difference between an observed value of y and the value predicted by the regression line. That is,

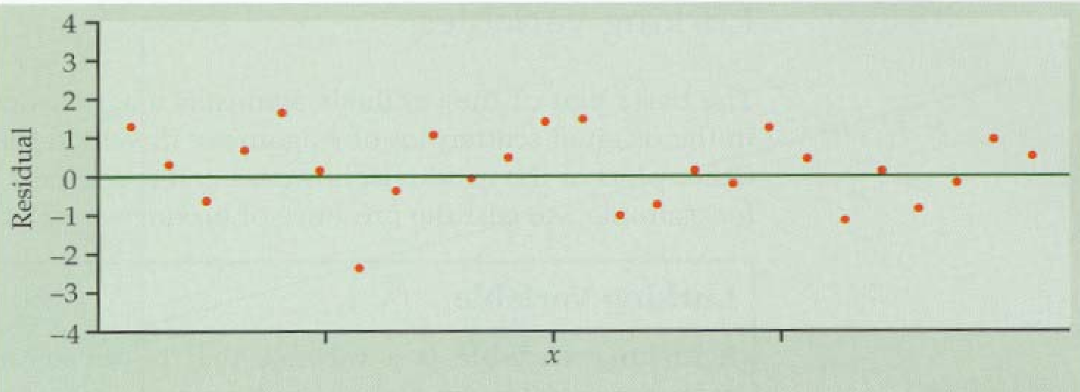
residual = observed y – predicted $e_i = Y_i - \hat{Y}_i$.

Residual Plots: Plot the residuals versus x variable(s)
It helps us assess the fit of a regression line. See figures on page 4(handout).

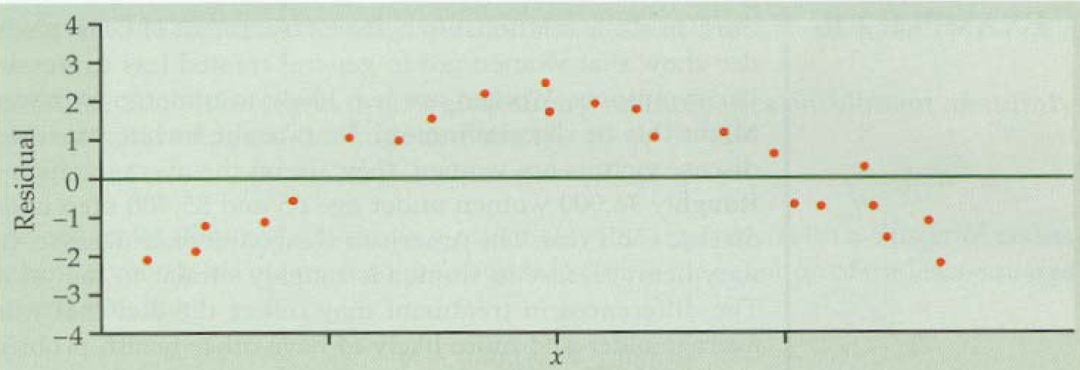
Outliers and influential Observations
See example 2.27 and 2.28 page 126-128.

Lurking variable: A variable that it's not included but has an important effect on the relationship. Plot the residuals versus time, if you can, to find that out. **See example 2.30 page 129.**

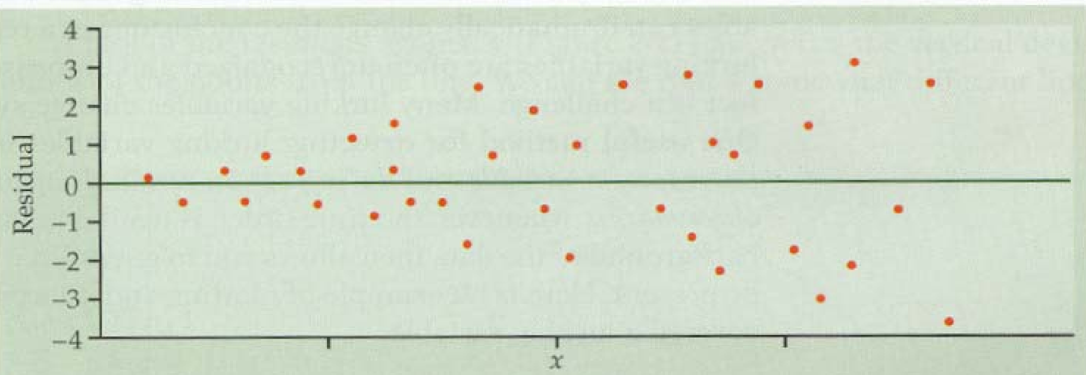
Homework: 2.101, 2.102, 2.109 pages 134-135.


Residual
4
3
2
1
0
−1
−2
−3
−4
x

(a)


Residual
4
3
2
1
0
−1
−2
−3
−4
x

(b)


Residual
4
3
2
1
0
−1
−2
−3
−4
x

(c)