Problem 1.11

The regression function relating production output by an employee after taking a training program (Y) to the production output before the training program (X) is $E\{Y\} = 20 + 0.95X$, where X ranges from 40 to 100. An observer concludes that the training program does not raise production output on the average because β_1, is not greater than 1.0. Comment.

Definition from the book: β_1, is the slope of the regression line. It indicates the change in the mean of the probability distribution of Y per unit increase in X.

Response 1.11: Based on the definition and $E\{Y\} = 20+0.95X$, a unit increase in X will result in the change in them mean of the probability distribution of Y by 0.95. If we let $X_i = 40$ and plug it into the $E\{Y\}$ and compare the difference, then we get

$$E\{Y_i\} = 20 + 0.95 * 40 = 58$$
$$E\{Y_i\} - X_i = 58 - 40 = 18$$

Next, if we let $X_i = 100$ and plug it into the $E\{Y\}$ and compare the difference, then we get

$$E\{Y_i\} = 20 + 0.95 * 100 = 115$$
$$E\{Y_i\} - X_i = 115 - 100 = 15$$

These two calculations show that since X ranges between 40 and 100, the training program raises production output despite β_1 being less than 1.0. Therefore, β_1 will raise the production until X stays within that range. Training becomes ineffective after the point when $X >= 400$, but because of the range restrictions on production output we wouldn't get to that point:

$$E\{Y_i\} - X_i = 20 + 0.95X - X = 0$$
$$0.05X = 20$$
$$x = 400$$

Problem 1.13

Computer programmers employed by a software developer were asked to participate in a month-long training seminar. During the seminar, each employee was asked to record the number of hours spent in class preparation each week. After completing the seminar, the productivity level of each participant was measured. A positive linear statistical relationship between participants' productivity levels and time spent in class preparation was found. The seminar leader concluded that increases in employee productivity are caused by increased class preparation time.

Given:

Y = participants' productivity levels

X = number of hours spent in class preparation each week.

a) Were the data used by the seminar leader observational or experimental data?

This is an observational data. Because the seminar leader did not exercise control over the explanatory variable through randomization, but only "observed" what the employees were reporting.

b) Comment on the validity of the conclusion reached by the seminar leader.

The seminar leader's conclusion cannot be considered valid. Because the positive linear statistical relationship between the productivity levels and the number of hours spent in class preparation each week does not imply cause-and-effect relationship. Thus higher productivity might not be a direct result of the higher number of hours spend preparing for the class.

c) Identify two or three alternative variables that might cause both the employee productivity scores and the employee class participation times to increase (decrease) simultaneously.

Some of the factors that could have simultaneous affect on productivity and hours spent in class preparation include:

- *The amount of coffee the employee consumes throughout the day. For instance, the employee who consumes more coffee might be able to have more energy to complete the tasks and to spend more time studying.*
- *The hours of sleep the employee gets. Those who get more sleep on a daily basis might be more likely to spend more time studying and to finish tasks faster compared to their sleep deprived colleagues.*
- *Family situation. For instance, if the employee has young kids who might be keeping them up at nights, their productivity might fall and they might have other responsibilities at home which would reduce their study time.*
- *The employees' motivation to receive a promotion. Those employees that want promotion and salary raise might be more motivated to spend more time studying and to complete tasks faster.*

d) How might the study be changed so that a valid conclusion about causal relationship between class preparation time and employee productivity can be reached?

The seminar leader might chose to add other variables to the equation or randomize the sample basing on different backgrounds or considerations.

Problem 1.16

Evaluate the following statement: "For the least squares method to be fully valid, it is required that the distribution of Y be normal."

This statement is false. Because none of the components of the model assume or require normality. The main goal is to find estimators β_0 and β_1 which would minimize the sum of the nsquared deviations.

Problem 1.18

According to (1.17), $\sum e_i = 0$ when regression model (1.1) is fitted to a set of n cases by the method of least squares. Is it also true that $\sum \epsilon_i = 0$? Comment.

Since the least squares method aims to minimize the sum of the nsquared deviations, $\sum e_i = 0$. *However,* ϵ_i *are random error values which we cannot control and therefore* $\sum \epsilon_i \neq 0$.

Exercise 1.32

Derive the expression for b_1 in (1.10a) from the normal equations in (1.9).

Formulas and identities used:

(1.10a) $b_1 = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2}$

(1.9a) $\sum Y_i = nb_0 + b_1 \sum X_i$

(1.9b) $\sum X_i Y_i = b_0 \sum X_i + b_1 \sum X_i^2$

$\sum a = na$

$\sum Y_i = n\bar{Y}$ or $\bar{Y} = \frac{\sum Y_i}{n}$

Prep-work:

1. From (1.9a):
$$b_0 = \frac{\sum Y_i - b_1 \sum X_i}{n} = \bar{Y} - b_1\bar{X}$$
$$b_0 = \bar{Y} - b_1\bar{X}$$

2. From (1.10) and the fact that $\bar{X}\&\bar{Y}$ are constants:
$$\begin{aligned}\sum(X_i - \bar{X})(Y_i - \bar{Y}) &= \sum(X_iY_i - \bar{X}Y_i - X_i\bar{Y} + \bar{X}\bar{Y}) = \\ &= \sum X_iY_i - \sum \bar{X}Y_i - \sum X_i\bar{Y} + \sum \bar{X}\bar{Y} = \\ &= \sum X_iY_i - \bar{X}\sum Y_i - \bar{Y}\sum X_i + n\bar{X}\bar{Y} = \\ &= \sum X_iY_i - \bar{X}\sum Y_i - \bar{Y}\sum X_i + \bar{X}\sum Y_i = \\ &= \sum X_iY_i - \bar{Y}\sum X_i \\ \sum(X_i - \bar{X})(Y_i - \bar{Y}) &= \sum X_iY_i - \bar{Y}\sum X_i.\end{aligned}$$

3. From (1.10):
$$\begin{aligned}\sum(X_i - \bar{X})^2 &= \sum(X_i^2 - 2X_i\bar{X} + \bar{X}^2) = \\ &= \sum X_i^2 - 2\bar{X}\sum X_i + \sum \bar{X}^2 = \\ &= \sum X_i^2 - 2\bar{X}\sum X_i + n\bar{X}^2 = \\ &= \sum X_i^2 - 2\bar{X}\sum X_i + \bar{X}\sum X_i = \\ &= \sum X_i^2 - \bar{X}\sum X_i \\ \sum(X_i - \bar{X})^2 &= \sum X_i^2 - \bar{X}\sum X_i.\end{aligned}$$

Derivation:

$$\begin{aligned}\sum X_i Y_i &= b_0 \sum X_i + b_1 \sum X_i^2 = \\ &= (\bar{Y} - b_1 \bar{X}) \sum X_i + b_1 \sum X_i^2 = \\ &= \bar{Y} \sum X_i - b_1 \bar{X} \sum X_i + b_1 \sum X_i^2 \\ \sum X_i Y_i - \bar{Y} \sum X_i &= b_1 \left(-\bar{X} \sum X_i + \sum X_i^2 \right) \\ b_1 &= \frac{\sum X_i Y_i - \bar{Y} \sum X_i}{\sum X_i^2 - \bar{X} \sum X_i}\end{aligned}$$

Using the *Prep-work 2 & 3* results we derive the expression for b_1 in (1.10a):

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

Exercise 1.36

Prove the result in (1.20) - that the sum of the residuals weighted by the fitted values is zero.

Formulas and identities used:

(1.20) $\sum_{i=1}^{n} \hat{Y}_i e_i = 0$

(1.17) $\sum_{i=1}^{n} e_i = 0$

(1.19) $\sum_{i=1}^{n} X_i e_i = 0$

$\hat{Y} = \bar{Y} + b_1 (X - \bar{X})$

$\bar{Y}$ and b_1 are constants

Proof:

$$\begin{aligned}\sum_{i=1}^{n} \hat{Y}_i e_i &= \sum_{i=1}^{n} (\bar{Y}_i + b_1 (X_i - \bar{X}_i) * e_i = \\ &= \sum_{i=1}^{n} \bar{Y} e_i + \sum_{i=1}^{n} b_1 X_i e_i - \sum_{i=1}^{n} b_1 \bar{X}_i e_i = \\ &= \bar{Y} \sum_{i=1}^{n} e_i + b_1 \sum_{i=1}^{n} X_i e_i - b_1 \sum_{i=1}^{n} \bar{X}_i e_i = \\ &= \bar{Y} * 0 + b_1 * 0 - b_1 * 0 = \\ &= 0\end{aligned}$$

$\therefore \sum_{i=1}^{n} \hat{Y}_i e_i = 0$