Problems (from NKNW4)

1.1. Refer to the sales volume example on page 4. Suppose that the number of units sold is measured accurately, but clerical errors are frequently made in determining the dollar sales. Would the relation between the number of units sold and dollar sales still be a functional one? Discuss.

1.2. The members of a health spa pay annual membership dues of $300 plus a charge of $2 for each visit to the spa. Let Y denote the dollar cost for the year for a member and X the number of visits by the member during the year. Express the relation between X and Y mathematically. Is it a functional relation or a statistical relation?

1.3 Experience with a certain type of plastic indicates that a relation exists between the hardness (measured in Brinell units) of items molded from the plastic (Y) and the elapsed time since termination of the molding process (X). It is proposed to study this relation by means of regression analysis. A participant in the discussion objects, pointing out that the hardening of the plastic "is the result of a natural chemical process that doesn't leave anything to chance, so the relation must be mathematical and regression analysis is not appropriate." Evaluate this objection.

1.4. In Table I.1, the lot size X is the same in production runs 1 and 24 but the work hours Y differ. What feature of regression model (1.1) is illustrated by this?

1.5 When asked to state the simple linear regression model, a student wrote it as follows: $E\{Y_i\} = \beta o + \beta_1 X_i + \varepsilon_i$. Do you agree?

1.6. Consider the normal error regression model (1.24). Suppose that the parameter values are $\beta_0 = 200$, $\beta_1 = 5.0$, and $\sigma = 4$.

a. Plot this normal error regression model in the fashion of Figure 1.6. Show the distributions of Y for X = 10, 20, and 40.

b. Explain the meaning of the parameters β_0 and β_1. Assume that the scope of the model includes X = 0.

1.7 In a simulation exercise, regression model (1.1) applies with $\beta_0 = 100$, $\beta_1 = 20$, and $\sigma^2 = 25$. An observation on Y will be made for X = 5.

a. Can you state the exact probability that Y will fall between 195 and 205? Explain.

b. If the normal error regression model (1.24) is applicable, can you now state the exact probability that Y will fall between 195 and 205? If so, state it.

1.8. In Figure 1.6, suppose another Y observation is obtained at X = 45. Would E{Y} for this new observation still be 104? Would the Y value for this new case again be 108?

1.9. A student in accounting enthusiastically declared: "Regression is a very powerful tool. We can isolate fixed and variable costs by fitting a linear regression model, even when we have no data for small lots." Discuss.

1.10. An analyst in a large corporation studied the relation between current annual salary (Y) and age (X) for the 46 computer programmers presently employed in the company. The analyst concluded that the relation is curvilinear, reaching a maximum at 47 years. Does this imply that the salary for a programmer increases until age 47 and then decreases? Explain.

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 + 0.9SX, 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.

1. 12. In a study of the relationship for senior citizens between physical activity and frequency of colds, participants were asked to monitor their weekly time spent in exercise over a fiveyear period and the frequency of colds. The study demonstrated that a negative statistical relation exists between time spent in exercise and frequency of colds. The investigator concluded that increasing the time spent in exercise is an effective strategy for reducing the frequency of colds for senior citizens.

    a. Were the data obtained in the study observational or experimental data?

    b. Comment on the validity of the conclusions reached by the investigator.

    c. Identify two or three other explanatory variables that might affect both the time spent in exercise and the frequency of colds for senior citizens simultaneously.

    How might the study be changed so that a valid conclusion about causal relationship between amount of exercise and frequency of colds can be reached?

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.

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

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

    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.

    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?

1.14. Refer to Problem 1.3. Four different elapsed times since termination of the molding process (treatments) are to be studied to see how they affect the hardness of a plastic. Sixteen

1.29. Refer to regression model (1.1). Assume that $X = 0$ is within the scope of the model. What is the implication for the regression function if $\beta_0 = 0$ so that the model is $Y_i = \beta_1 X_i + \varepsilon_i$? How would the regression function plot on a graph?

1.30. Refer to regression model (1.1). What is the implication for the regression function if $\beta 1 = 0$ so that the model is $Y_i = \beta_0 + \varepsilon_i$? How would the regression function plot on a graph?

1.31. Refer to **Plastic hardness** Problem 1.22. Suppose one test item was molded from a single batch of plastic and the hardness of this one item was measured at 16 different points in time. Would the error term in the regression model for this case still reflect the same effects as for the experiment initially described? Would you expect the error terms for the different points in time to be uncorrelated? Discuss.

1.32. Derive the expression for b1 in (1. lOa) from the normal equations in (1.9).

1.33. (Calculus needed.) Refer to the regression model $Y_i = \beta_0 + \varepsilon_i$ in Exercise 1.30. Derive the least squares estimator of b0 for this model.

1.34. Prove that the least squares estimator of β_0 obtained in Exercise 1.33 is unbiased.

1.35. Prove the result in (1.18)—that the sum of the Y observations is the same as the sum of the fitted values.

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

1.37. Refer to Table 1.1 for the Toluca Company example. When asked to present a point estimate of the expected work hours for lot sizes of 30 pieces, a person gave the estimate 202 because this is the mean number of work hours in the three runs of size 30 in the study. A critic states that this person's approach "throws away" most of the data in the study because cases with lot sizes other than 30 are ignored. Comment.

1.38. In **Airfreight breakage** Problem 1.21, the least squares estimates are $b_0 = 10.20$ and $b_1 = 4.00$, and $\Sigma e_i^2 = 17.60$. Evaluate the least squares criterion Q in (1.8) for the estimates (1) $b_0 = 9$, $b_1 = 3$; (2) $b_0 = 11$, $b_1 = 5$. Is the criterion Q larger for these estimates than for the least squares estimates?

1.39. Two observations on Y were obtained at each of three X levels, namely, at X = 5, X = 10, and X = 15.

a. Show that the least squares regression line fitted to the *three* points $(5, \bar{Y}_1)$, $(10, \bar{Y}_2)$, and $(15, \bar{Y}_3)$, where $\bar{Y}_1$, $\bar{Y}_2$, and $\bar{Y}_3$ denote the means of the Y observations at the three X levels, is identical to the least squares regression line fitted to the original six cases.

b. In this study, could the error term variance σ^2 be estimated without fitting a regression line? Explain.

1.40. In fitting regression model (1.1), it was found that observation Y_i fell directly on the fitted regression line (i.e., $Y_i = \hat{Y}_i$). If this case were deleted, would the least squares regression line fitted to the remaining *n*—1 cases be changed? *[Hint:* What is the contribution of case i to the least squares criterion Q in (1.8)?]