2.23. Refer to **Grade point average** Problem 1.19.

a. Set up the ANOVA table.

b. What is estimated by *MSR* in your ANOVA table? by *MSE*? Under what condition do *MSR* and *MSE* estimate the same quantity?

c. Conduct an F test of whether or not $\beta_1 = 0$. Control the α risk at .01. State the alternatives, decision rule, and conclusion.

d. What is the absolute magnitude of the reduction in the variation of Y when X is introduced into the regression model? What is the relative reduction? What is the name of the latter measure?

e. Obtain r and attach the appropriate sign.

f. Which measure, R^2 or r, has the more clear-cut operational interpretation? Explain.

*2.24. Refer to **Copier maintenance** Problem 1.20.

a. Set up the basic ANOVA table in the format of Table 2.2. Which elements of your table are additive? Also set up the ANOVA table in the format of Table 2.3. How do the two tables differ?

b. Conduct an F test to determine whether or not there is a linear association between time spent and number of copiers serviced; use $\alpha = .10$. State the alternatives, decision rule, and conclusion.

c. By how much, relatively, is the total variation in number of minutes spent on a call reduced when the number of copiers serviced is introduced into the analysis? Is this a relatively small or large reduction? What is the name of this measure?

d. Calculate r and attach the appropriate sign.

e. Which measure, r or R^2, has the more clear-cut operational interpretation?

*2.25. Refer to **Airfreight breakage** Problem 1.21.

a. Set up the ANOVA table. Which elements are additive?

b. Conduct an F test to decide whether or not there is a linear association between the number of times a carton is transferred and the number of broken ampules; control the α risk at .05. State the alternatives, decision rule, and conclusion.

c. Obtain the t^* statistic for the test in part (b) and demonstrate numerically its equivalence to the F^* statistic obtained in part (b).

d. Calculate R^2 and r. What proportion of the variation in Y is accounted for by introducing X into the regression model?

2.26. Refer to **Plastic hardness** Problem 1.22.

a. Set up the ANOVA table.

b. Test by means of an F test whether or not there is a linear association between the hardness of the plastic and the elapsed time. Use $\alpha = .01$. State the alternatives, decision rule, and conclusion.

c. Plot the deviations $Y_i - \hat{Y}_i$ against X_i on a graph. Plot the deviations $\hat{Y}_i - \bar{Y}$ against X_i on another graph, using the same scales as for the first graph. From your two graphs, does *SSE* or *SSR* appear to be the larger component of *SSTO*? What does this imply about the magnitude of R^2?

d. Calculate R^2 and r.

*2.27. Refer to **Muscle mass** Problem 1.27.

a. Conduct a test to decide whether or not there is a negative linear association between amount of muscle mass and age. Control the risk of Type I error at .05. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

c. The bivariate normal model (2.74) assumption is possibly inappropriate here. Compute the Spearman rank correlation coefficient, r_S.

d. Repeat part (b), this time basing the test of independence on the Spearman rank correlation computed in part (c) and test statistic (2.101). Use $\alpha = .05$. State the alternatives, decision rule, and conclusion.

e. How do your estimates and conclusions in parts (a) and (b) compare to those obtained in parts (c) and (d)?

2.48. Refer to **Crime rate** Problems 1.28, 2.30, and 2.31. Assume that the normal bivariate model (2.74) is appropriate.

a. Compute the Pearson product-moment correlation coefficient r_{12}.

b. Test whether crime rate and percentage of high school graduates are statistically independent in the population; use $\alpha = .01$. State the alternatives, decision rule, and conclusion.

c. How do your estimates and conclusions in parts (a) and (b) compare to those obtained in 2.31b and 2.30a, respectively?

2.49. Refer to **Crime rate** Problems 1.28 and 2.48. The bivariate normal model (2.74) assumption is possibly inappropriate here.

a. Compute the Spearman rank correlation coefficient r_S.

b. Test by means of the Spearman rank correlation coefficient whether an association exists between crime rate and percentage of high school graduates using test statistic (2.101) and a level of significance .01. State the alternatives, decision rule, and conclusion.

c. How do your estimates and conclusions in parts (a) and (b) compare to those obtained in Problems 2.48a and 2.48b, respectively?

Exercises

2.50. Derive the property in (2.6) for the k_i.

2.51. Show that b_0 as defined in (2.21) is an unbiased estimator of β_0.

2.52. Derive the expression in (2.22b) for the variance of b_0, making use of (2.31). Also explain how variance (2.22b) is a special case of variance (2.29b).

2.53. (Calculus needed.)

a. Obtain the likelihood function for the sample observations $Y_1, \ldots, Y_n$ given $X_1, \ldots, X_n$, if the conditions on page 83 apply.

b. Obtain the maximum likelihood estimators of β_0, β_1, and σ^2. Are the estimators of β_0 and β_1 the same as those in (1.27) when the X_i are fixed?

2.54. Suppose that normal error regression model (2.1) is applicable except that the error variance is not constant; rather the variance is larger, the larger is X. Does $\beta_1 = 0$ still imply that there is no linear association between X and Y? That there is no association between X and Y? Explain.

2.55. Derive the expression for *SSR* in (2.51).

2.56. In a small-scale regression study, five observations on Y were obtained corresponding to $X = 1$, 4, 10, 11, and 14. Assume that $\sigma = .6$, $\beta_0 = 5$, and $\beta_1 = 3$.

a. What are the expected values of *MSR* and *MSE* here?

b. For determining whether or not a regression relation exists, would it have been better or worse to have made the five observations at $X = 6$, 7, 8, 9, and 10? Why? Would the same answer apply if the principal purpose were to estimate the mean response for $X = 8$? Discuss.

Cited References

3.1. Barnett, V., and T. Lewis. *Outliers in Statistical Data.* 3rd ed. New York: John Wiley & Sons, 1994.

3.2. Looney, S. W., and T. R. Gulledge, Jr. "Use of the Correlation Coefficient with Normal Probability Plots," *The American Statistician* 39 (1985), pp. 75–79.

3.3. Shapiro, S. S., and M. B. Wilk. "An Analysis of Variance Test for Normality (Complete Samples)," *Biometrika* 52 (1965), pp. 591–611.

3.4. Levene, H. "Robust Tests for Equality of Variances," in *Contributions to Probability and Statistics,* ed. I. Olkin. Palo Alto, Calif.: Stanford University Press, 1960, pp. 278–92.

3.5. Brown, M. B., and A. B. Forsythe. "Robust Tests for Equality of Variances," *Journal of the American Statistical Association* 69 (1974), pp. 364–67.

3.6. Breusch, T. S., and A. R. Pagan. "A Simple Test for Heteroscedasticity and Random Coefficient Variation," *Econometrica* 47 (1979), pp. 1287–94.

3.7. Cook, R. D., and S. Weisberg. "Diagnostics for Heteroscedasticity in Regression," *Biometrika* 70 (1983), pp. 1–10.

3.8. Joglekar, G., J. H. Schuenemeyer, and V. LaRiccia. "Lack-of-Fit Testing When Replicates Are Not Available," *The American Statistician* 43 (1989), pp. 135–43.

3.9. Box, G. E. P., and D. R. Cox. "An Analysis of Transformations," *Journal of the Royal Statistical Society B* 26 (1964), pp. 211–43.

3.10. Draper, N. R., and H. Smith. *Applied Regression Analysis.* 3rd ed. New York: John Wiley & Sons, 1998.

3.11. Velleman, P. F., and D. C. Hoaglin. *Applications, Basics, and Computing of Exploratory Data Analysis.* Boston: Duxbury Press, 1981.

3.12. Cleveland, W. S. "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association* 74 (1979), pp. 829–36.

3.13. Altman, N. S. "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression," *The American Statistician* 46 (1992), pp. 175–85.

3.14. Haerdle, W. *Applied Nonparametric Regression.* Cambridge: Cambridge University Press, 1990.

Problems

3.1. Distinguish between (1) residual and semistudentized residual, (2) $E\{\varepsilon_i\} = 0$ and $\bar{e} = 0$, (3) error term and residual.

3.2. Prepare a prototype residual plot for each of the following cases: (1) error variance decreases with X; (2) true regression function is U shaped, but a linear regression function is fitted.

3.3. Refer to **Grade point average** Problem 1.19.

a. Prepare a box plot for the ACT scores X_i. Are there any noteworthy features in this plot?

b. Prepare a dot plot of the residuals. What information does this plot provide?

c. Plot the residual e_i against the fitted values $\hat{Y}_i$. What departures from regression model (2.1) can be studied from this plot? What are your findings?

d. Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality. Test the reasonableness of the normality assumption here using Table B.6 and $\alpha = .05$. What do you conclude?

e. Conduct the Brown-Forsythe test to determine whether or not the error variance varies with the level of X. Divide the data into the two groups, $X < 26$, $X \geq 26$, and use $\alpha = .01$. State the decision rule and conclusion. Does your conclusion support your preliminary findings in part (c)?

f. Information is given below for each student on two variables not included in the model, namely, intelligence test score (X_2) and high school class rank percentile (X_3). (Note that larger class rank percentiles indicate higher standing in the class, e.g., 1% is near the bottom of the class and 99% is near the top of the class.) Plot the residuals against X_2 and X_3 on separate graphs to ascertain whether the model can be improved by including either of these variables. What do you conclude?

i:	1	2	3	...	118	119	120
X_2:	122	132	119	...	140	111	110
X_3:	99	71	75	...	97	65	85

*3.4. Refer to **Copier maintenance** Problem 1.20.

a. Prepare a dot plot for the number of copiers serviced X_i. What information is provided by this plot? Are there any outlying cases with respect to this variable?

b. The cases are given in time order. Prepare a time plot for the number of copiers serviced. What does your plot show?

c. Prepare a stem-and-leaf plot of the residuals. Are there any noteworthy features in this plot?

d. Prepare residual plots of e_i versus $\hat{Y}_i$ and e_i versus X_i on separate graphs. Do these plots provide the same information? What departures from regression model (2.1) can be studied from these plots? State your findings.

e. Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality. Does the normality assumption appear to be tenable here? Use Table B.6 and $\alpha = .10$.

f. Prepare a time plot of the residuals to ascertain whether the error terms are correlated over time. What is your conclusion?

g. Assume that (3.10) is applicable and conduct the Breusch-Pagan test to determine whether or not the error variance varies with the level of X. Use $\alpha = .05$. State the alternatives, decision rule, and conclusion.

h. Information is given below on two variables not included in the regression model, namely, mean operational age of copiers serviced on the call (X_2, in months) and years of experience of the service person making the call (X_3). Plot the residuals against X_2 and X_3 on separate graphs to ascertain whether the model can be improved by including either or both of these variables. What do you conclude?

i:	1	2	3	...	43	44	45
X_2:	20	19	27	...	28	26	33
X_3:	4	5	4	...	3	3	6

*3.5. Refer to **Airfreight breakage** Problem 1.21.

a. Prepare a dot plot for the number of transfers X_i. Does the distribution of number of transfers appear to be asymmetrical?

b. The cases are given in time order. Prepare a time plot for the number of transfers. Is any systematic pattern evident in your plot? Discuss.

c. Obtain the residuals e_i and prepare a stem-and-leaf plot of the residuals. What information is provided by your plot?

d. Plot the residuals e_i against X_i to ascertain whether any departures from regression model (2.1) are evident. What is your conclusion?

e. Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality to ascertain whether the normality assumption is reasonable here. Use Table B.6 and $\alpha = .01$. What do you conclude?

f. Prepare a time plot of the residuals. What information is provided by your plot?

g. Assume that (3.10) is applicable and conduct the Breusch-Pagan test to determine whether or not the error variance varies with the level of X. Use $\alpha = .10$. State the alternatives, decision rule, and conclusion. Does your conclusion support your preliminary findings in part (d)?

3.6. Refer to **Plastic hardness** Problem 1.22.

a. Obtain the residuals e_i and prepare a box plot of the residuals. What information is provided by your plot?

b. Plot the residuals e_i against the fitted values $\hat{Y}_i$ to ascertain whether any departures from regression model (2.1) are evident. State your findings.

c. Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality. Does the normality assumption appear to be reasonable here? Use Table B.6 and $\alpha = .05$.

d. Compare the frequencies of the residuals against the expected frequencies under normality, using the 25th, 50th, and 75th percentiles of the relevant t distribution. Is the information provided by these comparisons consistent with the findings from the normal probability plot in part (c)?

e. Use the Brown-Forsythe test to determine whether or not the error variance varies with the level of X. Divide the data into the two groups, $X \leq 24$, $X > 24$, and use $\alpha = .05$. State the decision rule and conclusion. Does your conclusion support your preliminary findings in part (b)?

*3.7. Refer to **Muscle mass** Problem 1.27.

a. Prepare a stem-and-leaf plot for the ages X_i. Is this plot consistent with the random selection of women from each 10-year age group? Explain.

b. Obtain the residuals e_i and prepare a dot plot of the residuals. What does your plot show?

c. Plot the residuals e_i against $\hat{Y}_i$ and also against X_i on separate graphs to ascertain whether any departures from regression model (2.1) are evident. Do the two plots provide the same information? State your conclusions.

d. Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality to ascertain whether the normality assumption is tenable here. Use Table B.6 and $\alpha = .10$. What do you conclude?

e. Assume that (3.10) is applicable and conduct the Breusch-Pagan test to determine whether or not the error variance varies with the level of X. Use $\alpha = .01$. State the alternatives, decision rule, and conclusion. Is your conclusion consistent with your preliminary findings in part (c)?

3.8. Refer to **Crime rate** Problem 1.28.

a. Prepare a stem-and-leaf plot for the percentage of individuals in the county having at least a high school diploma X_i. What information does your plot provide?

b. Obtain the residuals e_i and prepare a box plot of the residuals. Does the distribution of the residuals appear to be symmetrical?

*3.13. Refer to **Copier maintenance** Problem 1.20.

a. What are the alternative conclusions when testing for lack of fit of a linear regression function?

b. Perform the test indicated in part (a). Control the risk of Type I error at .05. State the decision rule and conclusion.

c. Does the test in part (b) detect other departures from regression model (2.1), such as lack of constant variance or lack of normality in the error terms? Could the results of the test of lack of fit be affected by such departures? Discuss.

3.14. Refer to **Plastic hardness** Problem 1.22.

a. Perform the F test to determine whether or not there is lack of fit of a linear regression function; use $\alpha = .01$. State the alternatives, decision rule, and conclusion.

b. Is there any advantage of having an equal number of replications at each of the X levels? Is there any disadvantage?

c. Does the test in part (a) indicate what regression function is appropriate when it leads to the conclusion that the regression function is not linear? How would you proceed?

3.15. **Solution concentration.** A chemist studied the concentration of a solution (Y) over time (X). Fifteen identical solutions were prepared. The 15 solutions were randomly divided into five sets of three, and the five sets were measured, respectively, after 1, 3, 5, 7, and 9 hours. The results follow.

i:	1	2	3	...	13	14	15
X_i:	9	9	9	...	1	1	1
Y_i:	.07	.09	.08	...	2.84	2.57	3.10

a. Fit a linear regression function.

b. Perform the F test to determine whether or not there is lack of fit of a linear regression function; use $\alpha = .025$. State the alternatives, decision rule, and conclusion.

c. Does the test in part (b) indicate what regression function is appropriate when it leads to the conclusion that lack of fit of a linear regression function exists? Explain.

3.16. Refer to **Solution concentration** Problem 3.15.

a. Prepare a scatter plot of the data. What transformation of Y might you try, using the prototype patterns in Figure 3.15 to achieve constant variance and linearity?

b. Use the Box-Cox procedure and standardization (3.36) to find an appropriate power transformation. Evaluate SSE for $\lambda = -.2, -.1, 0, .1, .2$. What transformation of Y is suggested?

c. Use the transformation $Y' = \log_{10} Y$ and obtain the estimated linear regression function for the transformed data.

d. Plot the estimated regression line and the transformed data. Does the regression line appear to be a good fit to the transformed data?

e. Obtain the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show?

f. Express the estimated regression function in the original units.

*3.17. **Sales growth.** A marketing researcher studied annual sales of a product that had been introduced 10 years ago. The data are as follows, where X is the year (coded) and Y is sales in thousands

of units:

i:	1	2	3	4	5	6	7	8	9	10
X_i:	0	1	2	3	4	5	6	7	8	9
Y_i:	98	135	162	178	221	232	283	300	374	395

a. Prepare a scatter plot of the data. Does a linear relation appear adequate here?

b. Use the Box-Cox procedure and standardization (3.36) to find an appropriate power transformation of Y. Evaluate SSE for $\lambda = .3, .4, .5, .6, .7$. What transformation of Y is suggested?

c. Use the transformation $Y' = \sqrt{Y}$ and obtain the estimated linear regression function for the transformed data.

d. Plot the estimated regression line and the transformed data. Does the regression line appear to be a good fit to the transformed data?

e. Obtain the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show?

f. Express the estimated regression function in the original units.

3.18. **Production time.** In a manufacturing study, the production times for 111 recent production runs were obtained. The table below lists for each run the production time in hours (Y) and the production lot size (X).

i:	1	2	3	...	109	110	111
X_i:	15	9	7	...	12	9	15
Y_i:	14.28	8.80	12.49	...	16.37	11.45	15.78

a. Prepare a scatter plot of the data. Does a linear relation appear adequate here? Would a transformation on X or Y be more appropriate here? Why?

b. Use the transformation $X' = \sqrt{X}$ and obtain the estimated linear regression function for the transformed data.

c. Plot the estimated regression line and the transformed data. Does the regression line appear to be a good fit to the transformed data?

d. Obtain the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show?

e. Express the estimated regression function in the original units.

Exercises

3.19. A student fitted a linear regression function for a class assignment. The student plotted the residuals e_i against Y_i and found a positive relation. When the residuals were plotted against the fitted values $\hat{Y}_i$, the student found no relation. How could this difference arise? Which is the more meaningful plot?

3.20. If the error terms in a regression model are independent $N(0, \sigma^2)$, what can be said about the error terms after transformation $X' = 1/X$ is used? Is the situation the same after transformation $Y' = 1/Y$ is used?

3.21. Derive the result in (3.29).

3.22. Using (A.70), (A.41), and (A.42), show that $E\{MSPE\} = \sigma^2$ for normal error regression model (2.1).