2.10.1 PROBLEMS

2.1 Consider the following five observations. You are to do all the parts of this exercise using only a calculator.

x	y	$x - \overline{x}$	$(x - \overline{x})^2$	$y - \overline{y}$	$(x - \overline{x})(y - \overline{y})$
0	6				
1	2				
2	3				
3	1				
4	0				
$\sum x_i =$	$\sum y_i =$	$\sum(x_i - \overline{x}) =$	$\sum(x_i - \overline{x})^2 =$	$\sum(y_i - \overline{y}) =$	$\sum(x_i - \overline{x})(y_i - \overline{y}) =$

(a) Complete the entries in the table. Put the sums in the last row. What are the sample means $\overline{x}$ and $\overline{y}$?
(b) Calculate b_1 and b_2 using (2.7) and (2.8) and state their interpretation.
(c) Compute $\sum_{i=1}^{5} x_i^2$, $\sum_{i=1}^{5} x_i y_i$. Using these numerical values, show that

$$\sum(x_i - \overline{x})^2 = \sum x_i^2 - N\overline{x}^2 \quad \text{and} \quad \sum(x_i - \overline{x})(y_i - \overline{y}) = \sum x_i y_i - N\overline{xy}$$

(d) Use the least squares estimates from part (b) to compute the fitted values of y, and complete the remainder of the table below. Put the sums in the last row.

x_i	y_i	$\hat{y}_i$	$\hat{e}_i$	$\hat{e}_i^2$	$x_i\hat{e}_i$
0	6				
1	2				
2	3				
3	1				
4	0				
$\sum x_i =$	$\sum y_i =$	$\sum \hat{y}_i =$	$\sum \hat{e}_i =$	$\sum \hat{e}_i^2 =$	$\sum x_i\hat{e}_i =$

(e) On graph paper, plot the data points and sketch the fitted regression line $\hat{y}_i = b_1 + b_2 x_i$.
(f) On the sketch in part (e), locate the point of the means $(\overline{x}, \overline{y})$. Does your fitted line pass through that point? If not, go back to the drawing board, literally.
(g) Show that for these numerical values $\overline{y} = b_1 + b_2\overline{x}$.
(h) Show that for these numerical values $\overline{\hat{y}} = \overline{y}$, where $\overline{\hat{y}} = \sum\hat{y}_i/N$.
(i) Compute $\hat{\sigma}^2$.
(j) Compute $\widehat{\text{var}(b_2)}$.

2.2 A household has weekly income of \$2,000. The mean weekly expenditure for households with this income is $E(y|x = \$2{,}000) = \mu_{y|x=\$2{,}000} = \$200$, and expenditures exhibit variance $\text{var}(y|x = \$2{,}000) = \sigma^2_{y|x=\$2{,}000} = 100$.
(a) Assuming that weekly food expenditures are normally distributed, find the probability that a household with this income spends between \$180 and \$215 on food in a week. Include a sketch with your solution.
(b) Find the probability that a household with this income spends more than \$250 on food in a week. Include a sketch with your solution.
(c) Find the probability in part (a) if the variance of weekly expenditures is $\text{var}(y|x = \$2{,}000) = \sigma^2_{y|x=\$2{,}000} = 81$.
(d) Find the probability in part (b) if the variance of weekly expenditures is $\text{var}(y|x = \$2{,}000) = \sigma^2_{y|x=\$2{,}000} = 81$.

2.3* Graph the following observations of x and y on graph paper.

x	1	2	3	4	5	6
y	10	8	5	5	2	3

(a) Using a ruler, draw a line that fits through the data. Measure the slope and intercept of the line you have drawn.
(b) Use formulas (2.7) and (2.8) to compute, using only a hand calculator, the least squares estimates of the slope and the intercept. Plot this line on your graph.
(c) Obtain the sample means of $\bar{y} = \sum y_i/N$ and $\bar{x} = \sum x_i/N$. Obtain the predicted value of y for $x = \bar{x}$ and plot it on your graph. What do you observe about this predicted value?
(d) Using the least squares estimates from (b), compute the least squares residuals $\hat{e}_i$. Find their sum.
(e) Calculate $\sum x_i\hat{e}_i$.

2.4 We have defined the simple linear regression model to be $y = \beta_1 + \beta_2 x + e$. Suppose however that we knew, for a fact, that $\beta_1 = 0$.
(a) What does the linear regression model look like, algebraically, if $\beta_1 = 0$?
(b) What does the linear regression model look like, graphically, if $\beta_1 = 0$?
(c) If $\beta_1 = 0$ the least squares "sum of squares" function becomes $S(\beta_2) = \sum_{i=1}^{N}(y_i - \beta_2 x_i)^2$. Using the data,

x	1	2	3	4	5	6
y	4	6	7	7	9	11

plot the value of the sum of squares function for enough values of β_2 for you to locate the approximate minimum. What is the significance of the value of β_2 that minimizes $S(\beta_2)$? (*Hint*: Your computations will be simplified if you algebraically expand $S(\beta_2) = \sum_{i=1}^{N}(y_i - \beta_2 x_i)^2$ by squaring the term in parentheses and carrying the summation operator through.)
(d)◆ Using calculus, show that the formula for the least squares estimate of β_2 in this model is $b_2 = \sum x_i y_i / \sum x_i^2$. Use this result to compute b_2 and compare this value to the value you obtained geometrically.
(e) Using the estimate obtained with the formula in (d), plot the fitted (estimated) regression function. On the graph locate the point $(\bar{x}, \bar{y})$. What do you observe?
(f) Using the estimates obtained with the formula in (d), obtain the least squares residuals, $\hat{e}_i = y_i - b_2 x_i$. Find their sum.
(g) Calculate $\sum x_i\hat{e}_i$.

2.5 A small business hires a consultant to predict the value of weekly sales of their product if their weekly advertising is increased to $750 per week. The consultant takes a record of how much the firm spent on advertising per week and the corresponding weekly sales over the past six months. The consultant writes "Over the past six months the average weekly expenditure on advertising has been $500 and average weekly sales have been $10,000. Based on the results of a simple linear regression, I predict sales will be $12,000 if $750 per week is spent on advertising."
(a) What is the estimated simple regression used by the consultant to make this prediction?

(b) Sketch a graph of the estimated regression line. Locate the average weekly values on the graph.

2.6* A soda vendor at Louisiana State University football games observes that more sodas are sold the warmer the temperature at game time is. Based on 32 home games covering five years, the vendor estimates the relationship between soda sales and temperature to be $\hat{y} = -240 + 8x$, where y = the number of sodas she sells and x = temperature in degrees Fahrenheit,

(a) Interpret the estimated slope and intercept. Do the estimates make sense? Why, or why not?
(b) On a day when the temperature at game time is forecast to be 80°F, predict how many sodas the vendor will sell.
(c) Below what temperature are the predicted sales zero?
(d) Sketch a graph of the estimated regression line.

2.7 You have the results of a simple linear regression based on state-level data and the District of Columbia, a total of $N = 51$ observations.

(a) The estimated error variance $\hat{\sigma}^2 = 2.04672$. What is the sum of the squared least squares residuals?
(b) The estimated variance of b_2 is 0.00098. What is the standard error of b_2? What is the value of $\sum(x_i - \bar{x})^2$?
(c) Suppose the dependent variable y_i = the state's mean income (in thousands of dollars) of males who are 18 years of age or older and x_i the percentage of males 18 years or older who are high school graduates. If $b_2 = 0.18$, interpret this result.
(d) Suppose $\bar{x} = 69.139$ and $\bar{y} = 15.187$, what is the estimate of the intercept parameter?
(e) Given the results in (b) and (d), what is $\sum x_i^2$?
(f) For the state of Arkansas the value of $y_i = 12.274$ and the value of $x_i = 58.3$. Compute the least squares residual for Arkansas. (*Hint*: Use the information in parts (c) and (d).).

2.8◆ Professor E.Z. Stuff has decided that the least squares estimator is too much trouble. Noting that two points determine a line, Dr. Stuff chooses two points from a sample of size N and draws a line between them, calling the slope of this line the EZ estimator of β_2 in the simple regression model. Algebraically, if the two points are (x_1, y_1) and (x_2, y_2), the EZ estimation rule is

$$b_{EZ} = \frac{y_2 - y_1}{x_2 - x_1}$$

Assuming that all the assumptions of the simple regression model hold:

(a) Show that b_{EZ} is a "linear" estimator.
(b) Show that b_{EZ} is an unbiased estimator.
(c) Find the variance of b_{EZ}.
(d) Find the probability distribution of b_{EZ}.
(e) Convince Professor Stuff that the EZ estimator is not as good as the least squares estimator. No proof is required here.

2.10.2 Computer Exercises

2.9* The owners of a motel discovered that a defective product was used in its construction. It took seven months to correct the defects, during which 14 rooms in the

100-unit motel were taken out of service for 1 month at a time. The motel lost profits due to these closures, and the question of how to compute the losses was addressed by Adams (2008).[7] For this exercise use the data in *motel.dat*.

(a) The occupancy rate for the damaged motel is *MOTEL_PCT*, and the competitor occupancy rate is *COMP_PCT*. On the same graph, plot these variables against *TIME*. Which had the higher occupancy before the repair period? Which had the higher occupancy during the repair period?

(b) Plot *MOTEL_PCT* against *COMP_PCT*. Does there seem to be a relationship between these two variables? Explain why such a relationship might exist.

(c) Estimate a linear regression with $y = MOTEL\_PCT$ and $x = COMP\_PCT$. Discuss the result.

(d) Compute the least squares residuals from the regression results in (c). Plot these residuals against time. Does the model overpredict, underpredict, or accurately predict the motel's occupancy rate during the repair period?

(e) Consider a linear regression with $y = MOTEL\_PCT$ and $x = RELPRICE$, which is the ratio of the price per room charged by the motel in question relative to its competitors. What sign do you predict for the slope coefficient? Why? Does the sign of the estimated slope agree with your expectation?

(f) Consider the linear regression with $y = MOTEL\_PCT$ and $x = REPAIR$, which is an indicator variable, taking the value 1 during the repair period and 0 otherwise. Discuss the interpretation of the least squares estimates. Does the motel appear to have suffered a loss of occupancy, and therefore profits, during the repair period?

(g) Compute the average occupancy rate for the motel and competitors when the repairs were not being made (call these $\overline{MOTEL}_0$ and $\overline{COMP}_0$), and when they were being made ($\overline{MOTEL}_1$ and $\overline{COMP}_1$). During the nonrepair period, what was the difference between the average occupancies, $\overline{MOTEL}_0 - \overline{COMP}_0$? Does this comparison seem to support the motel's claims of lost profits during the repair period?

(h) Estimate a linear regression model with $y = MOTEL\_PCT – COMP\_PCT$ and $x = REPAIR$. How do the results of this regression relate to the result in part (g)?

2.10 The capital asset pricing model (CAPM) is an important model in the field of finance. It explains variations in the rate of return on a security as a function of the rate of return on a portfolio consisting of all publicly traded stocks, which is called the *market* portfolio. Generally the rate of return on any investment is measured relative to its opportunity cost, which is the return on a risk free asset. The resulting difference is called the *risk premium*, since it is the reward or punishment for making a risky investment. The CAPM says that the risk premium on security j is *proportional* to the risk premium on the market portfolio. That is,

$$r_j - r_f = \beta_j(r_m - r_f),$$

where r_j and r_f are the returns to security j and the risk-free rate, respectively, r_m is the return on the market portfolio, and β_j is the jth security's "*beta*" value. A stock's *beta* is important to investors since it reveals the stock's volatility. It measures the sensitivity of security j's return to variation in the whole stock market. As such, values of *beta* less than 1 indicate that the stock is "defensive" since its variation is

[7] A. Frank Adams (2008) "When a 'Simple' Analysis Won't Do: Applying Economic Principles in a Lost Profits Case," *The Value Examiner*, May/June 2008, 22–28. The authors thank Professor Adams for the use of his data.

less than the market's. A *beta* greater than 1 indicates an "aggressive stock." Investors usually want an estimate of a stock's *beta* before purchasing it. The CAPM model shown above is the "economic model" in this case. The "econometric model" is obtained by including an intercept in the model (even though theory says it should be zero) and an error term,

$$r_j - r_f = \alpha_j + \beta_j(r_m - r_f) + e$$

(a) Explain why the econometric model above is a simple regression model like those discussed in this chapter.
(b) In the data file *capm4.dat* are data on the monthly returns of six firms (Microsoft, GE, GM, IBM, Disney, and Mobil-Exxon), the rate of return on the market portfolio (*MKT*), and the rate of return on the risk free asset (*RISKFREE*). The 132 observations cover January 1998 to December 2008. Estimate the CAPM model for each firm, and comment on their estimated *beta* values. Which firm appears most aggressive? Which firm appears most defensive?
(c) Finance theory says that the intercept parameter α_j should be zero. Does this seem correct given your estimates? For the Microsoft stock, plot the fitted regression line along with the data scatter.
(d) Estimate the model for each firm under the assumption that $\alpha_j = 0$. Do the estimates of the *beta* values change much?

2.11 The file *br2.dat* contains data on 1080 houses sold in Baton Rouge, Louisiana, during mid-2005. The data include sale price, the house size in square feet, its age, whether it has a pool or fireplace or is on the waterfront. Also included is an indicator variable *TRADITIONAL* indicating whether the house style is traditional or not.[8] Variable descriptions are in the file *br2.def*.
(a) Plot house price against house size for houses with traditional style.
(b) For the traditional-style houses estimate the linear regression model $PRICE = \beta_1 + \beta_2 SQFT + e$. Interpret the estimates. Draw a sketch of the fitted line.
(c) For the traditional-style houses estimate the quadratic regression model $PRICE = \alpha_1 + \alpha_2 SQFT^2 + e$. Compute the marginal effect of an additional square foot of living area in a home with 2000 square feet of living space. Compute the elasticity of *PRICE* with respect to *SQFT* for a home with 2000 square feet of living space. Graph the fitted line. On the graph, sketch the line that is tangent to the curve for a 2000-square-foot house.
(d) For the regressions in (b) and (c) compute the least squares residuals and plot them against *SQFT*. Do any of our assumptions appear violated?
(e) One basis for choosing between these two specifications is how well the data are fit by the model. Compare the sum of squared residuals (*SSE*) from the models in (b) and (c). Which model has a lower *SSE*? How does having a lower *SSE* indicate a "better-fitting" model?
(f) For the traditional-style houses estimate the log-linear regression model $\ln(PRICE) = \gamma_1 + \gamma_2 SQFT + e$. Interpret the estimates. Graph the fitted line, and sketch the tangent line to the curve for a house with 2000 square feet of living area.

[8] The data file *br.dat* offers a wider range of style listings. Try this data set for a more detailed investigation of the effect of style.

(g) How would you compute the sum of squared residuals for the model in (f) to make it comparable to those from the models in (b) and (c)? Compare this sum of squared residuals to the *SSE* from the linear and quadratic specifications. Which model seems to fit the data best?

2.12* The file *stockton4.dat* contains data on 1500[9] houses sold in Stockton, CA during 1996–1998. Variable descriptions are in the file *stockton4.def*.

(a) Plot house selling price against house living area for all houses in the sample.
(b) Estimate the regression model $SPRICE = \beta_1 + \beta_2 LIVAREA + e$ for all the houses in the sample. Interpret the estimates. Draw a sketch of the fitted line.
(c) Estimate the quadratic model $SPRICE = \alpha_1 + \alpha_2 LIVAREA^2 + e$ for all the houses in the sample. What is the marginal effect of an additional 100 square feet of living area for a home with 1500 square feet of living area?
(d) In the same graph, plot the fitted lines from the linear and quadratic models. Which seems to fit the data better? Compare the sum of squared residuals (*SSE*) for the two models. Which is smaller?
(e) Estimate the regression model in (c) using only houses that are on large lots. Repeat the estimation for houses that are not on large lots. Interpret the estimates. How do the estimates compare?
(f) Plot house selling price against *AGE*. Estimate the linear model $SPRICE = \delta_1 + \delta_2 AGE + e$. Interpret the estimated coefficients. Repeat this exercise using the log-linear model $\ln(SPRICE) = \theta_1 + \theta_2 AGE + e$. Based on the plots and visual fit of the estimated regression lines, which of these two models would you prefer? Explain.
(g) Estimate a linear regression $SPRICE = \eta_1 + \eta_2 LGELOT + e$ with dependent variable *SPRICE* and independent variable the indicator *LGELOT* which identifies houses on larger lots. Interpret these results.

2.13 A longitudinal experiment was conducted in Tennessee beginning in 1985 and ending in 1989. A single cohort of students was followed from kindergarten through third grade. In the experiment children were randomly assigned within schools into three types of classes: small classes with 13–17 students, regular-sized classes with 22–25 students, and regular-sized classes with a full-time teacher aide to assist the teacher. Student scores on achievement tests were recorded as well as some information about the students, teachers, and schools. Data for the kindergarten classes are contained in the data file *star.dat*.

(a) Using children who are in either a regular-sized class or a small class, estimate the regression model explaining students' combined aptitude scores as a function of class size, $TOTALSCORE_i = \beta_1 + \beta_2 SMALL_i + e_i$. Interpret the estimates. Based on this regression result, what do you conclude about the effect of class size on learning?
(b) Repeat part (a) using dependent variables *READSCORE* and *MATHSCORE*. Do you observe any differences?
(c) Using children who are in either a regular-sized class or a regular-sized class with a teacher aide, estimate the regression model explaining student's combined aptitude scores as a function of the presence of a teacher aide, $TOTALSCORE = \gamma_1 + \gamma_2 AIDE + e$. Interpret the estimates. Based on this

[9] The data set *stockton3.dat* has 2,610 observations on these same variables.

regression result, what do you conclude about the effect on learning of adding a teacher aide to the classroom?

(d) Repeat part (c) using dependent variables *READSCORE* and *MATHSCORE*. Do you observe any differences?

2.14* Professor Ray C. Fair has for a number of years built and updated models that explain and predict the U.S. presidential elections. Visit his website at http://fairmodel.econ.yale.edu/vote2004/index2.htm. See in particular his paper entitled "A Vote Equation for the 2004 Election." The basic premise of the model is that the incumbent party's share of the two-party [Democratic and Republican] popular vote [incumbent means the party in power at the time of the election] is affected by a number of factors relating to the economy, and variables relating to the politics, such as how long the incumbent party has been in power, and whether the President is running for re-election. Fair's data, 33 observations for the election years from 1880 to 2008, are in the file *fair4.dat*. The dependent variable is $VOTE =$ percentage share of the popular vote won by the incumbent party. Consider the explanatory variable $GROWTH =$ growth rate in real per capita GDP in the first three quarters of the election year (annual rate). One would think that if the economy is doing well, and growth is high, the party in power would have a better chance of winning the election.

(a) Using the data for 1916–2008, plot a scatter diagram of *VOTE* against *GROWTH*. Does there appear to be positive association?

(b) Estimate the regression $VOTE = \beta_1 + \beta_2 GROWTH + e$ by least squares using the data from 1916 to 2008. Report and discuss the estimation result. Sketch, **by hand**, the fitted line on the data scatter from (a).

(c) Fit the regression in (b) using the data from 1916 to 2004. Predict the *VOTE* share for the incumbent party based on the actual 2008 value for *GROWTH*. How does the predicted vote for 2008 compare to the actual result?

(d) Economywide inflation may spell doom for the incumbent party in an election. The variable *INFLATION* is the growth in prices over the first 15 quarters of an administration. Using the data from 1916 to 2008, plot *VOTE* against *INFLATION*. Using the same sample, report and discuss the estimation results for the model $VOTE = \alpha_1 + \alpha_2 INFLATION + e$.

3.7.1 Problems

3.1 Using the regression output for the food expenditure model shown in Figure 2.9:

(a) Construct a 95% interval estimate for β_1 and interpret.

(b) Test the null hypothesis that β_1 is zero against the alternative that it is not at the 5% level of significance without using the reported *p*-value. What is your conclusion?

(c) Draw a sketch showing the *p*-value 0.0622 shown in Figure 2.9, the critical value from the *t*-distribution used in (b), and how the *p*-value could have been used to answer (b).

(d) Test the null hypothesis that β_1 is zero against the alternative that it is positive at the 5% level of significance. Draw a sketch of the rejection region and compute the *p*-value. What is your conclusion?

(e) Explain the differences and similarities between the "level of significance" and the "level of confidence."
(f) The results in (d) show that we are 95% confident that β_1 is positive. True, or false? If false, explain.

3.2 The general manager of an engineering firm wants to know whether a technical artist's experience influences the quality of his or her work. A random sample of 24 artists is selected and their years of work experience and quality rating (as assessed by their supervisors) recorded. Work experience (*EXPER*) is measured in years and quality rating (*RATING*) takes a value of 1 through 7, with 7 = excellent and 1 = poor. The simple regression model $RATING = \beta_1 + \beta_2 EXPER + e$ is proposed. The least squares estimates of the model, and the standard errors of the estimates, are

$$\begin{array}{cccc} \widehat{RATING} & = & 3.204 & + \; 0.076\,EXPER \\ \text{(se)} & & (0.709) & (0.044) \end{array}$$

(a) Sketch the estimated regression function. Interpret the coefficient of *EXPER*.
(b) Construct a 95% confidence interval for β_2, the slope of the relationship between quality rating and experience. In what are you 95% confident?
(c) Test the null hypothesis that β_2 is zero against the alternative that it is not using a two-tail test and the $\alpha = 0.05$ level of significance. What do you conclude?
(d) Test the null hypothesis that β_2 is zero against the one-tail alternative that it is positive at the $\alpha = 0.05$ level of significance. What do you conclude?
(e) For the test in part (c), the p-value is 0.0982. If we choose the probability of a Type I error to be $\alpha = 0.05$, do we reject the null hypothesis, or not, just based on an inspection of the p-value? Show, in a diagram, how this p-value is computed.

3.3* In an estimated simple regression model, based on 24 observations, the estimated slope parameter is 0.310 and the estimated standard error is 0.082.
(a) Test the hypothesis that the slope is zero against the alternative that it is not, at the 1% level of significance.
(b) Test the hypothesis that the slope is zero against the alternative that it is positive at the 1% level of significance.
(c) Test the hypothesis that the slope is zero against the alternative that it is negative at the 5% level of significance. Draw a sketch showing the rejection region.
(d) Test the hypothesis that the estimated slope is 0.5, against the alternative that it is not, at the 5% level of significance.
(e) Obtain a 99% interval estimate of the slope.

3.4 Consider a simple regression in which the dependent variable *MIM* = mean income of males who are 18 years of age or older, in thousands of dollars. The explanatory variable *PMHS* = percent of males 18 or older who are high school graduates. The data consist of 51 observations on the 50 states plus the District of Columbia. Thus *MIM* and *PMHS* are "state averages." The estimated regression, along with standard errors and t-statistics, is

$$\begin{array}{cccc} \widehat{MIM} = & (a) & + & 0.180PMHS \\ \text{(se)} & (2.174) & & (b) \\ (t) & (1.257) & & (5.754) \end{array}$$

(a) What is the estimated equation intercept? Show your calculation. Sketch the estimated regression function.

(b) What is the standard error of the estimated slope? Show your calculation.
(c) What is the *p*-value for the two-tail test of the hypothesis that the equation intercept is zero? Draw a sketch to illustrate.
(d) State the economic interpretation of the estimated slope. Is the sign of the coefficient what you would expect from economic theory?
(e) Construct a 99% confidence interval estimate of the slope of this relationship.
(f) Test the hypothesis that the slope of the relationship is 0.2 against the alternative that it is not. State in words the meaning of the null hypothesis in the context of this problem.

3.7.2 Computer Exercises

3.5 A life insurance company wishes to examine the relationship between the amount of life insurance held by a family and family income. From a random sample of 20 households, the company collected the data in the file *insur.dat*. The data are in units of thousands of dollars.

(a) Estimate the linear regression with dependent variable *INSURANCE* and independent variable *INCOME*. Write down the fitted model and draw a sketch of the fitted function. Identify the estimated slope and intercept on the sketch. Locate the point of the means on the plot.
(b) Discuss the relationship you estimated in (a). In particular,
  (i) What is your estimate of the resulting change in the amount of life insurance when income increases by $1,000?
  (ii) What is the standard error of the estimate in (i), and how do you use this standard error for interval estimation and hypothesis testing?
(c) One member of the management board claims that for every $1,000 increase in income, the amount of life insurance held will go up by $5,000. Choose an alternative hypothesis and explain your choice. Does your estimated relationship support this claim? Use a 5% significance level.
(d) Test the hypothesis that as income increases the amount of life insurance increases by the same amount. That is, test the hypothesis that the slope of the relationship is one.
(e) Write a short report (200–250 words) summarizing your findings about the relationship between income and the amount of life insurance held.

3.6* In Exercise 2.9 we considered a motel that had discovered that a defective product was used during construction. It took seven months to correct the defects, during which approximately 14 rooms in the 100-unit motel were taken out of service for one month at a time. The data are in *motel.dat*.

(a) In the linear regression model $MOTEL\_PCT = \beta_1 + \beta_2 COMP\_PCT + e$, test the null hypothesis $H_0 : \beta_2 \leq 0$ against the alternative hypothesis $H_0 : \beta_2 > 0$ at the $\alpha = 0.01$ level of significance. Discuss your conclusion. Include in your answer a sketch of the rejection region and a calculation of the *p*-value.
(b) Consider a linear regression with $y = MOTEL\_PCT$ and $x = RELPRICE$, which is the ratio of the price per room charged by the motel in question relative to its competitors. Test the null hypothesis that there is no relationship between these variables against the alternative that there is an inverse relationship between them, at the $\alpha = 0.01$ level of significance. Discuss your conclusion. Include in your answer a sketch of the rejection region, and a calculation of the *p*-value. In this exercise follow and **show** all the test procedure steps suggested in Chapter 3.4.

(c) Consider the linear regression $MOTEL\_PCT = \delta_1 + \delta_2 REPAIR + e$, where *REPAIR* is an indicator variable taking the value 1 during the repair period and 0 otherwise. Test the null hypothesis $H_0 : \delta_2 \geq 0$ against the alternative hypothesis $H_1 : \delta_2 < 0$ at the $\alpha = 0.05$ level of significance. Explain the logic behind stating the null and alternative hypotheses in this way. Discuss your conclusions.

(d) Using the model given in part (c), construct a 95% interval estimate for the parameter δ_2 and give its interpretation. Have we estimated the effect of the repairs on motel occupancy relatively precisely, or not? Explain.

(e) Consider the linear regression model with $y = MOTEL\_PCT - COMP\_PCT$ and $x = REPAIR$, that is $(MOTEL\_PCT - COMP\_PCT) = \gamma_1 + \gamma_2 REPAIR + e$. Test the null hypothesis that $\gamma_2 = 0$ against the alternative that $\gamma_2 < 0$ at the $\alpha = 0.01$ level of significance. Discuss the meaning of the test outcome.

(f) Using the model in part (e), construct and discuss the 95% interval estimate of γ_2.

3.7 Consider the capital asset pricing model (CAPM) in Exercise 2.10. Use the data in *capm4.dat* to answer each of the following:

(a) Test at the 5% level of significance the hypothesis that each stock's "*beta*" value is 1 against the alternative that it is not equal to 1. What is the economic interpretation of a *beta* equal to 1?

(b) Test at the 5% level of significance the null hypothesis that Mobil-Exxon's "*beta*" value is greater than or equal to 1 against the alternative that it is less than 1. What is the economic interpretation of a *beta* less than 1?

(c) Test at the 5% level of significance the null hypothesis that Microsoft's "*beta*" value is less than or equal to 1 against the alternative that it is greater than 1. What is the economic interpretation of a *beta* more than 1?

(d) Construct a 95% interval estimate of Microsoft's "*beta*." Assume that you are a stockbroker. Explain this result to an investor who has come to you for advice.

(e) Test (at a 5% significance level) the hypothesis that the intercept term in the CAPM model for each stock is zero, against the alternative that it is not. What do you conclude?

3.8 The file *br2.dat* contains data on 1080 houses sold in Baton Rouge, Louisiana during mid-2005. The data include sale price and the house size in square feet. Also included is an indicator variable *TRADITIONAL* indicating whether the house style is traditional or not.

(a) For the traditional-style houses estimate the linear regression model $PRICE = \beta_1 + \beta_2 SQFT + e$. Test the null hypothesis that the slope is zero against the alternative that it is positive, using the $\alpha = 0.01$ level of significance. Follow and show all the test steps described in Chapter 3.4.

(b) Using the linear model in (a), test the null hypothesis (H_0) that the expected price of a house of 2000 square feet is equal to, or less than, \$120,000. What is the appropriate alternative hypothesis? Use the $\alpha = 0.01$ level of significance. Obtain the *p*-value of the test and show its value on a sketch. What is your conclusion?

(c) Based on the estimated results from part (a), construct a 95% interval estimate of the expected price of a house of 2000 square feet.

(d) For the traditional-style houses, estimate the quadratic regression model $PRICE = \alpha_1 + \alpha_2 SQFT^2 + e$. Test the null hypothesis that the marginal effect of an additional square foot of living area in a home with 2000 square feet of living space is \$75 against the alternative that the effect is less than \$75. Use the

$\alpha = 0.01$ level of significance. Repeat the same test for a home of 4000 square feet of living space. Discuss your conclusions.

(e) For the traditional-style houses, estimate the log-linear regression model $\ln(PRICE) = \gamma_1 + \gamma_2 SQFT + e$. Test the null hypothesis that the marginal effect of an additional square foot of living area in a home with 2000 square feet of living space is \$75 against the alternative that the effect is less than \$75. Use the $\alpha = 0.01$ level of significance. Repeat the same test for a home of 4000 square feet of living space. Discuss your conclusions.

3.9* Reconsider the presidential voting data (*fair4.dat*) introduced in Exercise 2.14. Use the data from 1916 to 2008 for this exercise.

(a) Using the regression model $VOTE = \beta_1 + \beta_2 GROWTH + e$, test (at a 5% significance level) the null hypothesis that economic growth has no effect on the percentage vote earned by the incumbent party. Select an alternative hypothesis and a rejection region. Explain your choice.

(b) Using the regression model in part (a), construct a 95% interval estimate for β_2, and interpret.

(c) Using the regression model $VOTE = \beta_1 + \beta_2 INFLATION + e$, test the null hypothesis that inflation has no effect on the percentage vote earned by the incumbent party. Select an alternative hypothesis, a rejection region, and a significance level. Explain your choice.

(d) Using the regression model in part (c), construct a 95% interval estimate for β_2, and interpret.

(e) Test the null hypothesis that if $INFLATION = 0$ the expected vote in favor of the incumbent party is 50%, or more. Select the appropriate alternative. Carry out the test at the 5% level of significance. Discuss your conclusion.

(f) Construct a 95% interval estimate of the expected vote in favor of the incumbent party if $INFLATION = 2\%$. Discuss the interpretation of this interval estimate.

3.10 Reconsider Exercise 2.13, which was based on the experiment with small classes for primary school students conducted in Tennessee beginning in 1985. Data for the kindergarten classes is contained in the data file *star.dat*.

(a) Using children who are in either a regular-sized class or a small class, estimate the regression model explaining students' combined aptitude scores as a function of class size, $TOTALSCORE = \beta_1 + \beta_2 SMALL + e$. Test the null hypothesis that β_2 is zero, or negative, against the alternative that this coefficient is positive. Use the 5% level of significance. Compute the p-value of this test, and show its value in a sketch. Discuss the social importance of this finding.

(b) For the model in part (a), construct a 95% interval estimate of β_2 and discuss.

(c) Repeat part (a) using dependent variables *READSCORE* and *MATHSCORE*. Do you observe any differences?

(d) Using children who are in either a regular-sized class or a regular-sized class with a teacher aide, estimate the regression model explaining students' combined aptitude scores as a function of the presence or absence of a teacher aide, $TOTALSCORE = \gamma_1 + \gamma_2 AIDE + e$. Test the null hypothesis that γ_2 is zero or negative against the alternative that this coefficient is positive. Use the 5% level of significance. Discuss the importance of this finding.

(e) For the model in part (d), construct a 95% interval estimate of γ_2 and discuss.

(f) Repeat part (d) using dependent variables *READSCORE* and *MATHSCORE*. Do you observe any differences?

3.11 How much does experience affect wage rates? The data file *cps4_small.dat* contains 1000 observations on hourly wage rates, experience and other variables from the 2008 Current Population Survey (CPS).

(a) Estimate the linear regression $WAGE = \beta_1 + \beta_2 EXPER + e$ and discuss the results. Using your software plot a scatter diagram with *WAGE* on the vertical axis and *EXPER* on the horizontal axis. Sketch in by hand, or using your software, the fitted regression line.

(b) Test the statistical significance of the estimated slope of the relationship at the 5% level. Use a one-tail test.

(c) Repeat part (a) for the sub-samples consisting of (i) females, (ii) males, (iii) blacks, and (iv) white males. What differences, if any, do you notice?

(d) For each of the estimated regression models in (a) and (c), calculate the least squares residuals and plot them against *EXPER*. Are any patterns evident?

3.12 Is the relationship between experience and wages constant over one's lifetime? To investigate we will fit a quadratic model using the data file *cps4_small.dat*, which contains 1,000 observations on hourly wage rates, experience and other variables from the 2008 Current Population Survey (CPS).

(a) Create a new variable called $EXPER30 = EXPER - 30$. Construct a scatter diagram with *WAGE* on the vertical axis and *EXPER30* on the horizontal axis. Are any patterns evident?

(b) Estimate by least squares the quadratic model $WAGE = \gamma_1 + \gamma_2 (EXPER30)^2 + e$. Are the coefficient estimates statistically significant? Test the null hypothesis that $\gamma_2 \geq 0$ against the alternative that $\gamma_2 < 0$ at the $\alpha = 0.05$ level of significance. What conclusion do you draw?

(c) Using the estimation in part (b), compute the estimated marginal effect of experience upon wage for a person with 10 years' experience, 30 years' experience, and 50 years' experience. Are these slopes significantly different from zero at the $\alpha = 0.05$ level of significance?

(d) Construct 95% interval estimates of each of the slopes in part (c). How precisely are we estimating these values?

(e) Using the estimation result from part (b) create the fitted values $\widehat{WAGE} = \hat{\gamma}_1 + \hat{\gamma}_2 (EXPER30)^2$, where the ^ denotes least squares estimates. Plot these fitted values and *WAGE* on the vertical axis of the same graph against *EXPER30* on the horizontal axis. Are the estimates in part (c) consistent with the graph?

(f) Estimate the linear regression $WAGE = \beta_1 + \beta_2 EXPER30 + e$ and the linear regression $WAGE = \alpha_1 + \alpha_2 EXPER + e$. What differences do you observe between these regressions and why do they occur? What is the estimated marginal effect of experience on wage from these regressions? Based on your work in parts (b)–(d), is the assumption of constant slope in this model a good one? Explain.

(g) Use the larger data *cps4.dat* (4838 observations) to repeat parts (b), (c), and (d). How much has the larger sample improved the precision of the interval estimates in part (d)?

3.13* Is the relationship between experience and ln(wages) constant over one's lifetime? To investigate we will fit a log-linear model using the data file *cps4_small.dat*, which contains 1000 observations on hourly wage rates, experience and other variables from the 2008 Current Population Survey (CPS).

(a) Create a new variable called $EXPER30 = EXPER - 30$. Construct a scatter diagram with $\ln(WAGE)$ on the vertical axis and $EXPER30$ on the horizontal axis. Are any patterns evident?

(b) Estimate by least squares the quadratic model $\ln(WAGE) = \gamma_1 + \gamma_2(EXPER30)^2 + e$. Are the coefficient estimates statistically significant? Test the null hypothesis that $\gamma_2 \geq 0$ against the alternative that $\gamma_2 < 0$ at the $\alpha = 0.05$ level of significance. What conclusion do you draw?

(c) Using the estimation in part (b), compute the estimated marginal effect of experience upon wage for a person with 10 years of experience, 30 years of experience, and 50 years of experience. [Hint: If $\ln(y) = a + bx^2$ then $y = \exp(a + bx^2)$, and $dy/dx = \exp(a + bx^2) \times 2bx = 2bxy$]

(d) Using the estimation result from part (b) create the fitted values $\widehat{WAGE} = \exp\left(\hat{\gamma}_1 + \hat{\gamma}_2(EXPER30)^2\right)$, where the ^ denotes least squares estimates. Plot these fitted values and $WAGE$ on the vertical axis of the same graph against $EXPER30$ on the horizontal axis. Are the estimates in part (c) consistent with the graph?

3.14 Data on the weekly sales of a major brand of canned tuna by a supermarket chain in a large midwestern U.S. city during a mid-1990s calendar year are contained in the file *tuna.dat*. There are 52 observations on the variables. The variable $SAL1$ = unit sales of brand no. 1 canned tuna, $APR1$ = price per can of brand no. 1 canned tuna, $APR2$, $APR3$ = price per can of brands nos. 2 and 3 of canned tuna.

(a) Create the relative price variables $RPRICE2 = APR1/APR2$ and $RPRICE3 = APR1/APR3$. What do you anticipate the relationship between sales ($SAL1$) and the relative price variables to be? Explain your reasoning.

(b) Estimate the log-linear model $\ln(SAL1) = \beta_1 + \beta_2 RPRICE2 + e$. Interpret the estimate of β_2. Construct and interpret a 95% interval estimate of the parameter.

(c) Test the null hypothesis that the slope of the relationship in (b) is zero. Create the alternative hypothesis based on your answer to part (a). Use the 1% level of significance and draw a sketch of the rejection region. Is your result consistent with economic theory?

(d) Estimate the log-linear model $\ln(SAL1) = \gamma_1 + \gamma_2 RPRICE3 + e$. Interpret the estimate of γ_2. Construct and interpret a 95% interval estimate of the parameter.

(e) Test the null hypothesis that the slope of this relationship is zero. Create the alternative hypothesis based on your answer to part (a). Use the 1% level of significance and draw a sketch of the rejection region. Is your result consistent with economic theory?

3.15 What is the relationship between crime and punishment? This important question has been examined by Cornwell and Trumbull[2] using a panel of data from North Carolina. The cross sections are 90 counties, and the data are annual for the years 1981–1987. The data are in the file *crime.dat*.

(a) Using the data from 1987, estimate the log-linear regression relating the log of the crime rate to the probability of an arrest, $LCRMRTE = \beta_1 + \beta_2 PRBARR + e$. The probability of arrest is measured as the ratio of arrests to offenses. If we increase the probability of arrest by 10%, what will be the effect on the crime rate? What is a 95% interval estimate of this quantity?

[2] "Estimating the Economic Model of Crime with Panel Data," *Review of Economics and Statistics*, 76, 1994, 360–366. The data were kindly provided by the authors.

(b) Test the null hypothesis that there is no relationship between the crime rate and the probability of arrest against the alternative that there is an inverse relationship. Use the 1% level of significance.

(c) Repeat parts (a) and (b) using the probability of conviction (*PRBCONV*) as the explanatory variable. The probability of conviction is measured as the ratio of convictions to arrests.

4.7.1 Problems

4.1* (a) Supposing that a simple regression has quantities $\sum(y_i - \bar{y})^2 = 631.63$ and $\sum\hat{e}_i^2 = 182.85$, find R^2.

(b) Suppose that a simple regression has quantities $N = 20$, $\sum y_i^2 = 5930.94$, $\bar{y} = 16.035$, and $SSR = 666.72$, find R^2.

(c) Suppose that a simple regression has quantities $R^2 = 0.7911$, $SST = 552.36$, and $N = 20$, find $\hat{\sigma}^2$.

4.2* Consider the following estimated regression equation (standard errors in parentheses):

$$\hat{y} = 5.83 + 0.869x \quad R^2 = 0.756$$
$$\text{(se)} \ (1.23) \ (0.117)$$

Rewrite the estimated equation that would result if
(a) All values of x were divided by 20 before estimation
(b) All values of y were divided by 50 before estimation
(c) All values of y and x were divided by 20 before estimation

4.3 Using the data in Exercise 2.1 and only a calculator (show your work) compute
(a) The predicted value of y for $x_0 = 4$
(b) The se(f) corresponding to part (a)
(c) A 95% prediction interval for y given $x_0 = 4$
(d) A 95% prediction interval for y given $x = \bar{x}$. Compare the width of this interval to the one computed in part (c)

4.4 The general manager of an engineering firm wants to know whether a technical artist's experience influences the quality of his or her work. A random sample of 50 artists is selected and their years of work experience and quality rating (as assessed by their supervisors) recorded. Work experience (*EXPER*) is measured in years and quality rating (*RATING*) takes a value in the interval one to four, with 4 = very good and 1 = very poor. Two models are estimated by least squares. The estimates and standard errors are

Model 1 :

$$\widehat{RATING} = 3.4464 - 0.001459(EXPER - 35)^2 \quad N = 50$$
$$\text{(se)} \qquad (0.0375) \ (0.0000786)$$

Model 2 :

$$\widehat{RATING} = 1.4276 + 0.5343\ln(EXPER) \quad N = 49$$
$$\text{(se)} \qquad (0.1333) \ (0.0433)$$

(a) For each model, sketch the estimated regression function for *EXPER* = 10 to 40 years.
(b) Using each model, predict the rating of a worker with 10 years' experience.
(c) Using each model, find the marginal effect of another year of experience on the expected worker rating for a worker with 10 years' experience.
(d) Using each model, construct a 95% interval estimate for the marginal effect found in (c). Note that Model 2 has one fewer observations due to 1 worker having *EXPER* = 0.

4.5 Suppose you are estimating a simple linear regression model.
(a) If you multiply all the x values by 20, but not the y values, what happens to the parameter values β_1 and β_2? What happens to the least squares estimates b_1 and b_2? What happens to the variance of the error term?
(b) Suppose you are estimating a simple linear regression model. If you multiply all the y values by 50, but not the x values, what happens to the parameter values

β_1 and β_2? What happens to the least squares estimates b_1 and b_2? What happens to the variance of the error term?

4.6 The fitted least squares line is $\hat{y}_i = b_1 + b_2 x_i$.
(a) Algebraically, show that the fitted line passes through the point of the means, $(\bar{x}, \bar{y})$.
(b) Algebraically show that the average value of $\hat{y}_i$ equals the sample average of y. That is, show that $\bar{\hat{y}} = \bar{y}$, where $\bar{\hat{y}} = \Sigma\hat{y}_i/N$.

4.7 In a simple linear regression model suppose we know that the intercept parameter is zero, so the model is $y_i = \beta_2 x_i + e_i$. The least squares estimator of β_2 is developed in Exercise 2.4.
(a) What is the least squares predictor of y in this case?
(b) When an intercept is not present in a model, R^2 is often defined to be $R_u^2 = 1 - SSE/\Sigma y_i^2$, where SSE is the usual sum of squared residuals. Compute R_u^2 for the data in Exercise 2.4.
(c) Compare the value of R_u^2 in part (b) to the generalized $R^2 = r_{y\hat{y}}^2$, where $\hat{y}$ is the predictor based on the restricted model in part (a).
(d) Compute $SST = \Sigma(y_i - \bar{y})^2$ and $SSR = \Sigma(\hat{y}_i - \bar{y})^2$, where $\hat{y}$ is the predictor based on the restricted model in part (a). Does the sum of squares decomposition $SST = SSR + SSE$ hold in this case?

4.7.2 Computer Exercises

4.8 The first three columns in the file *wa_wheat.dat* contain observations on wheat yield in the Western Australian shires Northampton, Chapman Valley, and Mullewa, respectively. There are 48 annual observations for the years 1950–1997. For the Chapman Valley shire, consider the three equations

$$y_t = \beta_1 + \beta_2 t + e_t$$
$$y_t = \alpha_1 + \alpha_2 \ln(t) + e_t$$
$$y_t = \gamma_1 + \gamma_2 t^2 + e_t$$

(a) Using data from 1950–1996, estimate each of the three equations.
(b) Taking into consideration (i) plots of the fitted equations, (ii) plots of the residuals, (iii) error normality tests, and (iv) values for R^2, which equation do you think is preferable? Explain.

4.9* For each of the three functions in Exercise 4.8
(a) Find the predicted value and a 95% prediction interval for yield when $t = 48$. Is the actual value within the prediction interval?
(b) Find estimates of the slopes dy_t/dt at the point $t = 48$.
(c) Find estimates of the elasticities $(dy_t/dt)(t/y_t)$ at the point $t = 48$.
(d) Comment on the estimates you obtained in parts (b) and (c). What is their importance?

4.10 The file *london.dat* is a cross section of 1519 households drawn from the 1980–1982 British Family Expenditure Surveys. Data have been selected to include only households with one or two children living in Greater London. Self-employed and retired households have been excluded. Variable definitions are in the file *london.def*. The budget share of a commodity, say food, is defined as

$$WFOOD = \frac{\text{expenditure on food}}{\text{total expenditure}}$$

A functional form that has been popular for estimating expenditure functions for commodities is

$$WFOOD = \beta_1 + \beta_2 \ln(TOTEXP) + e$$

(a) Estimate this function for households with one child and households with two children. Report and comment on the results. (You may find it more convenient to use the files *lon1.dat* and *lon2.dat* that contain the data for the one and two children households, with 594 and 925 observations, respectively.)

(b) It can be shown that the expenditure elasticity for food is given by

$$\varepsilon = \frac{\beta_1 + \beta_2[\ln(TOTEXP) + 1]}{\beta_1 + \beta_2 \ln(TOTEXP)}$$

Find estimates of this elasticity for one- and two-child households, evaluated at average total expenditure in each case. Do these estimates suggest food is a luxury or a necessity? (*Hint*: Are the elasticities greater than one or less than one?)

(c) Analyze the residuals from each estimated function. Does the functional form seem appropriate? Is it reasonable to assume that the errors are normally distributed?

(d) Using the data on households with two children, *lon2.dat*, estimate budget share equations for fuel (*WFUEL*) and transportation (*WTRANS*). For each equation discuss the estimate of β_2 and carry out a two-tail test of statistical significance.

(e) Using the regression results from part (d), compute the elasticity ε for fuel and transportation first at the median of total expenditure (90), and then at the 95th percentile of total income (180). What differences do you observe? Are any differences you observe consistent with economic reasoning?

4.11* Reconsider the presidential voting data (*fair4.dat*) introduced in Exercises 2.14 and 3.9.

(a) Using the data from 1916 to 2008, estimate the regression model $VOTE = \beta_1 + \beta_2 GROWTH + e$. Based on these estimates, what is the predicted value of *VOTE* in 2008? What is the least squares residual for the 2008 election observation?

(b) Estimate the regression in (a) using the data from 1916–2004. Predict the value of *VOTE* in 2008 using the actual value of *GROWTH* for 2008, which was 0.22%. What is the prediction error in this forecast? Is it larger or smaller than the error computed in part (a)?

(c) Using the regression results from (b), construct a 95% prediction interval for the 2008 value of *VOTE* using the actual value of $GROWTH = 0.22\%$. Is the actual 2008 outcome within the prediction interval?

(d) Using the estimation results in (b), what value of *GROWTH* would have led to a prediction that the incumbent party [Republicans] would have won 50.1% of the vote?

4.12 In Chapter 4.6 we considered the demand for edible chicken, which the U.S. Department of Agriculture calls “broilers.” The data for this exercise are in the file *newbroiler.dat*.

(a) Using the 52 annual observations, 1950–2001, estimate the reciprocal model $Q = \alpha_1 + \alpha_2(1/P) + e$. Plot the fitted value of Q = per capita consumption of chicken, in pounds, versus P = real price of chicken. How well does the estimated relation fit the data?

(b) Using the estimated relation in part (a), compute the elasticity of per capita consumption with respect to real price when the real price is its median, \$1.31, and quantity is taken to be the corresponding value on the fitted curve. [Hint: The derivative (slope) of reciprocal model $y = a + b(1/x)$ is $dy/dx = -b(1/x^2)$]. Compare this estimated elasticity to the estimate found in Chapter 4.6 where the log-log functional form was used.

(c) Estimate the poultry demand using the linear-log functional form $Q = \gamma_1 + \gamma_2 \ln(P) + e$. Plot the fitted values of Q = per capita consumption of chicken, in pounds, versus P = real price of chicken. How well does the estimated relation fit the data?

(d) Using the estimated relation in part (c), compute the elasticity of per capita consumption with respect to real price when the real price is its median, \$1.31. Compare this estimated elasticity to the estimate from the log-log model and from the reciprocal model in part (b).

(e) Evaluate the suitability of the log-log, linear-log, and reciprocal models for fitting the poultry consumption data. Which of them would you select as best, and why?

4.12* The file *stockton2.dat* contains data on 880 houses sold in Stockton, CA, during mid