Regression Analysis

Assignment 1B

Problem 2.3 on page 79. This exercise uses the ELECTRICITY dataset and covers identifying the response and predictor variables (page 36), scatterplots (see page 39 and note that the least squares line is another name for the simple linear regression line—see page 42), fitting a simple linear regression model (page 42), and a slope hypothesis test (page 54). Make sure you read part (e) very carefully before answering.

Problem 2.3. The **ELECTRICITY** data file contains data from the CIA 2010 World Factbook (available at https://www.cia.gov/library/publications/the-world-factbook/) on electricity consumption in billions of kilowatt-hours (*Elec*) and gross domestic product (GDP) in billions of dollars (*Gdp*) for the 30 most populous countries. Here, GDP is based on purchasing power parities to account for between-country differences in price levels. The data file can be used to investigate the claim that there is a straight-line relationship between electricity consumption and GDP. For the purposes of this problem, assume that increases in electricity consumption tend to respond to increases in GDP (rather than the other way around).

a. Say which variable should be the predictor variable (*X*) and which the response variable (*Y*) for this problem, and consider the simple linear regression model $E(Y) = b_0 + b_1X$. *Before looking at the data*, say whether the population value of b_1 would be positive or negative under the claim that electricity consumption tends to increase in response to increases in GDP?
b. Plot the data in a scatterplot (make sure you put the appropriate variables on each axis) [computer help #15]. Add the least squares line to the scatterplot [computer help #26]. Briefly describe the dominant pattern in the points on the scatterplot and their position in relation to the least squares line.
c. Identify the countries with the two highest values of *Gdp* [computer help #18], remove them from the dataset [computer help #19], and redraw the scatterplot and least squares line [computer help #15 and #26]. How does your visual impression of the scatterplot change?

   Hint: Focus more on what you see overall, not so much on individual points. In particular, think about whether the overall trend changes and whether the variation of the points about the line changes.

d. Fit the appropriate simple linear regression model to the dataset that has the countries with the two highest values of *Gdp* removed [computer help #25]. Do the results provide evidence that the claim of a positive relationship could be true for the remaining 28 countries?

   Hint: This problem is asking you to do a hypothesis test. Use a significance level of 5%, make sure you write down the null and alternative hypotheses, and conclude with a statement either supporting or contradicting the claim once you have written up the results of the hypothesis test.

e. Based on the graph from part (c) and the model from part (d), which country has a particularly *high* electricity consumption and which country has a particularly *low* consumption *relative to the dominant pattern for the remaining countries*? [computer help #18].

[10 points]

Problem 2.5 on page 80. This exercise uses the CARS2 dataset and covers model evaluation (see Section 2.3 on pages 45-59). For part (b), don't worry about writing a two-page report—just briefly summarize the results and conclusions of applying the three numerical methods and one graphical method from Section 2.3 (for the graphical method, consider Figure 2.6 on page 45). Part (c) repeats one of the numerical methods but asks you to go further by interpreting the result in a practically meaningful way in the context of your chosen model (page 47).

Problem 2.5 The **CARS2** data file contains information for 127 new U.S. front-wheel drive passenger cars for the 2011 model year. These data come from a larger dataset (obtained from www.fueleconomy.gov), which is analyzed more fully in a case study in Section 6.2.

a. Transform city miles per gallon into "city gallons per hundred miles." In other words, create a new variable called *Cgphm* equal to 100/*Cmpg* [computer help #6]. To check that you've done the transformation correctly, confirm that the sample mean of *Cgphm* is 4.613.
b. Consider predicting *Cgphm* from either engine size in liters (variable *Eng*) or interior passenger and cargo volume in hundreds of cubic feet (variable *Vol*). Estimate and compare two alternative simple linear regression models: In the first, *Eng* is the predictor, while in the second, *Vol* is the predictor. Write a report about two pages long that compares the two models using *three* numerical methods for evaluating regression models and one graphical method [computer help #25 and #26]. Which of the two models would you recommend for explaining/predicting *Cgphm*?
c. Report the regression standard error (*s*) for the model you recommended in part (b) and use this number to say something about the predictive ability of your model.

[6 points]

Problem 2.7 on pages 80-81. This exercise also uses the CARS2 dataset and covers a slope confidence interval (page 58), a confidence interval for the population mean (page 68), and a prediction interval for an individual Y-value (page 69). Note that it is easier to obtain each of these intervals directly using statistical software rather than by trying to calculate them by hand.

Problem 2.7. Consider the **CARS2** data file from problem 2.5 again.

a. You should have found in problem 2.5 part (b) that, of the two simple linear regression models, the model with *Eng* as the predictor should be used to explain and predict *Cgphm*. You also should have found that the best estimate for the regression slope parameter for *Eng* is 0.8183. Find a 95% confidence interval for this slope parameter [computer help #27].
b. Use the model with *Eng* as the predictor to find a 95% confidence interval for the mean *Cgphm* when *Eng* is 3 [computer help #29].
c. Use the model with *Eng* as the predictor to find a 95% prediction interval for an individual *Cgphm* when *Eng* is 3 [computer help #30].

[6 points]