Regression Analysis

Assignment 2B

Problem 3.5 on pages 133-134. This exercise does not require use of statistical software. It covers the multiple linear regression model (page 90), the global usefulness F-test (see pages 101-104 and use the second F-statistic formula on page 103), and the individual regression parameter t-test (pages 109-113). Following the hint for part (d) should give you a deeper understanding of how to interpret regression parameter estimates (refer back to Figure 3.1 on page 86 for some graphical insight). Part (e) is a "big picture" question designed to get you thinking beyond the mechanics of simply implementing the techniques (in this case, prediction intervals for an individual Y-value on pages 127-129).

Problem 3.5. De Rose and Galarza (2000) used multiple linear regression to study *Att* = average attendance from the first few years of Major League Soccer (MLS, the professional soccer league in the U.S.).The 12 MLS teams at the time ranged in average attendance from 10,000 to 22,000 per game. De Rose and Galarza used the following predictor variables:

Pop = total population of metropolitan area within 40 miles (millions)
Teams = number of (male) professional sports teams in the four major sports
Temp = average temperature (April-September, ^{o}F).

The regression results reported in the study were:

Predictor variable	Parameter estimate	Two tail p-value
Intercept	28.721	0.001
Pop	1.350	0.001
Teams	-0.972	0.037
Temp	-0.238	0.012

a. Write out the estimated least squares (regression) equation for predicting *Att* (average attendance in thousands) from *Pop, Teams,* and *Temp*.
b. R^2 was 91.4%, suggesting that this model may be useful for predicting average attendance (for expansion teams, say). Test the global usefulness of the model using a significance level of 5%.

Hint : You will need to use the second formula for the global F-statistic on page 101 to solve this part. Also, you may find the following information useful: the 95th percentile of the F-distribution with 3 numerator degrees of freedom and 8 denominator degrees of freedom is 4.07.

c. Test, at a 5% significance level, whether the regression parameter estimate for *Teams* suggests that increasing the number of (male) professional sports teams in the four major sports (football, baseball, basketball, hockey) in a city is associated with a decrease in average MLS attendance in that city (all else being equal).

Hint: You'll need to do a lower-tail hypothesis test using the p-value method, but be careful because the p-values given in the table are two-tailed.

d. According to the model results, how much does average attendance differ for two cities with the same population and average temperature when one city has one fewer (male) professional sports teams in the four major sports?

Hint: Write out the equation from part (a) for predicted average attendance in thousands for one city (plug in Teams=1, say) and then do the same for the other city (plug in

Teams=2). The difference between the two equations gives you the answer to the problem. You should find that as long as you plug in values for Teams that differ by 1, you'll always get the same answer.

e. One purpose for the study was to predict attendance for future expansion teams. Since the study was published, some of the included cities are no longer represented in MLS and have been replaced by others. In one case, beginning with the 2006 season, the San Jose Earthquakes MLS franchise relocated to Houston, Texas, which was one of the potential cities considered in the study. A 95% prediction interval for average attendance for a potential Houston MLS team based on the model came to (10,980, 15,340). Briefly discuss how studies like this can help to inform decisions about future expansion teams for professional leagues like MLS.

[10 points]

Problem 3.6 on pages 134-135. This exercise uses the SMSA dataset and covers the multiple linear regression model (page 90), the nested model F-test (pages 104-109), individual regression parameter t-tests (pages 109-113), checking model assumptions (pages 118-123), model interpretation (pages 124-126), confidence intervals for the population mean (pages 126-127), and prediction intervals for an individual Y-value (pages 127-129). If you have a good memory, you'll recall that there are brief answers to even-numbered problems in Appendix E. Thus, the challenge for this exercise is not so much writing down the correct answers, but rather understanding how to obtain the correct answers.

Problem 3.6. Researchers at General Motors analyzed data on 56 U.S. Standard Metropolitan Statistical Areas (SMSAs) to study whether air pollution contributes to mortality. These data are available in the **SMSA** data file and were obtained from the "Data and Story Library" at http://lib.stat.cmu.edu/DASL/ (the original data source is the U.S. Department of Labor Statistics). The response variable for analysis is *Mort* = age adjusted mortality per 100,000 population (a mortality rate statistically modified to eliminate the effect of different age distributions in different population groups). The dataset includes predictor variables measuring demographic characteristics of the cities, climate characteristics, and concentrations of the air pollutant nitrous oxide (NO_x).

a. Fit the (complete) model $E(Mort) = b_0 + b_1Edu + b_2Nwt + b_3Jant + b_4Rain + b_5Nox + b_6Hum + b_7Inc,$, where *Edu* is median years of education, *Nwt* is percentage nonwhite, *Jant* is mean January temperature in degrees Fahrenheit, *Rain* is annual rainfall in inches, *Nox* is the natural logarithm of nitrous oxide concentration in parts per billion, *Hum* is relative humidity, and *Inc* is median income in thousands of dollars [computer help #31]. Report the least squares (regression) equation.
b. Do a nested model F-test (using a significance level of 5%) to see whether *Hum* and *Inc* provide significant information about the response, *Mort*, beyond the information provided by the other predictor variables. Use the fact that the 95th percentile of the F-distribution with 2 numerator degrees of freedom and 48 denominator degrees of freedom is 3.19 [computer help #8].
c. Do individual t-tests (using a significance level of 5%) for each predictor in the (reduced) model $E(Mort) = b_0 + b_1Edu + b_2Nwt + b_3Jant + b_4Rain + b_5Nox$. Use the fact that the 97.5th percentile of the t-distribution with 50 degrees of freedom is 2.01 [computer help #8].
d. Check the four model assumptions for the model from part (c) [computer help #35, #28, #15, #36, #14, and #22].

Hint: Ideally, you should look at six residual plots (five with each of the five predictors on the horizontal axis in turn and one with the predicted values on the horizontal axis) to check the zero mean, constant variance, and independence assumptions. You should also use a histogram and/or QQ-plot to check the normality assumption. The more comprehensive you can be in checking the assumptions, the more confident you can be about the validity of your model.

e. Write out the least squares equation for the model from part (c). Do the signs of the estimated regression parameters make sense in this context?
f. Based on the model from part (c), calculate a 95% confidence interval for *E(Mort)* for cities with the following characteristics: *Edu = 10, Nwt = 15, Jant = 35, Rain = 40,* and *Nox = 2 [computer help #29]*.
g. Based on the model from part (c), calculate a 95% prediction interval for *Mort** for a city with the following characteristics: *Edu = 10, Nwt = 15, Jant = 35, Rain = 40,* and *Nox = 2 [computer help #30]*.

[15 points]