Lesson 12: Simple Regression Analysis and Correlation

References

- Black, Chapter 12 Simple Regression Analysis and Correlation (pp. 424-448)
- Kabakoff, Chapter 8 Regression (pp. 167-175)
- Davies, Chapter 20 Simple Linear Regression (pp. 451-460)

Data set: newspapers.csv

Description: The data are from the Gale Directory of Publications, 1994. A sample of 34 newspapers are listed along with their Daily and Sunday circulations (in thousands).

Exercises:

1) Plot Sunday circulation versus Daily circulation. Does the scatter plot suggest a linear relationship between the two variables? Calculate the Pearson Product Moment Correlation Coefficient between Sunday and Daily circulation.

2) Fit a regression line with Sunday circulation as the dependent variable. Plot the regression line with the circulation data. (Use Lander pages 212 and 213 for reference.) Comment on the quality of the fit. What percent of the total variation in Sunday circulation is accounted for by the regression line?

3) Obtain 95% confidence intervals for the coefficients in the regression model. Use confint().

4) Determine a 95% prediction interval to predict Sunday circulation for all available values of Daily circulation. Use predict(model, interval="prediction", level=0.95). Then, define a new data frame using Daily = 500 and Sunday = NA. Predict an interval for Sunday circulation.

5) Use the tableware.csv data. Regress PRICE as a dependent variable against TIME. Comment on the quality of the fit. Is a simple linear regression model adequate or is something more needed?

6) Use the tableware.csv data. ANOVA can be accomplished using a regression model. Regress PRICE against the variables BOWL, CASS, DISH and TRAY as they are presented in the data file. What do the coefficients represent in this regression model? How is the effect of plate accounted for?

7) Use the tableware.csv data. Plot PRICE versus DIAM and calculate the Pearson product moment correlation coefficient. Include DIAM in the regression model in (6). Compare results between the two models. DIAM is referred to as a covariate. Does its inclusion improve upon the fit of the first model without DIAM?