Homework

# Linear Modeling Homework – Making Predictions

The textbook’s chapter on linear models (“Line Up, Please”) introduces linear predictive modeling using the workhorse tool known as multiple regression. The term “multiple regression” has an odd history, dating back to an early scientific observation of a phenomenon called “regression to the mean.” These days multiple regression is just an interesting name for using a simple linear modeling technique to measuring the connection between one or more predictor variables and an outcome variable In this exercise, we are going to use an open dataset to explore antelope population.

This is the first exercise of the semester where there is no sample R code to help you along. Because you have had so much practice with R by now, you can create and/or find all of the code you need to accomplish these steps:

1. Read in data from <http://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/mlr/excel/mlr01.xls>

This URL will enable you to download the dataset into excel.

The more general website can be found at

<http://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/mlr/frames/frame.html>

If you view this in a spreadsheet, you will find four columns of a small dataset. The first column shows the number of fawn in a given spring (fawn are baby antelope). The second column shows the population of adult antelope, the third shows the annual precipitation that year, and finally the last column shows how bad the winter was during that year.

2. You have the option of saving the file to your computer and read it into R, or read the data directly from the web into a dataframe.

3. You should inspect the data using the str() command to make sure all of the cases have been read in (n=8 years of observations) and that there are four variables.

4. Create bivariate plots of number of baby fawns versus adult antelope population, precipitation that year, and severity of the winter. Your code should produce three separate plots. Make sure the y-axis and x-axis are labeled. Keeping in mind that the number of fawns is the outcome (or dependent) variable, which axis should it go on in your plots?

5. Next, create three regression models of increasing complexity using lm(). In the first model, predict the number of fawns from the severity of the winter. In the second model, predict the number of fawns from two variables (one should be the severity of the winter). In the third model, predict the number of fawns from the three other variables. Which model works best? Which of the predictors are statistically significant in each model? If you wanted to create the most parsimonious model (i.e., the one that did the best job with the fewest predictors), what would it contain?