Homework 3

Stat 151A, Fall 2017

Due: October 19

1. **Ant Colonies** The following question is a fairly open-ended question with which you are to practice your regression skills. Put your code in an appendix. The output can stay in your text if you want, but you should always write a clear explanation of what the output means. As the restriction above suggests, your written answers should stand alone, so that if I did not know what the question was asking I could read your answer and understand what tests you did and what your conclusions were, in terms of the real-world variables of the original data set.

   Description of the Problem In this problem, we examine the foraging behavior of a species of ant known as a thatch ant (Formica planipilis). The researchers, led by Peter Nonacs of UCLA, attempted to identify if different colonies have different strategies for optimizing the tradeoff between collecting food and taking risks. Foraging for food is a dangerous activity, as an ant may find more food by being further away from the colony, but the ant faces more danger, and the colony risks the loss of the ant and any food it was carrying. In essence, two principal strategies are believed to exist:

   - a worker conservative strategy, where ants that are foraging further away from the colony are given more food so that they face less risk of starvation before returning.
   - an energy conservative strategy, where the distant workers are provided with less food, so that if they are lost, then there is less of a threat to the colony.

   Description of the Data Our table includes data on 649 randomly chosen ants, from 6 different colonies. For each ant, the following data is given:

   - Colony number, labeled 1-6.
   - Distance (meters): the distance from the colony's entrance the sample (i.e. ant) was taken.
   - Mass or weight (mg): How much the ant weighed in milligrams. This was relates to how much food (energy) the ant was carrying.
   - Headwidth (arbitrary units): A measure of the ant's maximum headwidth.

- Headwidth (mm): Same as above, but given in millimeters.
- Size: 5 intervals, relating to headwidth, indicating the worker class of the ant.

Scientific Questions The principle scientific questions that this data pose are:

- Do different colonies use different foraging strategies? (e.e. worker-conservative versus energy-conservative) Is there some difference across size classes?
- Are there differences across colonies in the distribution of sizes or distances of the member ants?
- What are the strategies that are in use? Are any colonies especially similar or different?

(a) (1 points) First, examine the data visually, using various plots, including boxplots and coplots.

(b) (3 points) Perform a regression of the mass on colony, distance, and size, and evaluate the appropriateness of your model using graphical techniques. If you find a transformation needed, justify your choice of transformation.

(c) (1 points) Interpret the coefficients relative to the scientific contributions and discuss what conclusions you can draw.

References

UCLA Datasets (2006), http://www.stat.ucla.edu/projects/datasets/ant-explanation.html

2. (1 points) Show that the ith standardized predicted residual, t_i, satisfies $t_i = r_i\sqrt{(n-p-2)/(n-p-1-r_i^2)}$ where r_i is the standardized residual.

3. In the Bodyfat dataset, consider the linear model

$$\text{BODYFAT} = \beta_0 + \beta_1\text{AGE} + \beta_2\text{WEIGHT} + \beta_3\text{HEIGHT} + \beta_4\text{THIGH} + e$$

In R, plot the following graphs:

(a) (.35 points) Residuals against fitted values.

(b) (.35 points) Standardized Residuals against fitted values.

(c) (.35 points) Residuals against Standardized Residuals.

(d) (.35 points) Predicted residuals against fitted values.

(e) (.35 points) Residuals against predicted residuals.

(f) (.35 points) Residuals against leverage.

(g) (.35 points) Predicted residuals against Standardized Predicted Residuals.

(h) (.35 points)Standardized residuals against Standardized Predicted residuals.

(i) (.35 points) Cooks Distance against the ID number of the subjects.

(j) (.35 points) Comment on these plots. Based on these plots, assess whether there are any outliers in the dataset; are there any influential observations.

(k) (.25) For each subject, calculate the p-value for testing whether the ith subject is an outlier based on the standardized predicted residual. Plot these p-values against the ID number of the subjects. How may of these p-values are less than 0.05? Does it make sense to rule all such subjects as outliers?

(l) (.25) Based on the analysis, does it make sense to fit the linear model with any of the subjects removed? If not, why not? If so, which ones; and in this case, report the summary for the linear model with the subjects removed.