Homework Five

Statistics 151a (Linear Models)

Due on 8 November 2018 by 11:59 PM

October 23, 2018

1. Again for the body fat dataset used in class, consider the following R code:

   ```
   g = lm(BODYFAT ~ AGE + WEIGHT + HEIGHT + THIGH, data = bodyfat)
   par(mfrow = c(2, 2))
   plot(g)
   ```

   When I run this code, R gives me four plots. Describe each of these plots and explain how to interpret them. (**0.75 points**)

2. Consider the bodyfat dataset and consider fitting a linear model for the response variable BODYFAT in terms of the explanatory variables AGE, WEIGHT, HEIGHT, ADIPOSITY, NECK, CHEST, ABDOMEN, HIP, THIGH, KNEE, ANKLE, BICEPS, FOREARM and WRIST.

   a) Using each of the following methods, perform variable selection to select a subset of the explanatory variables for modeling the response: (**0.25 points each = 1.5 points**)

      i. Backward elimination using the individual p-values.

      ii. Forward Selection using p-values.

      iii. Adjusted R^2.

      iv. AIC

      v. BIC

      vi. Mallow's C_p.

   b) Let $M_1, \dots, M_6$ denote the six models selected by each of the six variable selection methods of the previous part. Select one of these models by cross-validation. (**0.2 points**)

   c) Let M be the model selected in the previous part. Fit this model to the data. Perform regression diagnostics. Comment on the validity of the assumptions of the linear model. Identify influential observations and outliers. Delete them if necessary and re-fit the model. (**0.25 points for fitting the model from the previous part to the data+ 0.5 points for regression diagnositcs + 0.5 points for outliers and influential observations + 0.25 points for dealing with the outliers = 1.5 points**).

3. (**0.8 points**) Exercise 12.2 from the book.

4. (**0.75 points**) Show that in the partial regression plot, the residuals in the regression of $Res(y, X^{-j})$ on $Res(X_j, X^{-j})$ are exactly the same as $\hat{e}$. Note, you can use that the regression of $Res(y, X^{-j})$ on $Res(X_j, X^{-j})$ has intercept 0 and slope $\hat{\beta}_j$.

5. Determine if the statement below is true or false. (**0.25 points each question**)

   a) Dropping a variable from the model will cause the other variables to become more significant.

   b) If an observation with high leverage is dropped from the model fitting, then the estimate for the coefficients will become more precise.

   c) If AIC and Mallows C_p choose a best model and these both have the same number of parameters, then the models are identical (i.e., they include the same explanatory variables).

   d) R^2 can be used as a model selection criterion in the linear model.

6. **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) (**0.5 points**) First, examine the data visually, using various plots, including boxplots and coplots.

b) (**2 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

References