M358K - Homework 4 (Model selection in linear regression)

posted by: Monday 5th, 2018

due by: Monday November 19th at 11.59PM, 2018

Number of questions in this homework: 3.
Maximal points possible: 6 from writeup, 2 from code, 4 from presentation. This gives a total of **12 points**.

Instructions. This homework concerns the `mammals` dataset from the library openintro. You can load this dataset with the commands
```
library(openintro)
data(mammals)
```
For variable descriptions, type
```
?mammals
```

This dataset contains some missing values (denoted NA). Exclude all observations with an NA in one of the variables with the command:
```
mammals2 <- mammals[rowSums(is.na(mammals)) == 0,]
```
This command saves the data subset to a new dataframe, `mammals2`. We shall do all computations on `mammals2`.

Question 1. Basic multiple linear regression

1. We want to fit a linear regression model that can be used to predict TotalSleep. Explain why Dreaming, NonDreaming and Species are BAD variables to include in this regression model.

2. Treat Predation, Exposure and Danger as numericals. Run model1, the linear regression model with TotalSleep vs BodyWt, BrainWt, LifeS-

pan, Gestation, Predation, Exposure and Danger. Clearly show the R command that you use, and include the R's model summary.

3. Write down the equation that R gives you. Interpret all the coefficients and the p-values associated with the coefficients. Report the R^2 and adjusted R^2 of your model. What are the meaning of these values?

Question 2. Regression with categoricals vs numericals

1. Treat Predation, Exposure and Danger as categoricals. Run model2, the linear regression model with TotalSleep vs BodyWt, BrainWt, LifeSpan, Gestation, Predation, Exposure and Danger. Clearly show the R command that you use, and include the R's model summary.

2. Compare model1 and model 2: comment on the coefficients and the diagnostic plots. Say which, if any, of the (a) independence (b) normal distribution and (c) constant variance assumptions are violated.

Question 3. Model selection

1. Do variable selection with the `stepAIC` command, starting with model1. Call this model1.AIC. Compare model1.AIC against model1: comment on the coefficients and the diagnostic plots.

2. Do variable selection with the `stepAIC` command, starting with model2. Call this model2.AIC. Compare model2.AIC against model2: comment on the coefficients

3. Which model amongst the above 4 is the best? (Give a brief justification). For the better model, summarize the relationship between TotalSleep and other attributes of a mammal.

4. The species Homo Sapiens has the following attributes: BodyWt = 75, BrainWt = 1.4, LifeSpan = 77, Gestation = 268, Predation = 2, Exposure = 2, Danger = 2. Use your model to predict TotalSleep for this species. Is your prediction reasonable? Explain why or why not.