M358K - Homework 5 (Multiple linear regression)

posted by: Thursday, November 9th at 12AM, 2017

due by: Wednesday, November 22nd at 11.59PM, 2017

Instructions and grading scheme

Submission instructions. Please submit three files on Canvas with the following names:

- **homework5-writeup**: this is a word file containing your answers. All figures, tables etc must be inside this file! For how to save R figures, see instructions on Canvas.
- **homework5-code-final.R**: this is a text file containing the R codes you used to compute the numbers, tables, and figures CONTAINED IN your report ONLY. This file IS graded, see below.
- **homework5-code-draft.R**: this is a text file containing the R codes you used to explore the data. This file should give an idea on how you explored the dataset, and how you arrived at code-final.R. It can tables/figures which are EXCLUDED from your report, comments, etc. For instance, the R history of your various sessions is enough.

Acceptable word files are: anything that can be opened by LibreOffice, OpenOffice, Microsoft Word, or equivalent software. Example file extensions are .doc, .docx, .odt.

Grading scheme. For the write up: on each question you can earn 0/1/2 points.

- 2 = correct answer

- 1 = partially correct answer, or correct answer but with muddy or missing justification
- 0 = incorrect answer, unreasonable answer

For code-final.R: you can earn 0/5 points.

- 5 = code-final.R runs on the given dataset, gives all the tables and figures included in your report.
- 0 = no code file OR code file does not run at all OR code file does not produce the reports' tables/figures OR code file raises a plagiarism red flag, etc

For code-draft.R: you can earn 0/5 points.

- 5 = The grader is sufficiently convinced that you explored this dataset on your own.
- 0 = no code file OR code file raises a plagiarism red flag OR code file is irrelvant, does not give an indication on how you arrived at code-final.R

Bonus points for presentation:
Write-up: +2 points for nice report(grammatically correct sentences, no rambling discussions, discussions exceed expectations, extra analysis of the dataset)
R code: +2 points for neat layout. (code adequately commented, clearly laid out, easy to understand).

Bonus questions
These are extra challenging questions, and are additional opportunities to score points. On each bonus question you can get 0/4 points.

Questions

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`.

1. Do a pairwise plot of all variables. For each plot, briefly describe what you see.

2. 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.

3. Treat Predation, Exposure and Danger as numericals. Run model1, 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.

4. 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?

5. 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.

6. Compare model1 and model 2: comment on the coefficients and the diagnostic plots (Say which, if any, of the (a) independence (no mean trend) (b) normal distribution and (c) constant variance assumptions are violated.)

7. 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.

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

9. 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.

10. 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.

(Bonus) Ngoc played with this data and obtained an adjusted R^2 of 0.6848 and a reasonable fit (with no dropping of any observations). Present a model that has an adjusted R^2 at least as good. For your model, shows the diagnostic plot, and provide interpretations for the coefficients.