STAT 509 2017 Summer HW14

Instructor: Shiwen Shen

Lecture Day: June 6

1. (This problem is designed for you to learn some basic coding methods in R. You can download the book in the course webpage directly.) Read the first chapter ”Introduction” in **Practical Regression and Anova using R** and play the r code in the book chapter by yourself. (It is OK that you cannot understand everything!) You need to use `install.packages()` command to install the `faraway` package. For example,

```
install.packages("faraway")
library(faraway)
data(pima)
pima
```

   You can choose any USA CRAN mirror when you are installing, e.g. USA (NC). Reproduce the three plots in the page 12. Print out three plots and your code. *(Hint: follow book's code step by step, and you will get three plots naturally.)*

2. Diabetes and obesity are serious health concerns in the US and much of the developed world. Measuring the amount of body fat a person carries is one way to monitor weight control progress, but measuring it accurately involves either expensive X-ray equipment or a pool in which to dunk the subject. Instead body mass index (BMI) is often used as a proxy for body fat because it is easy to measure. In a study of 250 men at Bingham Young University, both BMI and body fat were measured. Researchers found the following summary statistics:

$$\sum_{i=1}^{n} x_i = 6322.28 \qquad \sum_{i=1}^{n} x_i^2 = 162674.18 \qquad \sum_{i=1}^{n} x_i y_i = 125471.10$$
$$\sum_{i=1}^{n} y_i = 4757.90 \qquad \sum_{i=1}^{n} y_i^2 = 107679.27$$

   where x denotes the BMI and y denotes the body fat.

   (a) Calculate the least squares estimates of the intercept and slope. *(Hint: use the results we have in class: $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$).*

   (b) Use the equation of the fitted line to predict what body fat would be observed, on average, for a man with a BMI of 30.

3. `teengamb` dataset (in faraway package) concerns a study of teenage gambling in Britain. Here is the description of this datset:

   - sex: 0 = male, 1 = female
   - status: socioeconomic status score based on parents' occupation
   - income: in pounds per week
   - verbal: verbal score in words out of 12 correctly defined
   - gamble: expenditure on gambling in pounds per year

Make a numerical and graphical summary of the data similar to the analysis the chapter 1.1.3. (Note: there is no missing value in `teengamb`), and comment on any features that you find interesting. Limit the output you present to a quantity that a busy reader would find sufficient to get a basic understanding of the data. (Hint: useful R functions are `library()`, `data()`, `summary()`, `hist()`, `plot()`. You can always type `help(subject)` to get detailed help on the subject, e.g. `help(plot)`.) Here is the R code for you to load the `teengamb` data into R:

```
library(faraway)
data(teengamb)
teengamb
```

4. For the `teengamb` dataset, fit a simple linear regression model with the expenditure on gambling as the dependent variable, and income as independent variable. Present a summary of the R output.

   (a) What are the values of $\hat{\beta}_0$ and $\hat{\beta}_1$?

   (b) Interpret the meaning of $\hat{\beta}_1$ in the context of the problem.

   (c) What is the function that we can use to predict the mean value of expenditure on gambling when the income is given?

   (d) Predict the value of expenditure on gambling when income is 2.

   (e) Calculate and present the residuals for all observations.

   (f) Calculate the mean and median of the residuals. (Hint: R code `mean()` and `median()`.)

   (g) Calculate the mean squared error (MSE) using residuals.

   (h) What is the p-value in testing the significance of β_1? What does it mean?