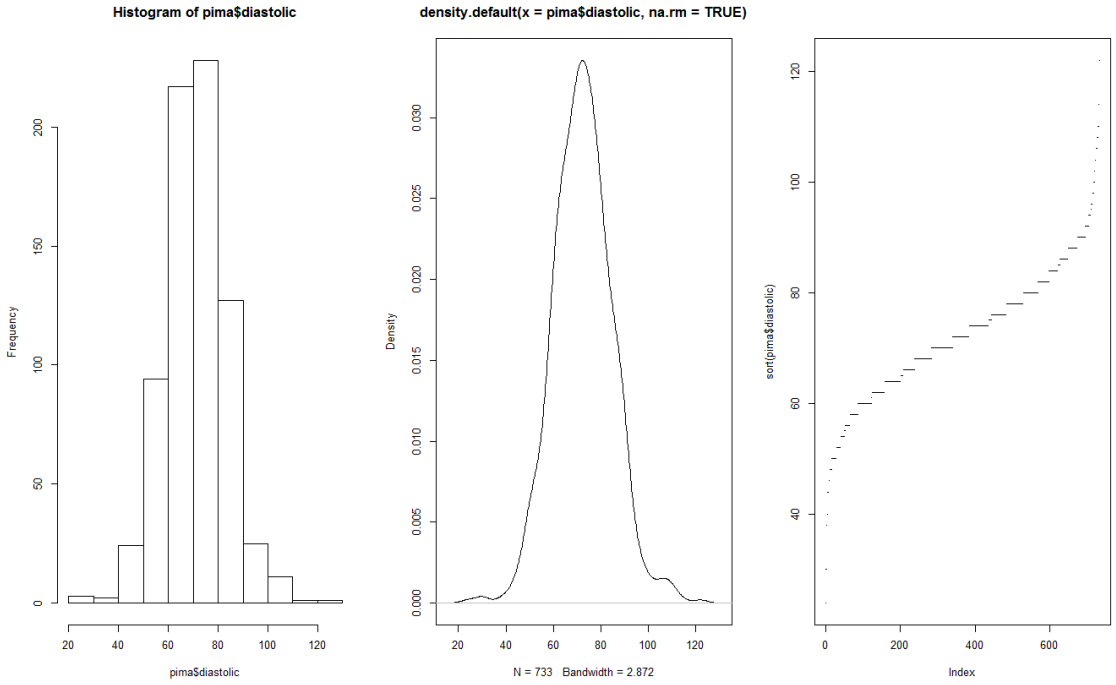
Homework 14 Solution

STAT 509 Statistics for Engineers
Summer 2017 Section 001
Instructor: Tahmidul Islam

1. (This problem is designed for you to learn some basic data manipulation in R. You can download the book directly from the course webpage.) 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.)*



The R code here is simply a copy-and-paste work from the book. The only thing different is that I use `par(mfrow=c(1,3))` to generate a frame including all three plots.

```
library(faraway)
data(pima)
pima
summary(pima)
sort(pima$diastolic)

pima$diastolic[pima$diastolic == 0] <- NA
```

```
pima$glucose[pima$glucose == 0] <- NA
pima$triceps[pima$triceps == 0] <- NA
pima$insulin[pima$insulin == 0] <- NA
pima$bmi[pima$bmi == 0] <- NA

pima$test <- factor(pima$test)
summary(pima$test)

par(mfrow=c(1,3))  # <---- start  a graphical frame with 1 by 3 plots #
hist(pima$diastolic)
plot(density(pima$diastolic,na.rm=TRUE))
plot(sort(pima$diastolic),pch=".")
```

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:* $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$*).*

Based on the summary statistics, we have

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{6322.28}{250} = 25.29, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{4757.90}{250} = 19.03$$

Therefore,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{125471.10 - 250(25.29)(19.03)}{162674.18 - 250(25.29)^2} = 1.86$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 19.03 - 1.86(25.29) = -28.01$$

(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.

Since $\hat{y} = -28.01 + 1.86x$, when $x = 30$ we have

$$\hat{y} = -28.01 + 1.86(30) = 27.79$$

Thus, the average body fat for a man with a BMI of 30 is 27.79.

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
```

   Using `summary(teengamb)`, we get a full numerical summary for the dataset.

```
      sex              status          income           verbal          gamble
 Min.   :0.0000   Min.   :18.00   Min.   : 0.600   Min.   : 1.00   Min.   :  0.0
 1st Qu.:0.0000   1st Qu.:28.00   1st Qu.: 2.000   1st Qu.: 6.00   1st Qu.:  1.1
 Median :0.0000   Median :43.00   Median : 3.250   Median : 7.00   Median :  6.0
 Mean   :0.4043   Mean   :45.23   Mean   : 4.642   Mean   : 6.66   Mean   : 19.3
 3rd Qu.:1.0000   3rd Qu.:61.50   3rd Qu.: 6.210   3rd Qu.: 8.00   3rd Qu.: 19.4
 Max.   :1.0000   Max.   :75.00   Max.   :15.000   Max.   :10.00   Max.   :156.0
```

   and we find that the binary variable, sex, is not summarized properly. Naturally, the first step is to change sex from a "quantitative variable" to a "categorical variable" by the following code:

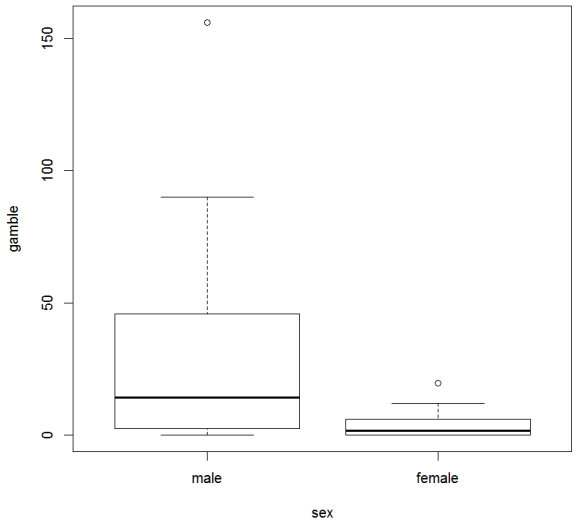
```
teengamb$sex <- factor(teengamb$sex)
levels(teengamb$sex) <- c("male", "female")
```

   Now, the summary looks much better:

```
     sex         status          income           verbal          gamble
 male  :28   Min.   :18.00   Min.   : 0.600   Min.   : 1.00   Min.   :  0.0
 female:19   1st Qu.:28.00   1st Qu.: 2.000   1st Qu.: 6.00   1st Qu.:  1.1
             Median :43.00   Median : 3.250   Median : 7.00   Median :  6.0
             Mean   :45.23   Mean   : 4.642   Mean   : 6.66   Mean   : 19.3
             3rd Qu.:61.50   3rd Qu.: 6.210   3rd Qu.: 8.00   3rd Qu.: 19.4
             Max.   :75.00   Max.   :15.000   Max.   :10.00   Max.   :156.0
```

   One question that we are interested in is to see the difference of expenditure on gambling between sex. First, let's take a look at the side-by-side boxplot, which is generated by the following code

```
plot(gamble ~ sex, data=teengamb)
```



Overall, the median level of expenditure of gambling of male is much higher than the one of female(comparing the thick line in the middle of two boxes). Besdies, the variability of the expenditure of gambling of male is greater than the variability of the one of female (comparing the length of two boxes). We can use the following code to see the detailed the summary statistics for different gender.

```
> ## split the gamble data into male part and female part
> MaleGamb <- teengamb$gamble[teengamb$sex=="male"]
> FemaleGamb <- teengamb$gamble[teengamb$sex=="female"]
>
> ## find the numeric summary for each gender
> summary(MaleGamb)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   2.775  14.250  29.780  42.180 156.000
> summary(FemaleGamb)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.100   1.700   3.866   6.000  19.600
>
> ## find the variance for each gender
> var(MaleGamb)
[1] 1393.095
> var(FemaleGamb)
[1] 26.53001
```

From R ouput, the median of male is approx. 8 times higher than the female and the variance of male is approx. 53 times higher than the female, which indicates our observation from the side-by-side boxplot is true.

Remark: you can present any point you are interested in, here I just give a simple example which is one of my interests.

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

```
> fit <- lm(gamble ~ income, data=teengamb)
> summary(fit)

Call:
lm(formula = gamble ~ income, data = teengamb)

Residuals:
    Min      1Q  Median      3Q     Max
-46.020 -11.874  -3.757  11.934 107.120

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -6.325      6.030  -1.049      0.3
income         5.520      1.036   5.330 3.05e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.95 on 45 degrees of freedom
Multiple R-squared:  0.387,     Adjusted R-squared:  0.3734
F-statistic: 28.41 on 1 and 45 DF,  p-value: 3.045e-06
```

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

Based on the R ouput, $\hat{\beta}_0 = -6.325$ and $\hat{\beta}_1 = 5.520$

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

Interpretation: On average the expenditure on gambling will vary (increase or decrease) 5.520 pounds per year when the home income varies (increases or decreases) 1 pound per week.

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

$$\widehat{\text{gamble}} = -6.325 + 5.520 \times \text{income}$$

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

$$\widehat{\text{gamble}} = -6.325 + 5.520(2) = 4.715$$

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

There are two ways to calculate it in R. First, the definition of residual; second, `residuals` function. For example,

```
## method 1: definition
teengamb$gamble - predict(fit)

## method 2: residuals function
residuals(fit)
```

and the calculated residuals are

```
         1          2          3          4          5          6
-4.7164115 -7.4766542 -4.7164115 -25.0188380  14.8835885 -12.7315249
         7          8          9         10         11         12
-22.5881100 -22.5169565  -3.0164115 -26.6983527 -10.1368968 -14.4977461
        13         14         15         16         17         18
-4.6205086  -1.1164115  -7.8368968   1.4438311 -46.0200512 -40.4802938
        19         20         21         22         23         24
-3.7573821 -12.9971395  -9.2368968  -6.2766542 -12.8971395 107.1197062
        25         26         27         28         29         30
 8.9414047   0.1438311  -9.2068809   3.8040737   3.6122679 -14.4381100
        31         32         33         34         35         36
28.0787356  20.8811620  13.5172798  -1.7164115  12.1438311  51.4823753
        37         38         39         40         41         42
31.0233458  19.3606768 -42.8802938  22.4917826   2.1835885  -6.7827202
        43         44         45         46         47
 3.0631032 -11.0170181  17.0533617  12.4438311  11.7233458
```

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

```
> resid <- residuals(fit)
> mean(resid)
[1] -5.203801e-16
> median(resid)
[1] -3.757382
```

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

```
> MSE <- sum(resid^2)/(47-2)
> MSE
[1] 622.4131
```

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

Based on the R output of the regression, we find the p-value to be 3.05×10^{-6}, which means that the variation of gamble variable can be explained linearly by income variable.