CMDA 3654: Assignment #2

Due on Friday, Sep 23, 2016

Jung Choi

"I have neither given nor received unauthorized assistance on this assignment."

September 23, 2016

1 IRIS DATA PLOT

1. (20 points) Consider the Iris data set from assignment 1 problem 3. Construct the following plots in R.
a) The ratio between the width and the length shows increasing in overall dimensions for each species. It shows fairly precise ratio for the Setosa(red) specie, however, as the data increases, other species does not seem to keep that consistency.

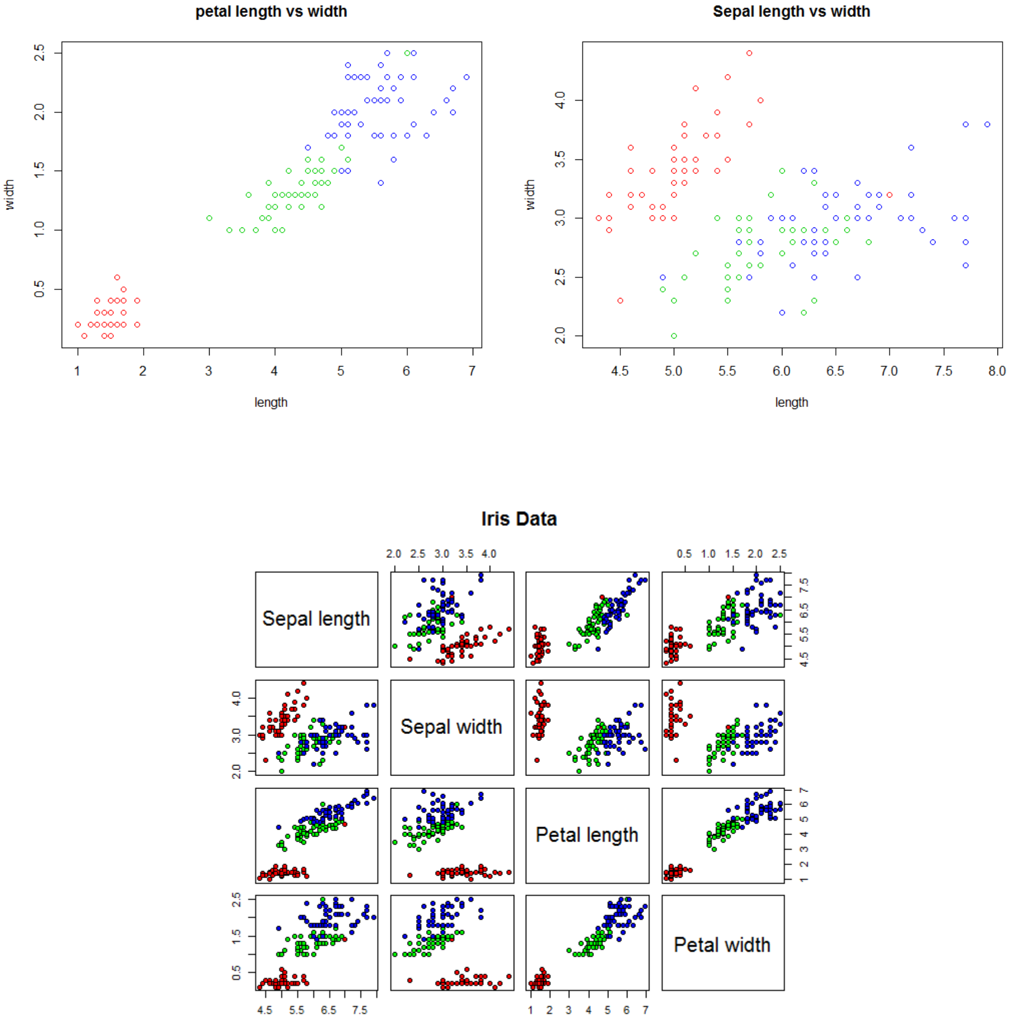

Figure 1: Petal and Sepal ratio

d) I would choose petal since the data shows more consistent ratio for each class which can represent proper linear regression model. The precise results for each class definitely shows less margin of error. Sepal on the other hand, there is no consistency with being too many outliers in the plot.

```
plot(data.iris$‘Petal length‘, data.iris$‘Petal width‘, type="p", main = "petal␣
    length␣vs␣width", xlab="length", ylab="width", pch =21, col=c("red","green","
    blue")[unclass(iris$Species)])
plot(data.iris$‘Sepal length‘, data.iris$‘Sepal width‘, type="p", main = "Sepal␣
    length␣vs␣width", xlab="length", ylab="width", pch =21, col=c("red","green","
    blue")[unclass(iris$Species)])

pairs(data.iris[1:4], main = "Iris␣Data", pch = 21, bg = c("red", "green", "blue")
    [unclass(iris$Species)])
```

2 BABYNAMES

2. (20 points) Consider the babynames data from assignment 1 problems 4,5 (a) Create a subset of the data with female babies named "Mary" from 1880-2014.

```
library("babynames")

#Setting up the data frame
data.baby=data.frame(babynames)

mary = subset(data.baby, data.baby$sex=="F" & data.baby$year>=1880 & data.baby$
    year<=2014 & data.baby$name=="Mary")

# (b) Create a subset of the data with female babies named "Sophia" from 1880
    2014.

sophia = subset(data.baby, data.baby$sex=="F" & data.baby$year>=1880 & data.baby$
    year<=2014 & data.baby$name=="Sophia")

# (c) Construct a plot of the proportion of female babies named "Mary" from 1880
    2014. On
# the same plot, add/overlay a plot of the proportion of female babies named "
    Sophia" from
# 1880 2014.
plot(x=1880:2014,y=mary$prop,col="red",ylab="Prop",xlab="Year")
points(x=1880:2014,y=sophia$prop,col="blue")
legend("topright", inset=0.05, title="Number␣of␣participants", c("Sophia","Mary"),
     fill=c("red","blue"))
```

d) Figure 2 illustrates the scatter plot of Mary and Sophia proportions across years from 1880 to 2014. This statistical model shows feasibility on how easily we can understand the correlations between the data set and perhaps, this could predict other possible outcomes in the future.

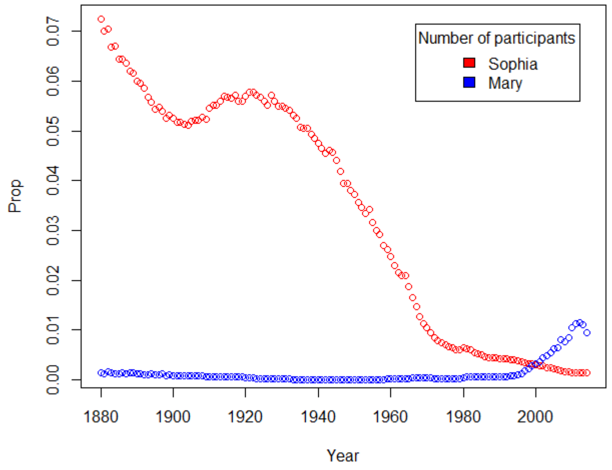

Figure 2: Proportion of female babies named "Mary" and "Sophia"

3 R CODE PROBLEM

3. Run the following code segments in R, describe what is wrong and how to correct it. Report only the R code with correction, do not report output.

(a) (10 points)
a = rnorm(100)
b = sqrt(a)
print(b).
ANSWER: b=sqrt(as.complex(a)). Since sqrt does not return a complex number when we take the sqrt of negative numbers from a. Then it would produce a warning message saying NaNs produced.The as.complex function is convenient way to simplify sqrt of negative numbers.

(b) (10 points)
a = c("1","2","3","4","5")
b = mean(a)
print(b)
ANSWER: b = mean(as.numeric(a)). You can not apply mathematical functions on string variables. So, as.numeric function will take string variables convert them into numeric variables.

4 Anscombe dataset

4. (20 points) Load the anscombe dataset in R. (Hint: data.anscombe = anscombe)
(a) Fit linear regression of (i) y1 on x1 (ii) y2 on x2 (iii) y3 on x3 and (iv) y4 on x4. Write down the four fitted regression lines. fit1 Coefficients:

```
(Intercept)           x1
     3.0001        0.5001
fit2 Coefficients:
(Intercept)           x2
      3.001         0.500
fit3 Coefficients:
(Intercept)           x3
     3.0025        0.4997
fit4 Coefficients:
(Intercept)           x4
     3.0017        0.4999
```

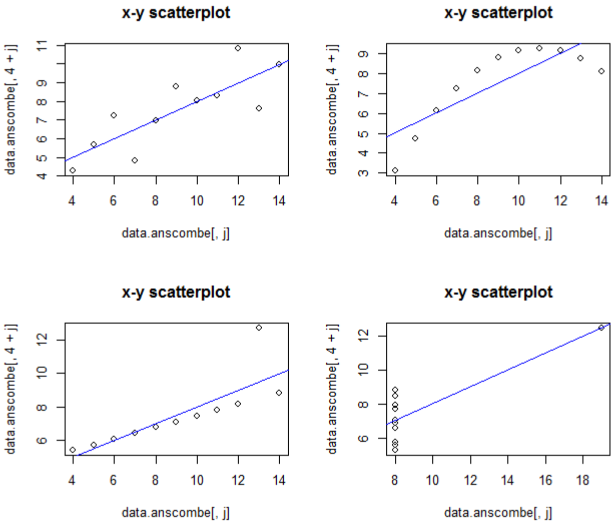

Figure 3: From left to right (i) y1 vs x1, (ii) y2 vs x2, (iii) y3 vs x3, (iv) y4 vs x4

(c) Use your judgment and describe the discrepancy between the plots and the regression lines. The dataset 1 show the best consistency between the plot and the regression line. Other such as dataset3 and dataset4 show outliers which does not show proper regression line relation to the given dataset.

5. First Anscombe dataset represents like what we would expect from scatter plot. It does show well fitted linear model. The second dataset shows negative skewed graph which does not have a linear correlation. However, the third dataset shows very precise linear correlation. There is an outlier that throws off the linear agression, but it can be adjusted by separating or removing the outlier. The last dataset have poor representation of both linear correlation and regression models. However, the single outlier makes it appear like it's showing proper linear

regression which does not show appropriate linear regression model.

```
data.anscombe = anscombe
fit1 = lm(data.anscombe$y1~ data.anscombe$x1) # lm( Y ~ X1 + X2 X3)
fit2 = lm(y2~x2, data=data.anscombe)
fit3 = lm(y3~x3, data=data.anscombe)
fit4 = lm(y4~x4, data=data.anscombe)
plot(y1~x1, data=anscombe)
points(y2~x2, data=anscombe)
points(y3~x3, data=anscombe)
points(y4~x4, data=anscombe)
par(mfrow=c(2,2))
for (j in 1:4){
  # plot(x[,j],y,xlab=names(data.h)[j],ylab="median value",main="x-y scatterplot")
  plot(data.anscombe[,j], data.anscombe[,4+j] ,main="x-y␣scatterplot")
  fit = lm(data.anscombe[,4+j]~ data.anscombe[,j], data=data.anscombe)
  abline(fit,col="blue")
}
round(cor(data.anscombe[,1],data.anscombe[,5]),2)
```