STAT 51200--FALL 2022
Applied Regression Analysis

Homework (Computer) Assignment # 01

1) Task #1

A) Search on the WWW and familiarize yourself with some of the web resources on R and of SAS tutorials, wiki, and the likes. One useful site (amongst many other sites) that you can find tutorials on R and on SAS is **https://www.tutorialspoint.com/.** See also the course's Canvas site for a 'quick-start' document.

B) Install R (or RStudio) directly on your computer, if you can, and verify that you can get access to this software and to SAS on IUAnyWare via **https://iuanyware.iu.edu/,.**

Getting Acquainted with R

2) Task #2 – Refer to the MHW Data – see details below.

A) Retrieve the data file MHW.csv from the class site on Canvas and save it on your local work directory. Change its name from MHW.csv to MHW.txt and make sure that you know the name your 'working directory' for R. This data set will serve us for several on-going examples and assignments/tasks.

B) Read in the data to an appropriate data frame:

- `MHW<- read.table("MHW.txt", header=TRUE, sep=",")`

C) Review the **structure** of this data frame by

- `str(MHW)`

D) Extract the **names** of the columns by

- `names(MHW)` # (or `colnames(MHW1)`)

E) Print the first 10 elements (rows) of this data frame.

- `MHW[1:10, ]`

F) Obtain a clearly labeled scatterplot of **'grain'** on **'straw'**.

- `plot(MHW[,3], MHW[,4], main="Scatterplot of Straw on Grain", xlab="grain", ylab="straw")`

#Alternatively, you may use,

- `attach(MHW)`
- `plot(grain, straw, main="Plot of straw on grain")`
- `detach()`

G) Define and calculate a new variable '**yield.ratio**' which is the ratio between **grain'** and **'straw'.** Attach and include it in the original data frame MHW. To verify that it has 5 columns, print now the first 8 rows of this amended data frame.

- `attach(MHW)`
- `yield.ratio<-grain/straw`
- `MHW1<-cbind(MHW, yield.ratio)`
- `MHW1[1:8,]`
- `detach()`

H) Save the amended MHW1 object (data frame) in R format:

- `save(MHW1, file="MHW1.RData")`

(it is conventional to give files with R type objects the .RData extension).

Task #3: **Visualize the data:**

A) Obtain a stem and leaf display of the grain and of the straw yields

- `attach(MHW1)`
- `stem(MHW1$grain) or stem(MHW1[ , 3])`
- `detach()`

B) Obtain a simple histogram of the grain and of the straw yields

- `hist(MHW1$grain)`

C) Obtain a fancy histogram of these data

- `hist(MHW1$grain, breaks = seq(2.6, 5.2, by = 0.1), col = "lightblue", border = "red", main = "Mercer-Hall Data", xlab = "Grain yield, lb. per plot")`

D) Add a 'sketch' of a non-parametric density estimate to this fancy histogram and save the resulting plots in a pdf file. Note the freq=TRUE option.

- `hist(MHW1$grain, freq=FALSE, breaks = seq(2.6, 5.2, by = 0.1), col = "lightblue", border = "red",`

```
main = "Mercer-Hall Data", xlab = "Grain yield,
lb. per plot")
```

- `lines(density(MHW1$grain, adj=1.5), lwd = 1.5, col = "brown")`

E) Obtain a side by side boxplot of the two yields, and label them appropriately.

- `boxplot(grain, straw)`

Task #4: **Summarize the data:**

A) Obtain the five-number summary**,**

- `summary(MHW1)`
- `apply(MHW1, 2, summary)`
- `summary(MHW1$grain)`

B) Obtain the various 'statistics'

- `min(MHW1$grain)`
- `max(MHW1$grain)`
- `mean(MHW1$grain)`
- `median(MHW1$grain)`
- `var(MHW1$grain)`
- `sd(MHW1$grain)`
- `quantile(MHW1$grain)`
- `IQR(MHW1$grain)`

C) Mark the 25% and the 75% quantiles on the fancy histogram you obtained in part E above,

- `qa<-quantile(MHW1$grain, probs = c(0.25, 0.75))`
- `abline(v=qa[1])`
- `abline(v=qa[2])`

D) Find the 'records' of the plots which attain the maximum and the minimum grain and straw yields. For example, you may use:

- `MHW1(which.max(MHW1$grain), ]`

Task #5: **Study the relationship:**

A) Obtain the marginal histogram of the each of the two yields and compare it to the theoretical normal density

```
attach(MHW1)
xx<-sort(grain)
hist(xx, nclass=30, freq=FALSE)
lines(xx, dnorm(xx, mean(xx), sqrt(var(xx))))
```

do the same for the variable straw. Do you think that the marginal normal distributions are appropriate?

B) Obtain again, a nicely labeled scatter-plot of grain versus straw. Draw a vertical and horizontal lines to indicate the respective means.

```
plot(grain, straw)
abline(v=mean(grain))
abline(h=mean(straw))
```

C) Preform a quick study of the 'linear relationship' between these two measured variables

```
lm.out<-lm(straw~ grain)
summary(lm.out)
```

D) Plot the "regression" line and add it to the scatter-plot you obtained in B.

```
abline(lm.out)
```

Task #6: **Getting Started:**

1) Read in the data to an appropriate INFILE statement:

```
data MHW;
infile 'MHW.txt'  delimiter=',' firstobs=2;
input r c grain straw;
run;
```

2) Print the content of the first 10 observations (rows)

```
Proc Print Data=MHW (obs=10);
run;
```

3) Obtain the means of the two variables of interest,

```
title ;
PROC MEANS DATA=MHW;
VAR grain straw;
RUN;
```

4) Obtain a clearly labeled scatterplot of **'grain'** on **'straw'**.

```
proc gplot data=mhw;
plot grain*straw;
run;
```

5) Define and calculate a new variable '**YieldRatio**' which is the ratio between **grain'** and **'straw'.** Attach and include it in the original data frame MHW. To verify that it has 5 columns, print now the first 5 rows of this amended data frame.

```
Data MHW1;
Set MHW;
YieldRatio=grain/straw;
Run;
proc print data=MHW1 (obs=5);
run;
```

6) Save the amended MHW data step to a file:

```
proc export data=MHW1;
outfile='Data\MHW1.txt' delimiter=' ';
run;
```

7) Label each plot in the field by its side/location in the field and print again the first 10 observations;

```
Data MHW2;
set MHW1;
IF r<= 10 and c<=12 then side='NW';
    ELSE IF r<=10 and c>12 then side='NE';
    ELSE IF r>10 and c<=12 then side='SW';
    ELSE IF r>10 and c>12 then side='SE';
run;
Proc print data=mhw2 (obs=10);
run;
proc sort data=mhw2;
by side;
     run;
```

8) Obtain the means of the three variables grouped by the plots' side;

```
proc means data=mhw2;
var grain straw YieldRatio;
by side;
run;
```

9) Obtain some standard descriptive statistics.

```
proc univariate data = MHW2 NORMAL PLOT;/*with added
   options*/
  title 'Some Descriptive Statistics';
  var grain straw YieldRatio;
  histogram grain straw;
run;
```

10) Do the same by according to each side;

```
proc univariate data = MHW2 NORMAL PLOT;/*with added
options*/
  title 'Some Descriptive Statistics';
  var grain straw YieldRatio;
  histogram grain straw;
  by side;
   run;
```

- **THE MHW-Data—An Ongoing Example Data Set**

In the early days of scientific agriculture, Mercer and Hall [1] were trying to determine the optimum plot size for agricultural yield trials: Plots that are too *small* will be too variable; and plots that are too *large* waste resources (land, labor, seed); if the land area is limited, the number of treatments will be unnecessarily small.

So, they performed a very simple experiment: an apparently homogeneous field was selected, prepared as uniformly as possible and planted to the same variety of wheat. They attempted to treat all parts of the field exactly the same in all respects during subsequent farm operations. When the wheat had matured, the field was divided into 500 equally-size plots. Each plot was harvested separately. Both grain and straw were air-dried, then hand-threshed and weighed to a precision of 0.01 lb (= 4.54 g). The reported values are thus air-dry weight, in lb per plot.

The field was a square of 1 acre, which is 0.40469 hectare or 4,046.9 m2, which was divided into a 20 rows by 25 columns, giving 500 plots, each of 1/500 acre, which is about 8.09 m2 (3.30 m long x 2.45 m wide). We can assume that the rows ran W to E, with 25 plots in each row, beginning at 1 on the W and running to 25 at the E, so that columns run N to S with 20 plots in each, running from 1 at the N to 20 at the S. Thus the NW corner (1,1) is plot 1, the NE corner (1, 25) is plot 481, the SE corner (25, 20) is plot 500, and the SW corner (1, 20) is plot 20.

The CSV Data File: The data has been prepared as the comma-separated values ("CSV") file mhw.csv in a plain-text editor. The first line gives the four field names:
"**r**","**c**","**grain**","**straw**" standing for:

r : Row number in the field
c : Column number in the field
grain : Grain yield, lbs per plot
straw : Straw yield, lbs per plot

Each of the 500 lines in the data file represents a plot; the four fields are separated by commas. For example, the first line is: 1,1,3.63,6.37.

[1] W B Mercer and A D Hall. The experimental error of field trials. *The Journal of Agricultural Science (Cambridge)*, 4:107–132, 1911.