# STAT537: Statistics for Research I: HW#7

Due on Sep. 20, 2016

*Dr. Schmidhammer TR 11:10am – 12:25pm*

**Wenqiang Feng**

# Contents

# Problem 1

Homework on Correlation Coefficients

*Solution.* 1. **Generate a scatterplot for these data, with grain size on the horizontal axis, and yield on the vertical axis.**

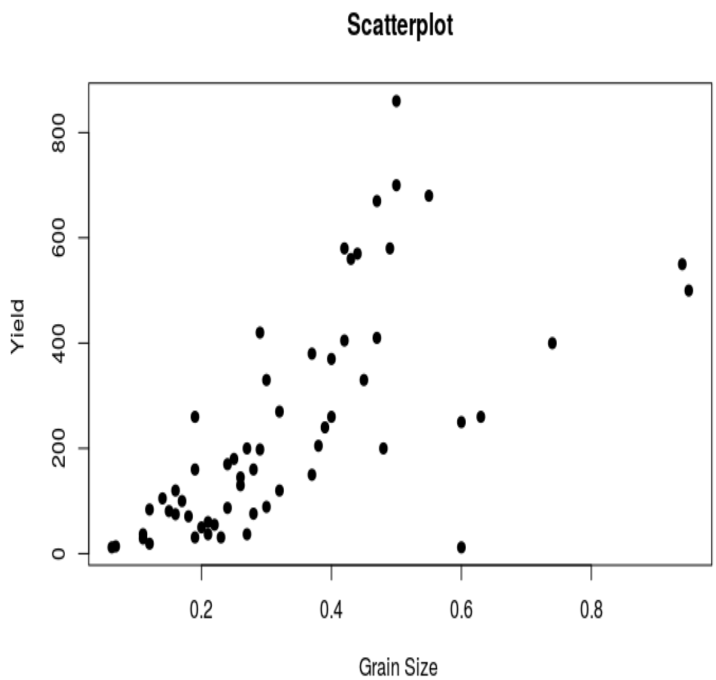


Figure 1: Scatterplot for these data, with grain size on the horizontal axis, and yield on the vertical axis.

2. **Compute the Pearson Product Moment correlation coefficient r and Spearmans rho**

   (a) Pearson Product Moment correlation coefficient $r = 0.667871$

   ```
   > cor(GrainSize, Yield, method="pearson")
   [1] 0.667871
   ```

   (b) Spearman's rho $\rho = 0.7634203$

   ```
   > cor(GrainSize, Yield, method="spearman")
   [1] 0.7634203
   ```

3. **Test the hypothesis:**

   (a) Pearson Product Moment correlation coefficient r: Since the p-value $= 7.543e-09 < 0.05$, so reject $H_0$. Hence we can say that we have enough evidence to believe $H_1$, i.e. we have enough evidence to believe that the Grain Size values are correlated with the Yield value at 95% level.

   ```
   > out
   ```

```
Pearson's product-moment correlation

data:  GrainSize and Yield
t = 6.7748, df = 57, p-value = 7.543e-09
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4967473 0.7890091
sample estimates:
     cor
0.667871
```

(b) Spearmans rho: Since the p-value $= 2.059e - 12 < 0.05$, so reject $H_0$. Hence we can say that we have enough evidence to believe $H_1$, i.e. we have enough evidence to believe that the Grain Size values are correlated with the Yield value at 95% level.

```
Spearman's rank correlation rho

data:  GrainSize and Yield
S = 8095.8, p-value = 2.059e-12
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7634203
```

4. **Construct a 95% Confidence Interval for $\rho$.**

(a) 95 percent confidence interval: [0.630624, 0.8527845]

```
> CIrho(out2$estimate,N)
           rho    2.5 %    97.5 %
[1,] 0.7634203 0.630624 0.8527845
```

□

# Problem 2

Homework on Simple Linear Regression

*Solution.* 1. **Generate a scatterplot for these data, with grain size on the horizontal axis, and yield on the vertical axis.**

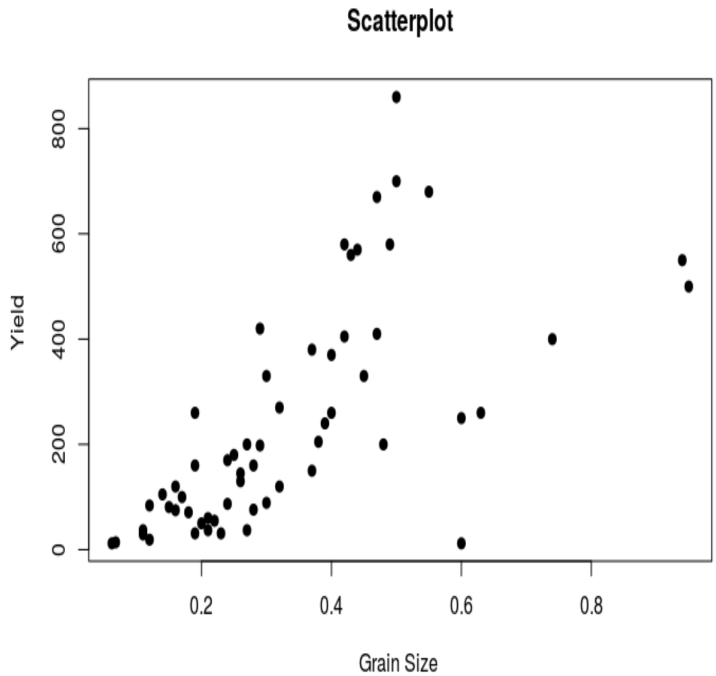


Figure 2: Scatterplot for these data, with grain size on the horizontal axis, and yield on the vertical axis.

2. **Find the least squares estimates of $\beta_0$ and $\beta_1$ in the model:** $\beta_0 = -9.294$ and $\beta_1 = 744.979$.

```
> fit

Call:
lm(formula = Yield ~ GrainSize)

Coefficients:
(Intercept)    GrainSize
     -9.294      744.979
```

3. **Test the hypothesis:** From the following summary, we can see that the p-value for $\beta_1$ is $7.543e-09 < 0.05$, Hence reject $H_0$. Therefore we can say that we have enough evidence to believe $H_1$, i.e. $\beta_1 \neq 0$.

```
> summary(fit)

Call:
lm(formula = Yield ~ GrainSize)

Residuals:
    Min      1Q  Median      3Q     Max
-425.69 -100.43  -28.70   55.03  496.80

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)    -9.294     42.255  -0.220    0.827
GrainSize     744.979    109.964   6.775 7.54e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 159.4 on 57 degrees of freedom
Multiple R-squared:  0.4461,Adjusted R-squared:  0.4363
F-statistic:  45.9 on 1 and 57 DF,  p-value: 7.543e-09
```

4. Compute the residuals for these data. Do any residuals exceed $\pm 3s_\epsilon$?

   (a) Residual:

```
> res= fit$residuals
> res
           1            2            3            4            5            6
 -24.8948554  -27.3647300  -43.6538518  -35.6538518  -61.1036427    3.8963573
           7            8            9           10           11           12
   9.9967755  -21.4530154  -34.9028063   10.0971937  -17.3525972  -53.8023882
          13           14           15           16           17           18
-101.2521791   27.7478209  127.7478209  -89.7019700 -110.1517609  -87.1517609
          19           20           21           22           23           24
 -99.6015518 -131.0513427  -82.5011336    0.4988664    3.0490755  -54.4007154
          25           26           27           28           29           30
 -39.4007154 -154.8505063    8.1494937 -123.3002972  -39.3002972   -8.7500881
          31           32           33           34           35           36
-125.1998790  213.2499119  115.8001210 -109.0994609   40.9005391 -116.3484154
          37           38           39           40           41           42
 113.6515846  -68.7982063  -41.2479972  -28.6977881   81.3022119  101.4026301
          43           44           45           46           47           48
 276.4026301  248.9528391  251.5030482    4.0532573   69.1536755  329.1536755
          49           50           51           52           53           54
-148.2961154  224.2540937  336.8043028  496.8043028  279.5553483 -425.6936063
          55           56           57           58           59
-187.6936063 -200.0429790 -141.9906790 -140.9864972 -198.4362881
```

   (b) Since

$$s_\epsilon^2 = \frac{\sum_1^n (y_i - \hat{y}_i)^2}{n-2}$$

   hence

$$s_\epsilon = \sqrt{\frac{\sum_1^n (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum_1^n \text{residual}^2}{n-2}} = 159.3766.$$

   Therefore $\pm 3s_\epsilon = [-478.1297, 478.1297]$.

   (c) Check: From the following check table, we can see that the 52-th residual (496.8043028 ) is not in the range.

```
> check = checkRange(res,-3*s_eps,3*s_eps)
 [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
 [8]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[15]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

```
[22]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[29]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[36]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[43]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[50]  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE
[57]  TRUE  TRUE  TRUE
```

□

# Appendix

## R code for HW#7

Listing 1: Source code for problem 1

```r
rm(list = ls())
# set the path or enverionment
setwd("/home/feng/Dropbox/UTK_Course/Stat537/HW#7/HW#7/code")


#install.packages("readxl") # CRAN version
library(readxl)
#install.packages("moments")
library(moments)
rawdata = read_excel("Data.xlsx",sheet = 1)
attach(rawdata)

plot(GrainSize, Yield, main="Scatterplot",
     xlab="Grain Size ", ylab="Yield ", pch=19)

#corrlation
cor(GrainSize, Yield, method="pearson")
cor(GrainSize, Yield, method="spearman")

#install.packages("Hmisc")
library(Hmisc)
out1<-cor.test(GrainSize,Yield,method ="pearson",conf.level=0.95)
out1
out2<-cor.test(GrainSize,Yield,method ="spearman",conf.level=0.95)
out2
#install.packages("mada")
library(mada)
N = dim(rawdata)[1]
CIrho(out2$estimate,N)

# regression

fit <- lm(Yield ~ GrainSize)
fit
summary(fit)

res= fit$residuals
s_eps = sqrt(sum(res^2)/(length(res)-2))
```

```
s_eps
range =c(-3*s_eps,3*s_eps)
range
i=length(res)


checkRange <- function(data, lower, upper) {
  n = length(data)
  result = logical(length = n)
  for (i in 1:n){
  result[i]=  data[i] >= lower && data[i] <= upper
  }
  print(result)
}

check = checkRange(res,-3*s_eps,3*s_eps)

a= -3*s_eps <= res
b= res <= 3*s_eps
a&&b
```