Statistical Methods – COSC 6323 - HomeWork-6

By Dinesh Narlakanti (2083649)

INTRODUCTION

More than 2,50,000 participants from Texas A&M University, University of Houston and University of California, Irvine. The data in this document has information about the participants who recorded their perinasal perspiration values, heart beat rate(Chest and Wrist) and breathe rate while performing different tasks i.e; Resting Baseline(RB), Single task(ST), Priming(PM), Relaxing Video(RV), Dual Task(DT) and Presentation Session(PR).

This report purely concentrates on:
i) Improving the agreement between two heart rate channels(chest and wrist).
ii) Computing p value, correlation coefficient r and determination coefficient r^2. of bivariate relationship between the heart rate channels.
iii) Strategy to clean the data(remove outliers) and rerunning the regression.

GETTING STARTED WITH THE HOMEWORK

Step-1 Installing required packages and importing data. Also, removing the NAs from chest and wrist heart rate channels and aggregating by participant id and treatment.

```
psy_data <- read.csv("C:/Users/ndine/Downloads/Physiological Data - QC1.csv")
psy_data <- psy_data[!is.na(psy_data$Chest_HR_QC),]
psy_data <- psy_data[!is.na(psy_data$Wrist_HR_QC),]

chest <- aggregate(psy_data$Chest_HR_QC, by =
                     list(psy_data$Participant_ID, psy_data$Treatment),
                   FUN = mean)
wrist <- aggregate(psy_data$Wrist_HR_QC, by =
                     list(psy_data$Participant_ID, psy_data$Treatment),
                   FUN = mean)
aggregated <- inner_join(chest, wrist, by = c('Group.1', 'Group.2'))
colnames(aggregated) <- c('Participant_ID', 'Treatment', 'Chest HR', 'Wrist HR')
```

Step-2: Running linear regression model and printing summary that contains p value, intercept coefficients and determination coefficient of it.

```
relation <- lm(aggregated$`Wrist HR` ~ aggregated$`Chest HR`)

summary1 <- summary(relation)
summary1
```

```
##
## Call:
## lm(formula = aggregated$`Wrist HR` ~ aggregated$`Chest HR`)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.185  -4.686  -1.556   3.033  42.482
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           44.41398    4.55097   9.759  < 2e-16 ***
```

```
## aggregated$`Chest HR`  0.45252    0.05627   8.042 1.14e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.471 on 180 degrees of freedom
## Multiple R-squared:  0.2643, Adjusted R-squared:  0.2603
## F-statistic: 64.68 on 1 and 180 DF,  p-value: 1.139e-13
```

Step-3: Running cor.test to find the correlation coefficient.

```
cor1 <-cor.test(aggregated$`Chest HR`, aggregated$`Wrist HR`)
cor1
```

```
##
##  Pearson's product-moment correlation
##
## data:  aggregated$`Chest HR` and aggregated$`Wrist HR`
## t = 8.0424, df = 180, p-value = 1.139e-13
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3984941 0.6137071
## sample estimates:
##      cor
## 0.514148
```

Step-4: Finding the absolute value of the difference of two heart rate channels.

```
aggregated$difference <- abs(aggregated$`Chest HR` - aggregated$`Wrist HR`)
```

Step-5: Applying strategy to remove the outliers

```
Q <- quantile(aggregated$difference, probs=c(.25, .75), na.rm = FALSE)

iqr <- IQR(aggregated$difference)

eliminated <- subset(aggregated, aggregated$difference > (Q[1] - 1.5*iqr) &
                       aggregated$difference < (Q[2]+1.5*iqr))
```

Step-6: Rerunning the regression and core.test

```
relation2 <- lm(eliminated$`Wrist HR` ~ eliminated$`Chest HR`)
summary2 <- summary(relation2)
summary2
```

```
##
## Call:
## lm(formula = eliminated$`Wrist HR` ~ eliminated$`Chest HR`)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.0684  -2.8808  -0.1053   2.0959  20.0771
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             22.76620    3.21181   7.088 3.77e-11 ***
## eliminated$`Chest HR`  0.70656    0.03961  17.836  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 5.201 on 165 degrees of freedom
## Multiple R-squared:  0.6585, Adjusted R-squared:  0.6564
## F-statistic: 318.1 on 1 and 165 DF,  p-value: < 2.2e-16
```

```
cor2<- cor.test(eliminated$`Wrist HR`, eliminated$`Chest HR`)
cor2
```

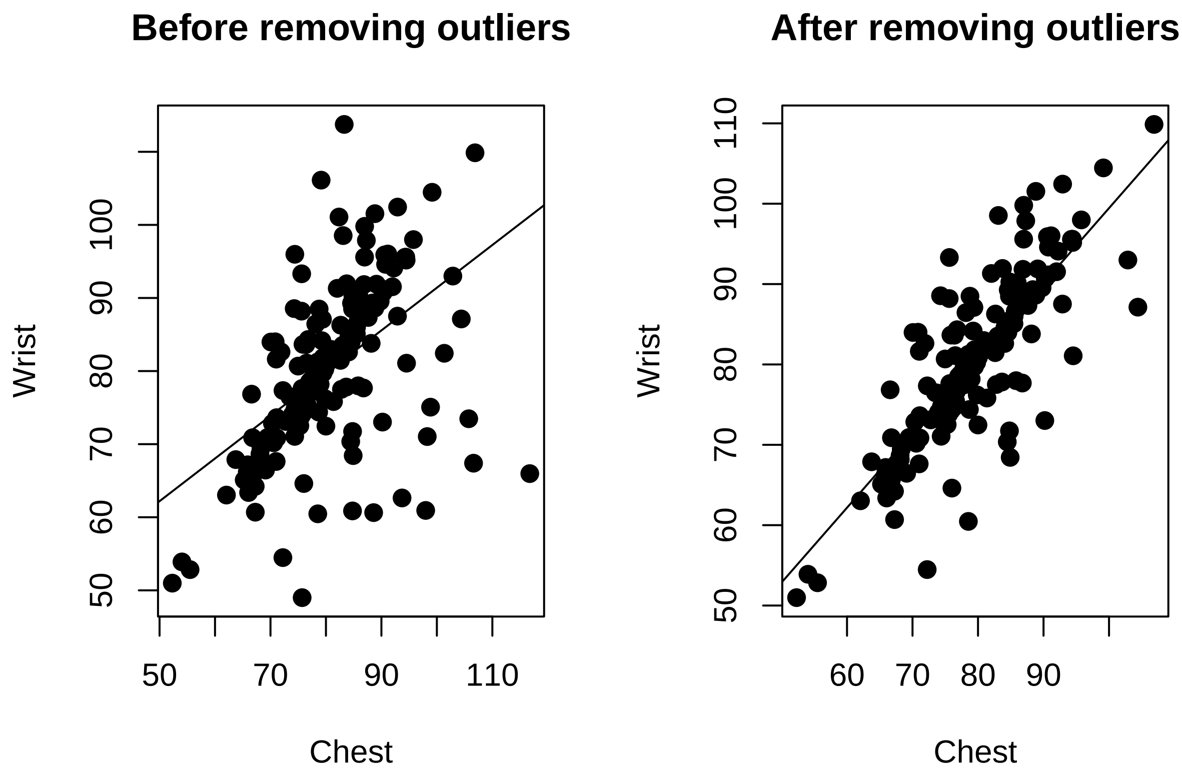
```
##
##  Pearson's product-moment correlation
##
## data:  eliminated$`Wrist HR` and eliminated$`Chest HR`
## t = 17.836, df = 165, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7523142 0.8576424
## sample estimates:
##       cor
## 0.8114682
```

Step-7: Plotting the graphs before and after removing outliers

```
par(mfrow=c(1,2))

plot(aggregated$`Wrist HR`,aggregated$`Chest HR`,main = "Before removing outliers",
     abline(lm(aggregated$`Chest HR` ~ aggregated$`Wrist HR`)), cex = 1.3,
     pch = 16,
     xlab = "Chest",
     ylab = "Wrist")

plot(eliminated$`Wrist HR`,eliminated$`Chest HR`,main = "After removing outliers",
     abline(lm(eliminated$`Chest HR` ~ eliminated$`Wrist HR`)), cex = 1.3,
     pch = 16,
     xlab = "Chest",
     ylab = "Wrist")
```

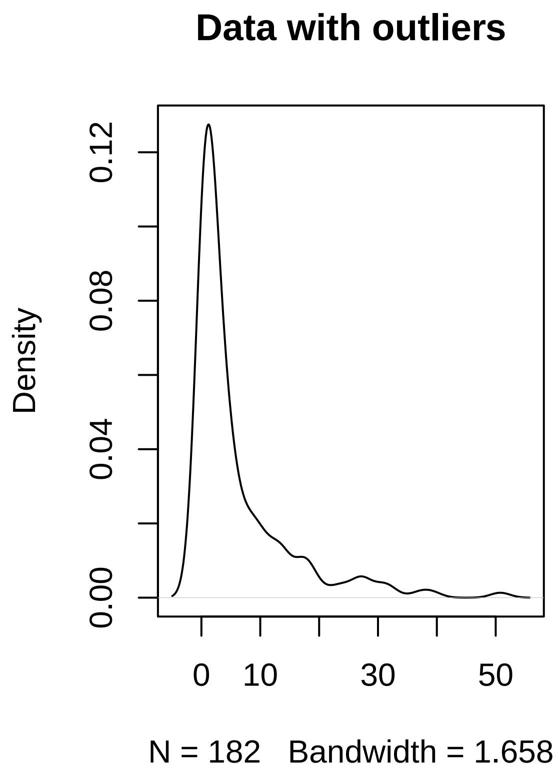


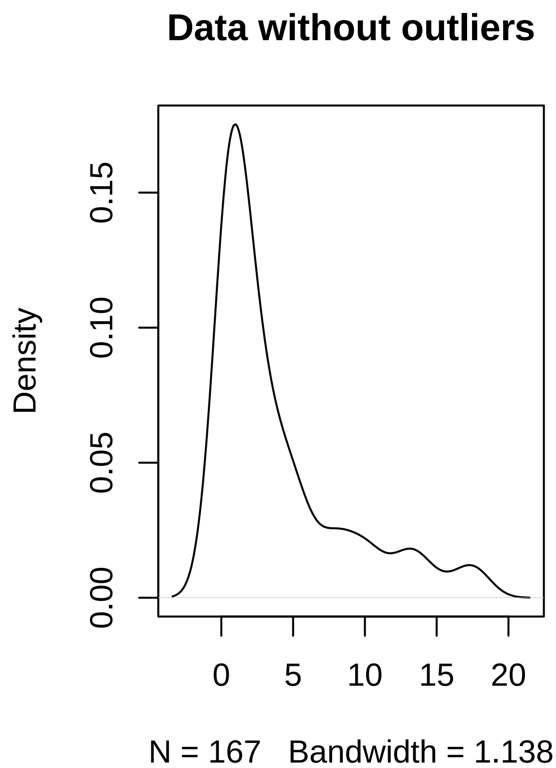
Step-8: Plotting the pdf of the difference before and after removing outliers

```
par(mfrow=c(1,2))

PDF <- density(aggregated$difference)
plot(PDF, main = "Data with outliers")

PDF2 <- density(eliminated$difference)
plot(PDF2, main = "Data without outliers" )
```





ANALYSIS OF THE RESULTS
1) Before removing Outliers
P-VALUE:

```
summary1$coefficients[8]
```

```
## [1] 1.139497e-13
```

R-Value:

```
cor1$estimate
```

```
##      cor
## 0.514148
```

R^2 Value:

```
summary1$r.squared
```

```
## [1] 0.2643482
```

2) After removing Outliers
P-VALUE:

```
summary2$coefficients[8]
```

```
## [1] 2.447296e-40
```

R-Value:

```
cor2$estimate
```

```
##       cor
```

```
## 0.8114682
```

R^2 Value:

```
summary2$r.squared
```

```
## [1] 0.6584807
```

3) Outliers mean that they are in slight disturbance from the remaining data. We can observe the outliers from the graph that has title 'Data with outliers'
4) Identified and removed outliers by using IQR strategy.
5) Yes, things got improved after removing the outliers. The notable improvements found are:
i. Correlation coefficient increased from 0.514 to 0.8114
ii. Determination coefficient increased from 0.2643 to 0.6585
iii. P-valued decreased from 1.139e-13 to $<$ 2.2e-16
6) As p-value is less than 0.05, we can reject the null hypothesis and conclude that both variables are significant to each other.
7) We can conclude that model is statistically significant by looking at the coefficients and p values. As p value decreases the model becomes significant.
8) Higher the t-value, the more significant is the model. Here, t-value increased from 8.042 to 17.836 after outliers are removed.
9) Standard Error and F-statistic can also be used to measure the good fitness of model.