Report_2

Your Name

16 October 2019

Reading the Data Files

The NHANES data set consists of four data files: adult.csv, youth.csv, lab.csv and exam.csv. The adult file contains information about subjects who were over 17 years old. The youth file contains information about younger subjects. The lab and exam files contain additional data about both adults and youth.

Reading the adult.csv file

The following command reads the adult.csv file into R.

```
dataDir = "/home/ralphw/su2/math207/nhanes/"
adult = read.csv(paste(dataDir,"adult.csv", sep=''), header=TRUE)
```

The size of this data set is

```
dim(adult)
```

```
## [1] 20050  1238
```

This means that the file has 20,050 rows and 1238 columns. So, there are 20,050 adults in the data set and 1238 pieces of information about each in this file. The lab.csv and exam.csv files include additional information about these subjects.

Reading the youth.csv file

The following command reads the youth.csv file into R.

```
youth = read.csv(paste(dataDir, "youth.csv", sep=''), header=TRUE)
```

The size of this data set is

```
dim(youth)
```

```
## [1] 13944   687
```

This means that the file has 13,944 rows and 687 columns.

Reading the exam.csv file

The following command reads the exam.csv file into R.

```
exam = read.csv(paste(dataDir, "exam.csv", sep=''), header=TRUE)
```

The exam of this data set is

```
dim(exam)
```

```
## [1] 31311  2368
```

There are 31,311 rows and 2368 columns in the exam file.

Reading the lab.csv file

The following command reads the lab.csv file into R.

```
lab = read.csv(paste(dataDir, "lab.csv", sep=''), header=TRUE)
```

The size of this data set is

```
dim(lab)
```

```
## [1] 29314   356
```

There are 29,314 rows and 356 columns in the lab file.

Regression Examples

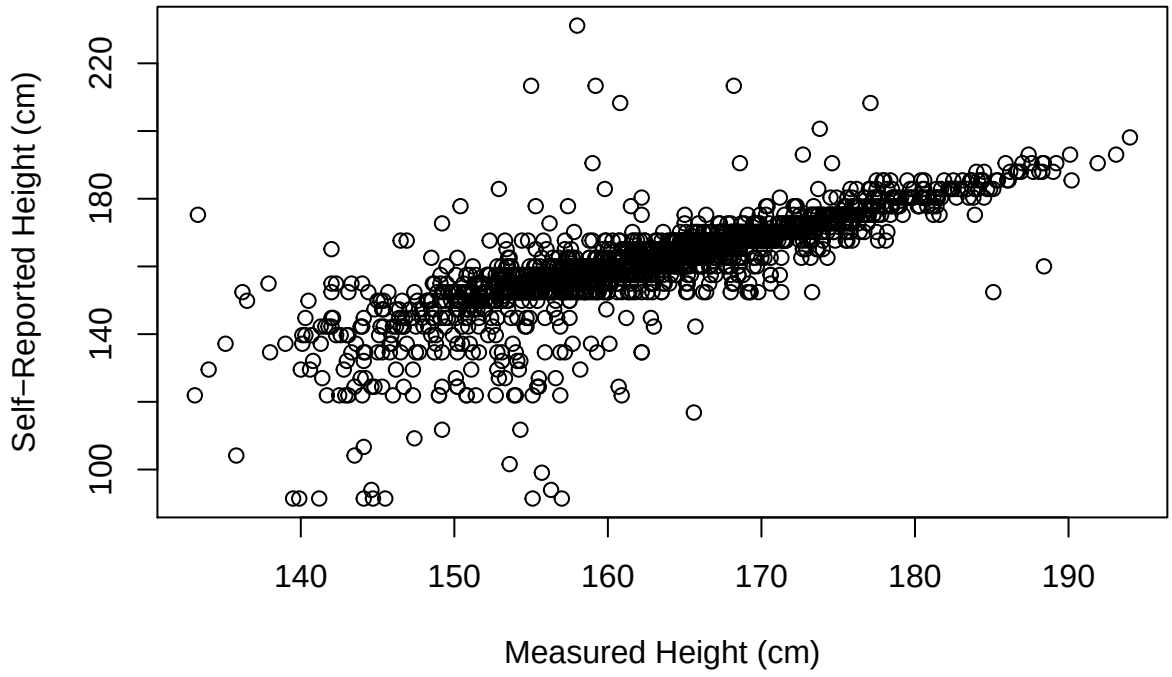
Measured Height vs Self-Reported Height

Extract the subject ID, measured height and self-reported height. Then remove the NA (blank) values and remove the missing values (error codes 88888 and 888). Finally, convert self-reported height from inches to cm. See page 199 of the exam-acc.pdf code book.

```
x1 <- data.frame(SEQN=exam$SEQN, BMPHT=exam$BMPHT, BMPSRHIS=exam$BMPSRHIS)
x2 <- na.omit(x1)
x3 <- x2[(x2$BMPH != 888888) & (x2$BMPSRHIS != 888), ]
x3$BMPSRHIS <- 2.54 * x3$BMPSRHIS
```

Plot the measured and reported heights.

```
plot(x3$BMPHT, x3$BMPSRHIS, xlab="Measured Height (cm)",
     ylab="Self-Reported Height (cm)")
```



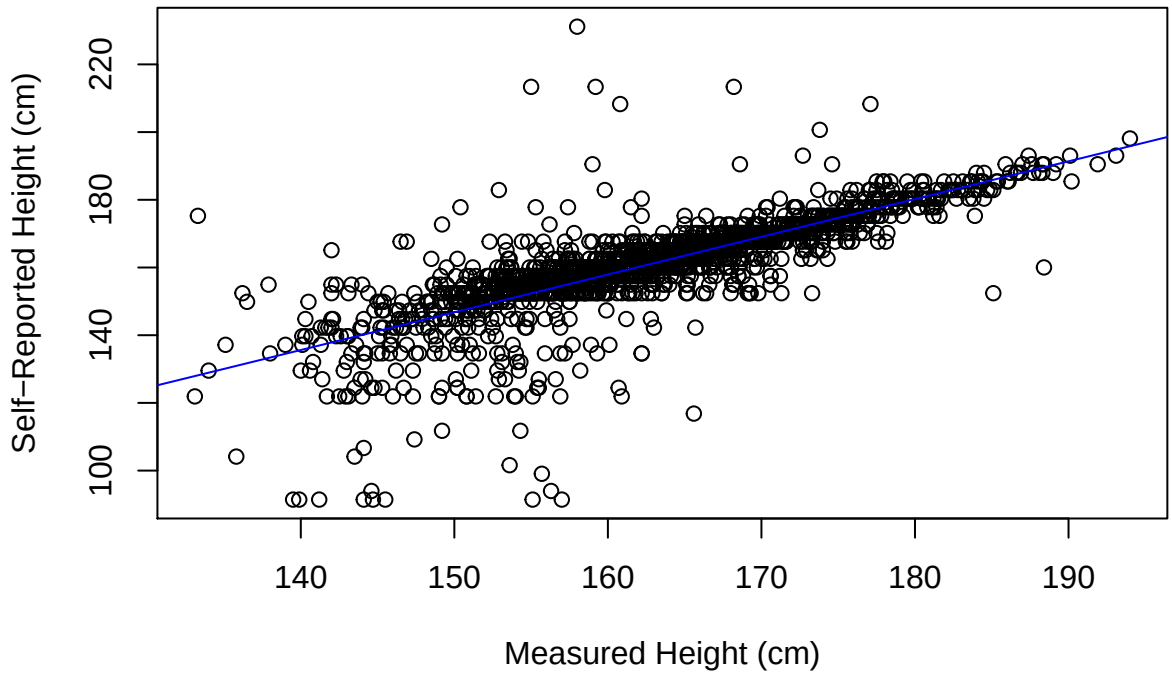
The correlation appears to be strong and positive.

```
cor(x3$BMPHT, x3$BMPSRHIS)
```

```
## [1] 0.7819654
```

Now compute the regression line and add it to the plot.

```
model = lm(x3$BMPSRHIS ~ x3$BMPHT)
plot(x3$BMPHT, x3$BMPSRHIS, xlab="Measured Height (cm)",
     ylab="Self-Reported Height (cm)")
abline(model, col='blue')
```




```
summary(model)
```

```
##
## Call:
## lm(formula = x3$BMPSRHIS ~ x3$BMPHT)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.133  -1.515   0.740   2.896  75.452
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -20.4829     3.2745  -6.255 4.86e-10 ***
## x3$BMPHT      1.1150     0.0201  55.469  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.477 on 1955 degrees of freedom
## Multiple R-squared:  0.6115, Adjusted R-squared:  0.6113
## F-statistic:  3077 on 1 and 1955 DF,  p-value: < 2.2e-16
```

From the output we determine that the regression line is

$$\text{self-reported height} = 1.11 \times \text{measured height} - 20.48$$

This equation suggests that taller people in the study tended to overestimate their heights.

Measured Height vs Recumbent Length

Do a similar regression calculation to compare BMPHT and BMPRECUM.

Measured Height vs Weight

Do a similar regression calculation to compare BMPHT and BMPWT.

Measured Height vs Head Circumference

Do a similar regression calculation to compare BMPHT and BMPHEAD.

Age vs Body Mass Index

Do a similar regression calculation to compare HSAGEIR and BMPBMI.

White Blood Cell Count vs Red Blood Cell Count

Do a similar regression calculation to compare WCPSI and RCPSI in the lab data set. See pages 70 and 71 of the lab-acc.pdf code book.