NHANES Regression Examples

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

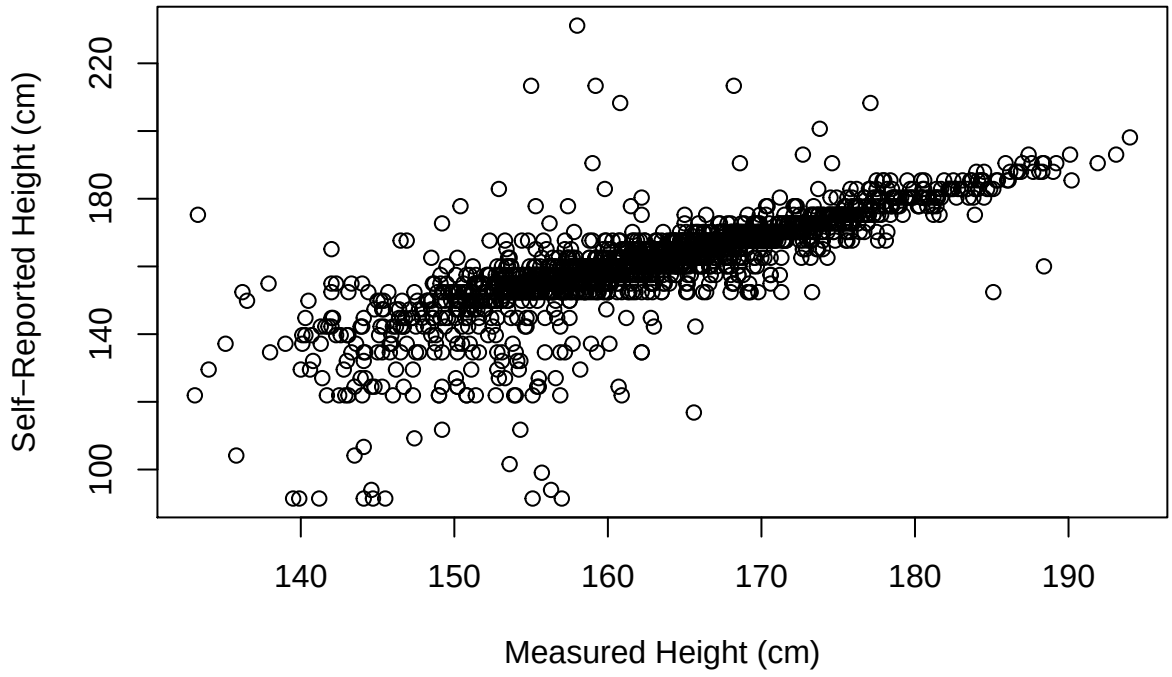
Measured Height vs Self-Reported Height

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$BMPSRHIS)
x2 <- na.omit(x1)
x3 <- x2[(x2$BMPHT != 88888) & (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)")
```



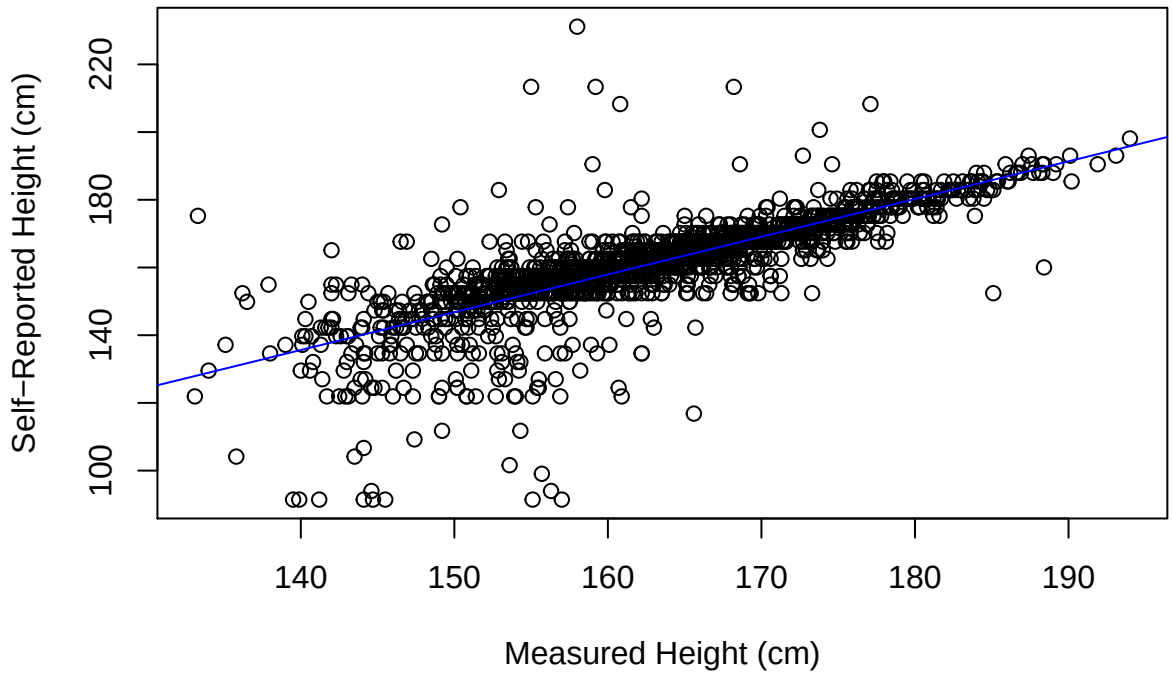
The correlation appears to be strong and positive.

```
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.

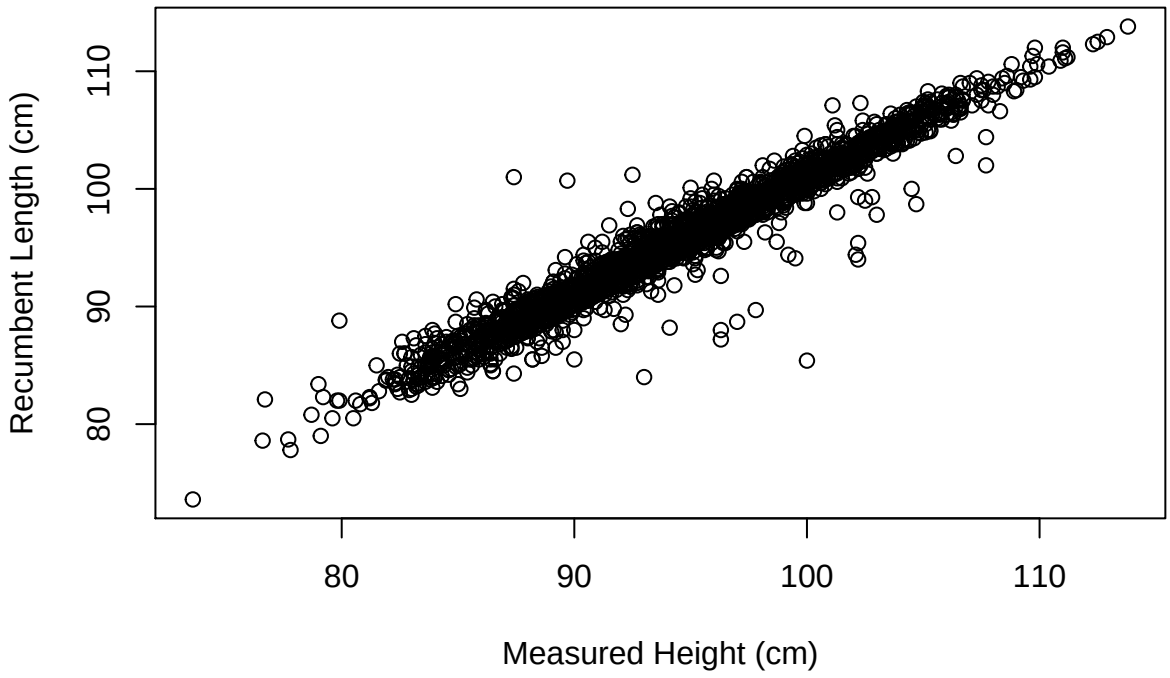
Measured Height vs Recumbent Length

Do a similar regression calculation to compare BMPHT and BMPRECUM.

```
x1 <- data.frame(SEQN=exam$SEQN, BMPHT=exam$BMPHT, BMPRECUM=exam$BMPRECUM)
x2 <- na.omit(x1)
x3 <- x2[(x2$BMPHT != 88888) & (x2$BMPRECUM != 88888), ]
```

Plot the measured and reported heights.

```
plot(x3$BMPHT, x3$BMPRECUM, xlab="Measured Height (cm)",
     ylab="Recumbent Length (cm)")
```



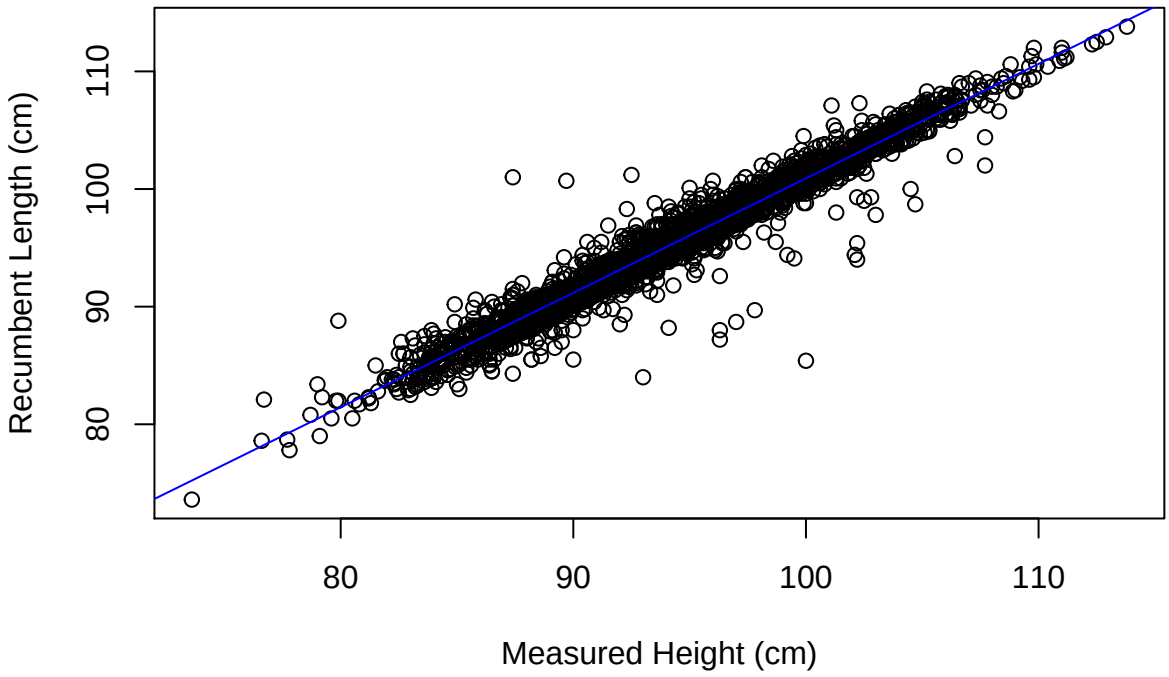
The correlation appears to be strong and positive.

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

```
## [1] 0.9740492
```

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

```
model = lm(x3$BMPRECUM ~ x3$BMPHT)
plot(x3$BMPHT, x3$BMPRECUM, xlab="Measured Height (cm)",
     ylab="Recumbent Length (cm)")
abline(model, col='blue')
```




```
summary(model)
```

```
##
## Call:
## lm(formula = x3$BMPRECUM ~ x3$BMPHT)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.5070  -0.6464   0.0209   0.6859  12.3536
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.60107    0.44705   8.055 1.26e-15 ***
## x3$BMPHT     0.97306    0.00472 206.166  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.382 on 2295 degrees of freedom
## Multiple R-squared:  0.9488, Adjusted R-squared:  0.9487
## F-statistic: 4.25e+04 on 1 and 2295 DF,  p-value: < 2.2e-16
```

From the output we determine that the regression line is

$$\text{recumbent length} = 0.973 \times \text{measured height} + 3.601$$

This equation suggests that taller people in the study tended to have longer recumbent length.

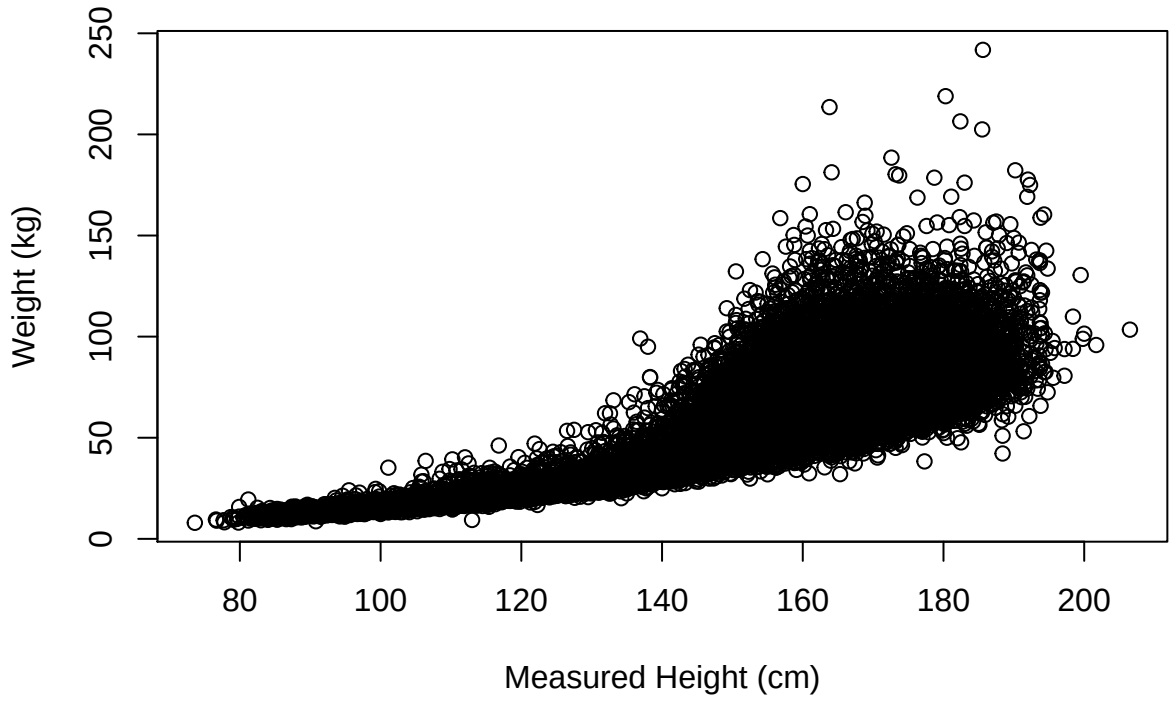
Measured Height vs Weight

Do a similar regression calculation to compare BMPHT and BMPWT.

```
x1 <- data.frame(SEQN=exam$SEQN, BMPHT=exam$BMPHT, BMPWT=exam$BMPWT)
x2 <- na.omit(x1)
x3 <- x2[(x2$BMPHT != 88888) & (x2$BMPWT != 888888), ]
```

Plot the measured and reported heights.

```
plot(x3$BMPHT, x3$BMPWT, xlab="Measured Height (cm)",
     ylab="Weight (kg)")
```



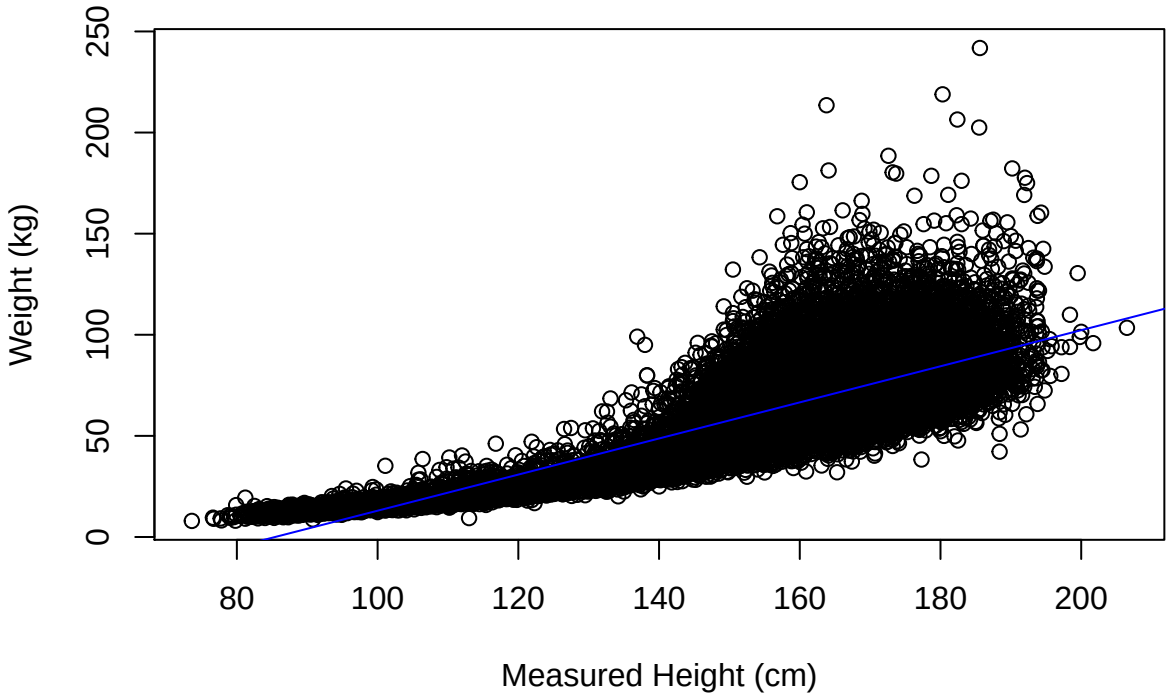
The correlation appears to be strong and positive.

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

```
## [1] 0.8463355
```

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

```
model = lm(x3$BMPWT ~ x3$BMPHT)
plot(x3$BMPHT, x3$BMPWT, xlab="Measured Height (cm)",
     ylab="Weight (kg)")
abline(model, col='blue')
```




```
summary(model)
```

```
##
## Call:
## lm(formula = x3$BMPWT ~ x3$BMPHT)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.714 -10.186  -1.661   7.302 152.386
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -76.282199   0.519422  -146.9   <2e-16 ***
## x3$BMPHT      0.892760   0.003368   265.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.72 on 27828 degrees of freedom
## Multiple R-squared:  0.7163, Adjusted R-squared:  0.7163
## F-statistic: 7.026e+04 on 1 and 27828 DF,  p-value: < 2.2e-16
```

From the output we determine that the regression line is

$$\text{weight} = 0.8928 \times \text{measured height} - 76.28$$

This equation suggests that taller people in the study tended to have longer recumbent length.

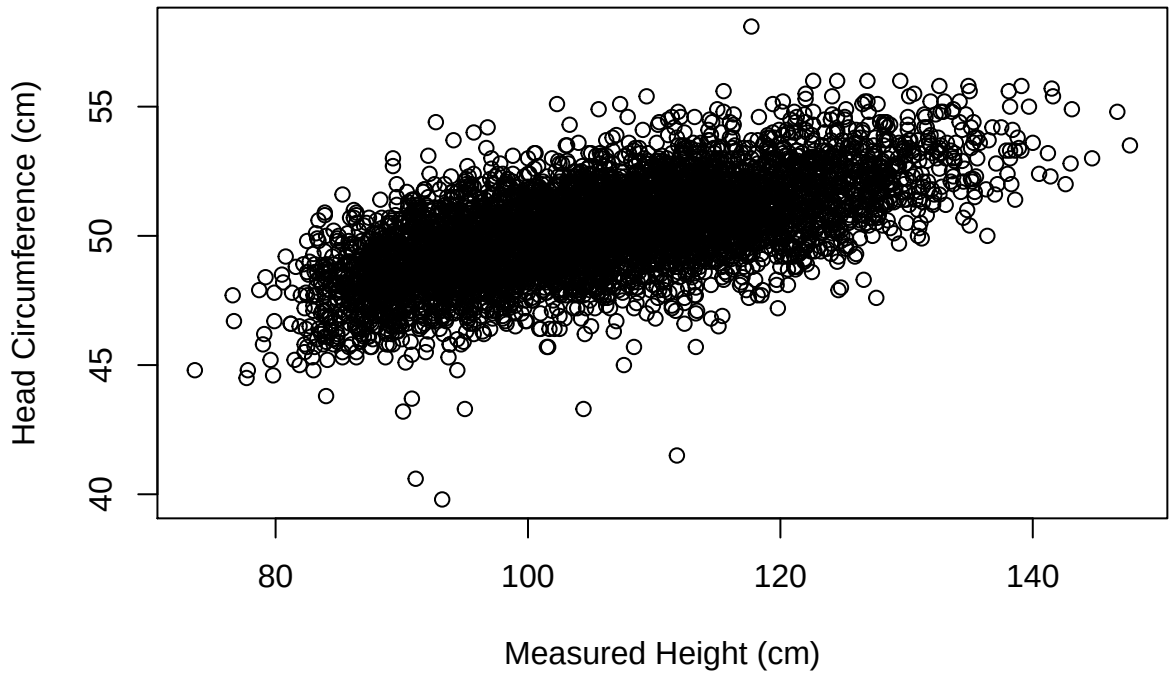
Measured Height vs Head Circumference

Do a similar regression calculation to compare BMPHT and BMPHEAD.

```
x1 <- data.frame(SEQN=exam$SEQN, BMPHT=exam$BMPHT, BMPHEAD=exam$BMPHEAD)
x2 <- na.omit(x1)
x3 <- x2[(x2$BMPHT != 88888) & (x2$BMPHEAD != 8888), ]
```

Plot the measured and reported heights.

```
plot(x3$BMPHT, x3$BMPHEAD, xlab="Measured Height (cm)",
     ylab="Head Circumference (cm)")
```



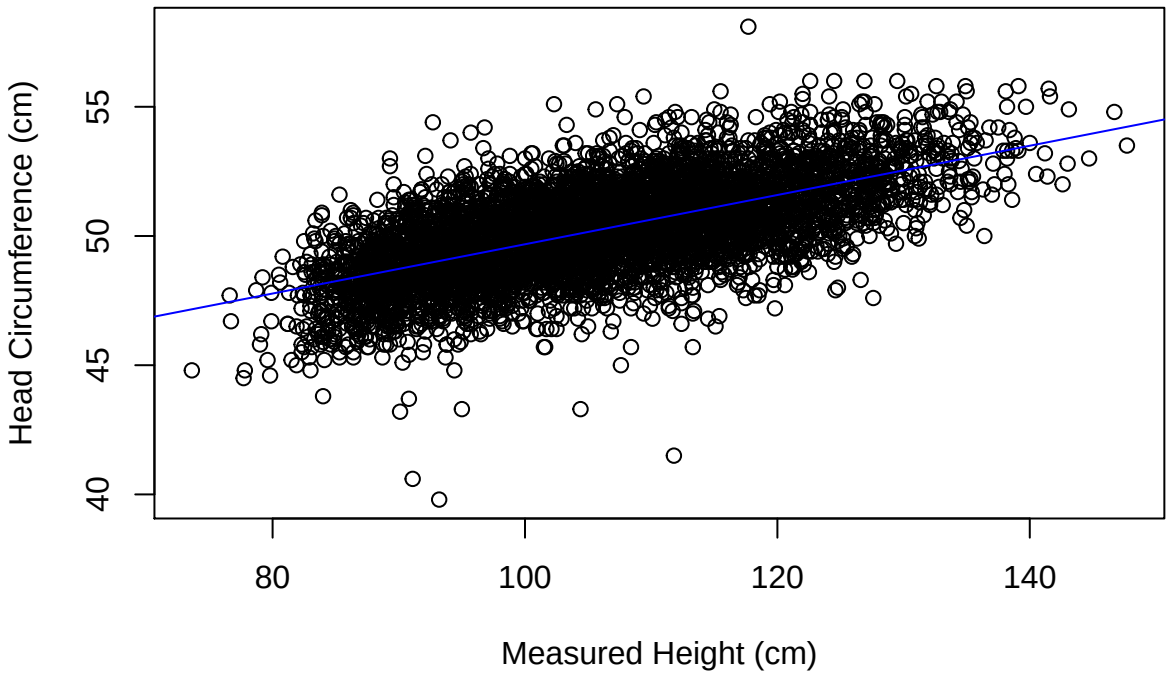
The correlation appears to be positive but not as strong as those in the previous examples.

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

```
## [1] 0.6392132
```

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

```
model = lm(x3$BMPHEAD ~ x3$BMPHT)
plot(x3$BMPHT, x3$BMPHEAD, xlab="Measured Height (cm)",
     ylab="Head Circumference (cm)")
abline(model, col='blue')
```




```
summary(model)
```

```
##
## Call:
## lm(formula = x3$BMPHEAD ~ x3$BMPHT)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3048 -0.9018  0.0011  0.9284  6.7328
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 40.147793   0.166256   241.5   <2e-16 ***
## x3$BMPHT     0.095322   0.001563    61.0   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.405 on 5386 degrees of freedom
## Multiple R-squared:  0.4086, Adjusted R-squared:  0.4085
## F-statistic:  3721 on 1 and 5386 DF,  p-value: < 2.2e-16
```

From the output we determine that the regression line is

$$\text{head circumference} = 0.0953 \times \text{measured height} + 40.15$$

This equation suggests that taller people in the study tended to have greater head circumferences.

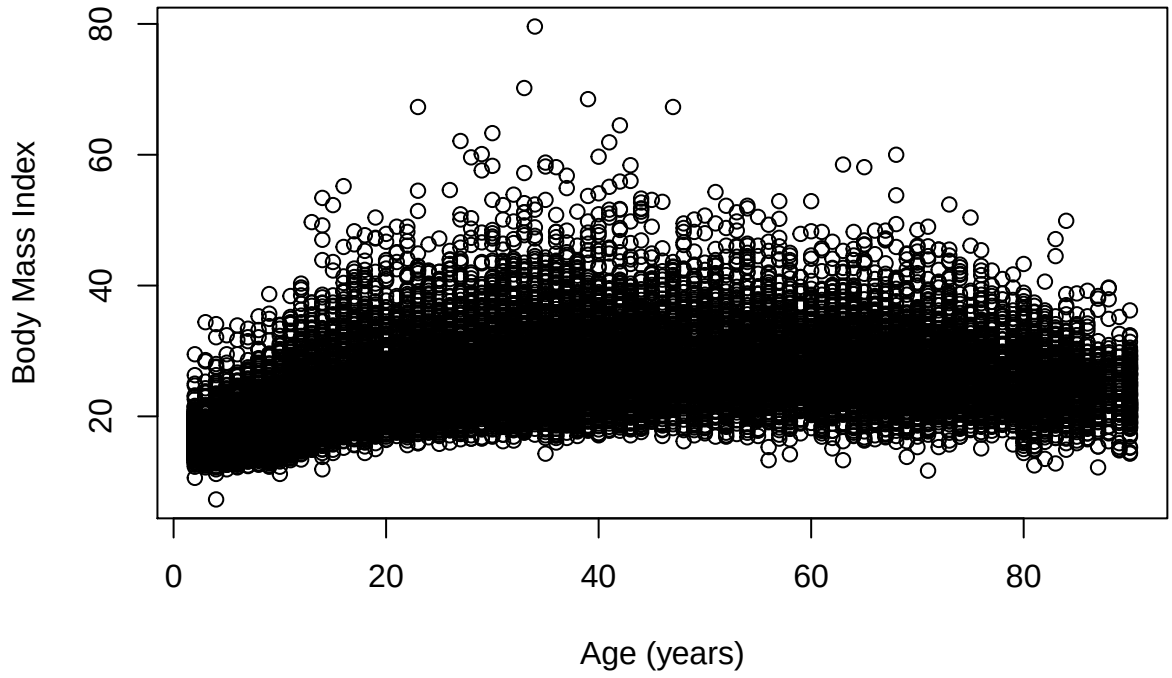
Age vs Body Mass Index

Do a similar regression calculation to compare HSAGEIR and BMPBMI.

```
x1 <- data.frame(SEQN=exam$SEQN, HSAGEIR=exam$HSAGEIR, BMPBMI=exam$BMPBMI)
x2 <- na.omit(x1)
x3 <- x2[(x2$HSAGEIR != 88888) & (x2$BMPBMI != 8888), ]
```

Plot the ages and body mass indices.

```
plot(x3$HSAGEIR, x3$BMPBMI, xlab="Age (years)",
     ylab="Body Mass Index")
```



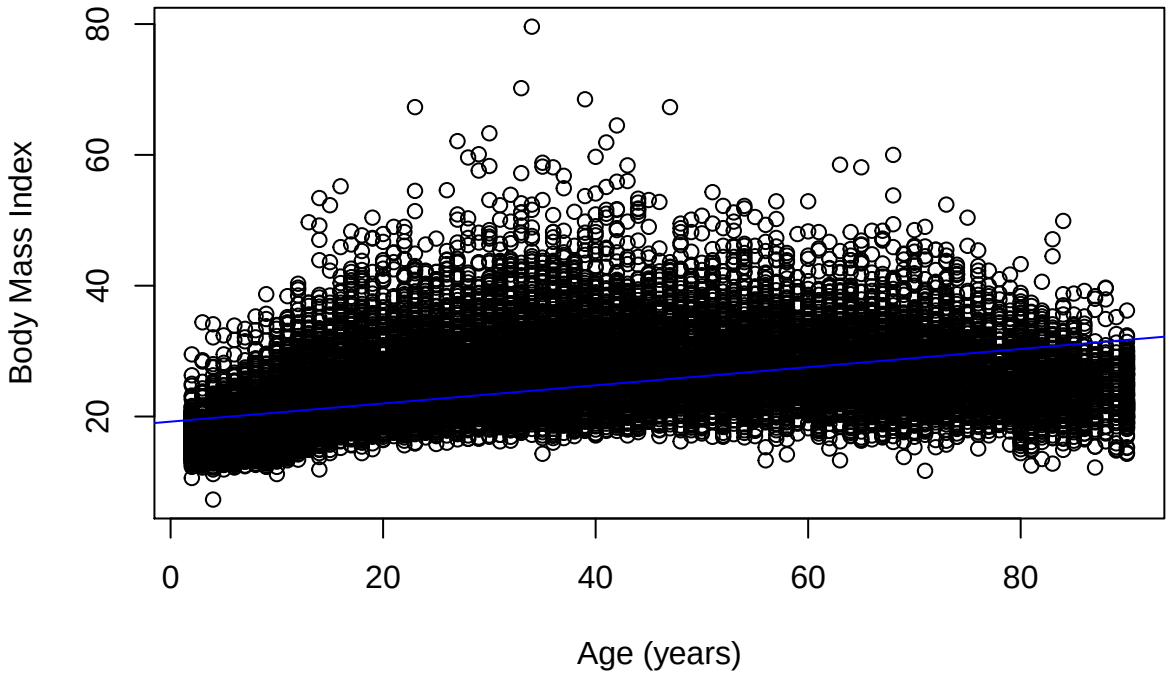
The correlation appears to be positive but not as strong as those in the previous examples.

```
cor(x3$HSAGEIR, x3$BMPBMI)
```

```
## [1] 0.515556
```

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

```
model = lm(x3$BMPBMI ~ x3$HSAGEIR)
plot(x3$HSAGEIR, x3$BMPBMI, xlab="Age (years)",
     ylab="Body Mass Index")
abline(model, col='blue')
```




```
summary(model)
```

```
##
## Call:
## lm(formula = x3$BMPBMI ~ x3$HSAGEIR)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.096  -3.913  -1.210   2.807  55.662
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19.216804   0.057429   334.6   <2e-16 ***
## x3$HSAGEIR   0.138844   0.001383   100.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.808 on 27828 degrees of freedom
## Multiple R-squared:  0.2658, Adjusted R-squared:  0.2658
## F-statistic: 1.007e+04 on 1 and 27828 DF,  p-value: < 2.2e-16
```

From the output we determine that the regression line is

$$\text{bmi} = 0.139 \times \text{age} + 19.22$$

This equation suggests that older people in the study tended to have higher bmi.

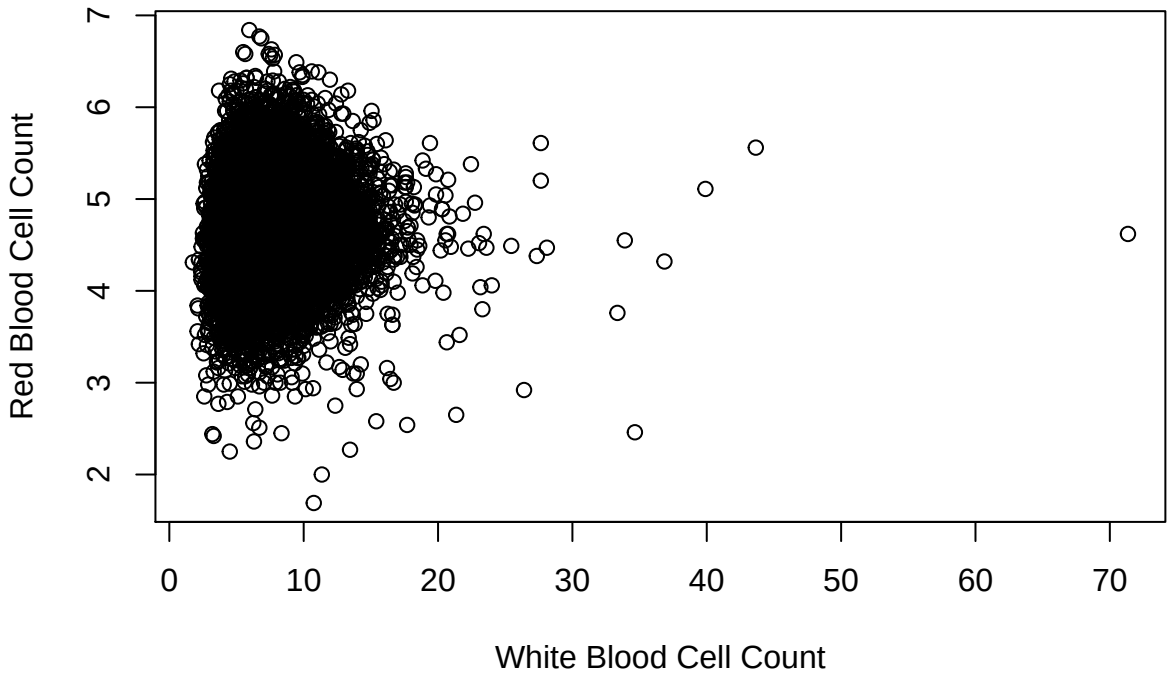
White Blood Cell Count vs Red Blood Cell Count

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.

```
x1 <- data.frame(SEQN=lab$SEQN, WCPSI=lab$WCPSI, RCPSI=lab$RCPSI)
x2 <- na.omit(x1)
x3 <- x2[(x2$WCPSI != 88888) & (x2$RCPSI != 8888), ]
```

Plot the white and red blood cell counts.

```
plot(x3$WCPSI, x3$RCPSI, xlab="White Blood Cell Count",
     ylab="Red Blood Cell Count")
```



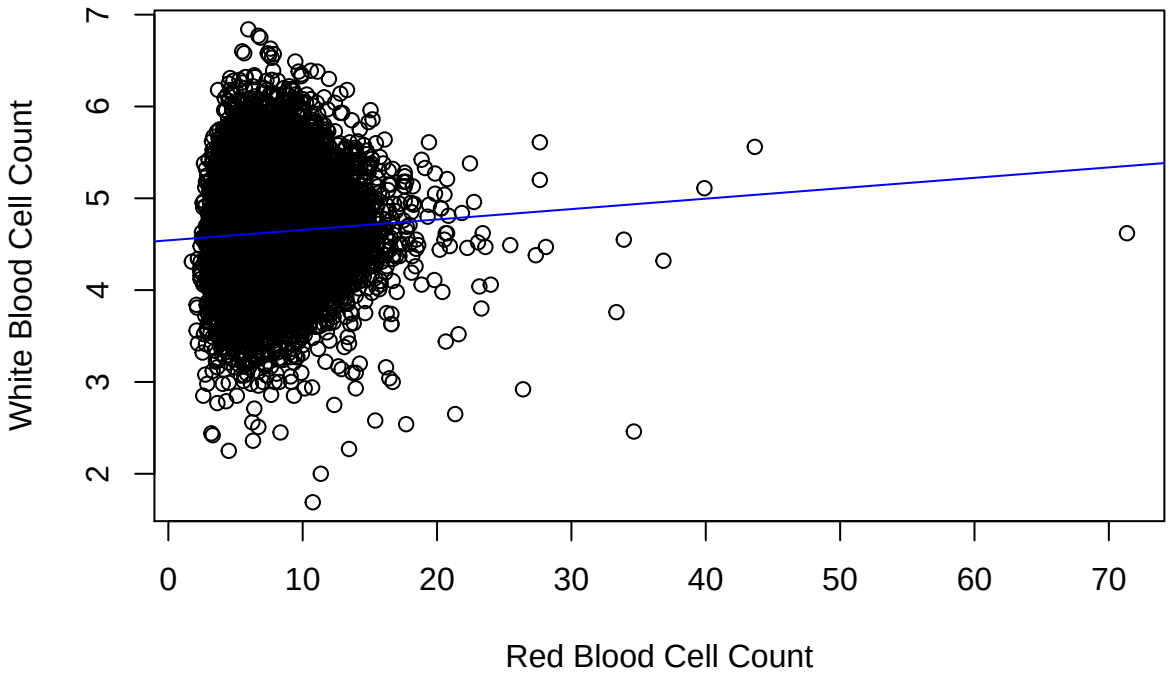
The correlation appears to be quite weak.

```
cor(x3$WCPSI, x3$RCPSI)
```

```
## [1] 0.05855297
```

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

```
model = lm(x3$RCPSI ~ x3$WCPSI)
plot(x3$WCPSI, x3$RCPSI, xlab="Red Blood Cell Count",
     ylab="White Blood Cell Count")
abline(model, col='blue')
```




```
summary(model)
```

```
##
## Call:
## lm(formula = x3$RCPSI ~ x3$WCPSI)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.97422 -0.30115 -0.02337  0.28364  2.23024
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.542256   0.009239 491.655   <2e-16 ***
## x3$WCPSI    0.011345   0.001191   9.524   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4543 on 26368 degrees of freedom
## Multiple R-squared:  0.003428,   Adjusted R-squared:  0.003391
## F-statistic: 90.71 on 1 and 26368 DF,  p-value: < 2.2e-16
```

From the output we determine that the regression line is

$$\text{red cell count} = 0.0113 \times \text{white cell count} + 4.542$$

This analysis suggests that white blood cell count is not a good predictor of red blood cell count.