PH245_hw2

PH245 HW2 Xiaoying Liu

1.

```
library(ggplot2)
library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 3.4.4
```

```
## Loading required package: Matrix
```

```
## Loading required package: foreach
```

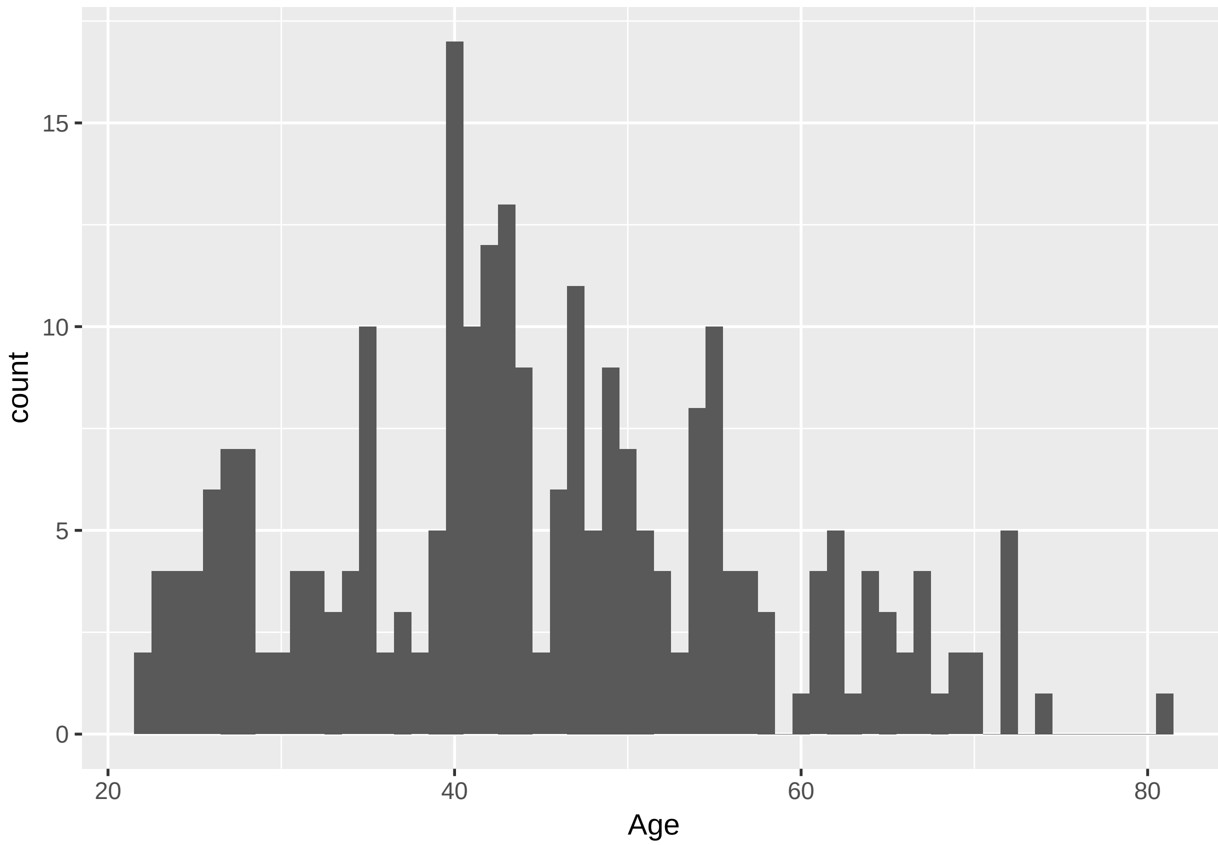
```
## Warning: package 'foreach' was built under R version 3.4.3
```

```
## Loaded glmnet 2.0-16
```

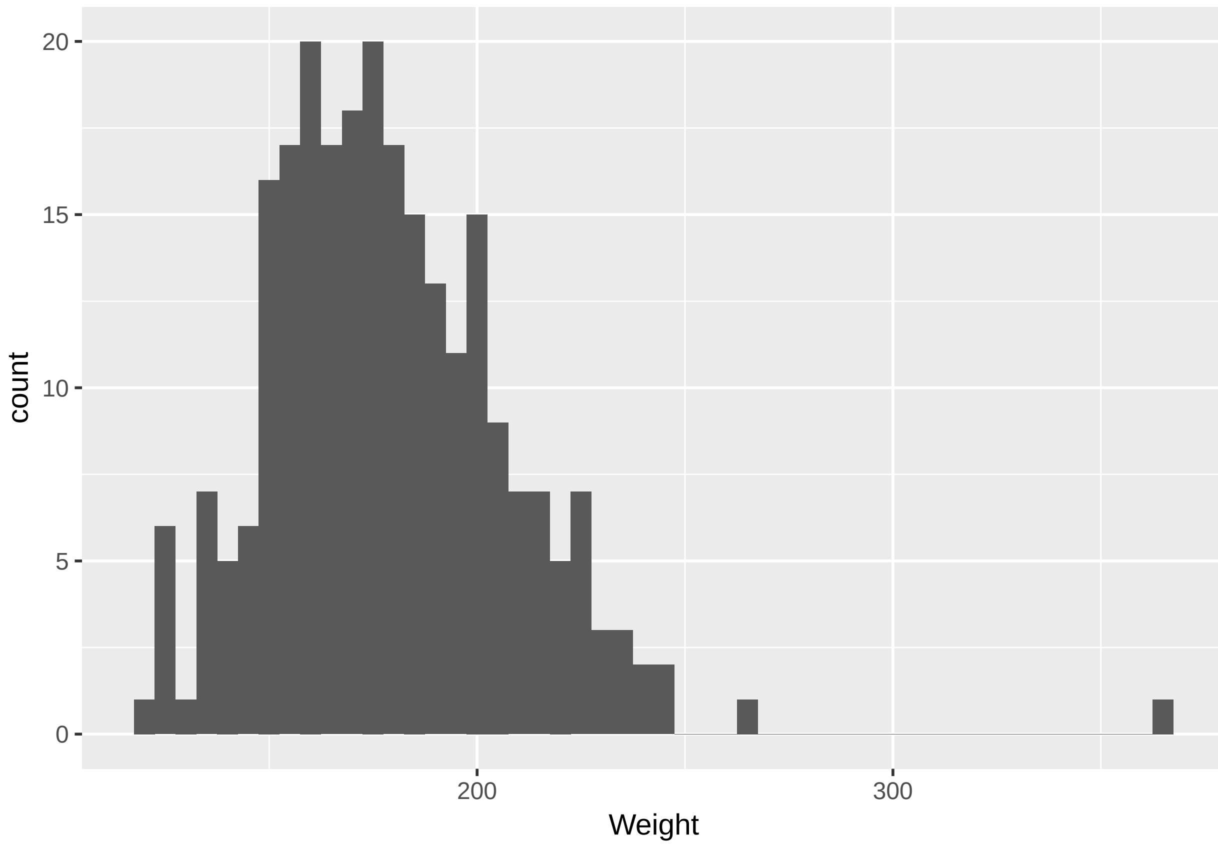
```
data=read.table(file='Data-HW2-Bodyfat.txt', header=F)
colnames(data)=c('Case Number', "BroznekBF",
                    "SiriBF", "Density",
                    "Age", "Weight", "Height", "AdipositIndex",
                    "FatFreeWeight", "NeckCirc", "ChestCirc",
                    "AbdomenCirc", "HipCirc", "ThighCirc",
                    "KneeCirc", "AnkleCirc",
                    "ExtendedBicepsCirc", "ForearmCirc",
                    "WristCirc")
#head(data)
```

#EDA

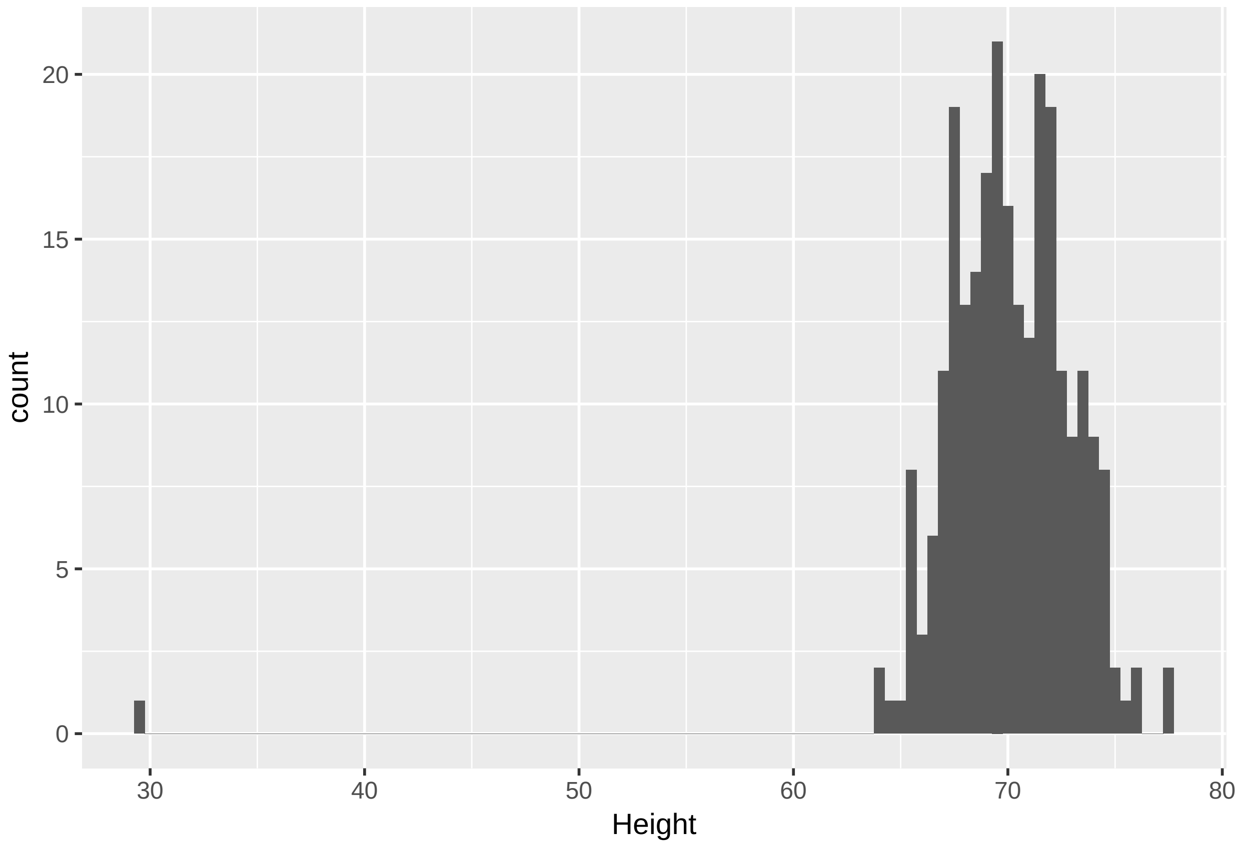
```
ggplot(data=data,aes(x=Age))+geom_histogram(binwidth=1)
```




```
ggplot(data=data,aes(x=Weight))+geom_histogram(binwidth=5)
```




```r
ggplot(data=data,aes(x=Height))+geom_histogram(binwidth=0.5)
```




```r
print(nrow(data))
```

```
## [1] 252
```

```r
cor(data)
```

```
##                      Case Number   BroznekBF      SiriBF     Density
## Case Number          1.000000000  0.11095086  0.11182544 -0.10960539
## BroznekBF            0.110950863  1.00000000  0.99974434 -0.98808673
## SiriBF               0.111825441  0.99974434  1.00000000 -0.98778240
## Density             -0.109605390 -0.98808673 -0.98778240  1.00000000
## Age                  0.341253503  0.28917352  0.29145844 -0.27763721
## Weight               0.033727935  0.61315611  0.61241400 -0.59406188
## Height               0.040943134 -0.08910641 -0.08949538  0.09788114
## AdiposityIndex       0.047717462  0.72799418  0.72748388 -0.71473204
## FatFreeWeight       -0.040092608  0.02013209  0.01937491 -0.00574871
## NeckCirc             0.071112330  0.49148893  0.49059185 -0.47296636
## ChestCirc            0.120514823  0.70288516  0.70262034 -0.68259865
## AbdomenCirc          0.121719735  0.81370622  0.81343228 -0.79895463
## HipCirc             -0.023736967  0.62569993  0.62520092 -0.60933143
## ThighCirc           -0.080708189  0.56128438  0.55960753 -0.55309098
## KneeCirc             0.047938697  0.50778587  0.50866524 -0.49504035
## AnkleCirc           -0.070644290  0.26678256  0.26596977 -0.26489003
## ExtendedBicepsCirc  -0.015676890  0.49303089  0.49327113 -0.48710872
## ForearmCirc          0.001959724  0.36327744  0.36138690 -0.35164842
## WristCirc            0.081845381  0.34757276  0.34657486 -0.32571598
##                             Age      Weight      Height AdiposityIndex
## Case Number          0.34125350  0.03372794  0.04094313     0.04771746
```

```
## BroznekBF           0.28917352  0.61315611 -0.08910641     0.72799418
## SiriBF              0.29145844  0.61241400 -0.08949538     0.72748388
## Density            -0.27763721 -0.59406188  0.09788114    -0.71473204
## Age                 1.00000000 -0.01274609 -0.17164514     0.11885126
## Weight             -0.01274609  1.00000000  0.30827854     0.88735216
## Height             -0.17164514  0.30827854  1.00000000    -0.02489094
## AdiposityIndex      0.11885126  0.88735216 -0.02489094     1.00000000
## FatFreeWeight      -0.23790534  0.79219519  0.48779841     0.54719009
## NeckCirc            0.11350519  0.83071622  0.25370988     0.77785691
## ChestCirc           0.17644968  0.89419052  0.13489181     0.91179865
## AbdomenCirc         0.23040942  0.88799494  0.08781291     0.92388010
## HipCirc            -0.05033212  0.94088412  0.17039426     0.88326922
## ThighCirc          -0.20009576  0.86869354  0.14843561     0.81270609
## KneeCirc            0.01751569  0.85316739  0.28605321     0.71365983
## AnkleCirc          -0.10505810  0.61368542  0.26474369     0.50031664
## ExtendedBicepsCirc -0.04116212  0.80041593  0.20781557     0.74638418
## ForearmCirc        -0.08505555  0.63030143  0.22864922     0.55859425
## WristCirc           0.21353062  0.72977489  0.32206533     0.62590659
##                    FatFreeWeight    NeckCirc  ChestCirc AbdomenCirc
## Case Number          -0.04009261  0.07111233  0.1205148  0.12171973
## BroznekBF             0.02013209  0.49148893  0.7028852  0.81370622
## SiriBF                0.01937491  0.49059185  0.7026203  0.81343228
## Density              -0.00574871 -0.47296636 -0.6825987 -0.79895463
## Age                  -0.23790534  0.11350519  0.1764497  0.23040942
## Weight                0.79219519  0.83071622  0.8941905  0.88799494
## Height                0.48779841  0.25370988  0.1348918  0.08781291
## AdiposityIndex        0.54719009  0.77785691  0.9117986  0.92388010
## FatFreeWeight         1.00000000  0.67911804  0.5929571  0.49565221
## NeckCirc              0.67911804  1.00000000  0.7848350  0.75407737
## ChestCirc             0.59295714  0.78483505  1.0000000  0.91582767
## AbdomenCirc           0.49565221  0.75407737  0.9158277  1.00000000
## HipCirc               0.70348104  0.73495788  0.8294199  0.87406618
## ThighCirc             0.67668053  0.69569734  0.7298586  0.76662393
## KneeCirc              0.70362435  0.67240498  0.7194964  0.73717888
## AnkleCirc             0.58294600  0.47789242  0.4829879  0.45322269
## ExtendedBicepsCirc    0.64929534  0.73114592  0.7279075  0.68498272
## ForearmCirc           0.55027717  0.62366027  0.5801727  0.50331609
## WristCirc             0.67335898  0.74482640  0.6601623  0.61983243
##                        HipCirc   ThighCirc    KneeCirc   AnkleCirc
## Case Number        -0.02373697 -0.08070819  0.04793870 -0.07064429
## BroznekBF           0.62569993  0.56128438  0.50778587  0.26678256
## SiriBF              0.62520092  0.55960753  0.50866524  0.26596977
## Density            -0.60933143 -0.55309098 -0.49504035 -0.26489003
## Age                -0.05033212 -0.20009576  0.01751569 -0.10505810
## Weight              0.94088412  0.86869354  0.85316739  0.61368542
## Height              0.17039426  0.14843561  0.28605321  0.26474369
## AdiposityIndex      0.88326922  0.81270609  0.71365983  0.50031664
## FatFreeWeight       0.70348104  0.67668053  0.70362435  0.58294600
## NeckCirc            0.73495788  0.69569734  0.67240498  0.47789242
## ChestCirc           0.82941992  0.72985855  0.71949640  0.48298789
## AbdomenCirc         0.87406618  0.76662393  0.73717888  0.45322269
## HipCirc             1.00000000  0.89640979  0.82347262  0.55838682
## ThighCirc           0.89640979  1.00000000  0.79917030  0.53979705
## KneeCirc            0.82347262  0.79917030  1.00000000  0.61160820
```

```
## AnkleCirc           0.55838682  0.53979705  0.61160820  1.00000000
## ExtendedBicepsCirc  0.73927252  0.76147745  0.67870883  0.48485454
## ForearmCirc         0.54501412  0.56684218  0.55589819  0.41904999
## WristCirc           0.63008954  0.55868478  0.66450729  0.56619459
##                    ExtendedBicepsCirc  ForearmCirc   WristCirc
## Case Number                -0.01567689  0.001959724  0.08184538
## BroznekBF                   0.49303089  0.363277442  0.34757276
## SiriBF                      0.49327113  0.361386903  0.34657486
## Density                    -0.48710872 -0.351648418 -0.32571598
## Age                        -0.04116212 -0.085055552  0.21353062
## Weight                      0.80041593  0.630301433  0.72977489
## Height                      0.20781557  0.228649220  0.32206533
## AdiposityIndex              0.74638418  0.558594251  0.62590659
## FatFreeWeight               0.64929534  0.550277173  0.67335898
## NeckCirc                    0.73114592  0.623660267  0.74482640
## ChestCirc                   0.72790748  0.580172731  0.66016232
## AbdomenCirc                 0.68498272  0.503316087  0.61983243
## HipCirc                     0.73927252  0.545014120  0.63008954
## ThighCirc                   0.76147745  0.566842179  0.55868478
## KneeCirc                    0.67870883  0.555898191  0.66450729
## AnkleCirc                   0.48485454  0.419049991  0.56619459
## ExtendedBicepsCirc          1.00000000  0.678255131  0.63212642
## ForearmCirc                 0.67825513  1.000000000  0.58558825
## WristCirc                   0.63212642  0.585588251  1.00000000
```

#(a)

```r
#response variable
siriBF=data$SiriBF

#predictor variable
age=data[,5]
weight=data[,6]
height=data[,7]
circumferences=data[,10:19]
predictors=cbind(age, weight, height, circumferences)

fittingData=cbind(siriBF,predictors)

fittingDataNoOutliers = fittingData[-c(seq(1, nrow(fittingData))[fittingData$weight > 300],
                                      seq(1, nrow(fittingData))[fittingData$height < 40]
                                     ),]

#fitting
fit = lm(formula=siriBF~., data=fittingDataNoOutliers)

summary(fit)
```

```
##
## Call:
## lm(formula = siriBF ~ ., data = fittingDataNoOutliers)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.9900  -3.1244  -0.1674   3.0248   9.8648
```

```
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.68516   23.37412   0.072 0.942587
## age                   0.07189    0.03217   2.234 0.026389 *
## weight               -0.01762    0.06714  -0.263 0.793153
## height               -0.24675    0.19114  -1.291 0.197989
## NeckCirc             -0.38682    0.23486  -1.647 0.100887
## ChestCirc            -0.11919    0.10825  -1.101 0.272004
## AbdomenCirc           0.90452    0.09140   9.897  < 2e-16 ***
## HipCirc              -0.15878    0.14586  -1.089 0.277446
## ThighCirc             0.17299    0.14683   1.178 0.239926
## KneeCirc             -0.04580    0.24560  -0.186 0.852230
## AnkleCirc             0.18502    0.21985   0.842 0.400862
## ExtendedBicepsCirc    0.17968    0.17039   1.054 0.292732
## ForearmCirc           0.27605    0.20692   1.334 0.183454
## WristCirc            -1.80162    0.53304  -3.380 0.000848 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.255 on 236 degrees of freedom
## Multiple R-squared:  0.7505, Adjusted R-squared:  0.7368
## F-statistic: 54.61 on 13 and 236 DF,  p-value: < 2.2e-16
```

#(b)

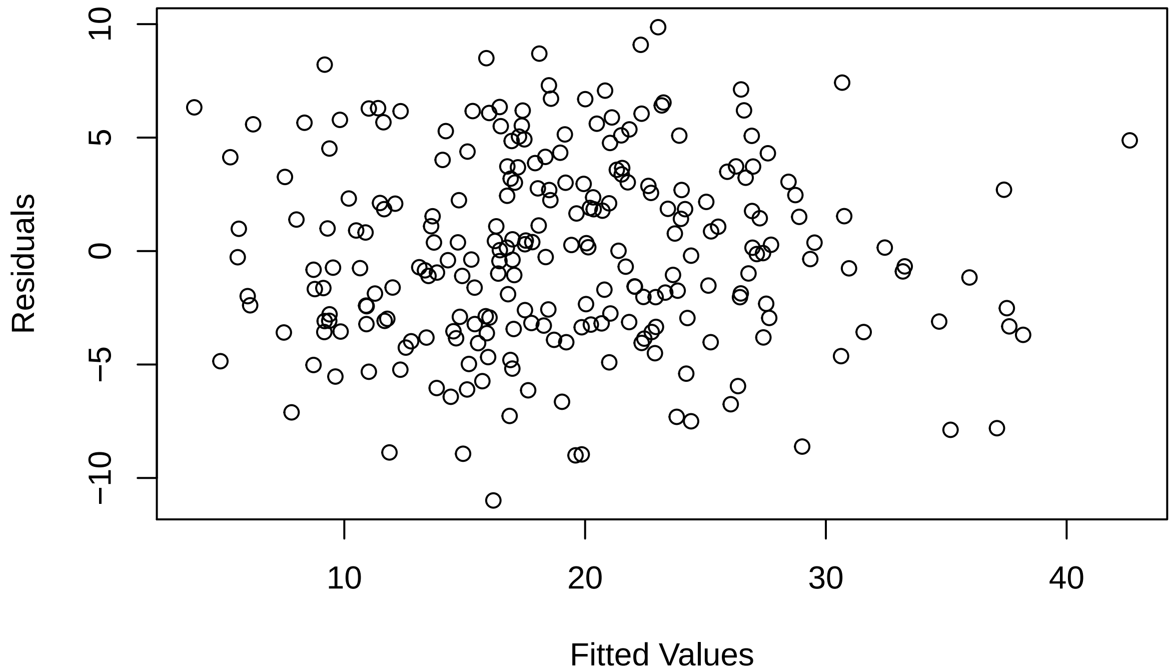
```
#Age Coefficient Estimate: .07189
#Interpretation: For every increase in age by 1 year, there is a .07189 increase in body fat percentage
#via Siri's equation.
#P-Value: .026389
#Hypothesis Test with alpha=.05: We would reject our null hypothesis that the coefficient estimate
#of age is 0
```

#(c)

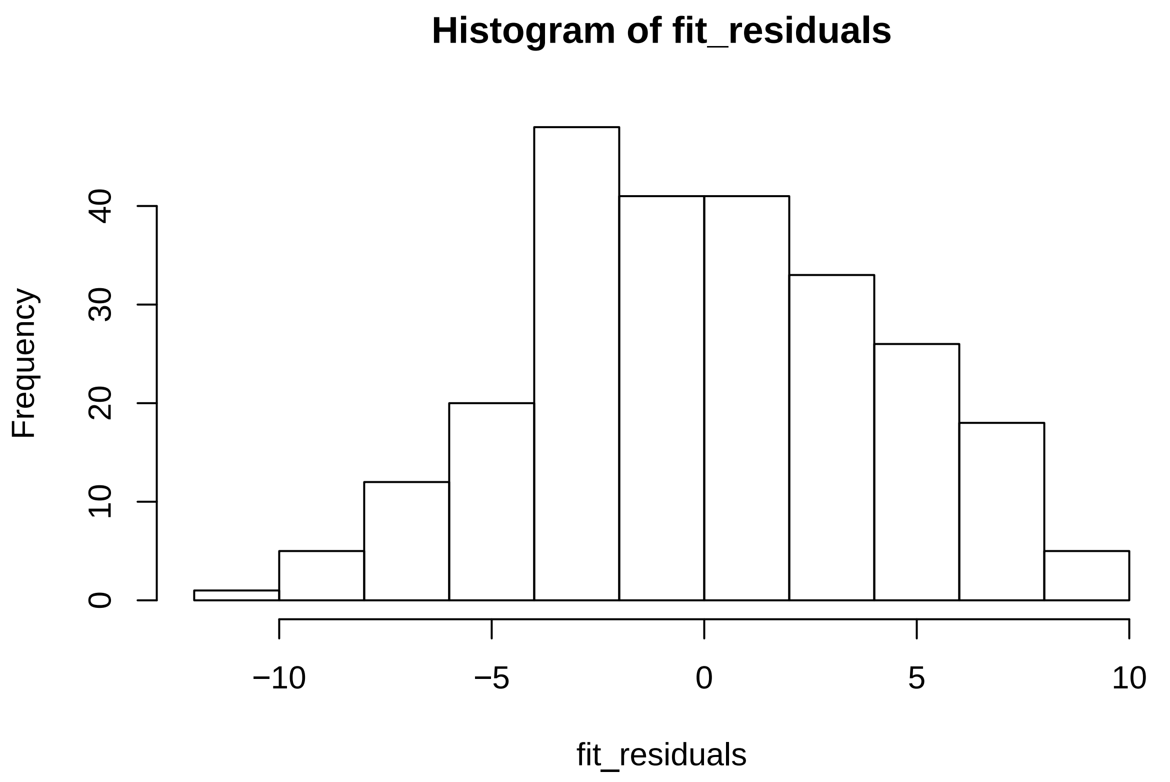
```
#Abdomen Circumference Coefficient Estimate: 0.90452
#Interpretation: For every increase in Abdomen Circumference by 1 centimeter, there is a 0.90452 increa
#in observed body fat percentage via Siri's equation.
#P-Value: nearly 0
#Hypothesis Test with alpha=.05: We would reject our null hypothesis that the coefficient estimate
#of Abdomen Cicumference is 0
```

#(d)

```
fit_values = fitted.values(fit)
fit_residuals = residuals(fit)
plot(x=fit_values, y=fit_residuals, xlab='Fitted Values', ylab='Residuals')
```




```r
hist(fit_residuals)
```




```r
#The residual plot appears to be fine -- points seem to be randomly scatteredaround the line y=0.
#There doesn't seem to be any sort of particular shape indicating bias.

#key assumptions
#1.There must be linear relationships between our response and predictor variables.
#2.Residuals should be normally distributed - The histogram shows a nearly normal distribution.
```

```
#3.There is no multicollinearity. From EDA, eight is heavily correlated with many of the circumferences
#and many of the circumferences seem to be correlated with each other (i.e. hip and thigh)
#4.Homoscedasticity. there doesn't seem to be any sort of variance in residual across fitted values
#and around the line y=0. There also doesn't seem to be any bias in the shape of a particular shape in
```

#(e)

```
#In class, we fit the model with 3 predictor variables(age, weight, and height),and all 4 terms are
#assumed to be statistically significant to body fat percentage. However, in our full model, only Age,
#Abdomen circumference and Wrist cricumference are statistically significant to body fat percentage.
#With larger number of predictors, the coefficient of any given predictor is likely to grow smaller
#since it contributes less to the response variable.

#Weight has the smallest p value in reduced model, but weight is one of the least significant predictor
#in the full model. Since weight being highly correlated with many of the circumference values,
#when these circumference values are added into the model, the coefficient of weight may decrease
#because the it captures the essence of circumferences in class model but not in full model.


#In terms of adjusted $R^{2}$, this statistic provides a measure of how well the model is fitting the a
#The adjusted $R^{2}$ helps to explain how much of the variance in our response variable is
#due to our predictor variables. Our class model captures less of the variance than our more full-featu
```

#(f)

```
#We are looking at the magnitude of the differences ($Residuals^{2}$).

#Null hypothesis: mean ( $Residuals_{Reduced}^{2}$ ) = mean ( $Residuals_{Full}^{2}$ ).
#The variance in the observed residuals is due to random chance and both models are equally accurate.

#Alternative hypothesis: mean ( $Residuals_{Reduced}^{2}$ ) < mean ( $Residuals_{Full}^{2}$ ).
#The variance in the observed residuals is not due to random chance and the full model,
#with greater accuracy than the reduced model (smaller residuals), is preferred.


# Find (Residuals of Full)^2
full_squared_residuals = fit_residuals**2
head(full_squared_residuals)
```

```
##         1         2         3         4         5         6
## 13.100785  9.515232 45.093203  2.581479  3.114926 13.610265
```

```
# Find the (Residuals of Reduced)^2
reducedFittingData = cbind(siriBF, data[,5:7]) # Relevant Dataset: Response + Reduced Predictors

reducedFittingDataNoOutliers = reducedFittingData[
    -c(seq(1, nrow(reducedFittingData))[reducedFittingData$Weight > 300],
       seq(1, nrow(reducedFittingData))[reducedFittingData$Height < 40]
      ),]

reducedFit = lm(formula=siriBF~., data=reducedFittingDataNoOutliers)
reduced_squared_residuals = residuals(reducedFit) ** 2
head(reduced_squared_residuals)
```

```
##           1           2           3           4           5           6
##   0.8719114  36.9942272 107.2414618  27.5486664 148.3193303   5.0311087
```

```r
# Run a T-Test on the two sets of squared residuals to determine whether the observed variance
#in the two sets of residuals is significant
ttest = t.test(full_squared_residuals, reduced_squared_residuals)

#Show the results of the T-Test
print("Null Hypothesis:")
```

```
## [1] "Null Hypothesis:"
```

```r
ttest$null.value
```

```
## difference in means
##                   0
```

```r
print("CI of the difference:")
```

```
## [1] "CI of the difference:"
```

```r
ttest$conf.int
```

```
## [1] -16.309653  -6.500809
## attr(,"conf.level")
## [1] 0.95
```

```r
print(paste("T-Statistic:", ttest$statistic))
```

```
## [1] "T-Statistic: -4.57134819786503"
```

```r
print(paste("P-value", ttest$p.value))
```

```
## [1] "P-value 6.41913102273647e-06"
```

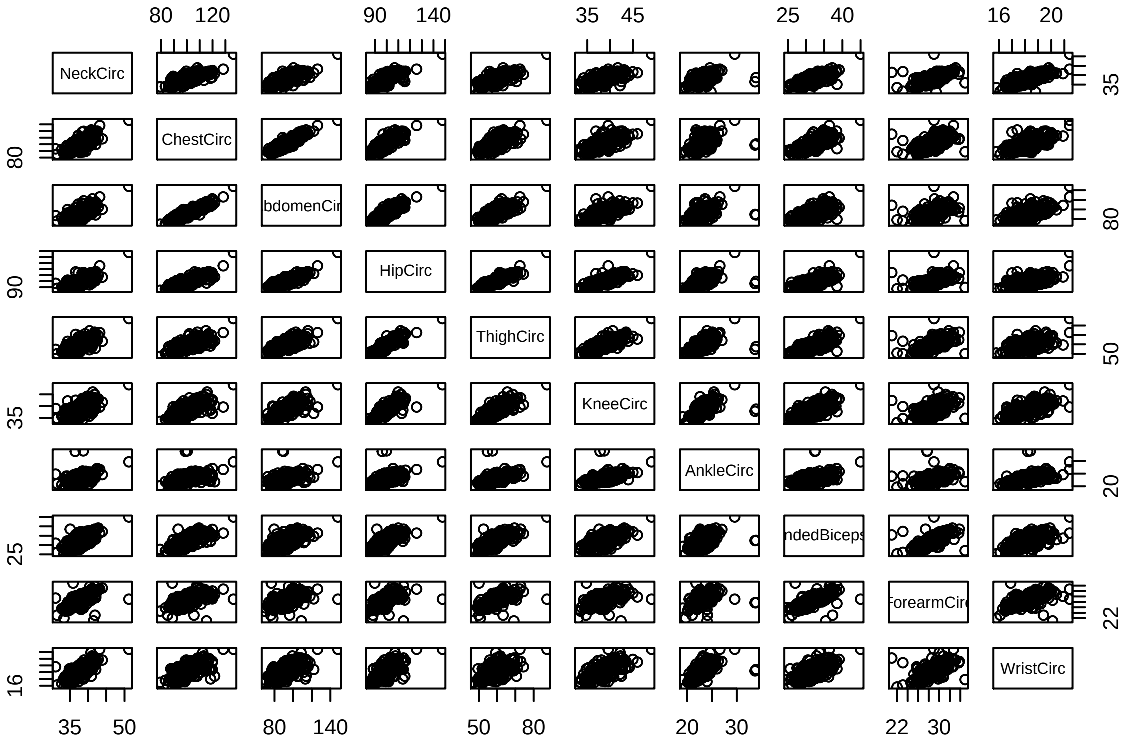
```r
#Interpreting the T-Test: in our T-Test, we generated a 95% confidence interval [-16.31, -6.5]
#indicating that we are 95% confident that the true value of the difference between our
#two residual means lies in that range. With a p-value of nearly 0, we reject our null hypothesis
#that the variance in the observed residuals is random.

#What we've tested and found is that the squared residuals of the reduced model are larger than
#the squared residuals of our full model in a statistically significant way.

#Thus, our full model is preferred over the reduced model.
```

#(g)

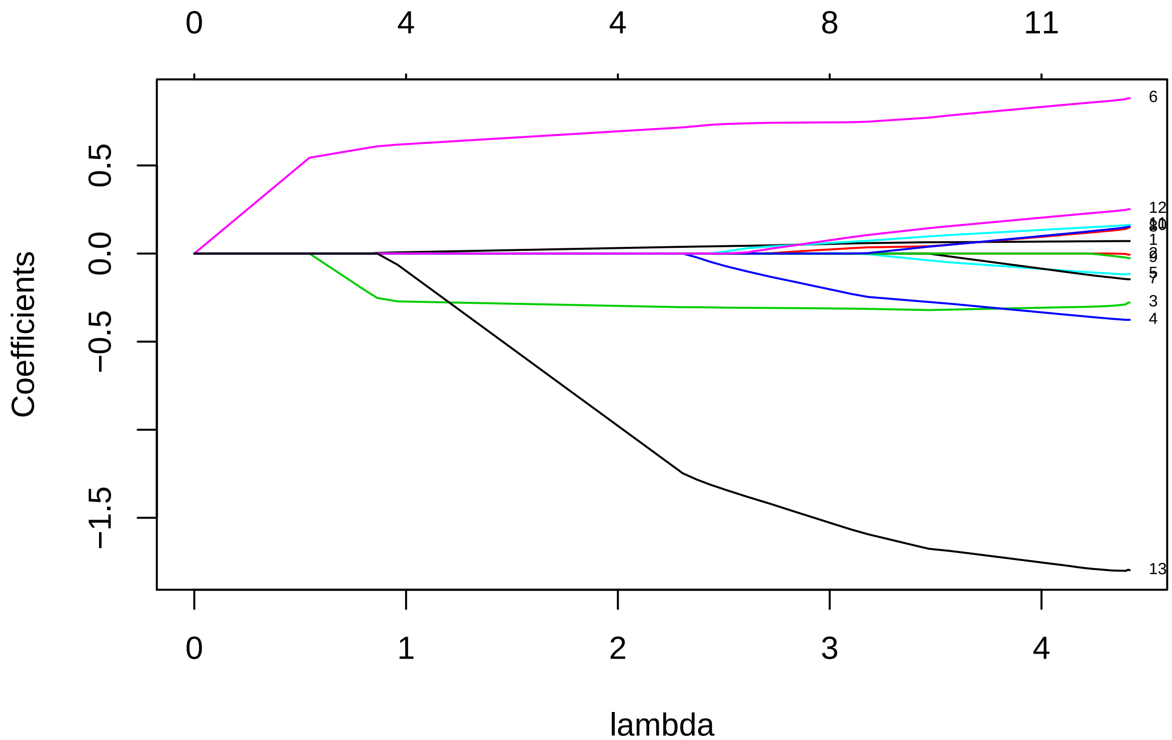
```r
plot(data[,10:19])
```




```
#Observing scatter plot,there are pretty high correlations among all of the variables. This matches
#our intuition that these circumferences strongly correlated as a human being.
#LASSO regularization can zero out some relatively insignificant parameters,
#so that there is less multicollinearity among our predictor variables.
```

#(h)

```
# Using cross-validation to obtain the best lambda value
lassoModel = cv.glmnet(x=as.matrix(fittingDataNoOutliers[,2:ncol(fittingDataNoOutliers)]),
                       y=as.matrix(fittingDataNoOutliers[,1]),
                       alpha=1)
plot(lassoModel$glmnet.fit, xlab="lambda", label=TRUE)
```




```
coef(lassoModel, s=lassoModel$lambda.min)
```

```
## 14 x 1 sparse Matrix of class "dgCMatrix"
##                               1
## (Intercept)         -0.06613773
## age                  0.05516630
## weight               .
## height              -0.31174824
## NeckCirc            -0.20866382
## ChestCirc            .
## AbdomenCirc          0.74471834
## HipCirc              .
## ThighCirc            0.02472615
## KneeCirc             .
## AnkleCirc            .
## ExtendedBicepsCirc   0.06159666
## ForearmCirc          0.07938202
## WristCirc           -1.53653272
```

```
print(paste("Optimal Lambda: ", lassoModel$lambda.min))
```

```
## [1] "Optimal Lambda:  0.113730492416577"
```