The purpose of this project was to try and create a statistical model to be able to reasonably determine body fat percentage. The motivation for this experiment was that there is a severe lack of education in the area of health in this country in combination with a lot of unhealthy lifestyles. While we didn’t have the time energy resources or data to dive deeper into the issue we did have data for determining a key health metric: body fat percentage. Body composition is one of the most important metrics because it can indicate in a simplified high level manner if a person's health habits are adequate or not. However to directly calculate body fat percentage it requires experts and special equipment that isn’t available to everyone (especially those who are likely to lack meaningful knowledge on the issue). To overcome this obstacle we used a study of 252 males body fat percentage and various other physical measurements to create a model for calculating this. The other measurements are much easier and can be performed with common household measurement tools (primarily a tape measure and scale).

In the course of this experiment a lot of approaches were considered or tried that were covered in this course. Initially what was considered were unsupervised approach in order to see if any clear relationships in the data. After trying various levels of k, it w as found that k=3 provided a pretty good trade off of simplicity and variance explained. The other unsupervised approaches proved to not be helpful. With PCA the dimensionality reduction comes at the loss of interpretability and the hierarchical clustering produced trees that didn’t seem to yield any apparent insights. The clusters was used in the last part of the experiment for classification.

Next we attempted to use best subsets to try and reduce the dimensionality of our data. This is important because of the nature of the desired outcome. Because we want this to be a quick and simple way to determine body fat percentage the more measurements it takes the less likely it will be that anyone will actually use the model. After performing a best subsets analysis there were a handful of features that stood out in the data set, namely, abdomen2circumfrence, wrist, forearm, neck, weight, and age.

After finding the best subsets we performed a linear regression in order to try and predicate the specific values of body fat percentage at different data points. In performing the linear regression on all of the features we saw that the same top feature of best subset occured as the most significant variables in the linear model. We took this as a strong indication that these were the best features to use in constructing our model moving forward. We used these features for both approaches to determining high risk candidates.

The first approach that we used to attempt to classify high risk individuals was by the k = 3 means clustering performed earlier on. We used these categories to separate high risk to low risk. We used the bottom two means to be the low risk class and the highest as the high risk class. From there we divided the data set into a training and validation set. We trained a logistic regression on the training data and set the prediction threshold at 50%. From there we applied the model to the test data to see the classification accuracy. We ended up getting an error rate of around 48% which isn’t great. We were thinking that the method would prove more effective. This model was clearly ruled out for our end goal of a high risk classifier.

The second approach that we took used a linear regression model to determine the classifications. To start we used the same training and validation sets that we had for the previous approach. Then we fit the linear regression on the training data. To classify the samples we took the top third of body fat percentage individuals and classified them as high risk individuals. From there we used these labels in combination with the linear regression to predict if a subject was high risk. If the model projected the individual’s body fat percentage to be above 15 then it was classified high risk. In stark contrast to the logistic regression this model had a 95% accuracy rate. Based on this we determine this model to be sufficient for our end goal.

Note that both of these approaches assumed that there were high risk body fat percentage individuals in the data set and they were the top of the data set in body fat percentage. This assumption likely doesn’t hold. We arrived at this conclusion because the threshold for the top third of the data set occurred at 15 which relatively speaking isn’t not typically deemed as “high risk”. Nonetheless it was instructive for our model and a proof of concept that could be extended further.

R Code

> bf = load.table("BodyFat.data", header=TRUE)

Error in load.table("BodyFat.data", header = TRUE) :

could not find function "load.table"

In addition: Warning message:

R graphics engine version 12 is not supported by this version of RStudio. The Plots tab will be disabled until a newer version of RStudio is installed.

> bf = table.load("BodyFat.data",header=TRUE)

Error in table.load("BodyFat.data", header = TRUE) :

could not find function "table.load"

> bf = read.table("BodyFat.data", header=TRUE)

> bf$Density = NULL

> pairs(bf)

> plot(bf)

> plot(bf$Percentbodyfat~.,data=bf)

> pca.out = prcomp(bf, scale=TRUE)

> biplot(pca.out,scale=0)

> pca.out$sdev

[1] 2.9082686 1.2558666 1.0204164 0.8209298 0.7914850

[6] 0.6426275 0.5516337 0.5110329 0.4421779 0.4196986

[11] 0.3613245 0.2774343 0.2051750 0.1530840

> names(pca.out)

[1] "sdev" "rotation" "center" "scale" "x"

3 41.41250 24.36964 35.74286 30.44821 19.10357

Within cluster sum of squares by cluster:

[1] 41601.45 34112.22 58440.12

(between\_SS / total\_SS = 61.8 %)

Available components:

[1] "cluster" "centers" "totss"

[4] "withinss" "tot.withinss" "betweenss"

[7] "size" "iter" "ifault"

> k.out = kmeans(bf, 5)

> k.out

K-means clustering with 5 clusters of sizes 67, 58, 41, 62, 24

Cluster means:

Percentbodyfat Age.years. Weight.lbs. Height.inches.

1 17.25970 37.61194 175.4164 70.90672

2 11.39138 41.27586 145.5991 68.62069

3 18.90244 59.02439 166.1280 69.03659

4 24.79839 46.80645 200.4435 70.80645

5 29.01667 44.79167 235.5229 71.92708

Neck.cm. Chest.cm. Abdomen2circum.cm. Hip.cm. Thigh.cm.

1 37.56567 98.57612 89.49552 99.50896 59.78507

2 35.45690 91.54655 80.73276 92.29138 53.94138

3 37.70000 99.28537 90.93659 96.77561 56.83171

4 39.65968 106.96613 100.43871 104.35484 62.60161

5 41.50000 116.28333 112.07500 113.25833 67.69583

Knee.cm. Ankle.cm. Biceps.cm. Forearm.cm. Wrist.cm.

1 38.35373 23.17313 32.00597 28.72388 18.05075

2 36.03793 21.80690 29.49828 26.81724 17.30690

3 37.91707 22.79756 31.00976 28.00488 18.33415

4 40.20323 23.77742 34.43226 30.06452 18.78226

5 42.40417 24.81250 36.30833 30.46667 19.35417

Within cluster sum of squares by cluster:

[1] 15549.249 20348.414 9796.512 24130.771 30558.550

(between\_SS / total\_SS = 71.5 %)

Available components:

[1] "cluster" "centers" "totss"

[4] "withinss" "tot.withinss" "betweenss"

[7] "size" "iter" "ifault"

> k.out = kmeans(bf, 2)

> k.out

K-means clustering with 2 clusters of sizes 161, 91

Cluster means:

Percentbodyfat Age.years. Weight.lbs. Height.inches.

1 15.41491 44.53416 161.9016 69.60404

2 25.76044 45.50549 209.0418 71.11264

Neck.cm. Chest.cm. Abdomen2circum.cm. Hip.cm. Thigh.cm.

1 36.79876 96.1795 86.52981 96.13913 56.81491

2 40.10330 109.0418 103.21758 106.56703 63.99011

Knee.cm. Ankle.cm. Biceps.cm. Forearm.cm. Wrist.cm.

1 37.33913 22.55963 30.77640 27.82609 17.84658

2 40.80440 24.06264 34.92198 30.14615 18.90769

Within cluster sum of squares by cluster:

[1] 90850.54 87203.61

(between\_SS / total\_SS = 49.4 %)

Available components:

[1] "cluster" "centers" "totss"

[4] "withinss" "tot.withinss" "betweenss"

[7] "size" "iter" "ifault"

> k.out = kmeans(bf, 3)

> plot(bf,col=(k.out$cluster+1))

> plot(Percentbodyfat~Weight.lbs, data=bf,col=(k.out$cluster+1))

Error in eval(predvars, data, env) : object 'Weight.lbs' not found

> plot(Percentbodyfat~Weightlbs, data=bf,col=(k.out$cluster+1))

Error in eval(predvars, data, env) : object 'Weightlbs' not found

> names(bf)

[1] "Percentbodyfat" "Age.years."

[3] "Weight.lbs." "Height.inches."

[5] "Neck.cm." "Chest.cm."

[7] "Abdomen2circum.cm." "Hip.cm."

[9] "Thigh.cm." "Knee.cm."

[11] "Ankle.cm." "Biceps.cm."

[13] "Forearm.cm." "Wrist.cm."

> plot(Percentbodyfat~Weight.lbs., data=bf,col=(k.out$cluster+1))

> hc.complete=hclust(dist(x),methood="complete")

Error in hclust(dist(x), methood = "complete") :

unused argument (methood = "complete")

> hc.complete=hclust(dist(x),method="complete")

Error in as.matrix(x) : object 'x' not found

> hc.complete=hclust(dist(bf),method="complete")

> plot(hc.complete)

> hc.ave = hclust(dist(bf),method="average")

> plot(hc.ave)

> hc.sing = hclust(dist(bf),method="single")

> plot(hc.sing)

> plot(Percentbodyfat~Height.inches, data=bf,col=(k.out$cluster+1))

Error in eval(predvars, data, env) : object 'Height.inches' not found

> plot(Percentbodyfat~Height.inches., data=bf,col=(k.out$cluster+1))

> plot(Percentbodyfat~Abdomen2circum.cm., data=bf,col=(k.out$cluster+1))

> plot(Percentbodyfat~Hip.cm., data=bf,col=(k.out$cluster+1))

> plot(Percentbodyfat~Chest.cm., data=bf,col=(k.out$cluster+1))

> library(leaps)

> regfit.full = regsubsets(Percentbodyfat~.,data=bf)

> summary(regfit.full)

Subset selection object

Call: regsubsets.formula(Percentbodyfat ~ ., data = bf)

1 subsets of each size up to 8

Selection Algorithm: exhaustive

Age.years. Weight.lbs. Height.inches. Neck.cm.

1 ( 1 ) " " " " " " " "

2 ( 1 ) " " "\*" " " " "

3 ( 1 ) " " "\*" " " " "

4 ( 1 ) " " "\*" " " " "

5 ( 1 ) " " "\*" " " "\*"

6 ( 1 ) "\*" "\*" " " " "

7 ( 1 ) "\*" "\*" " " "\*"

8 ( 1 ) "\*" "\*" " " "\*"

Chest.cm. Abdomen2circum.cm. Hip.cm. Thigh.cm.

1 ( 1 ) " " "\*" " " " "

2 ( 1 ) " " "\*" " " " "

3 ( 1 ) " " "\*" " " " "

4 ( 1 ) " " "\*" " " " "

5 ( 1 ) " " "\*" " " " "

6 ( 1 ) " " "\*" " " "\*"

7 ( 1 ) " " "\*" " " "\*"

8 ( 1 ) " " "\*" "\*" "\*"

Knee.cm. Ankle.cm. Biceps.cm. Forearm.cm. Wrist.cm.

1 ( 1 ) " " " " " " " " " "

2 ( 1 ) " " " " " " " " " "

3 ( 1 ) " " " " " " " " "\*"

4 ( 1 ) " " " " " " "\*" "\*"

5 ( 1 ) " " " " " " "\*" "\*"

6 ( 1 ) " " " " " " "\*" "\*"

7 ( 1 ) " " " " " " "\*" "\*"

8 ( 1 ) " " " " " " "\*" "\*"

> plot(Percentbodyfat~Wrist.cm., data=bf,col=(k.out$cluster+1))

> plot(Percentbodyfat~Forearm.cm., data=bf,col=(k.out$cluster+1))

> plot(Percentbodyfat~Neck.cm., data=bf,col=(k.out$cluster+1))

> save.image("~/Desktop/ProjectWorkspace.RData")

> lm.fit = lm(Percentbodyfat~.,data=bf)

Warning message:

R graphics engine version 12 is not supported by this version of RStudio. The Plots tab will be disabled until a newer version of RStudio is installed.

> summary lm.fit

Error: unexpected symbol in "summary lm.fit"

> summary(lm.fit)

Call:

lm(formula = Percentbodyfat ~ ., data = bf)

Residuals:

Min 1Q Median 3Q Max

-11.1687 -2.8639 -0.1014 3.2085 10.0068

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -18.18849 17.34857 -1.048 0.29551

Age.years. 0.06208 0.03235 1.919 0.05618 .

Weight.lbs. -0.08844 0.05353 -1.652 0.09978 .

Height.inches. -0.06959 0.09601 -0.725 0.46925

Neck.cm. -0.47060 0.23247 -2.024 0.04405 \*

Chest.cm. -0.02386 0.09915 -0.241 0.81000

Abdomen2circum.cm. 0.95477 0.08645 11.044 < 2e-16 \*\*\*

Hip.cm. -0.20754 0.14591 -1.422 0.15622

Thigh.cm. 0.23610 0.14436 1.636 0.10326

Knee.cm. 0.01528 0.24198 0.063 0.94970

Ankle.cm. 0.17400 0.22147 0.786 0.43285

Biceps.cm. 0.18160 0.17113 1.061 0.28966

Forearm.cm. 0.45202 0.19913 2.270 0.02410 \*

Wrist.cm. -1.62064 0.53495 -3.030 0.00272 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.305 on 238 degrees of freedom

Multiple R-squared: 0.749, Adjusted R-squared: 0.7353

F-statistic: 54.65 on 13 and 238 DF, p-value: < 2.2e-16

> plot(Percentbodyfat~Age.years., data=bf,col=(k.out$cluster+1))

> reducedbf = data.frame(bf$Percentbodyfat,bf$Age.years.,bf$Abdomen2circum.cm.,bf$Wrist.cm.,bf$Forearm.cm.,bf$Neck.cm.,bf$Weight.lbs.,bf$Age.years.)

> reduced.lm = lm(Percentbodyfat~.,data=reducedbf)

Error in eval(predvars, data, env) : object 'Percentbodyfat' not found

> View(reducedbf)

> reduced.lm = lm(bf.Percentbodyfat~.,data=reducedbf)

> summary(reduced)

Error in summary(reduced) : object 'reduced' not found

> summary(reduced.lm)

Call:

lm(formula = bf.Percentbodyfat ~ ., data = reducedbf)

Residuals:

Min 1Q Median 3Q Max

-10.6385 -3.1238 -0.2273 3.1358 9.7974

Coefficients: (1 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) -25.89272 8.18676 -3.163 0.00176

bf.Age.years. 0.04605 0.02869 1.605 0.10979

bf.Abdomen2circum.cm. 0.94526 0.06838 13.824 < 2e-16

bf.Wrist.cm. -1.58567 0.51241 -3.095 0.00220

bf.Forearm.cm. 0.57210 0.18552 3.084 0.00228

bf.Neck.cm. -0.40039 0.22181 -1.805 0.07228

bf.Weight.lbs. -0.09386 0.03146 -2.984 0.00314

bf.Age.years..1 NA NA NA NA

(Intercept) \*\*

bf.Age.years.

bf.Abdomen2circum.cm. \*\*\*

bf.Wrist.cm. \*\*

bf.Forearm.cm. \*\*

bf.Neck.cm. .

bf.Weight.lbs. \*\*

bf.Age.years..1

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.314 on 245 degrees of freedom

Multiple R-squared: 0.7406, Adjusted R-squared: 0.7343

F-statistic: 116.6 on 6 and 245 DF, p-value: < 2.2e-16

> reducedbf = data.frame(bf$Percentbodyfat,bf$Age.years.,bf$Abdomen2circum.cm.,bf$Wrist.cm.,bf$Forearm.cm.,bf$Neck.cm.,bf$Weight.lbs.)

> reduced.lm = lm(bf.Percentbodyfat~.,data=reducedbf)

> summary(reduced.lm)

Call:

lm(formula = bf.Percentbodyfat ~ ., data = reducedbf)

Residuals:

Min 1Q Median 3Q Max

-10.6385 -3.1238 -0.2273 3.1358 9.7974

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -25.89272 8.18676 -3.163 0.00176

bf.Age.years. 0.04605 0.02869 1.605 0.10979

bf.Abdomen2circum.cm. 0.94526 0.06838 13.824 < 2e-16

bf.Wrist.cm. -1.58567 0.51241 -3.095 0.00220

bf.Forearm.cm. 0.57210 0.18552 3.084 0.00228

bf.Neck.cm. -0.40039 0.22181 -1.805 0.07228

bf.Weight.lbs. -0.09386 0.03146 -2.984 0.00314

(Intercept) \*\*

bf.Age.years.

bf.Abdomen2circum.cm. \*\*\*

bf.Wrist.cm. \*\*

bf.Forearm.cm. \*\*

bf.Neck.cm. .

bf.Weight.lbs. \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.314 on 245 degrees of freedom

Multiple R-squared: 0.7406, Adjusted R-squared: 0.7343

F-statistic: 116.6 on 6 and 245 DF, p-value: < 2.2e-16

> plot(reduced.lm)

> ?sample

> .25 \* 252

[1] 63

> test = sample(1:252,size=63,replace=TRUE)

> reduced.train = lm(bf.Percentbodyfat~.,data=reducedbf[-test,])

> predicted = predict(reduced.train,newdata=reducedbf[test,])

> mse = mean((reducedbf[test,]$bf.Percentbodyfat-predicted)^2)

> mse

[1] 24.86519

> sqrt(25)

[1] 5

> highrisk = ifelse(k.out$cluster == 3,1,0)

> reducedbf$highrisk = highrisk

> glm.fit = glm(highrisk~.-bf.Percentbodyfat,data=reducedbf[-test,], family=binomial)

> glm.pred = predict(glm.fit,reducedbf[test,],type="response")

> glm.prob = predict(glm.fit,reducedbf[test,],type="response")

> glm.pred[glm.probs>.5]="Yes"

Error in glm.pred[glm.probs > 0.5] = "Yes" : object 'glm.probs' not found

> glm.pred = predict(glm.fit,reducedbf[test,],type="response")

> glm.pred[glm.probs>.5]="Yes"

Error in glm.pred[glm.probs > 0.5] = "Yes" : object 'glm.probs' not found

> glm.pred[glm.pred>.5]="Yes"

> table(glm.pred,reducedbf[test,])

> glm.pred = predict(glm.fit,reducedbf[test,],type="response")

> glm.prob = predict(glm.fit,reducedbf[test,],type="response")

> glm.pred=rep("No",252)

> glm.pred[glm.prob>.5]="Yes"

> glm.pred=rep(0,252)

> glm.pred[glm.prob>.5]=1

> table(glm.pred,reducedbf[test,]$highrisk)

Error in table(glm.pred, reducedbf[test, ]$highrisk) :

all arguments must have the same length

> glm.pred=rep(0,63)

> glm.pred[glm.prob>.5]=1

> table(glm.pred,reducedbf[test,]$highrisk)

glm.pred 0 1

0 30 26

1 4 3

> mean(glm.pred == reducedbf[test,]$highrisk)

[1] 0.5238095

> plot(bf.Percentbodyfat~bf.Abdomen2circum.cm.,data=reducedbf)

> plot(bf.Percentbodyfat~bf.Abdomen2circum.cm.,data=reducedbf,col=(highrisk[test]+1))

> plot(bf.Percentbodyfat~bf.Abdomen2circum.cm.,data=reducedbf,col=(highrisk[test]+2))

> plot(bf.Percentbodyfat~bf.Abdomen2circum.cm.,data=reducedbf,col=(highrisk[test]+1))

> plot(bf.Percentbodyfat~bf.Abdomen2circum.cm.,data=reducedbf,col=(k.out$cluster+1))

> plot(bf.Percentbodyfat~bf.Abdomen2circum.cm.,data=reducedbf,col=(highrisk[test]+1))

> plot(bf.Percentbodyfat~bf.Abdomen2circum.cm.,data=reducedbf,col=(k.out$cluster+1))

> plot(bf.Percentbodyfat~bf.Abdomen2circum.cm.,data=reducedbf,col=(highrisk[test]+1))

> highrisk = ifelse(k.out$cluster == 1,1,0)

> glm.fit = glm(highrisk~.-bf.Percentbodyfat,data=reducedbf[-test,], family=binomial)

> glm.prob = predict(glm.fit,reducedbf[test,],type="response")

> glm.pred=rep(0,63)

> glm.pred[glm.prob>.5]=1

> table(glm.pred,reducedbf[test,]$highrisk)

glm.pred 0 1

0 30 26

1 4 3

> mean(glm.pred == reducedbf[test,]$highrisk)

[1] 0.5238095

> plot(bf.Percentbodyfat~bf.Abdomen2circum.cm.,data=reducedbf,col=(highrisk[test]+1))

> plot(bf.Percentbodyfat~bf.Abdomen2circum.cm.,data=reducedbf,col=(highrisk+1))

> plot(bf.Percentbodyfat~bf.Abdomen2circum.cm.,data=reducedbf,col=(highrisk))

> highrisk = ifelse(k.out$cluster == 0,1,0)

> plot(bf.Percentbodyfat~bf.Abdomen2circum.cm.,data=reducedbf,col=(highrisk))

> highrisk = ifelse(k.out$cluster == 3,1,0)

> plot(bf.Percentbodyfat~bf.Abdomen2circum.cm.,data=reducedbf,col=(highrisk))

> plot(bf.Percentbodyfat~bf.Abdomen2circum.cm.,data=reducedbf,col=(k.out$cluster))

> abline(reduced.lm)

Warning message:

In abline(reduced.lm) :

only using the first two of 7 regression coefficients