bodyfat <- read.csv("/home/thomas/git/datascience/MSDataSci/MATH550/data/bodyfat.csv", header=TRUE)

attach(bodyfat)

**# 1**

c <- cor(bodyfat)

round(c,3)

**bodyfat age weight height neck chest abdomen hip thigh knee ankle biceps forearm wrist**

**bodyfat 1.000 0.291 0.612 -0.025 0.491 0.703 0.813 0.625 0.560 0.509 0.266 0.493 0.361 0.347**

**age 0.291 1.000 -0.013 -0.245 0.114 0.176 0.230 -0.050 -0.200 0.018 -0.105 -0.041 -0.085 0.214**

**weight 0.612 -0.013 1.000 0.487 0.831 0.894 0.888 0.941 0.869 0.853 0.614 0.800 0.630 0.730**

**height -0.025 -0.245 0.487 1.000 0.321 0.227 0.190 0.372 0.339 0.501 0.393 0.319 0.322 0.398**

**neck 0.491 0.114 0.831 0.321 1.000 0.785 0.754 0.735 0.696 0.672 0.478 0.731 0.624 0.745**

**chest 0.703 0.176 0.894 0.227 0.785 1.000 0.916 0.829 0.730 0.719 0.483 0.728 0.580 0.660**

**abdomen 0.813 0.230 0.888 0.190 0.754 0.916 1.000 0.874 0.767 0.737 0.453 0.685 0.503 0.620**

**hip 0.625 -0.050 0.941 0.372 0.735 0.829 0.874 1.000 0.896 0.823 0.558 0.739 0.545 0.630**

**thigh 0.560 -0.200 0.869 0.339 0.696 0.730 0.767 0.896 1.000 0.799 0.540 0.761 0.567 0.559**

**knee 0.509 0.018 0.853 0.501 0.672 0.719 0.737 0.823 0.799 1.000 0.612 0.679 0.556 0.665**

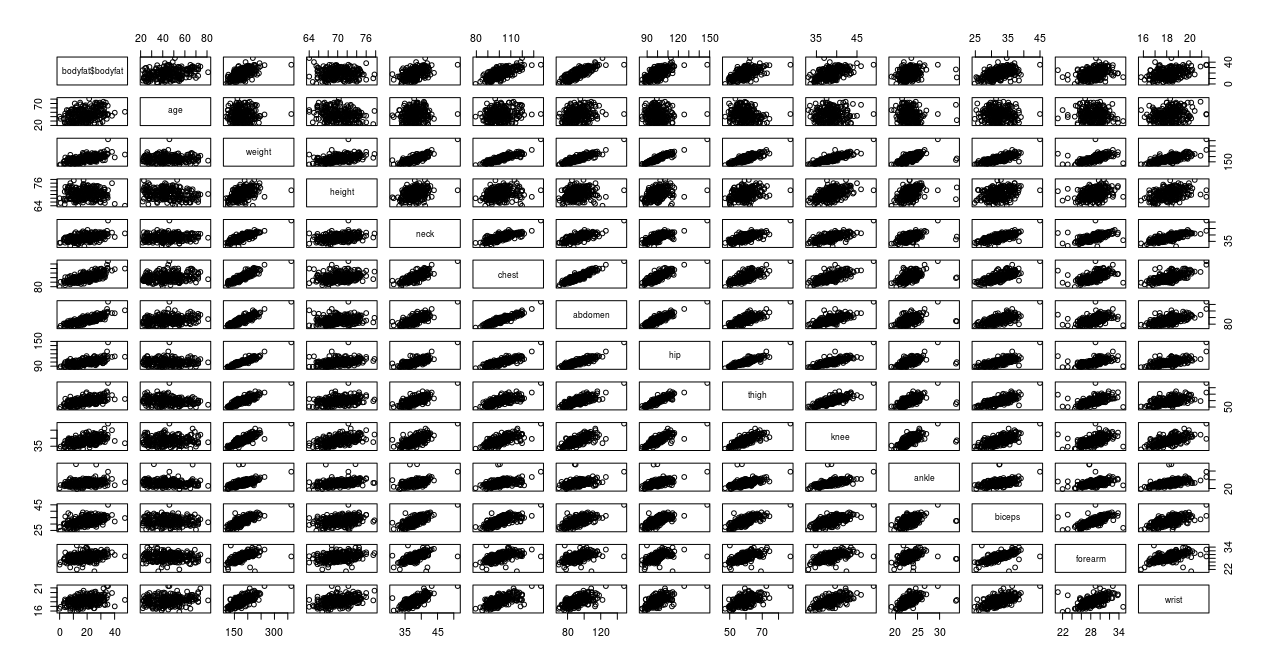
**ankle 0.266 -0.105 0.614 0.393 0.478 0.483 0.453 0.558 0.540 0.612 1.000 0.485 0.419 0.566**

**biceps 0.493 -0.041 0.800 0.319 0.731 0.728 0.685 0.739 0.761 0.679 0.485 1.000 0.678 0.632**

**forearm 0.361 -0.085 0.630 0.322 0.624 0.580 0.503 0.545 0.567 0.556 0.419 0.678 1.000 0.586**

**wrist 0.347 0.214 0.730 0.398 0.745 0.660 0.620 0.630 0.559 0.665 0.566 0.632 0.586 1.000**

pairs(bodyfat$bodyfat~age+weight+height+neck+chest+abdomen+hip+thigh+knee+ankle+biceps+forearm+wrist)



**The data at a glance looks like it has a lot of multicollinearity between several of its predictors. This make me question how effective it would be to include every variable in the analysis since there is a chance that the predictors will take up the same span. However, even if we drop several predictors in our analysis, it might not have a bad effect on the R value in the end. This is because it visually looks like several of the predictors are strongly correlated with body fat already. I also do not see any need to take any transforms of the predictors since they all look linear.**

**# 2**

model <- lm(bodyfat ~., data=bodyfat)

summary(model)

C**all:**

**lm(formula = bodyfat ~ ., data = bodyfat)**

**Residuals:**

**Min 1Q Median 3Q Max**

**-11.1966 -2.8824 -0.1111 3.1901 9.9979**

**Coefficients:**

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

**(Intercept) -21.35323 22.18616 -0.962 0.33680**

**age 0.06457 0.03219 2.006 0.04601 \***

**weight -0.09638 0.06185 -1.558 0.12047**

**height -0.04394 0.17870 -0.246 0.80599**

**neck -0.47547 0.23557 -2.018 0.04467 \***

**chest -0.01718 0.10322 -0.166 0.86792**

**abdomen 0.95500 0.09016 10.592 < 2e-16 \*\*\***

**hip -0.18859 0.14479 -1.302 0.19401**

**thigh 0.24835 0.14617 1.699 0.09061 .**

**knee 0.01395 0.24775 0.056 0.95516**

**ankle 0.17788 0.22262 0.799 0.42505**

**biceps 0.18230 0.17250 1.057 0.29166**

**forearm 0.45574 0.19930 2.287 0.02309 \***

**wrist -1.65450 0.53316 -3.103 0.00215 \*\***

**---**

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

**Residual standard error: 4.309 on 238 degrees of freedom**

**Multiple R-squared: 0.7486, Adjusted R-squared: 0.7348**

**F-statistic: 54.5 on 13 and 238 DF, p-value: < 2.2e-16**

**The summary above confirms a few assumptions made in #1. There are Several variables that are much more correlated to body fat than others. Also, because the data looked like it had a lot of multicollinearity, it may be possible to toss out many of the poorly correlating predictors with much effect on the model predictability. Tossing out a few predictors may also improve the F-statistic.**

**Below I present a new model with several predictors removed after performing backwards BIC step. This process showed that it is possible to reduce the model to four predictors with a very minimal impact on its predictability by using BIC as a criterion for which predictors to select.**

**# 3**

n <- length(model$residuals)

backBIC <- step(model,direction="backward", data=bodyfat, k=log(n))

**# last step of backward BIC**

**Step: AIC=762.72**

**bodyfat ~ weight + abdomen + forearm + wrist**

**Df Sum of Sq RSS AIC**

**<none> 4658.2 762.72**

**- forearm 1 127.8 4786.1 764.01**

**- wrist 1 218.2 4876.4 768.73**

**- weight 1 566.4 5224.7 786.11**

**- abdomen 1 5948.9 10607.1 964.56**

# using the four variables given by back BIC:

model\_new <- lm(bodyfat$bodyfat~weight+abdomen+forearm+wrist)

summary(model\_new)

**Call:**

**lm(formula = bodyfat$bodyfat ~ weight + abdomen + forearm + wrist)**

**Residuals:**

**Min 1Q Median 3Q Max**

**-10.5626 -3.1235 -0.1461 3.1313 9.0867**

**Coefficients:**

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

**(Intercept) -34.85407 7.24500 -4.811 2.62e-06 \*\*\***

**weight -0.13563 0.02475 -5.480 1.05e-07 \*\*\***

**abdomen 0.99575 0.05607 17.760 < 2e-16 \*\*\***

**forearm 0.47293 0.18166 2.603 0.009790 \*\***

**wrist -1.50556 0.44267 -3.401 0.000783 \*\*\***

**---**

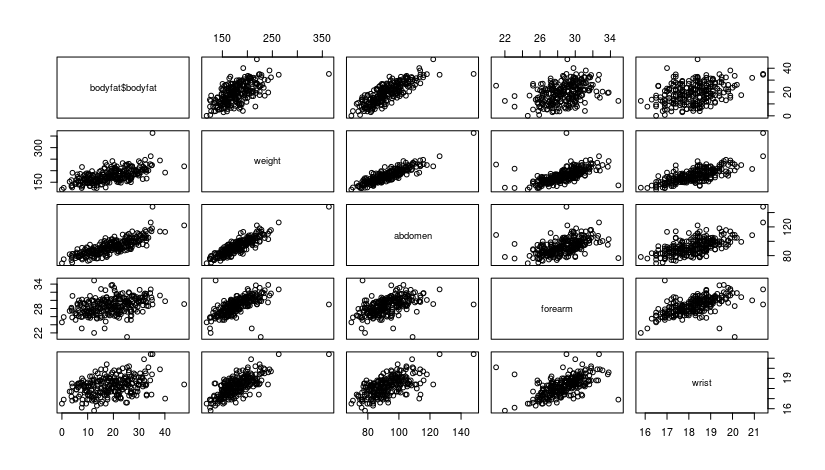
**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**Residual standard error: 4.343 on 247 degrees of freedom**

**Multiple R-squared: 0.735, Adjusted R-squared: 0.7307**

**F-statistic: 171.3 on 4 and 247 DF, p-value: < 2.2e-16**

pairs(bodyfat$bodyfat~weight+abdomen+forearm+wrist)



**# 4**

# Residual Plots vs Predictior

StanRes1 <- rstandard(model\_new)

par(mfrow=c(2,2))

plot(weight,StanRes1, ylab="Standardized Residuals",pch=19,col=10)

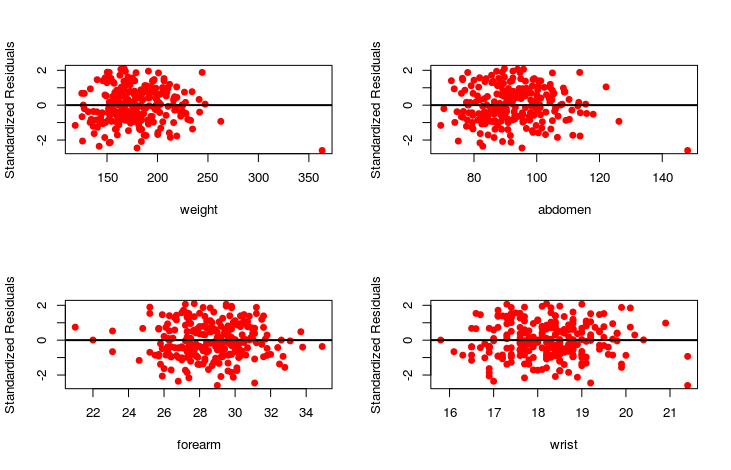
abline(h=0,lwd=2)

plot(abdomen,StanRes1, ylab="Standardized Residuals",pch=19,col=10)

abline(h=0,lwd=2)

plot(forearm,StanRes1, ylab="Standardized Residuals",pch=19,col=10)

abline(h=0,lwd=2)

plot(wrist,StanRes1, ylab="Standardized Residuals",pch=19,col=10)

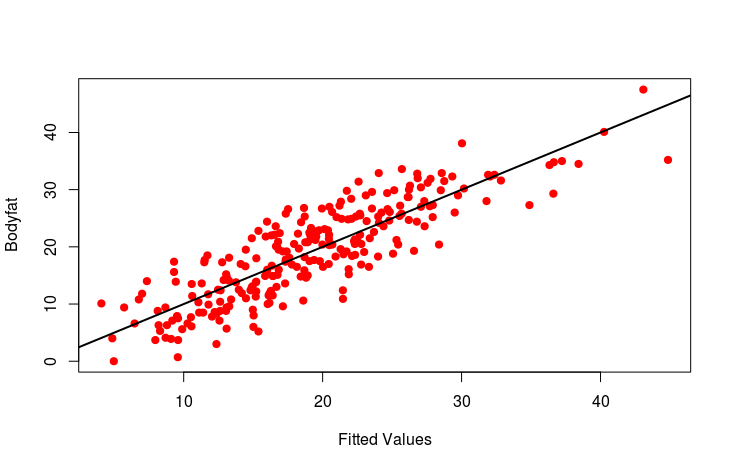
abline(h=0,lwd=2)

# Residual Plots vs Fits

par(mfrow=c(1,1))

plot(model\_new$fitted.values,bodyfat$bodyfat,xlab="Fitted Values", ylab="Bodyfat",col=10,pch=19)

abline(lsfit(model\_new$fitted.values,bodyfat$bodyfat),lwd=2)

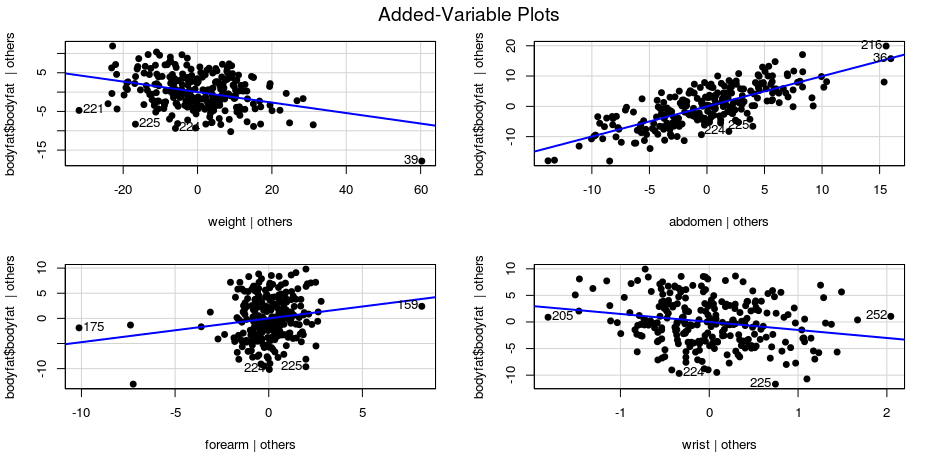


**# 5 Added Variable Plots**

library(car)

par(mfrow=c(2,2))

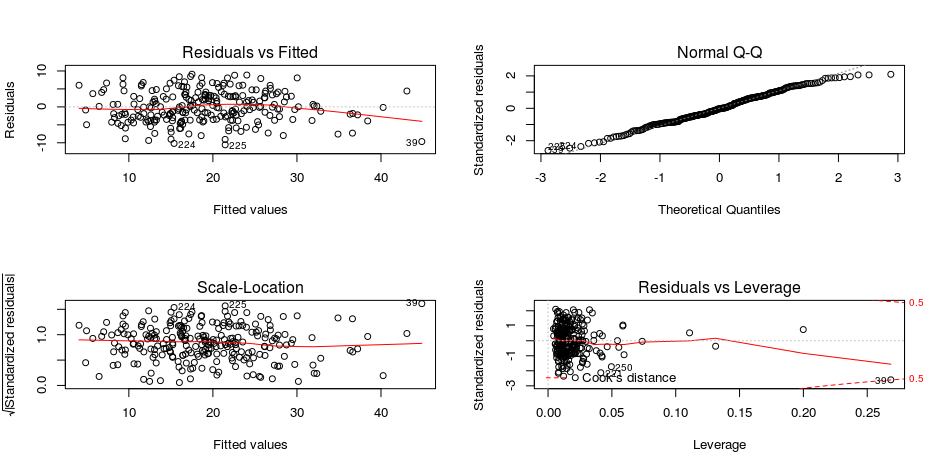
avPlots(model\_new,terms=~.,ask=FALSE,pch=19)



**# 6 Response Vs Fitted**

par(mfrow = c(2,2))

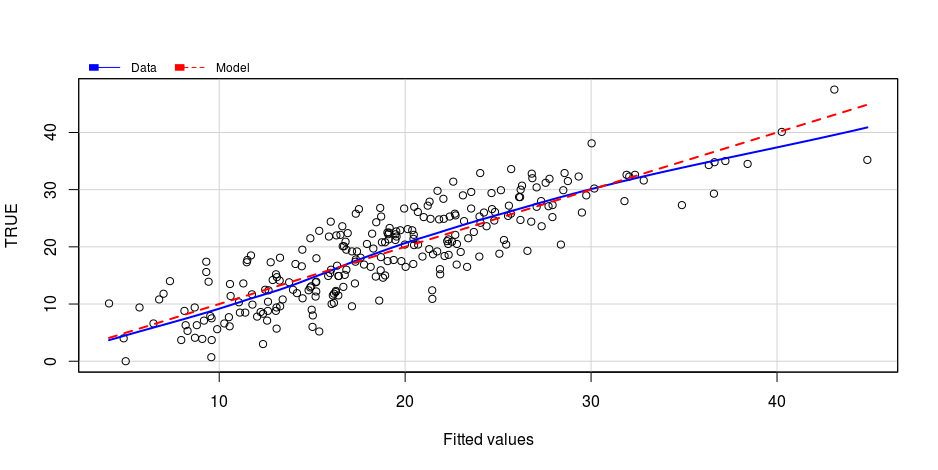
plot(model\_new)



**# 7 Marginal Model Plots**

par(mfrow = c(1,1))

mmp(model\_new)



**# 8 Variance Inflation Factors**

vif(model\_new)

**weight abdomen forearm wrist**

**7.040774 4.864380 1.793374 2.273047**

**# 9 Conclusion**

**The four variables selected by backwards step BIC produced a more elegant model than using every potential predictor given. These variables were weight, abdomen, forearm, and wrist.**

**The standardized residuals for each were flat.**

**The added variable plots for each indicated that each predictor was responsible for a visually significant portion of the variability in the model.**

**The response plots look good with the exception of point number 39 in the leverage plot. It had a high leverage and a larger cooks distance. Looking back at the added variable plots, we can see number 39 separated from the rest of the points in the weight plot as a visually high leverage point. It also has a lower than predicted level of body fat for the amount of weight given. Perhaps this is an outlier because not all body weight is due to fat – much is also due to muscle.**