Final Project

Yixuan Jiao, Landi Guo, Fengdi Zhang

2022-12-13

Abstract

Introduction

Although fat is an essential part of our body and an important source of stored energy, excessive amounts of body fat is associated with type 2 diabetes, heart diseases, and stroke. More than 25 percent body fat is considered obese for adult male, while more than 32 percent body fat is considered obese for adult women. National data from the 2017-2020 National Health and Nutrition Examination Survey revealed that 41.9 percent of adults in the U.S. have obesity, and the adult obesity rate is over 35 percent in nineteen states. As obesity becomes one of the most common medical conditions in the U.S., it is important to find an accurate and easy way to find out one's body fat. Unfortunately, body fat is not always straightforward to measure. We are given a dataset containing percentage of body fat, age, height, weight, and other body circumference measurements for 252 men. How can we possibly estimate body fat for men in a more convenient way? This project aims to build a multiple linear regression model using scale and tape measurements to predict body fat for men.

Methods

Exploratory Analysis

Exploratory analysis is conducted to check for patterns, distributions, and anomalies in the dataset. This dataset contains 252 observations which are all male, and 16 variables of interest. The first three variables are body fat measured in three different ways (Brozek's, Siri's, body density). The rest of the variables are age(years), weight(lbs), heights(inches), neck(neck circumference in cm), chest(chest circumference in cm), abdomen(abdomen circumference in cm), hip(hip circumference in cm), thigh(thigh circumference in cm), knee(knee circumference in cm), ankle(ankle circumference in cm), bicep(extended biceps circumference in cm), forearm(forearm circumference in cm), and wrist(wrist circumference in cm). These are simple measurements that can potentially be used to predict body fats.

In the rest of the analysis, percent body fat using Brozek's equation is chosen as the outcome. Firstly, one entry with 0 body fat is removed from the dataset. Mean and range are summarized for the remaining observations(Table1). Then marginal distributions for each variable and pairwise relationship between each pair of variables are plotted(Fig1). The distributions for all variables are symmetric, therefore no transformation is needed. To confirm the normality of `bodyfat_brozek`, formal Shapiro's test is conducted. The results of the test align with the histogram that `bodyfat_brozek` is normally distributed (Fig2). Pairwise scatterplot shows that all variables are linearly correlated with `bodyfat_brozek`. Additionally, there are many variables that are highly correlated with other variables, which require further investigation.

Model Building

Variance inflation factor (VIF) for each variable is calculated for checking collinearity. Variables with VIF > 5 suggest that the coefficients might be misleading due to collinearity and high collinearity. `weight`, `hip`, `abdomen`, `chest`, and `thigh` have VIF > 5, but according to the p-values of the complete linear model, only `abdomen` is significant. Therefore, all these variables except `abdomen` are excluded. Re-calculate the VIFs and no more collinearity is found. Different model selection procedures are conducted on the remaining variables to generate candidate models. Procedures include automatic procedure (stepwise), criterion based procedure (Cp value and Adjusted R^2), and LASSO. Interactions between main effects are considered by conducting a two-way ANOVA test. A 10-fold cross validation is used to compare the candidate models based on predictive ability and select a final "best" model. Diagnostic plots are also generated for comparison and final model selection.

Result

The candidate model from stepwise regression includes `age`, `height`, `neck`, `abdomen`, `forearm`, and `wrist` as predictors to predict `bodyfat_brozek`. The candidate models from LASSO and criterion approach include all the same predictors as the model from stepwise regression, but also include an extra predictor `bicep`. Among the predictors in the candidate models, two-way ANOVA test(table?) reveals interaction between neck and abdomen has significant p-value = 0.000477 for model 1 and p-value = 0.000379 for model 2, suggesting it should be included in the model under 5% significance. Therefore, the final candidate models are as following: Candidate model 1: `bodyfat_brozek` ~ `age` + `height` + `neck` + `abdomen` + `forearm` + `wrist` + `neck:abdomen` Candidate model 2: `bodyfat_brozek` ~ `age` + `height` + `neck` + `abdomen` + `forearm` + `wrist` + `bicep` + `neck:abdomen` The RMSE for 10-fold cross validation is 3.95 for model 1, and 3.98 for model 2. Therefore, model 1 has better predictivity than model 2. The adjusted R-squared for is 0.7385 for both model 1 and model 2, which means that `bicep` is not adding value to the model. Diagnosis plots(figure?) are also suggest that model 1 fits the underlying assumptions of linear regression better than model 2, because *add reasons.* In conclusion, the final model chosen is candidate model 1.

Appendix

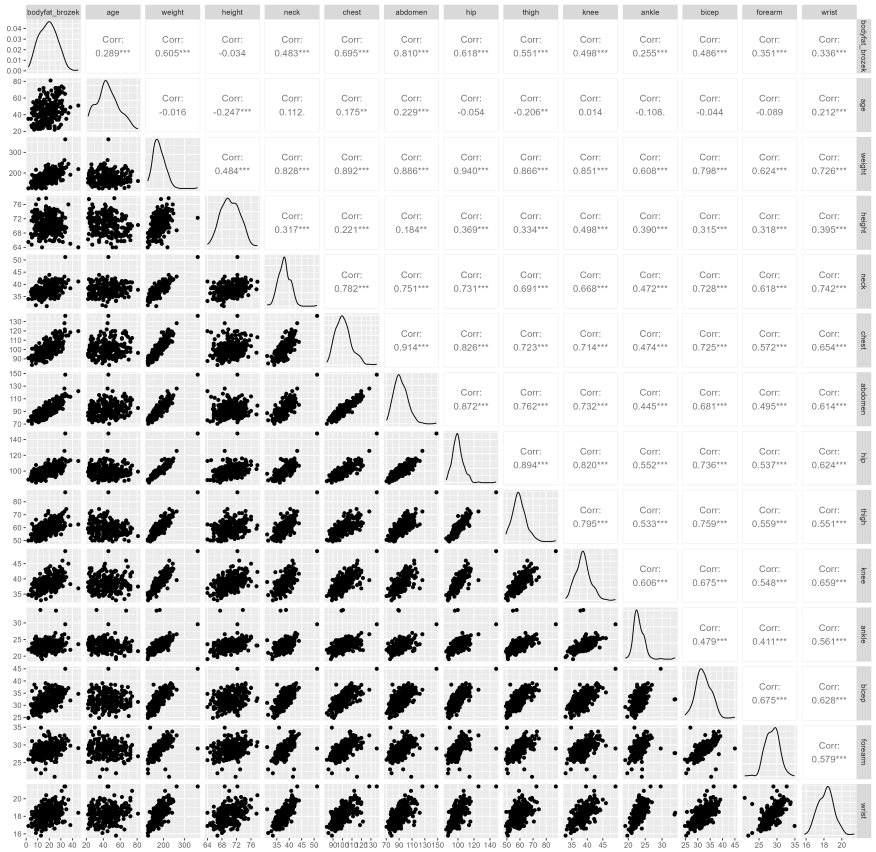


Fig1: Distributions for each Variable and Pairwise Relationship between each pair of Variables

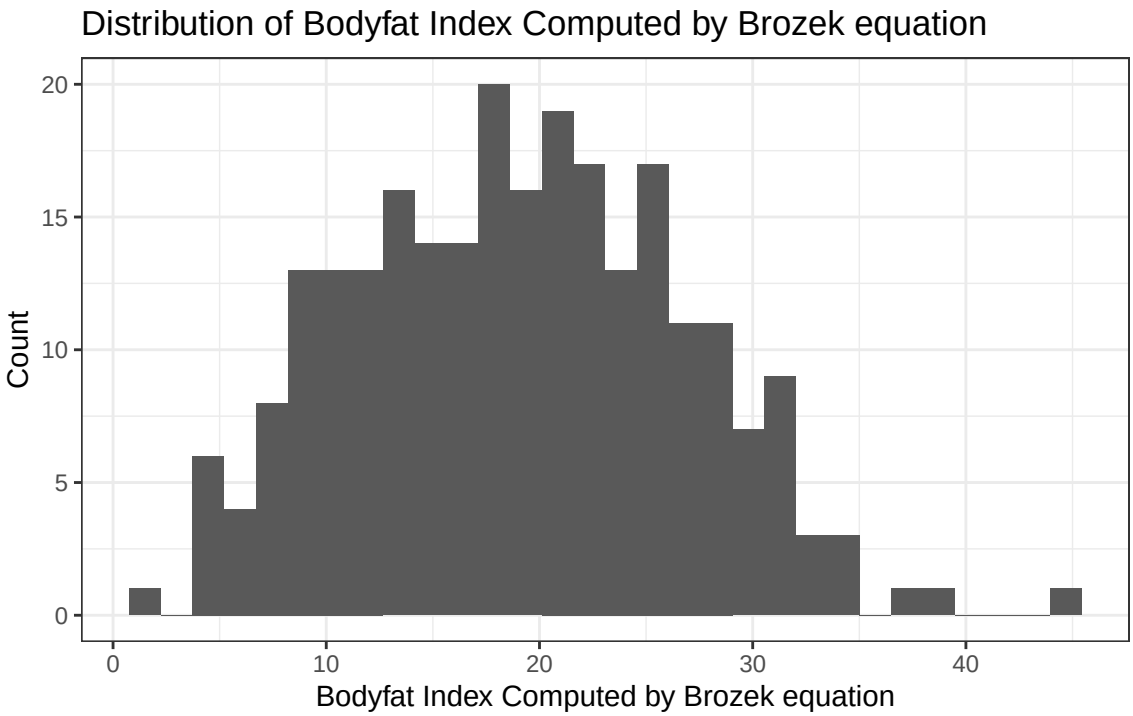


Fig2: Distribution of Bodyfat Index Computed by Brozek equation

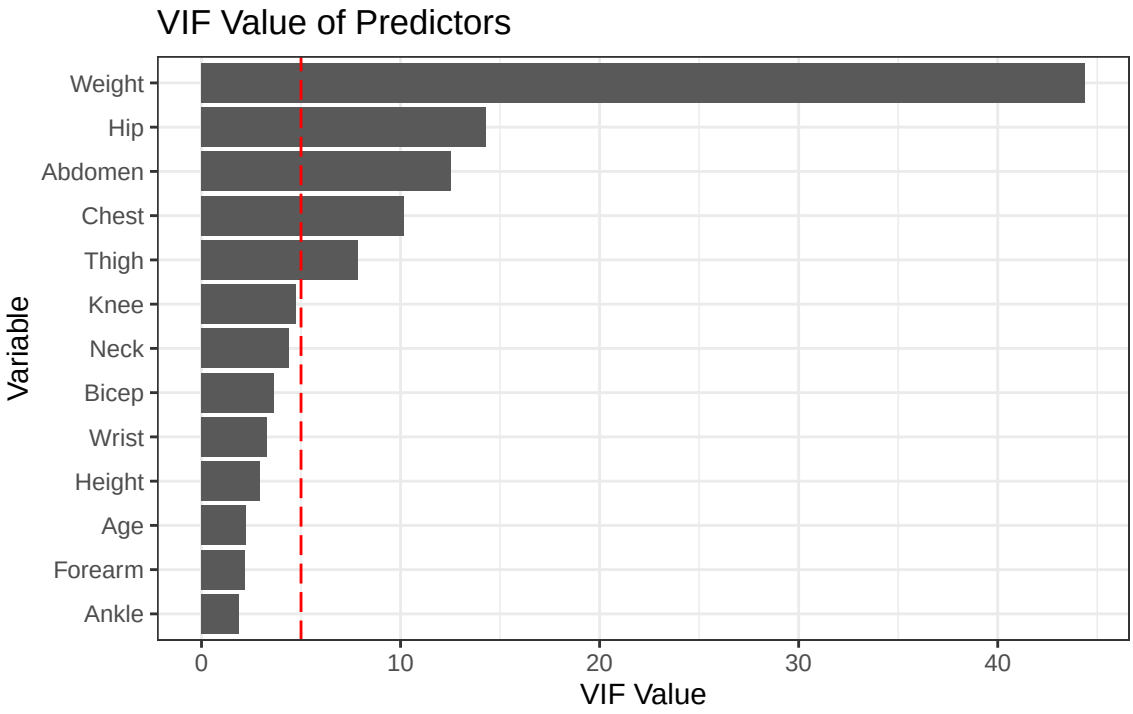
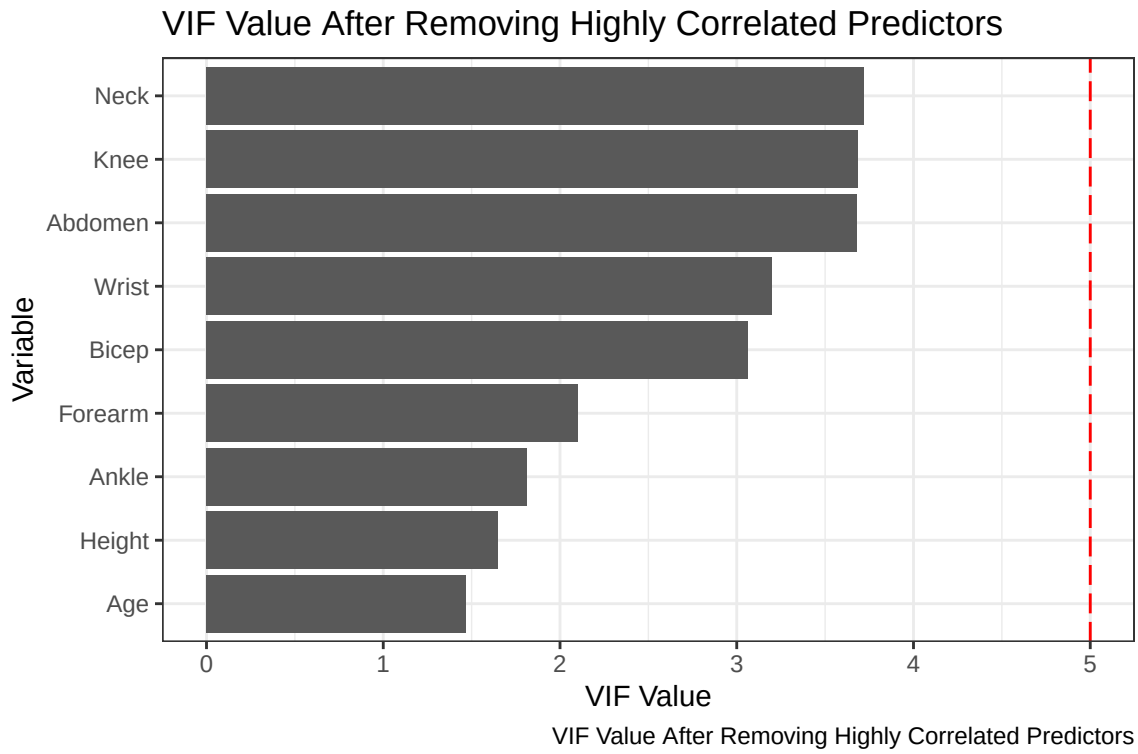
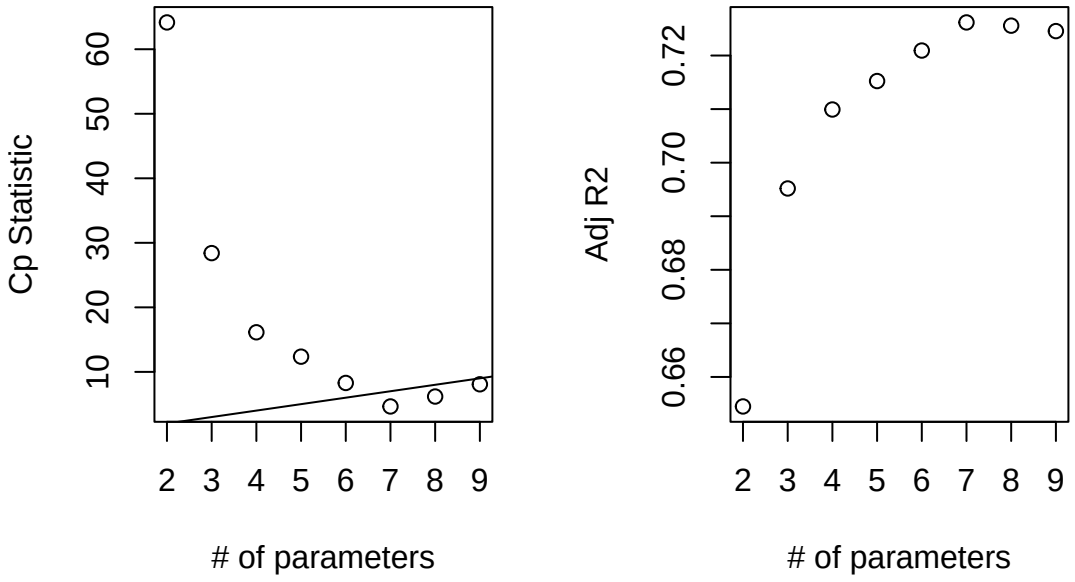


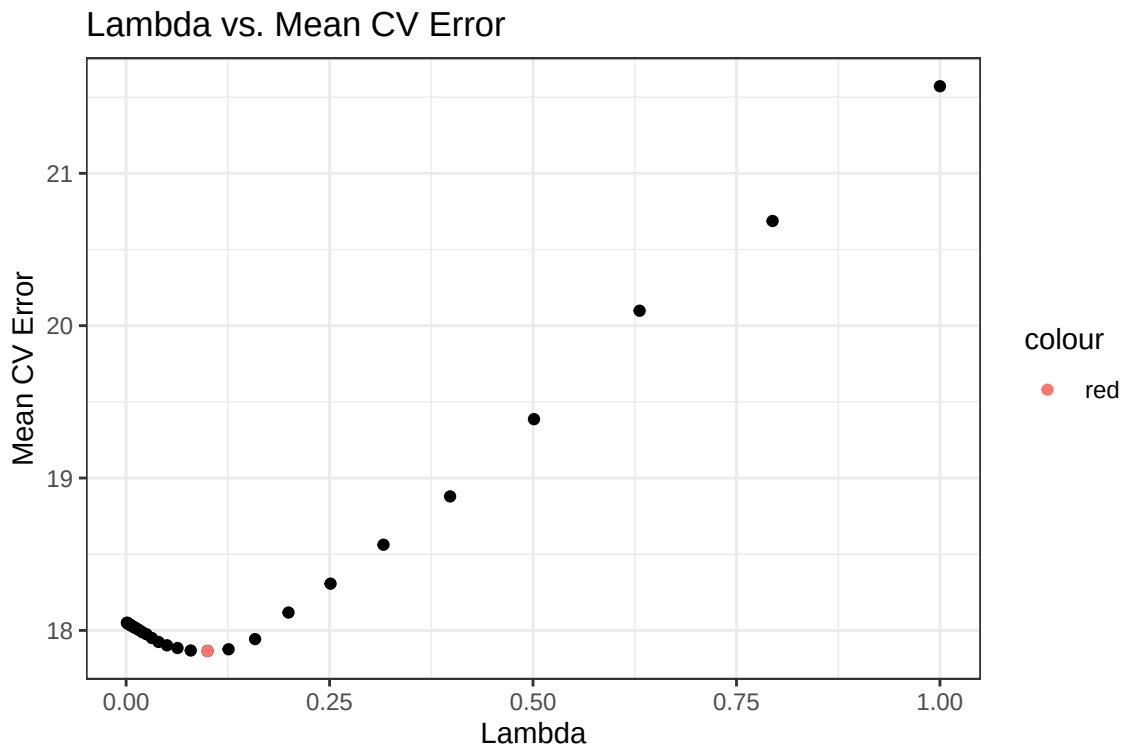
Fig3: VIF Value of Predictors






```
##
## Call:  cv.glmnet(x = as.matrix(df_body_fat[c(-1)]), y = df_body_fat$bodyfat_brozek,      lambda = lam
##
## Measure: Mean-Squared Error
```

```
##
##     Lambda Index Measure    SE Nonzero
## min 0.1000    11   17.86 1.453       7
## 1se 0.3981     5   18.88 1.344       4
```




```
lasso_fit <- glmnet(as.matrix(df_body_fat[c(-1)]), df_body_fat$bodyfat_brozek, lambda = cv_object$lambda
coef(lasso_fit)
```

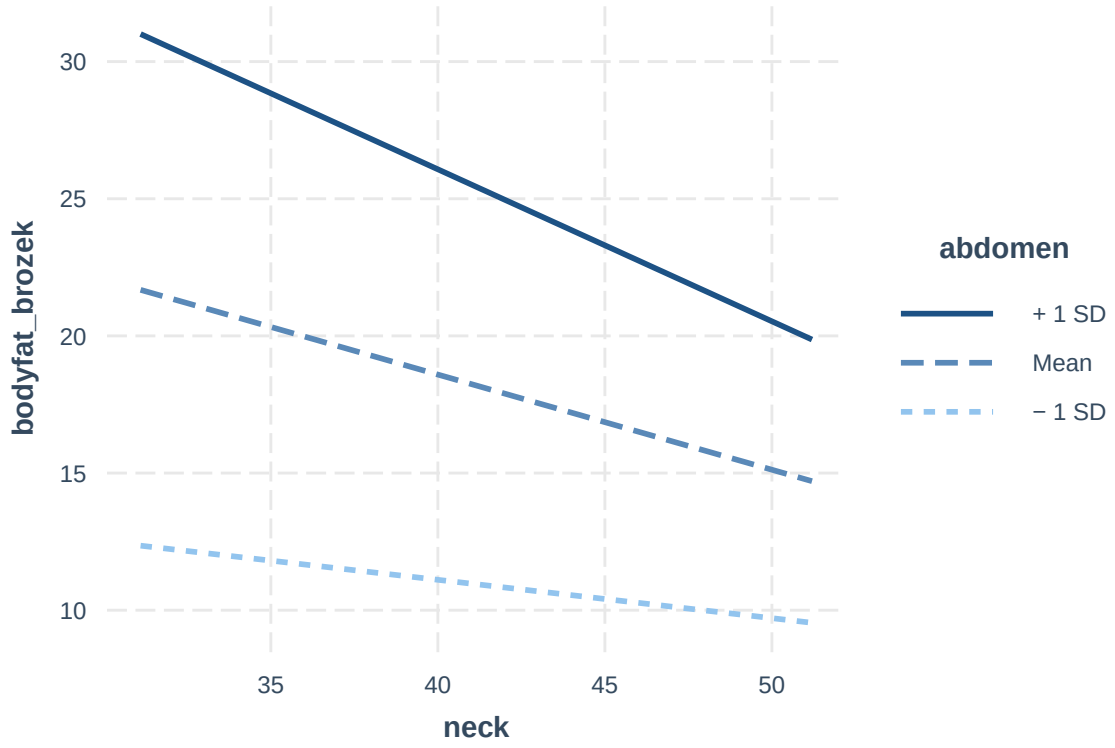
```
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                       s0
## (Intercept)  4.17107658
## age          0.05751790
## height      -0.26632895
## neck        -0.35880068
## abdomen      0.67802424
## knee         .
## ankle        .
## bicep        0.01014336
## forearm      0.30061461
## wrist       -1.48831650
```

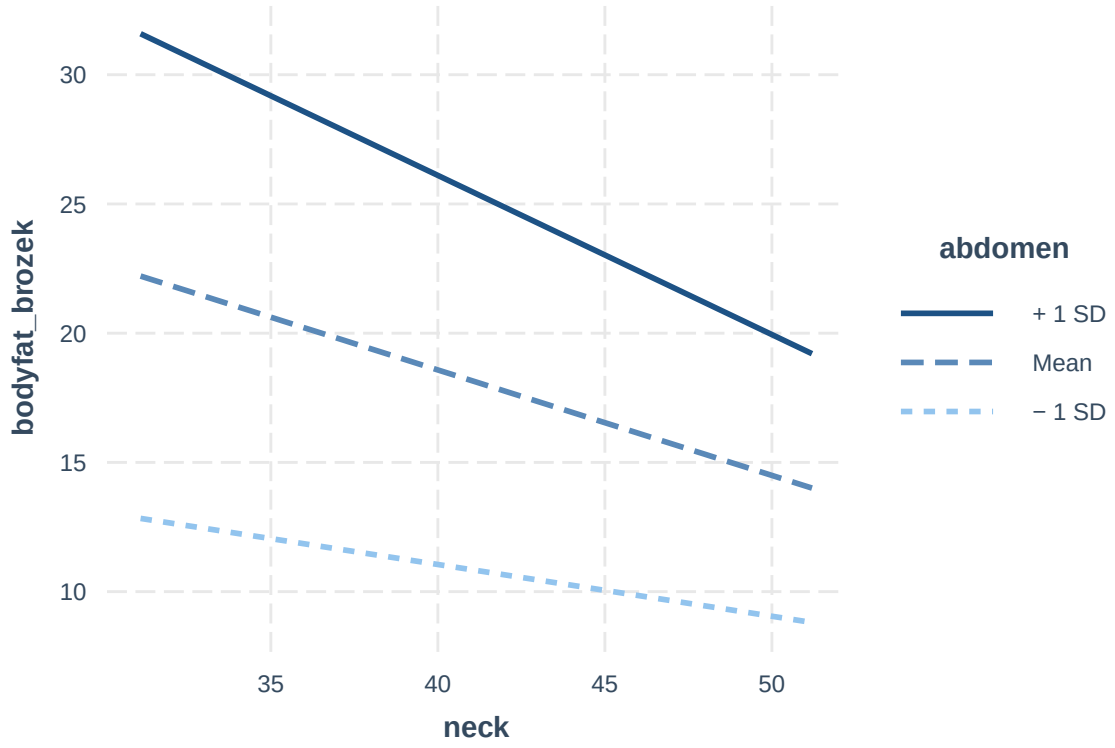
```
train = trainControl(method = "cv", number = 10)
model_caret = train(bodyfat_brozek ~ age + height + neck + abdomen +
    forearm + bicep + wrist + neck:abdomen, data = df_body_fat,
trControl = train,
method = 'lm',
na.action = na.pass)
print(model_caret)
```

```
## Linear Regression
##
## 251 samples
##   7 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 227, 226, 225, 227, 225, 227, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   3.982511  0.7372084  3.293905
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
train = trainControl(method = "cv", number = 10)
model_caret = train(bodyfat_brozek ~ age + height + neck + abdomen +
    forearm + wrist + neck:abdomen, data = df_body_fat,
trControl = train,
method = 'lm',
na.action = na.pass)
print(model_caret)
```

```
## Linear Regression
##
## 251 samples
##   6 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 225, 225, 226, 227, 227, 225, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   3.950642  0.7453959  3.247606
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```





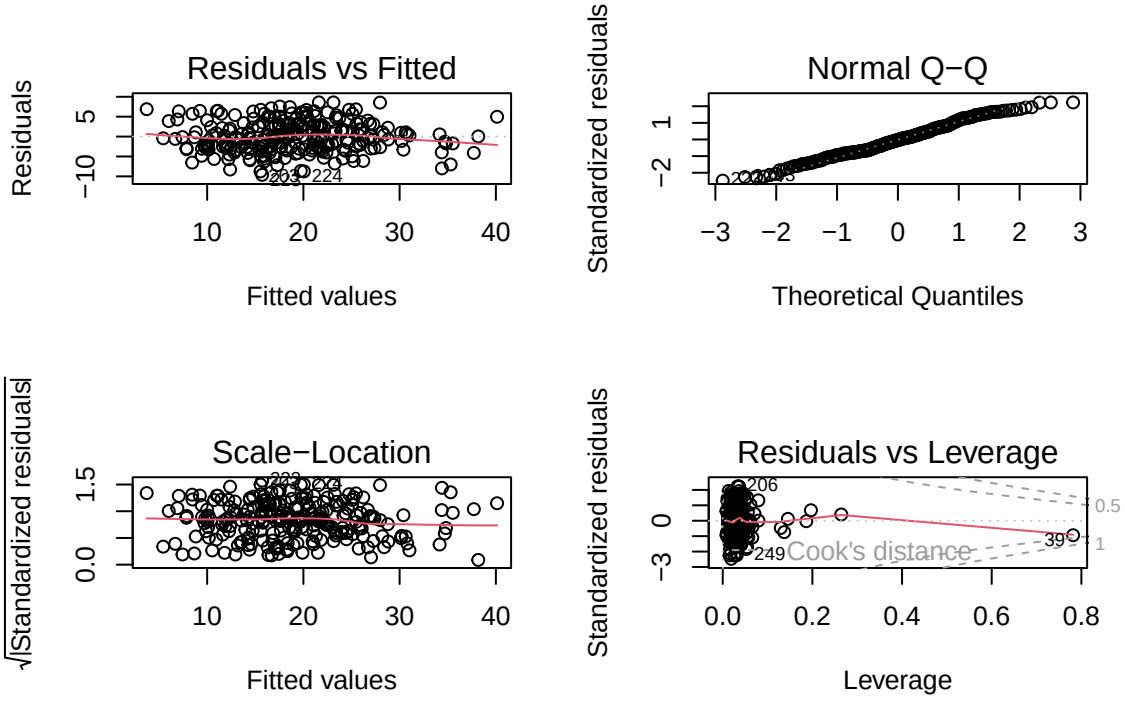

```
aov(update(stepwise.fit, . ~ . + neck*abdomen)) %>% summary
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## age            1   1229    1229  79.830  < 2e-16 ***
```

```
## height         1     22      22   1.452 0.229403
## neck           1   3268    3268 212.312  < 2e-16 ***
## abdomen        1   5982    5982 388.581  < 2e-16 ***
## forearm        1     55      55   3.548 0.060805 .
## wrist          1    230     230  14.924 0.000144 ***
## neck:abdomen   1    193     193  12.541 0.000477 ***
## Residuals    243   3741      15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
aov(update(stepwise.fit, . ~ . + bicep + neck*abdomen)) %>% summary
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## age            1   1229    1229  79.811  < 2e-16 ***
## height         1     22      22   1.452 0.229461
## neck           1   3268    3268 212.264  < 2e-16 ***
## abdomen        1   5982    5982 388.492  < 2e-16 ***
## forearm        1     55      55   3.547 0.060839 .
## wrist          1    230     230  14.920 0.000144 ***
## bicep          1      8       8   0.489 0.485247
## neck:abdomen   1    200     200  12.994 0.000379 ***
## Residuals    242   3726      15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```




```
##
## Call:
```

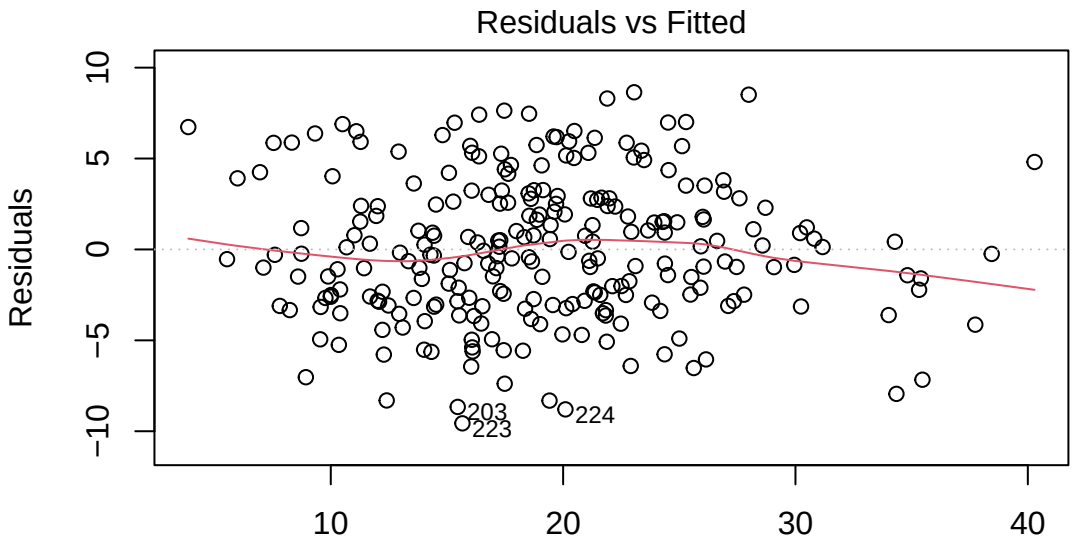
Table 1: Discriptive Statistics

Characteristic	**N = 251**
bodyfat_brozek	19 (13, 25)
age	43 (36, 54)
height	70.00 (68.38, 72.25)
neck	38.00 (36.40, 39.45)
abdomen	91 (85, 99)
knee	38.50 (37.05, 39.95)
ankle	22.80 (22.00, 24.00)
bicep	32.10 (30.25, 34.35)
forearm	28.70 (27.30, 30.00)
wrist	18.30 (17.60, 18.80)

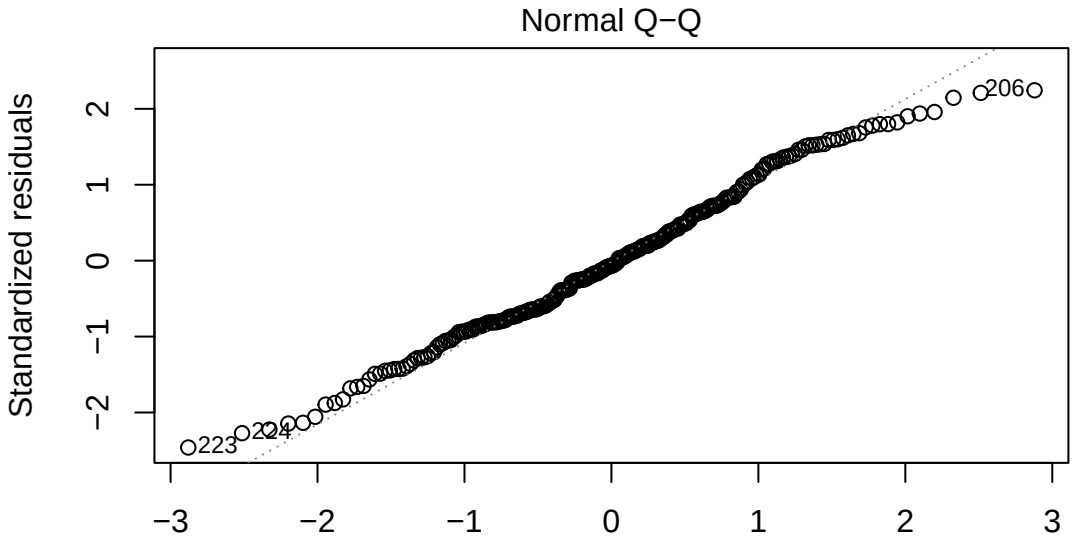
Table 2: Model Parameter for Stepwise Procedure

Term	Estimate	Standard Error	Test Statisitcs	P-value
(Intercept)	7.21	7.70	0.94	0.34950
age	0.07	0.02	2.93	0.00371
height	-0.27	0.11	-2.39	0.01777
neck	-0.53	0.20	-2.69	0.00771
abdomen	0.71	0.04	19.17	0.00000
forearm	0.47	0.17	2.73	0.00685
wrist	-1.72	0.46	-3.77	0.00020

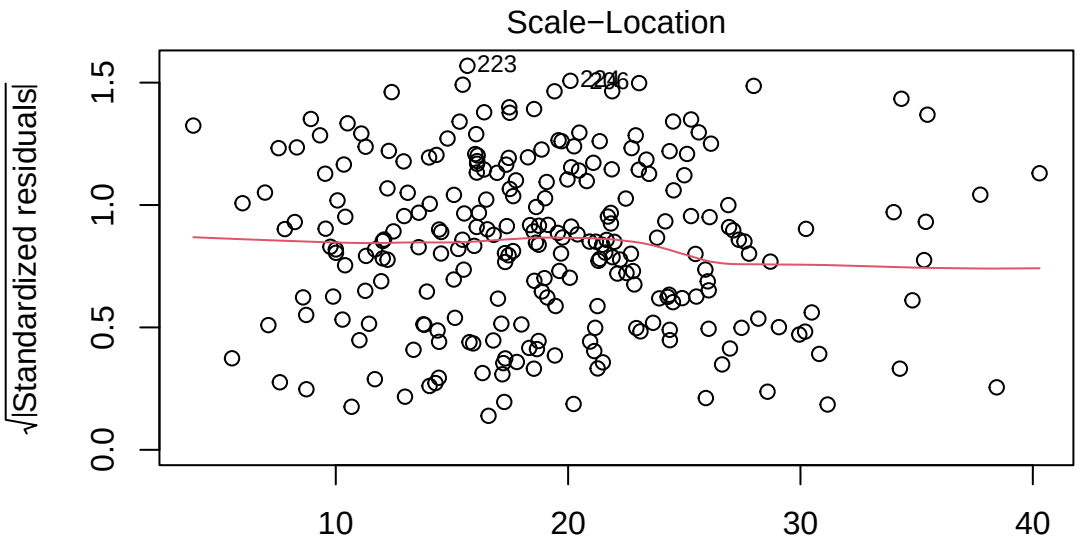
```
## lm(formula = bodyfat_brozek ~ age + height + neck + abdomen +
##     forearm + wrist + neck:abdomen, data = df_body_fat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.565 -2.836 -0.246  2.758  8.644
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -61.049723  20.691982  -2.950 0.003484 **
## age            0.053396   0.023067   2.315 0.021458 *
## height        -0.316458   0.112170  -2.821 0.005180 **
## neck           1.390707   0.575455   2.417 0.016399 *
## abdomen        1.479768   0.219648   6.737 1.16e-10 ***
## forearm        0.278790   0.176236   1.582 0.114970
## wrist         -1.605981   0.447615  -3.588 0.000403 ***
## neck:abdomen  -0.019415   0.005483  -3.541 0.000477 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.923 on 243 degrees of freedom
## Multiple R-squared:  0.7459, Adjusted R-squared:  0.7385
## F-statistic: 101.9 on 7 and 243 DF,  p-value: < 2.2e-16
```



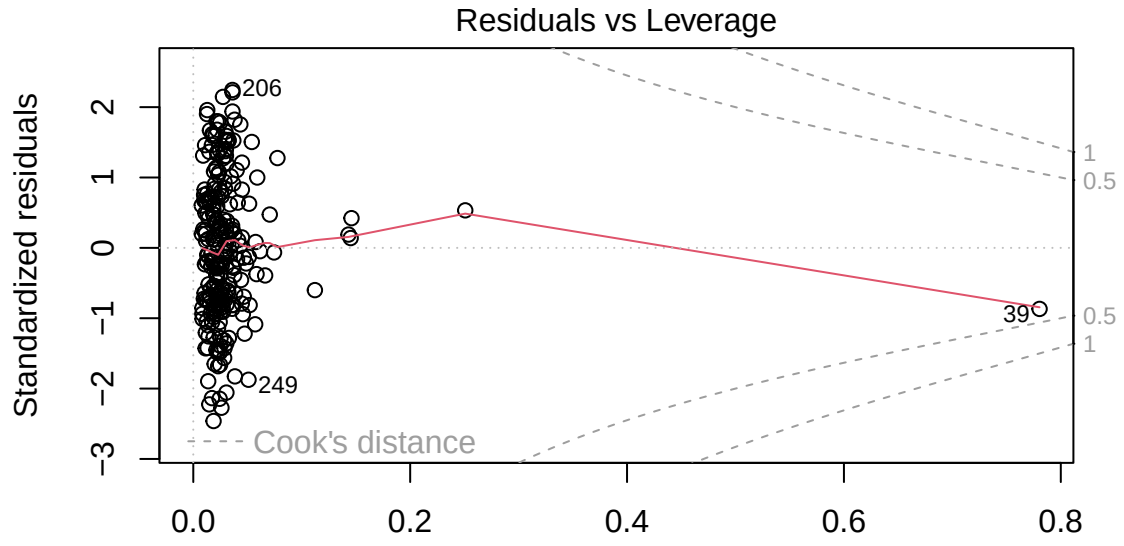
lm(bodyfat_brozek ~ age + height + neck + abdomen + forearm + wrist + neck:



lm(bodyfat_brozek ~ age + height + neck + abdomen + forearm + wrist + neck:



lm(bodyfat_brozek ~ age + height + neck + abdomen + forearm + wrist + neck:



lm(bodyfat_brozek ~ age + height + neck + abdomen + forearm + wrist + neck: