1. INTRODUCTION

1.1. Peak oxygen consumption

Peak oxygen consumption (VO_2peak, or VO_2max) is the maximum amount of oxygen a person can inhale and is measured in mL/kg/min. Importantly, it is a measure of heart health. Low VO_2peak is more closely associated with risk of heart disease and death compared to other risk factors such as smoking habits and hypertension[1]. Understanding which factors contribute to VO_2peak, and predicting an individual's VO_2peak based on these factors, is imperative to characterizing an individual's risk for cardiovascular disease and mortality. This study seeks to answer the following questions: (1) which variables are most associated with VO_2peak, and (2) what type of relationship do these variables exhibit with VO_2peak?

1.2. The data

VO_2peak data on 435 patients between the ages of 18 and 44, along with their sex, body composition (e.g., total body fat percent), and brain MRI data, along with 31 other variables, were obtained from the University of Illinois Urbana-Champaign. After accounting for missing data, 427 individuals remained in the study, with an approximately even split between males (n=219) and females (n=208).

2. MODEL BUILDING[2]

2.1.1. Identification of significant variables

To identify variables associated with VO_2peak, Pearson's correlation coefficient (r), which quantifies the extent of linear association between two variables ($0 < r < 1$, with $r = 1$ suggesting a perfect linear relationship), was calculated for each independent variable and VO_2peak. The tests revealed moderate positive correlations between VO2peak and height, lean mass, and tNAA/Cr ratio (a marker of brain health); negative correlations were observed between VO2peak and age, BMI, and percent body fat.

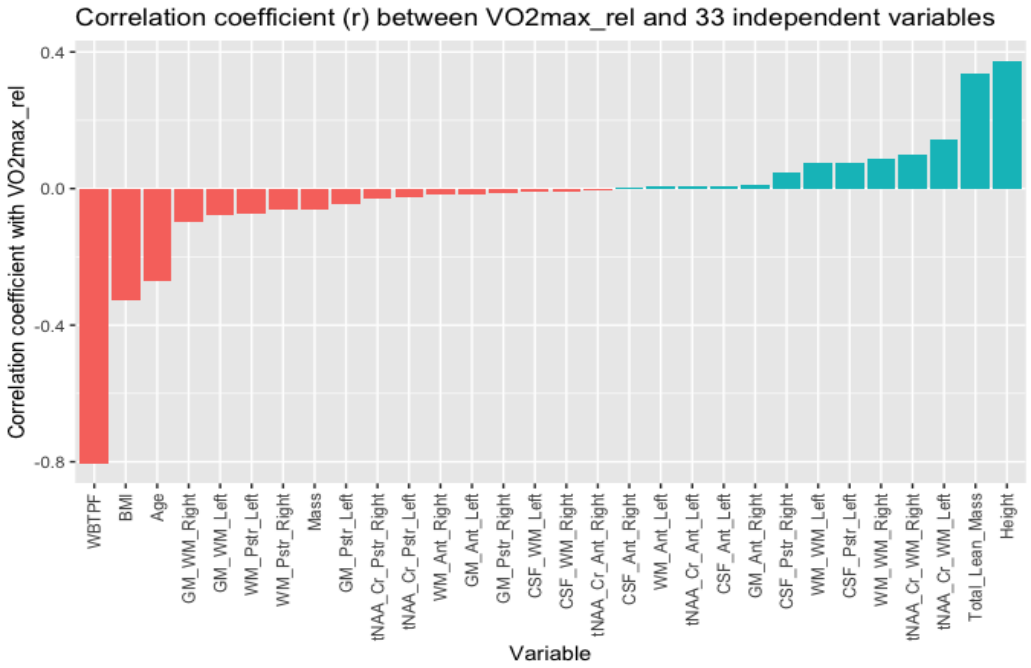


Figure 1: Correlation coefficients between VO_2peak and 33 independent variables. WBTPF: Whole body total percent fat; GM: grey matter; WM: white matter.

[1] Kokkinos P, Faselis C, Samuel IBH, et al: Cardiorespiratory Fitness and Mortality Risk Across the Spectra of Age, Race, and Sex. J Am Coll Cardiol 80:598-609, 2022

[2] A summary of all models built is found in Table 1.

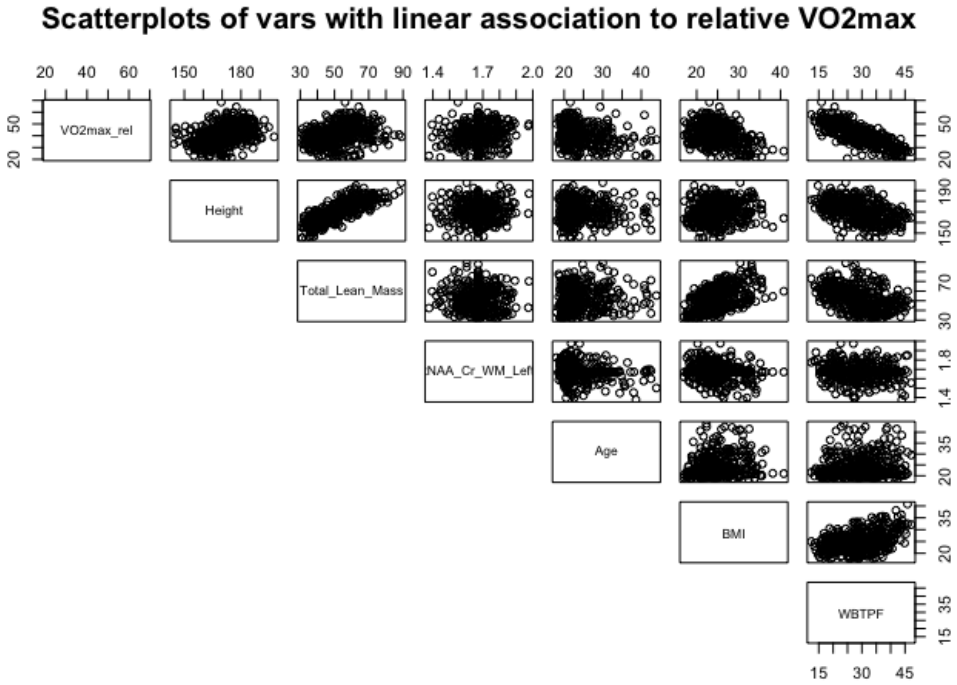


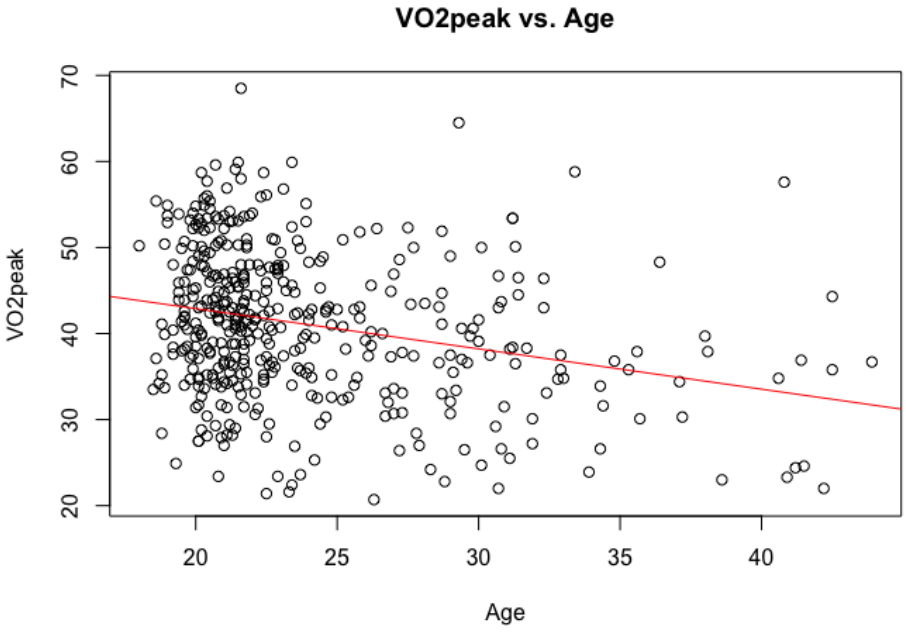
Figure 2: Scatterplot of variables with linear associations to relative VO_2peak (VO_2max).

Variable screening procedures, which systematically identify significant variables in a model, were employed to find more independent variables. Each approach (forward selection, backward elimination, stepwise selection) yielded different results.

2.1.2. Simple models

Prior studies on VO_2peak note that it is negatively associated with age and sex; **Model 1** simply predicted VO_2peak in terms of age (Figure 3).

Figure 3: Model 1 predicts VO_2peak in terms of age. The circles represent raw data points; the red line represents average predicted VO_2peak based on age.



This model suggests that age is a significant predictor of VO_2peak; in particular, it suggests that, on average, VO_2peak decreases by between 0.30 ml/kg/min and 0.61 ml/kg/min for every one-year increase in age. However, this model was only had an R^2=0.07 (i.e., it was only able to predict about 7% of the variation in VO_2peak). In other words, 93% of the variation in VO_2peak was unexplained by age. Another variable is known to be strongly associated with VO_2peak: sex. **Model 2** predicted VO_2peak in terms of age and sex. Although sex was found to be a significant predictor of VO_2peak, and increased R^2 to 0.34, R^2 is still quite low for Model 2.

2.1.3. Leveraging correlation coefficient data

To improve R^2, other variables were explored. In **Model 3**, any variable with a correlation coefficient greater than 0.1 or less than -0.1 was included. The resulting model predicted about 68% of the variation in VO_2peak. This model performs well, but other significant variables likely exist.

2.1.4. Implementing stepwise regression principles

Model 4, containing all 31 independent variables in the dataset, was built, and had an R^2 of 0.69. This is not much better than Model 3; further, it likely contains variables that are not significant. Thus, three separate iterative processes were utilized to identify significant and nonsignificant variables. Forward selection characterized VO_2peak in terms of total percent body fat and age (**Model 5**) with an R^2=0.67. Backward elimination utilized age, total percent body fat, mass, lean mass, grey/white matter fraction, and cerebrospinal fluid (CSF) white matter to predict VO_2peak (**Model 6**) with a similar R^2; the stepwise selection model included age, percent body fat, CSF white matter, and white matter fraction (**Model 7**).

2.1.5. Improving prediction

The best model obtained from the algorithmic process in terms of prediction (i.e., judging only based on R^2) was Model 7, thus, it will be explored further.

Residual analysis indicates that the model does a good job of predicting VO_2peak in terms of the independent variables. Figure 4 plots residuals (actual VO_2peak minus predicted VO_2peak) vs. predicted VO_2peak. The uniform distribution of residuals observed (i.e., residuals that fit into a rectangular shape) suggest that the model does a good job of predicting VO_2peak regardless of the values of the independent variables.

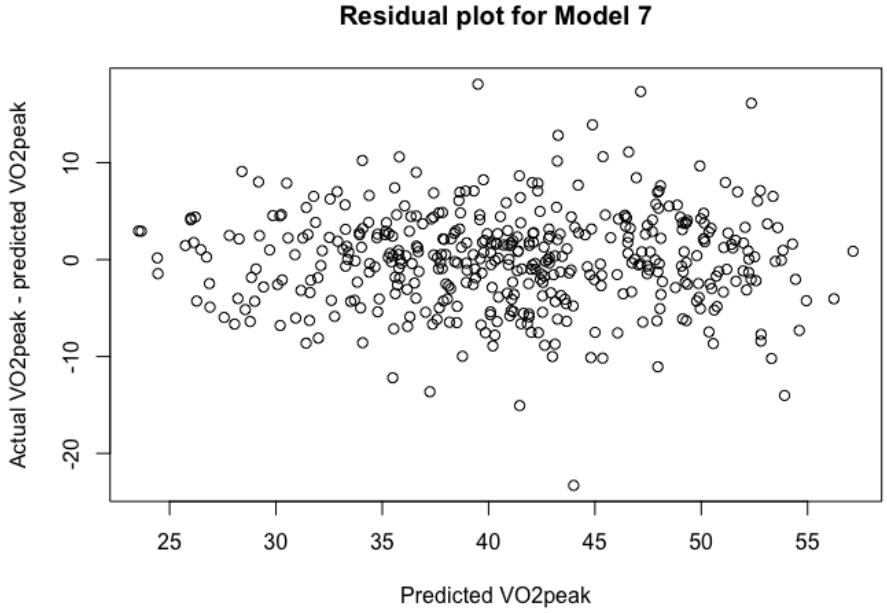


Figure 4: Residual plot of Model 7. The residuals seem uniformly distributed, which suggests that this model does a good job of estimating mean VO_2peak based on its independent variables.

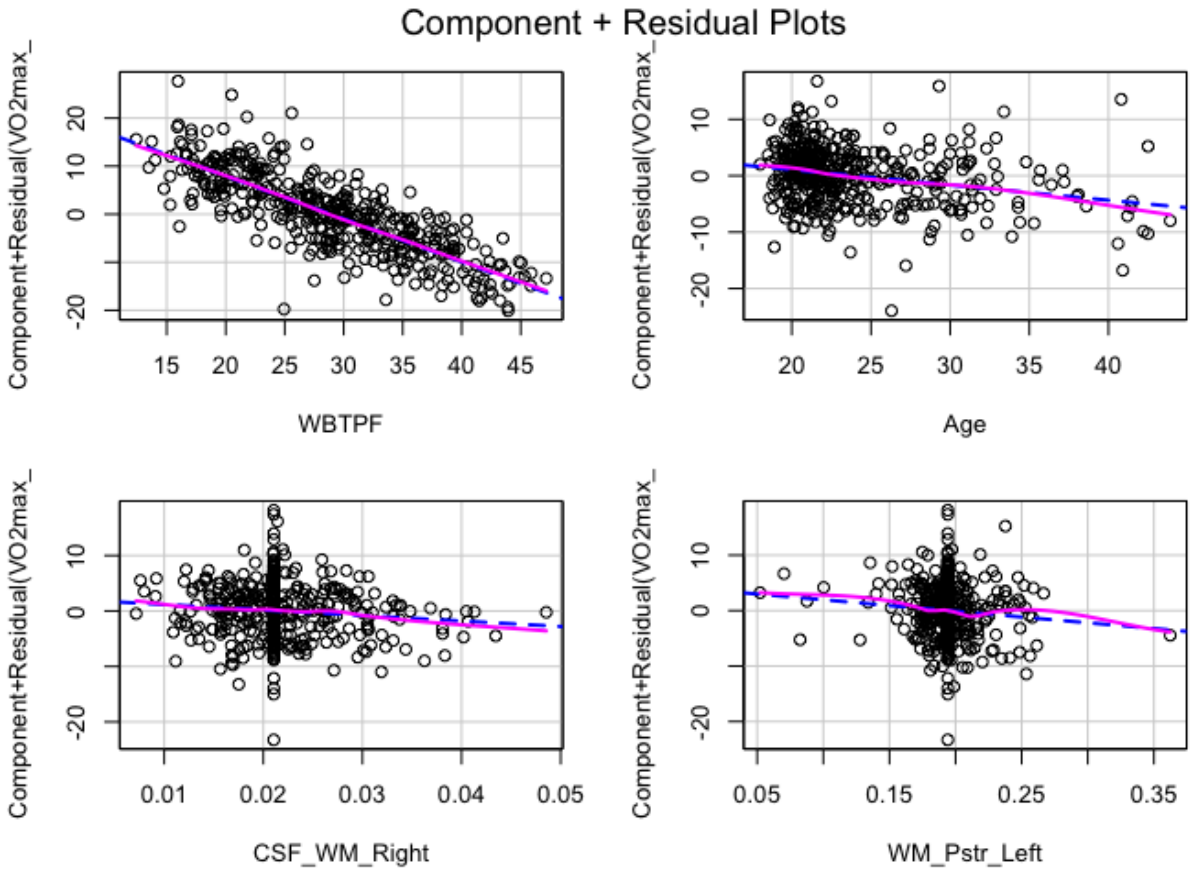


Figure 5: Partial residual plots of the independent variables in Model 7. Purple lines overlaying blue lines (in the absence of multicollinearity, which is the case here [see appendix for more details]) suggest uniform (i.e., good) residual behavior. WBTPF: Total percent body fat; CSF_WM_Right: white matter cerebrospinal fluid fraction; WM_Pstr_Left: left white matter fraction.

Figure 5 shows plots of partial residuals for Model 7. These partial residuals describe the relationship between residuals and a single independent variable after removing the effects of other independent variables. The blue line represents the expected association for the independent variable and residual assuming that they follow a linear relationship; the purple line shows the association between the independent variable and partial residual based on the model. From Figure 5, since the purple lines seem to follow the blue lines quite well, this suggests that Model 7's independent variables properly characterize their relationship to VO_2max.

Attempts to improve Model 7 were made by adding a variety of interaction and quadratic terms. After including interaction terms between age/sex, percent body fat/sex, and age/percent body fat, as well as a percent body fat^2 term, R^2 improved to 0.69 (**Model 8**). This improvement was nearly statistically significant[3]. Given that Model 8 has the highest R^2, among all the models build, it is the best one for prediction. However, its complexity makes its interpretation difficult.

3. THE FINAL MODEL AND ITS INTERPRETATION

When multiple models perform similarly, the best model is the simplest one. Based on this principle, Model 5 is the best model. Its R^2 is similar to the R^2 generated from other, more complex models (only 0.02 less than Model 8), yet it only includes two terms: age and percent

[3] p=0.0502 from nested F-test.

body fat. The addition of interaction/higher-order terms did not improve Model 5's R^2 significantly. Based on this analysis, these two variables, age and percent body fat, are the best predictors of VO_2peak.

A residual plot of actual minus predicted VO_2peak vs. predicted VO_2peak for Model 8 suggests uniform distribution of residuals. In addition, residual plots of Model 8 suggest a linear relationship between age and VO_2peak and percent body fat and VO_2peak.

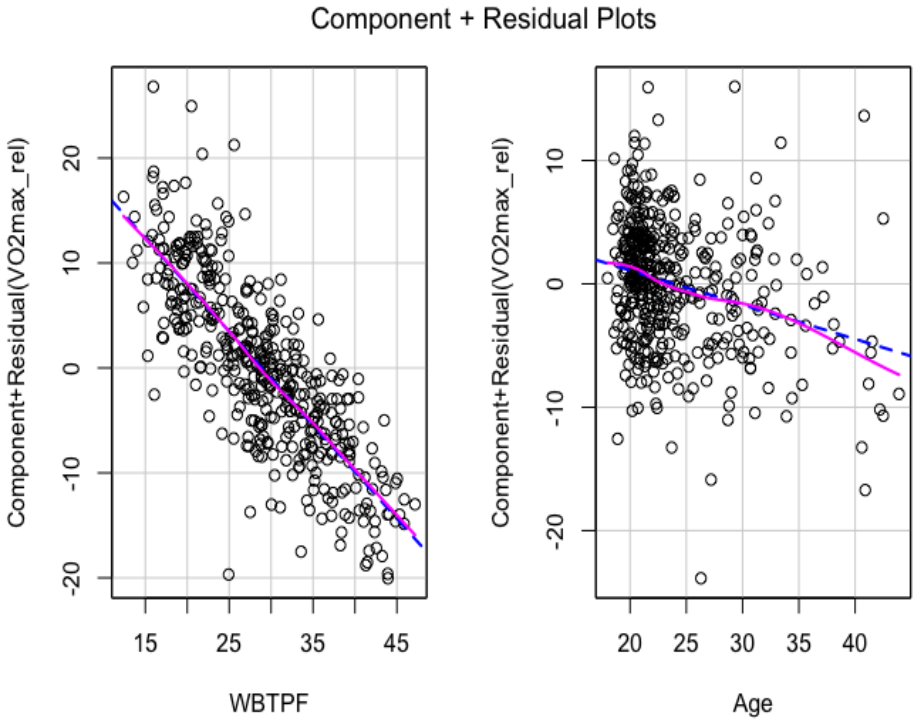


Figure 6: Partial residual plots of the variables in Model 5. These plots suggest that a linear relationship exists between total percent body fat and VO_2peak, and between age and VO_2peak. WBTPF: Whole body total percent fat.

Based on the final model generated, for every percent increase in body fat, on average, VO_2peak decreases by between 0.82 and 0.95 ml/kg/min. Further, for every one-year increase in age, on average, VO_2peak decreases by between 0.18 and 0.38 ml/kg/min. This model was not significantly improved when another term, sex, was added. Age and percent body fat are much stronger predictors of VO_2peak than sex.

In summary, the two factors most closely associated with VO_2peak were age and percent body fat. Based on this analysis, these two variables have a linear relationship with VO_2peak. Future investigations on treatments that seek to improve VO_2peak may benefit from testing whether reducing percent body fat improves VO_2peak (i.e., whether the relationship between these two variables is causal).

Table 1: Summary of models built. R^2 represents the total amount of variation in VO_2peak that the model can explain. E.g., an R^2 of 0.67 suggests that 67% of the variation in VO_2peak can be explained by the independent variables in the model.

Model	**Independent vars**	**R^2**	**Benefits**	**Pitfalls**
1	Sex	0.07	Simplicity	Very poor R^2
2	Sex, age	0.35	Simplicity	Poor R^2
3	Age, BMI, % body fat, tNAA/Cr ratio, lean mass, height	0.67	Fairly simple, good R^2	More significant variables probably exist
4	All 31 variables	0.67	Good R^2	Likely includes many insignificant variables
5	Age, % body fat	0.67	Good R^2, very simple model	Could be missing more significant variables
6	Age, % body fat, mass, lean mass, grey matter fraction, white matter fraction, CSF white matter	0.68	Good R^2	Contains many independent variables (16 total, since grey matter/white matter/CSF data are split into multiple columns)
7	Age, % body fat, CSF white matter (right), white matter fraction (left posterior)	0.68	Good R^2	Other models have similar R^2 with fewer independent variables
8	Age, percent body fat, CSF white matter (right), white matter fraction (left posterior), sex, (percent body fat)^2, percent body fat*age, percent body fat*sex, age*sex	0.69	Highest R^2	High model complexity makes interpretation difficult

4. APPENDIX

4.1. The data

The data gathered were obtained from the University of Illinois Urbana-Champaign. A CSV can be accessed here. The file contains VO_2peak data on 435 patients, along with 31 other variables of note, including sex, age, height, BMI, grey matter fraction, white matter fraction, tNAA/Cr ratio, mass, body fat composition, and others.

4.2. Resources used

The main resource used was in-class notes. I used them to learn how to perform statistical tests, variable screening procedures, create base R plots, filter variables, and interpret the results of statistical tests. I also used the textbook for further learning.

I used the book *R for Data Science* to learn how to create plots using ggplot2 (which are supplemented by dplyr and forcats). I also used the book *Machine Learning in R* to learn how to impute missing data. I also turned to Stack Overflow to address coding questions (e.g., I found the QuantPsyc package on a Stack Overflow forum to compute standardized beta coefficients).

4.3. GitHub repository

https://github.com/roberttnovo/Stat214FinalProject/tree/main

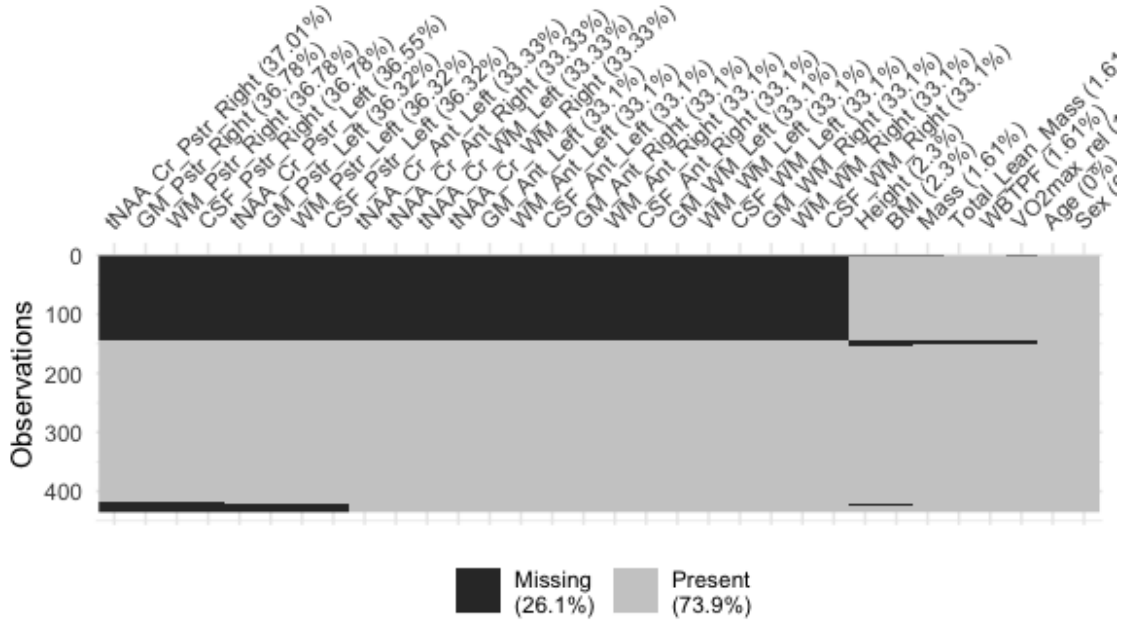
4.4. Video explanation link

YouTube: *https://www.youtube.com/watch?v=x9pA03OIQV4&feature=youtu.be*

4.5. R code used

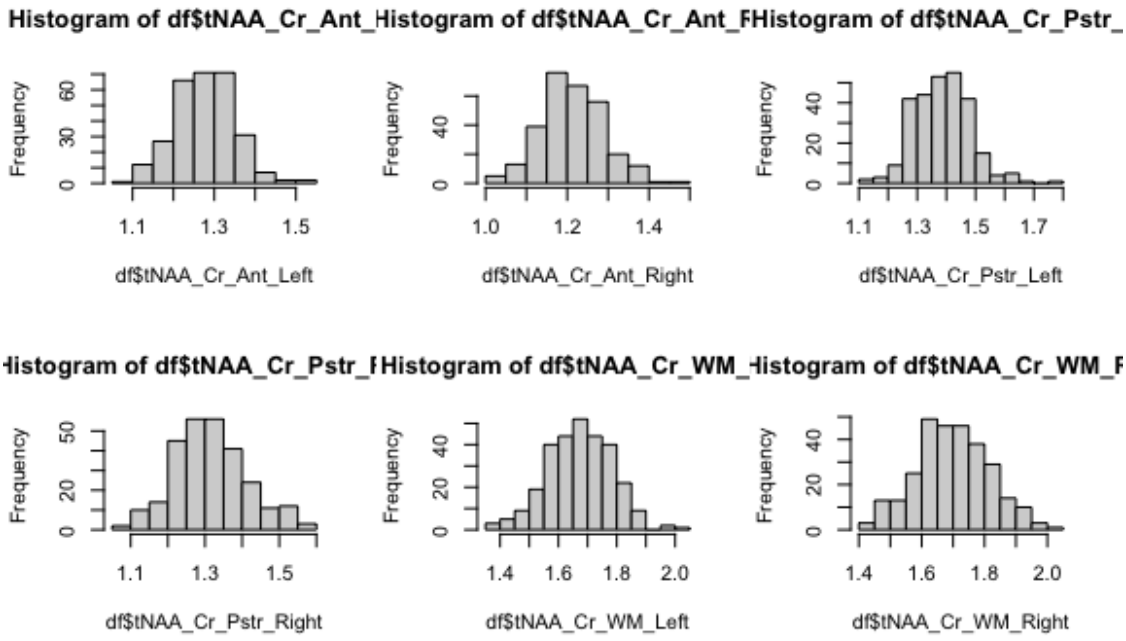
```
# Load libraries -----------------------------------------------------------

library(visdat) # For visualizing missing data
library(ggplot2) # For more cool visualizations
library(dplyr) # For doing some data transformations
library(forcats) # For more visuals
library(corrr) # For looking at correlation coefficient data easily
library(olsrr) # For stepwise regression
library(QuantPsyc) # For standardized beta
library(car) # For VIF calculations

# Load data ----------------------------------------------------------------

data_raw <- read.csv("/Users/robertnovo/Desktop/stat214/Project/Baseline_Data
_Insight1b.csv")


# Data cleaning ------------------------------------------------

# Which columns will I remove from the model?
colnames(data)

# SubjectNum: R already has row numbers, so I don't need that column.
# I also won't use the confidence column.
# Also, I don't need the Tech column, that just tells me who did the MRI scan
s.
df <- subset(data_raw, select = -c(SubjectNum, Confidence, Tech)) # df will b
e the data frame I use to do my analysis

# Volume of interest (VOI) is just a feature of the MRI scan,
```

```
# not saying anything about an individual person, so I will drop it.
# Also, scanfrac is another feature of the MRI scan that doesn't say anything
# about the patient themself, so I will drop it too.
df <- subset(df, select = -c(VOI_RL, VOI_AP, scanfrac_Ant_Right, scanfrac_Pst
r_Right,
                             scanfrac_Ant_Left, scanfrac_Pstr_Left,
                             scanfrac_WM_Left, scanfrac_WM_Right))

# The readme.txt tells me that VO2max_absolute is simply relative VO2max * Ma
ss/1000,
# and that Fat_Free_VO2max is VO2max_absolute/Total_Lean_Mass. Since I want t
o predict
# VO2max_rel (relative VO2max), I will drop these two variables
df <- subset(df, select = -c(VO2max_abs, Fat_Free_VO2max))

# Missing values ------------------------------------------------------------
# Let's see what data is missing
vis_miss(df, cluster = TRUE, sort_miss = TRUE)
```

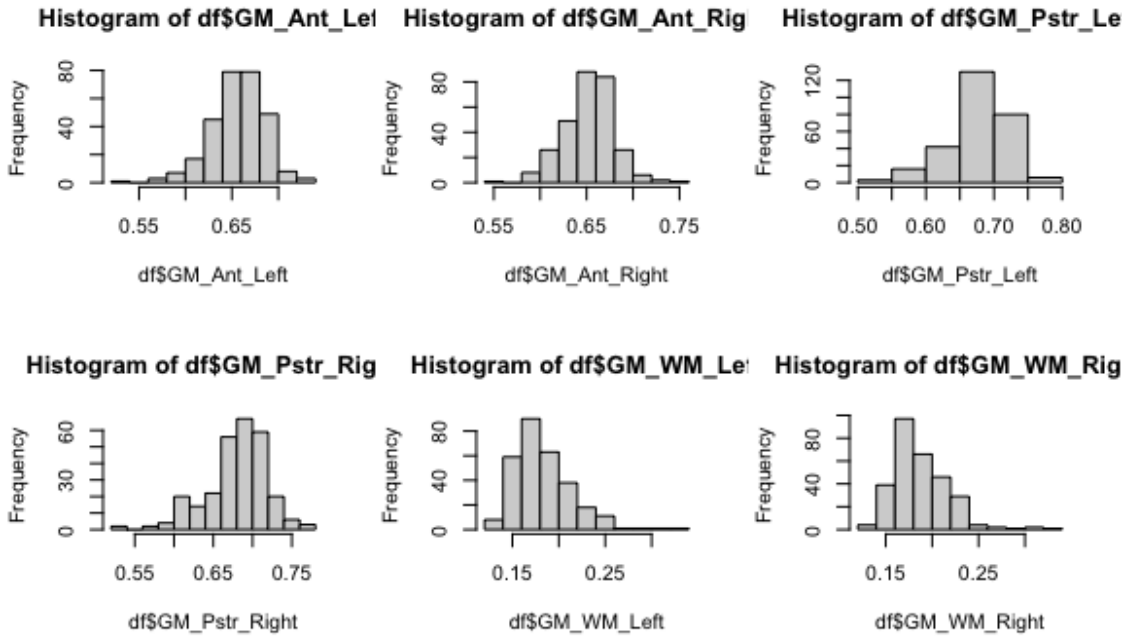



```
# It seems like all columns that start with tNAA or GM have the most missing
data

# Let's look at tNAA_Cr (NAA/Cr ratio) distribution
par(mfrow = c(2, 3))
hist(df$tNAA_Cr_Ant_Left)
hist(df$tNAA_Cr_Ant_Right)
hist(df$tNAA_Cr_Pstr_Left)
hist(df$tNAA_Cr_Pstr_Right)
```

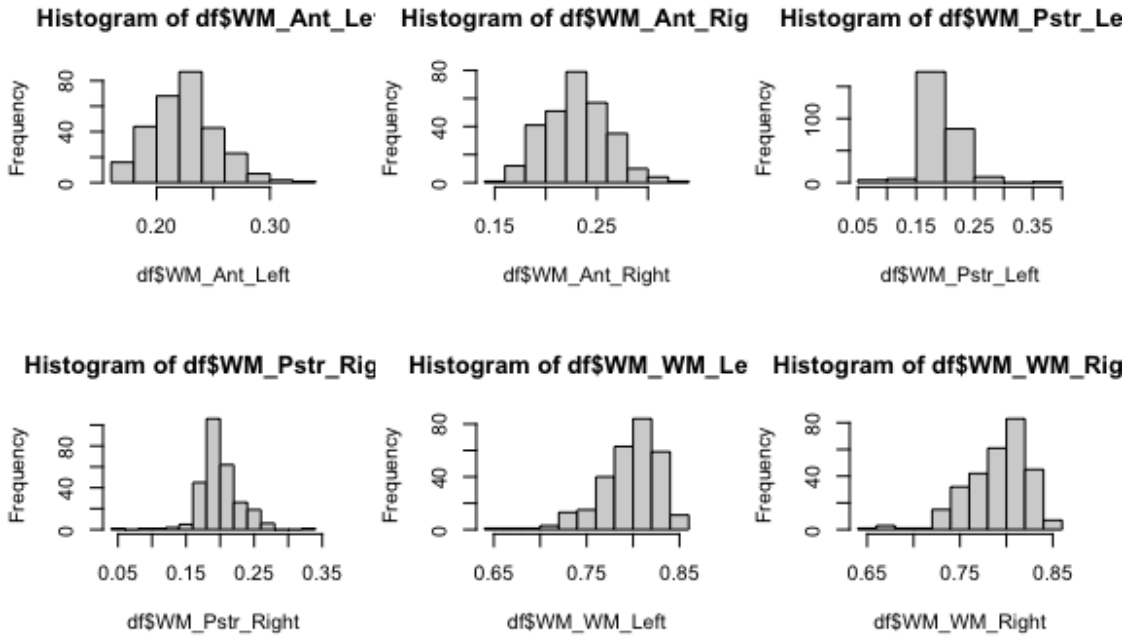
```
hist(df$tNAA_Cr_WM_Left)
hist(df$tNAA_Cr_WM_Right)
```



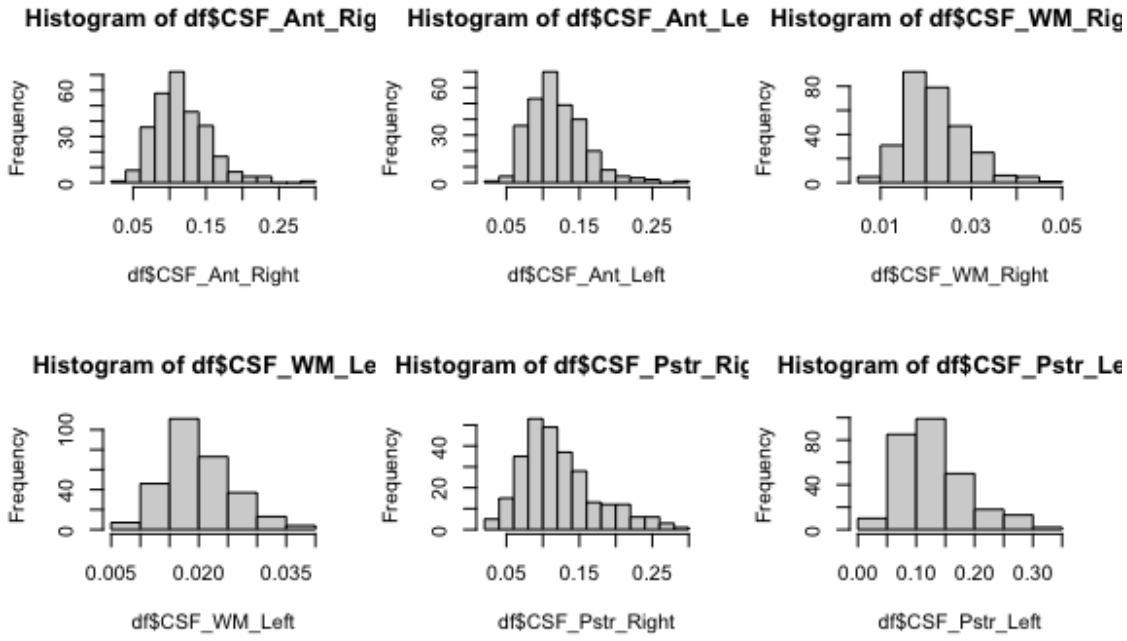

```
# The data seem normally distributed with low SD so I will impute using the m
ean

df$tNAA_Cr_Ant_Left[is.na(df$tNAA_Cr_Ant_Left)] <- mean(df$tNAA_Cr_Ant_Left,
na.rm = TRUE)
df$tNAA_Cr_Ant_Right[is.na(df$tNAA_Cr_Ant_Right)] <- mean(df$tNAA_Cr_Ant_Righ
t, na.rm = TRUE)
df$tNAA_Cr_Pstr_Left[is.na(df$tNAA_Cr_Pstr_Left)] <- mean(df$tNAA_Cr_Pstr_Lef
t, na.rm = TRUE)
df$tNAA_Cr_Pstr_Right[is.na(df$tNAA_Cr_Pstr_Right)] <- mean(df$tNAA_Cr_Pstr_R
ight, na.rm = TRUE)
df$tNAA_Cr_WM_Left[is.na(df$tNAA_Cr_WM_Left)] <- mean(df$tNAA_Cr_WM_Left, na.
rm = TRUE)
df$tNAA_Cr_WM_Right[is.na(df$tNAA_Cr_WM_Right)] <- mean(df$tNAA_Cr_WM_Right,
na.rm = TRUE)

# Let's look at GM (gray matter fraction)
par(mfrow = c(2, 3))
hist(df$GM_Ant_Left)
hist(df$GM_Ant_Right)
hist(df$GM_Pstr_Left)
hist(df$GM_Pstr_Right)
hist(df$GM_WM_Left)
hist(df$GM_WM_Right)
```



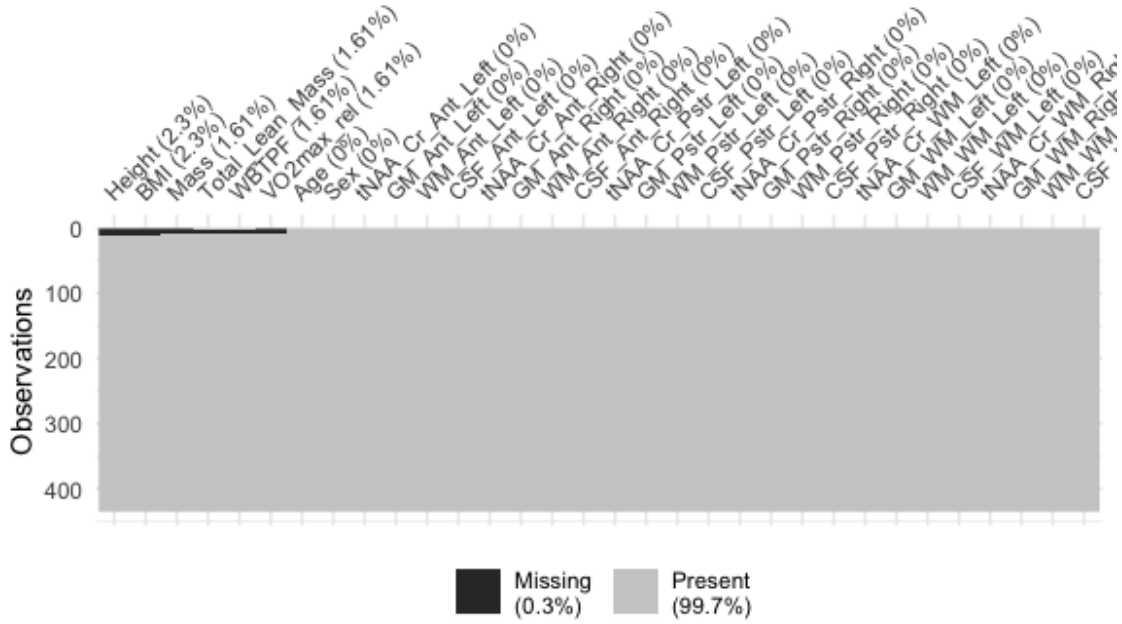

```
# These data seem to be more skewed, so I will impute using median
df$GM_Ant_Left[is.na(df$GM_Ant_Left)] <- median(df$GM_Ant_Left, na.rm = TRUE)
df$GM_Ant_Right[is.na(df$GM_Ant_Right)] <- median(df$GM_Ant_Right, na.rm = TR
UE)
df$GM_Pstr_Left[is.na(df$GM_Pstr_Left)] <- median(df$GM_Pstr_Left, na.rm = TR
UE)
df$GM_Pstr_Right[is.na(df$GM_Pstr_Right)] <- median(df$GM_Pstr_Right, na.rm =
 TRUE)
df$GM_WM_Left[is.na(df$GM_WM_Left)] <- median(df$GM_WM_Left, na.rm = TRUE)
df$GM_WM_Right[is.na(df$GM_WM_Right)] <- median(df$GM_WM_Right, na.rm = TRUE)

# White matter fraction
par(mfrow = c(2, 3))
hist(df$WM_Ant_Left)
hist(df$WM_Ant_Right)
hist(df$WM_Pstr_Left)
hist(df$WM_Pstr_Right) # the three above seem normally disributed; these thre
e
hist(df$WM_WM_Left) # seem right-skewed
hist(df$WM_WM_Right)
```




```
# 3 imputed using mean; 3 using median
df$WM_Ant_Left[is.na(df$WM_Ant_Left)] <- mean(df$WM_Ant_Left, na.rm = TRUE)
df$WM_Ant_Right[is.na(df$WM_Ant_Right)] <- mean(df$WM_Ant_Right, na.rm = TRU
E)
df$WM_Pstr_Left[is.na(df$WM_Pstr_Left)] <- mean(df$WM_Pstr_Left, na.rm = TRU
E)
df$WM_Pstr_Right[is.na(df$WM_Pstr_Right)] <- median(df$WM_Pstr_Right, na.rm =
 TRUE)
df$WM_WM_Left[is.na(df$WM_WM_Left)] <- median(df$WM_WM_Left, na.rm = TRUE)
df$WM_WM_Right[is.na(df$WM_WM_Right)] <- median(df$WM_WM_Right, na.rm = TRUE)

# CSF
par(mfrow = c(2, 3))
hist(df$CSF_Ant_Right) # left skew
hist(df$CSF_Ant_Left) # left skew
hist(df$CSF_WM_Right) # left skew
hist(df$CSF_WM_Left) # left skew
hist(df$CSF_Pstr_Right) # left skew
hist(df$CSF_Pstr_Left) # left skew
```




```
# These data are all left skewed; will impute with median
df$CSF_Ant_Right[is.na(df$CSF_Ant_Right)] <- median(df$CSF_Ant_Right, na.rm =
 TRUE)
df$CSF_Ant_Left[is.na(df$CSF_Ant_Left)] <- median(df$CSF_Ant_Left, na.rm = TR
UE)
df$CSF_WM_Right[is.na(df$CSF_WM_Right)] <- median(df$CSF_WM_Right, na.rm = TR
UE)
df$CSF_WM_Left[is.na(df$CSF_WM_Left)] <- median(df$CSF_WM_Left, na.rm = TRUE)
df$CSF_Pstr_Right[is.na(df$CSF_Pstr_Right)] <- median(df$CSF_Pstr_Right, na.r
m = TRUE)
df$CSF_Pstr_Left[is.na(df$CSF_Pstr_Left)] <- median(df$CSF_Pstr_Left, na.rm =
 TRUE)

# Visualize missing data again
vis_miss(df, cluster = TRUE, sort_miss = TRUE)
```

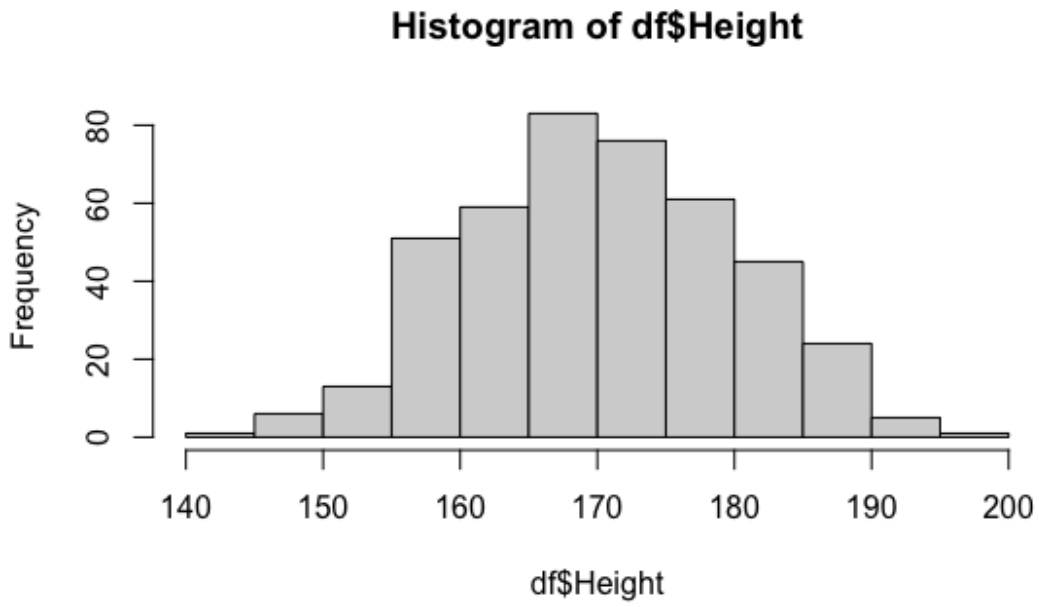



```
# Now it's just a few missing values in height, bmi, mass, lean mass, wbtpf,
# and vo2max
sapply(df, function(y) sum(length(which(is.na(y))))) # See how many NAs in ea
ch column
```

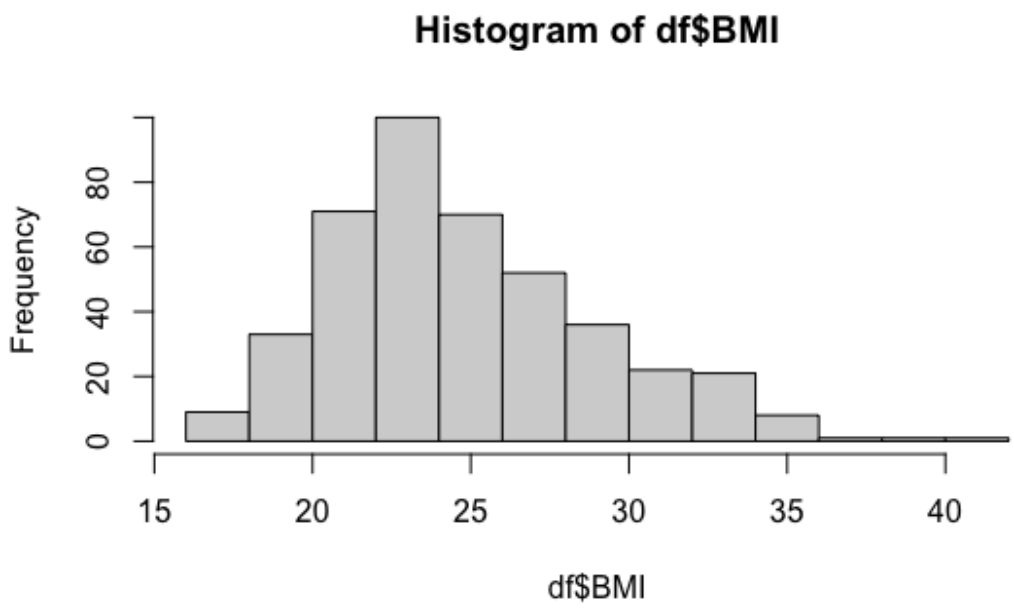
```
##                Age                Sex             Height               Mas
s
##                  0                  0                 10
7
##    Total_Lean_Mass                BMI              WBTPF         VO2max_re
l
##                  7                 10                  7
7
##   tNAA_Cr_Ant_Left        GM_Ant_Left        WM_Ant_Left       CSF_Ant_Lef
t
##                  0                  0                  0
0
##  tNAA_Cr_Ant_Right       GM_Ant_Right       WM_Ant_Right      CSF_Ant_Righ
t
##                  0                  0                  0
0
##  tNAA_Cr_Pstr_Left       GM_Pstr_Left       WM_Pstr_Left      CSF_Pstr_Lef
t
##                  0                  0                  0
0
## tNAA_Cr_Pstr_Right      GM_Pstr_Right      WM_Pstr_Right     CSF_Pstr_Righ
t
##                  0                  0                  0
0
```

```
##    tNAA_Cr_WM_Left          GM_WM_Left          WM_WM_Left         CSF_WM_Lef
t
##                  0                   0                   0
0
##   tNAA_Cr_WM_Right        GM_WM_Right        WM_WM_Right       CSF_WM_Righ
t
##                  0                   0                   0
0
```

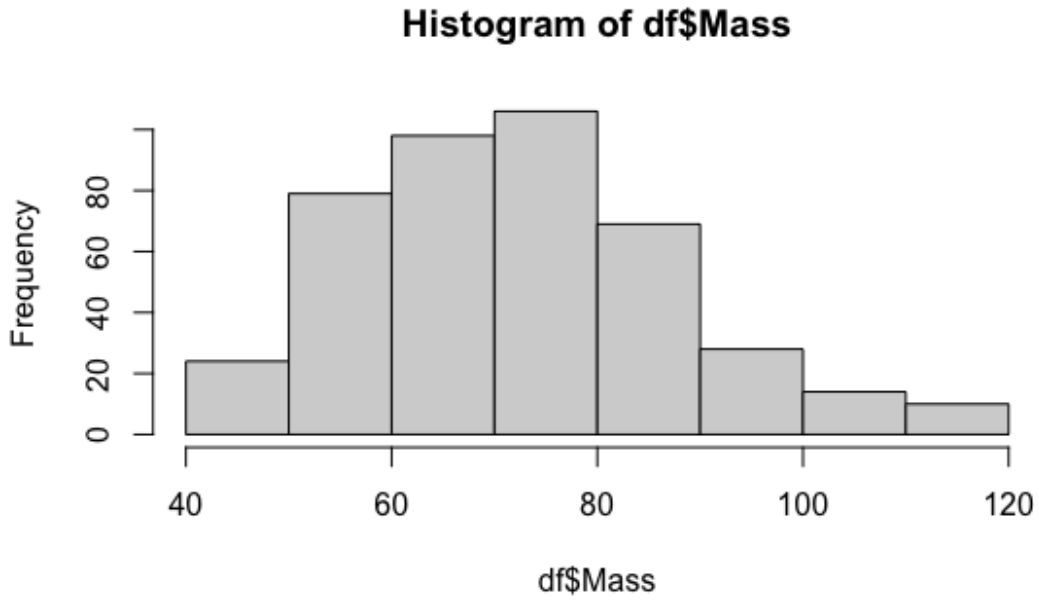
```
# Height
hist(df$Height) # Normal
```



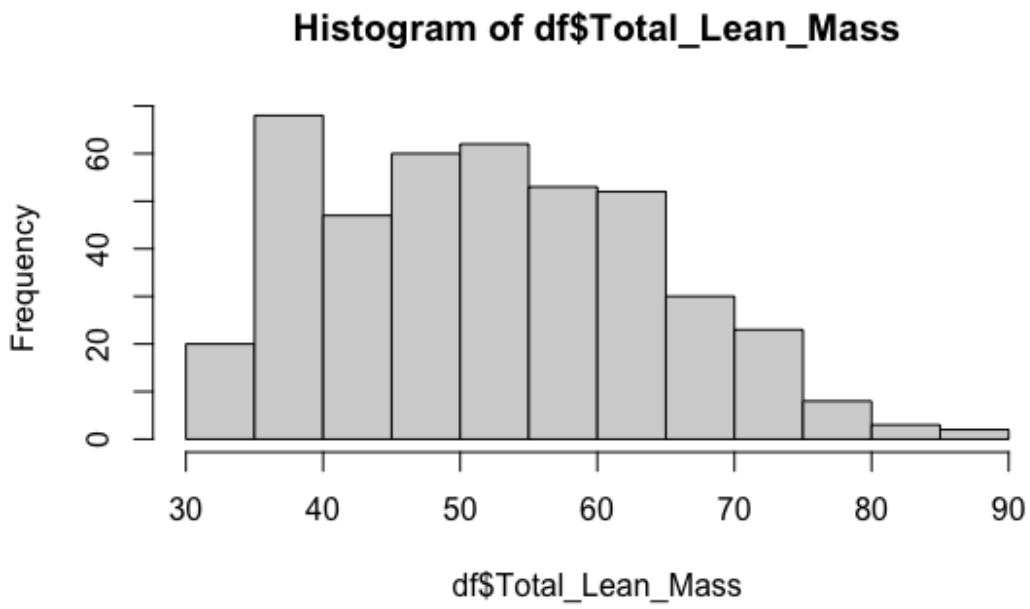

```
df$Height[is.na(df$Height)] <- mean(df$Height, na.rm = TRUE)

# BMI
hist(df$BMI) # Left skew
```



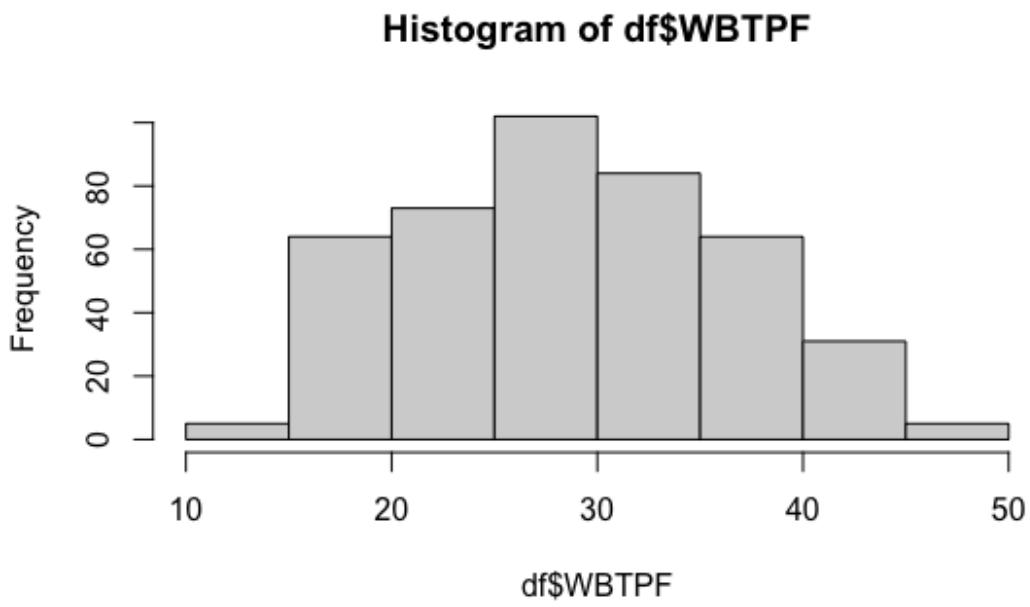

```r
df$BMI[is.na(df$BMI)] <- mean(df$BMI, na.rm = TRUE)

# Mass
hist(df$Mass) # Left skew
```



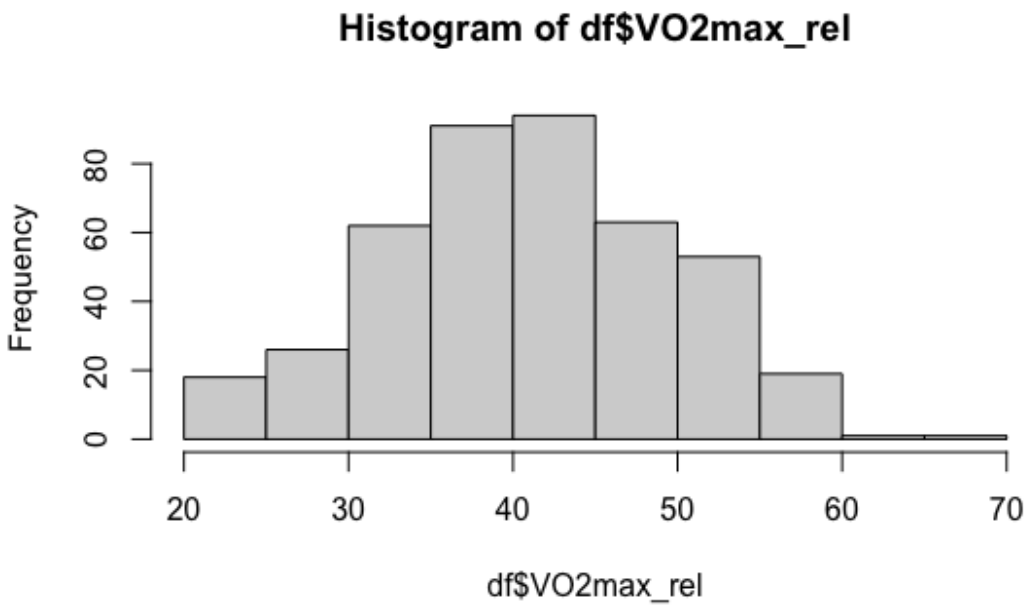

```r
df$Mass[is.na(df$Mass)] <- median(df$Mass, na.rm = TRUE)

# Total lean mass, wbtpf, vo2max
hist(df$Total_Lean_Mass) # Uniform
```



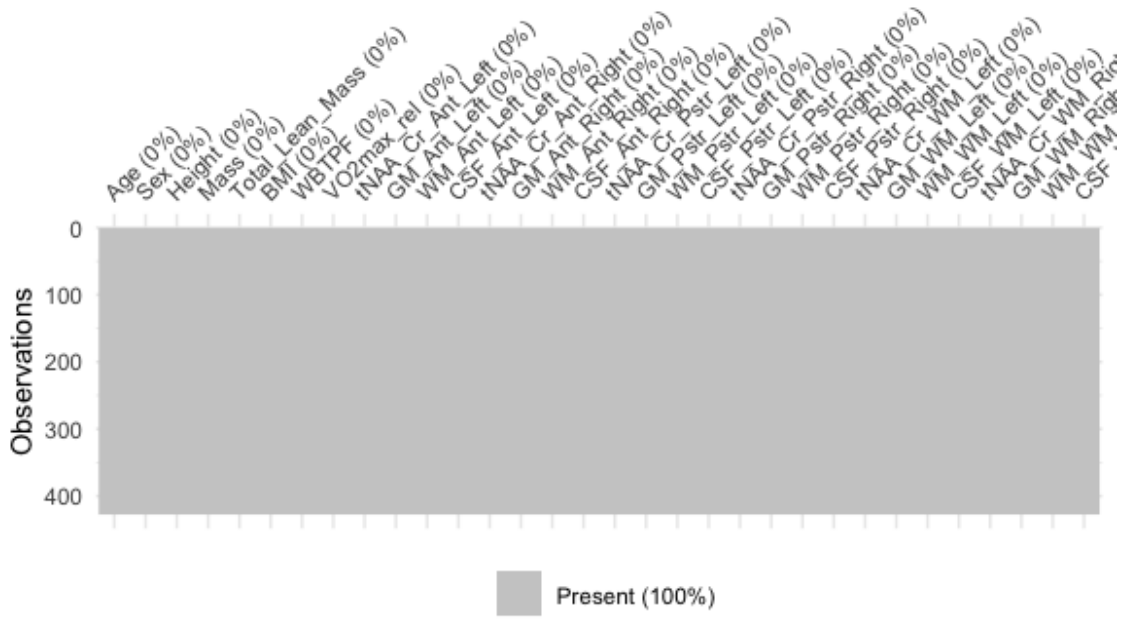

```
hist(df$WBTPF) # Somewhat unform
```




```
hist(df$VO2max_rel) # Normal
```




```
# These variables are so important to the analysis that I don't want to imput
e them
# There's only 7 missing values for each of these three variables, and they'r
e all
# in the same row. They make up a tiny portion of the data (7/435 observation
s)
# so I'm just going to drop them from the dataframe

df <- df %>%
  filter(!is.na(Total_Lean_Mass) & !is.na(WBTPF) & !is.na(VO2max_rel))

# Visualize missing data again
vis_miss(df, cluster = TRUE, sort_miss = TRUE)
```

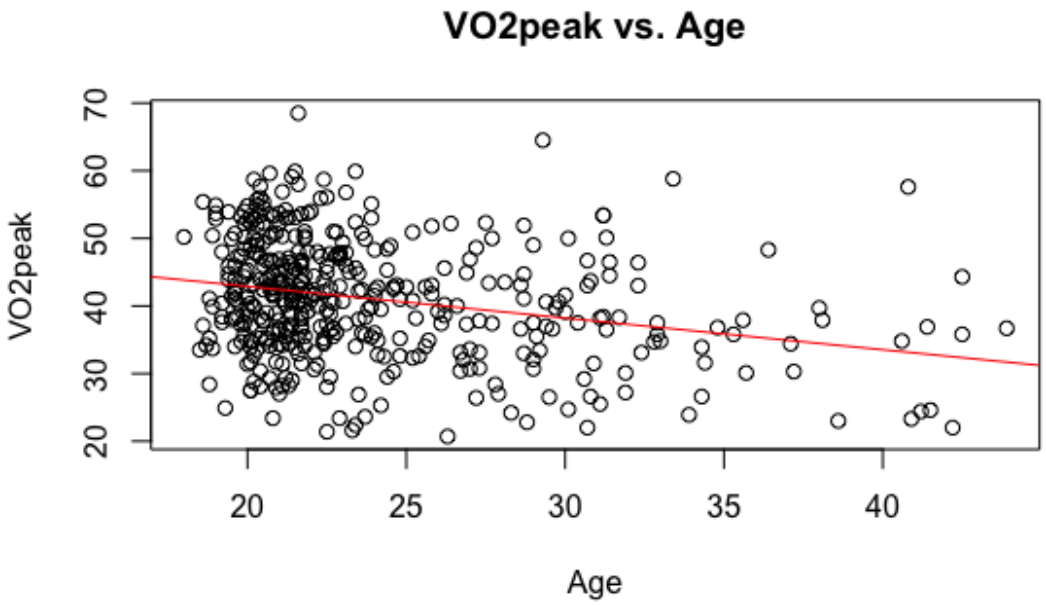



```
# No more missing data (yay)

par(mfrow = c(1, 1)) # Fix graph display

# Hypothesizing E(y) ------------------------------------------------------

# I know VO2peak is closely associated with age, so let's make that our base
# model
base_model <- lm(VO2max_rel ~ Age, data = df)
summary(base_model)

##
## Call:
## lm(formula = VO2max_rel ~ Age, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.3205  -5.8977   0.0846   5.8622  26.3572
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 52.27634    1.99447  26.211  < 2e-16 ***
## Age         -0.46915    0.08155  -5.753 1.68e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.399 on 425 degrees of freedom
## Multiple R-squared:  0.07225,    Adjusted R-squared:  0.07007
## F-statistic:  33.1 on 1 and 425 DF,  p-value: 1.678e-08
```

```
confint(base_model)
```

```
##                   2.5 %     97.5 %
## (Intercept) 48.3560942 56.1965888
## Age         -0.6294311 -0.3088637
```

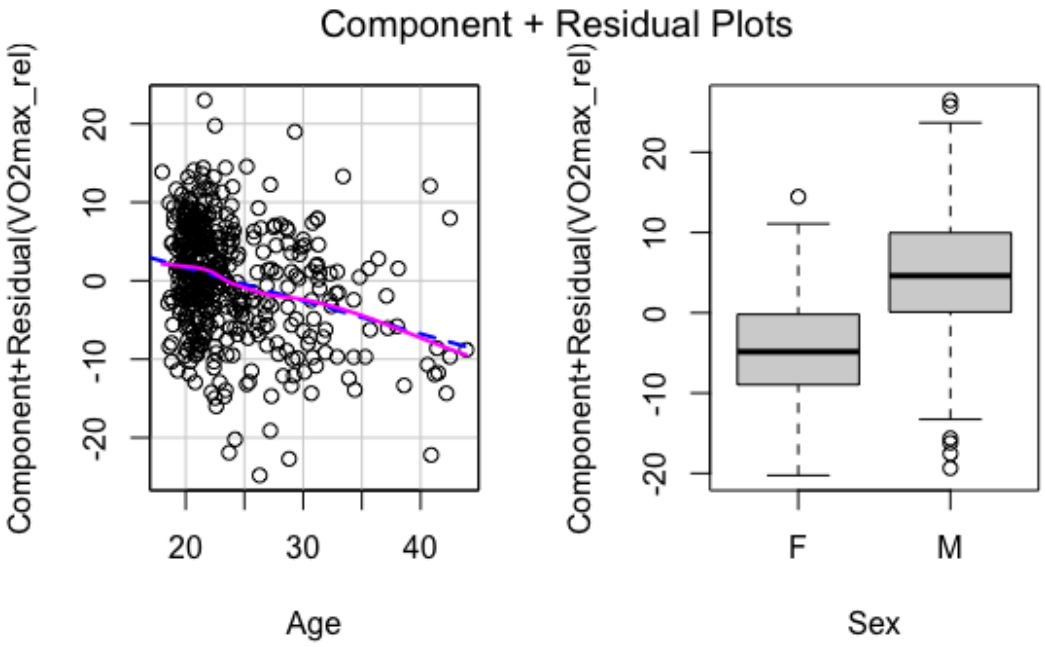
```
plot(df$Age, df$VO2max_rel, xlab = 'Age', ylab = 'VO2peak', main = 'VO2peak v
s. Age')
abline(base_model, col = 'red')
```




```
# Lets add sex
base_model <- lm(VO2max_rel ~ Age + Sex, data = df)
summary(base_model)
```

```
##
## Call:
## lm(formula = VO2max_rel ~ Age + Sex, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.8125  -4.3391   0.0242   4.9790  22.0055
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  46.4409     1.7288  26.864  < 2e-16 ***
## Age          -0.4217     0.0685  -6.156 1.73e-09 ***
## SexM          9.1620     0.6830  13.414  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

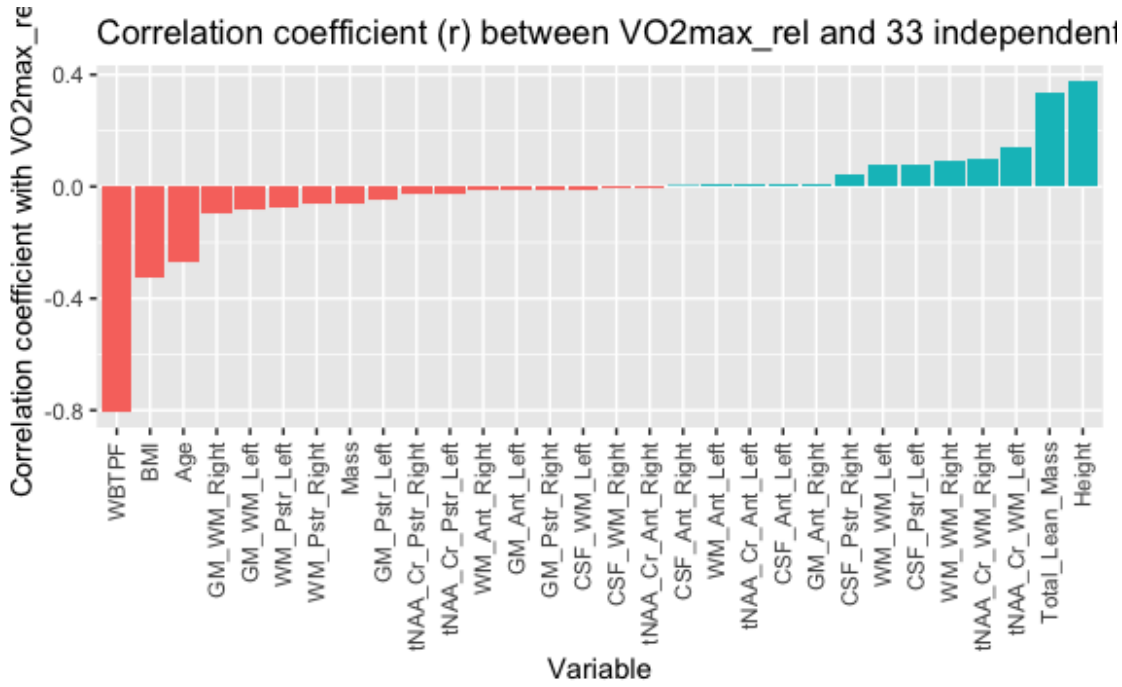
```
## Residual standard error: 7.045 on 424 degrees of freedom
## Multiple R-squared:  0.3487, Adjusted R-squared:  0.3456
## F-statistic: 113.5 on 2 and 424 DF,  p-value: < 2.2e-16

crPlots(base_model)
```



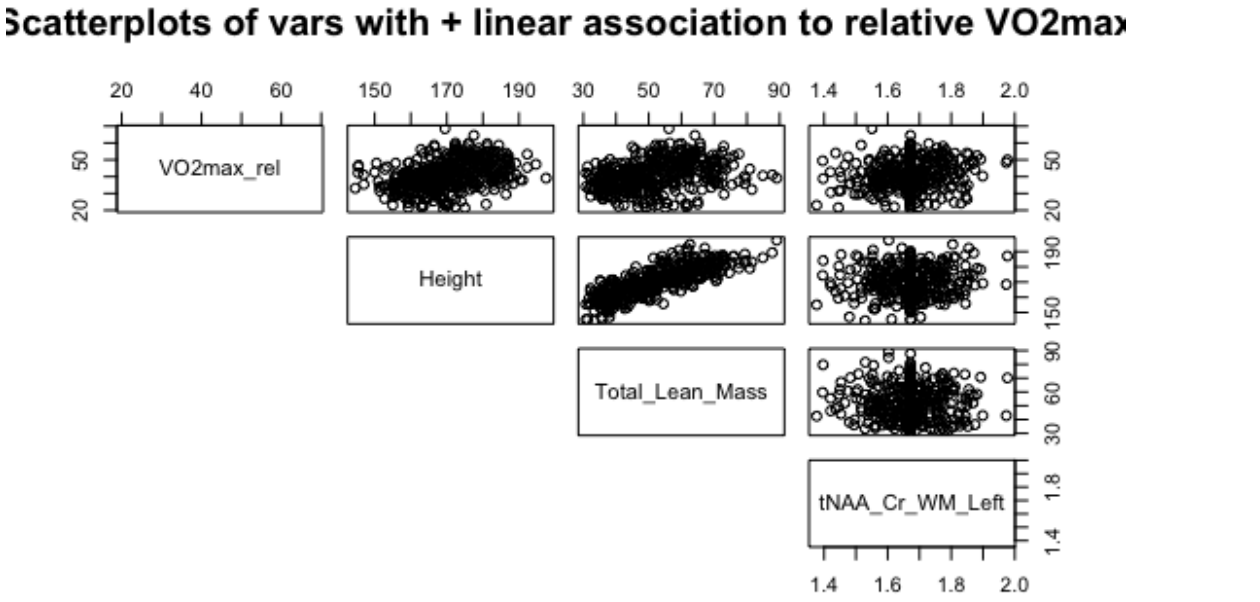

```
# First, I want to see the correlation between relativeVO2max and
# all other variables
# There's 33 variables in this dataframe, so I'll use a package, corrr, to
# make my analysis easier
correlation <- df %>%
  correlate() %>%
  focus(VO2max_rel) %>%
  arrange(desc(VO2max_rel)) %>%
  print(n = Inf)

## # A tibble: 30 × 2
##    term               VO2max_rel
##    <chr>                   <dbl>
##  1 Height                 0.375
##  2 Total_Lean_Mass        0.338
##  3 tNAA_Cr_WM_Left        0.143
##  4 tNAA_Cr_WM_Right       0.0989
##  5 WM_WM_Right            0.0887
##  6 CSF_Pstr_Left          0.0754
##  7 WM_WM_Left             0.0747
##  8 CSF_Pstr_Right         0.0458
##  9 GM_Ant_Right           0.0114
## 10 CSF_Ant_Left           0.00686
## 11 tNAA_Cr_Ant_Left       0.00568
```

```
## 12 WM_Ant_Left            0.00523
## 13 CSF_Ant_Right          0.00394
## 14 tNAA_Cr_Ant_Right     -0.00406
## 15 CSF_WM_Right          -0.00879
## 16 CSF_WM_Left           -0.0108
## 17 GM_Pstr_Right         -0.0133
## 18 GM_Ant_Left           -0.0157
## 19 WM_Ant_Right          -0.0157
## 20 tNAA_Cr_Pstr_Left     -0.0266
## 21 tNAA_Cr_Pstr_Right    -0.0275
## 22 GM_Pstr_Left          -0.0438
## 23 Mass                  -0.0620
## 24 WM_Pstr_Right         -0.0627
## 25 WM_Pstr_Left          -0.0750
## 26 GM_WM_Left            -0.0794
## 27 GM_WM_Right           -0.0971
## 28 Age                   -0.269
## 29 BMI                   -0.329
## 30 WBTPF                 -0.804
```

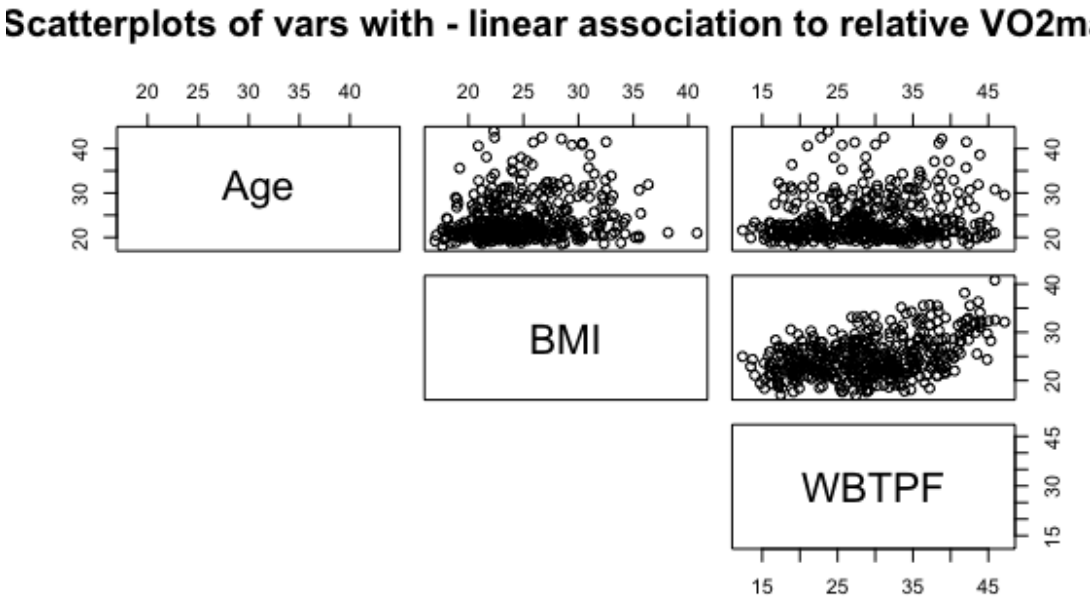
```r
# Let's visualize this using ggplot
ggplot(correlation, aes(x = fct_reorder(term, VO2max_rel), y = VO2max_rel)) +
  geom_col(aes(fill = VO2max_rel > 0), show.legend = FALSE) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  ggtitle("Correlation coefficient (r) between VO2max_rel and 33 independent
variables ") +
  xlab("Variable") +
  ylab("Correlation coefficient with VO2max_rel")
```



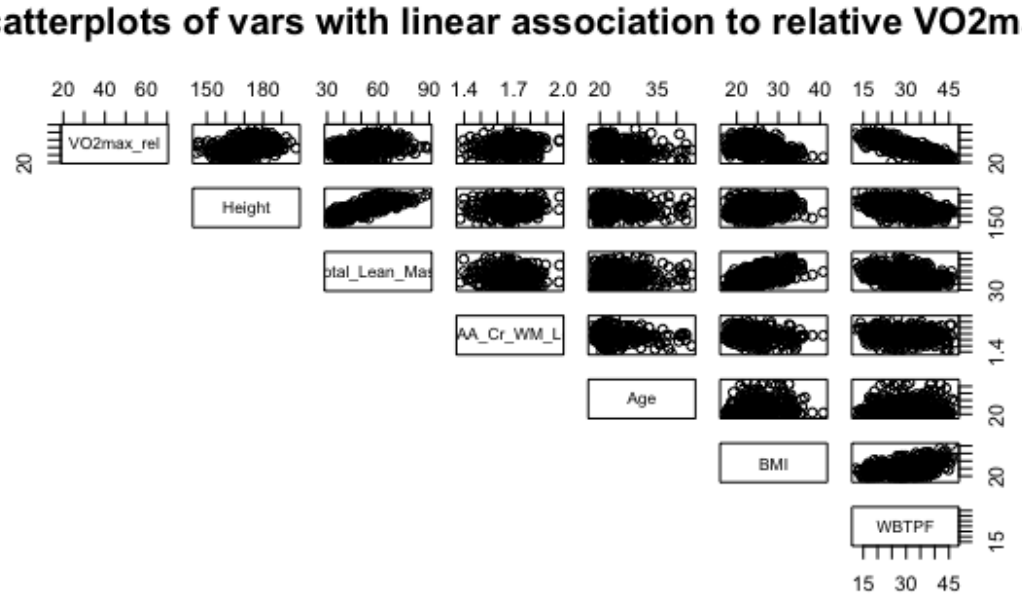
```
# Height, followed by lean mass, and white matter fraction, seem to have the
# strongest positive association with VO2peak
# Whole body total percent fat (WBTPF) seems to have a very strong negative
# association with VO2peak
pairs(df[, c('VO2max_rel', 'Height', 'Total_Lean_Mass',
             'tNAA_Cr_WM_Left')], lower.panel = NULL, main = 'Scatterplots of
 vars with + linear association to relative VO2max')
```




```
pairs(df[, c('Age', 'BMI','WBTPF')], lower.panel = NULL,
      main = 'Scatterplots of vars with - linear association to relative VO2m
ax')
```

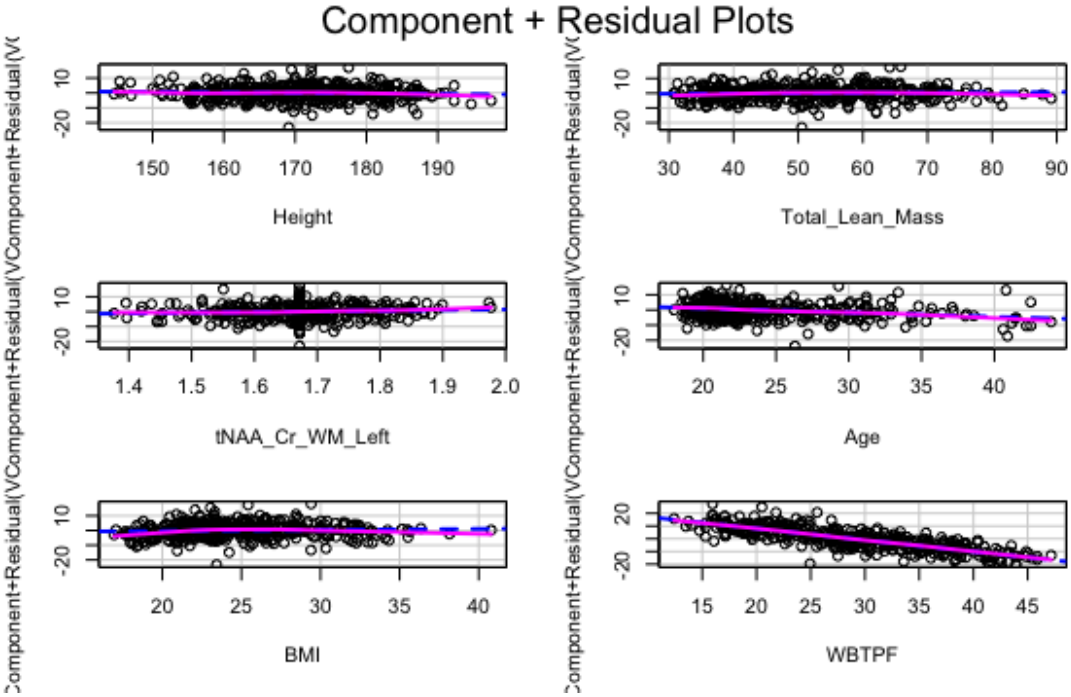



```
pairs(df[, c('VO2max_rel', 'Height', 'Total_Lean_Mass',
             'tNAA_Cr_WM_Left', 'Age', 'BMI','WBTPF')], lower.panel = NULL,
      main = 'Scatterplots of vars with linear association to relative VO2max
')
```




```
# For now, my hypothesized model will be:
model_one <- lm(VO2max_rel ~ Height + Total_Lean_Mass + tNAA_Cr_WM_Left +
```

```
               Age + BMI + WBTPF, data = df)
summary(model_one)

##
## Call:
## lm(formula = VO2max_rel ~ Height + Total_Lean_Mass + tNAA_Cr_WM_Left +
##     Age + BMI + WBTPF, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.2050  -3.0257   0.0513   2.9886  17.6969
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     69.41239   12.20313   5.688 2.41e-08 ***
## Height          -0.03323    0.08206  -0.405    0.686
## Total_Lean_Mass  0.01890    0.12789   0.148    0.883
## tNAA_Cr_WM_Left  4.47716    2.81726   1.589    0.113
## Age             -0.27632    0.04997  -5.530 5.65e-08 ***
## BMI              0.06963    0.26698   0.261    0.794
## WBTPF           -0.90995    0.10606  -8.579  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.999 on 420 degrees of freedom
## Multiple R-squared:  0.6752, Adjusted R-squared:  0.6706
## F-statistic: 145.5 on 6 and 420 DF,  p-value: < 2.2e-16

crPlots(model_one)
```



```r
# This model seems pretty good, with an F-stat < 2.2*10^-16, and
# an R squared of 0.6707

# Let's add sex
model_one_sex <- lm(VO2max_rel ~ Height + Total_Lean_Mass + tNAA_Cr_WM_Left +
                    Age + BMI + WBTPF + Sex, data = df)
summary(model_one_sex)

##
## Call:
## lm(formula = VO2max_rel ~ Height + Total_Lean_Mass + tNAA_Cr_WM_Left +
##     Age + BMI + WBTPF + Sex, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.5886  -3.0401   0.0628   2.9475  17.5469
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     70.33743   12.24238   5.745 1.77e-08 ***
## Height          -0.04339    0.08275  -0.524   0.6003
## Total_Lean_Mass  0.00725    0.12848   0.056   0.9550
## tNAA_Cr_WM_Left  4.74503    2.83135   1.676   0.0945 .
## Age             -0.27287    0.05010  -5.446 8.79e-08 ***
## BMI              0.05448    0.26747   0.204   0.8387
## WBTPF           -0.88234    0.10991  -8.027 1.01e-14 ***
## SexM             0.89372    0.93218   0.959   0.3382
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.999 on 419 degrees of freedom
## Multiple R-squared:  0.6759, Adjusted R-squared:  0.6705
## F-statistic: 124.8 on 7 and 419 DF,  p-value: < 2.2e-16

# Not much better

# Stepwise regression ----------------------------------------------------

# Let's see if stepwise regression can identify any other relevant variables

model_all <- lm(VO2max_rel ~ ., data = df)
summary(model_all)

##
## Call:
## lm(formula = VO2max_rel ~ ., data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.5269  -3.0750   0.3206   2.8089  17.0433
##
```

```
## Coefficients: (3 not defined because of singularities)
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -129.49437  510.02933  -0.254  0.79971
## Age                  -0.27666    0.05285  -5.235 2.67e-07 ***
## SexM                  0.91437    0.95648   0.956  0.33967
## Height                0.12353    0.10930   1.130  0.25905
## Mass                 -0.53426    0.21473  -2.488  0.01325 *
## Total_Lean_Mass       0.47187    0.23195   2.034  0.04257 *
## BMI                   0.65312    0.36564   1.786  0.07482 .
## WBTPF                -0.52611    0.18445  -2.852  0.00457 **
## tNAA_Cr_Ant_Left     -1.22857    9.57313  -0.128  0.89795
## GM_Ant_Left         -79.85174  542.50310  -0.147  0.88306
## WM_Ant_Left         -49.11803  539.14108  -0.091  0.92746
## CSF_Ant_Left       -127.31086  531.22886  -0.240  0.81072
## tNAA_Cr_Ant_Right    -5.56968    9.05352  -0.615  0.53878
## GM_Ant_Right        -78.01120   61.60892  -1.266  0.20617
## WM_Ant_Right        -94.65128   61.81495  -1.531  0.12651
## CSF_Ant_Right              NA         NA      NA       NA
## tNAA_Cr_Pstr_Left     8.98227    6.21606   1.445  0.14924
## GM_Pstr_Left        287.77893  952.57855   0.302  0.76273
## WM_Pstr_Left        252.20296  948.74622   0.266  0.79051
## CSF_Pstr_Left       294.14239  940.30070   0.313  0.75458
## tNAA_Cr_Pstr_Right   -8.80963    5.67812  -1.552  0.12158
## GM_Pstr_Right      -181.74907  400.50190  -0.454  0.65022
## WM_Pstr_Right      -156.68162  397.12779  -0.395  0.69340
## CSF_Pstr_Right     -181.24157  388.93681  -0.466  0.64148
## tNAA_Cr_WM_Left      10.77339    5.72086   1.883  0.06041 .
## GM_WM_Left           49.27870  163.19569   0.302  0.76284
## WM_WM_Left           50.61236  161.79653   0.313  0.75459
## CSF_WM_Left                NA         NA      NA       NA
## tNAA_Cr_WM_Right     -4.59074    5.90676  -0.777  0.43750
## GM_WM_Right         159.63978  132.55023   1.204  0.22916
## WM_WM_Right         168.04618  128.86239   1.304  0.19296
## CSF_WM_Right               NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.988 on 398 degrees of freedom
## Multiple R-squared:  0.6935, Adjusted R-squared:  0.6719
## F-statistic: 32.16 on 28 and 398 DF,  p-value: < 2.2e-16

# R^2=0.6729

# We'll do forward selection first
ols_step_forward_p(model_all, penter = 0.05)

##
##                             Selection Summary

## ------------------------------------------------------------------------
```

```
##         Variable                   Adj.

## Step    Entered     R-Square    R-Square     C(p)         AIC         RMSE

## ----------------------------------------------------------------------------
##    1    WBTPF         0.6471      0.6462    35.3172    2620.4509    5.1801

##    2    Age           0.6720      0.6705     4.9061    2591.1336    4.9995

## ----------------------------------------------------------------------------

# The only variables entered were WBTPF and Age
model_forward <- lm(VO2max_rel ~ WBTPF + Age, data = df)
summary(model_forward)

##
## Call:
## lm(formula = VO2max_rel ~ WBTPF + Age, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.2057  -3.2876   0.1243   3.0059  18.3147
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 73.45994    1.41009   52.10  < 2e-16 ***
## WBTPF       -0.89142    0.03201  -27.84  < 2e-16 ***
## Age         -0.27847    0.04902   -5.68  2.5e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.999 on 424 degrees of freedom
## Multiple R-squared:  0.672,  Adjusted R-squared:  0.6705
## F-statistic: 434.4 on 2 and 424 DF,  p-value: < 2.2e-16

# R^2=0.6705

# Backward selection----
ols_step_backward_p(model_all, prem = 0.05)

##
##
##                                Elimination Summary

## ----------------------------------------------------------------------------
----------
##         Variable                           Adj.

## Step         Removed          R-Square    R-Square     C(p)         AIC
     RMSE
## ----------------------------------------------------------------------------
```

```
----------
##    1    tNAA_Cr_Ant_Left        0.6935      0.6728    27.0165    2618.2148
    4.9821
##    2    WM_Pstr_Left            0.6934      0.6735    25.0837    2616.2870
    4.9763
##    3    WM_Pstr_Right           0.6934      0.6743    23.1837    2614.3942
    4.9707
##    4    tNAA_Cr_Ant_Right       0.6927      0.6743    22.0600    2613.3329
    4.9700
##    5    tNAA_Cr_WM_Right        0.6921      0.6746    20.7942    2612.1177
    4.9684
##    6    CSF_Pstr_Right          0.6914      0.6746    19.7643    2611.1526
    4.9683
##    7    tNAA_Cr_Pstr_Left       0.6908      0.6748    18.5103    2609.9468
    4.9667
##    8    GM_Pstr_Right           0.6903      0.6751    17.1087    2608.5826
    4.9643
##    9    Sex                     0.6895       0.675    16.2675    2607.8114
    4.9654
##   10    Height                  0.6883      0.6746    15.7008    2607.3264
    4.9681
##   11    tNAA_Cr_Pstr_Right       0.687       0.674    15.3932    2607.1083
    4.9724
##   12    tNAA_Cr_WM_Left         0.6857      0.6735    15.0823    2606.8793
    4.9766
##   13    BMI                     0.6843      0.6728    14.9041    2606.7813
    4.9816
##   14    GM_Pstr_Left            0.6829      0.6722    14.7426    2606.6923
    4.9867
##   15    CSF_Pstr_Left           0.6814      0.6714    14.6910    2606.7081
    4.9924
## ------------------------------------------------------------------------
----------
```

```
# Age, mass, lean mass, wbtpf, gm_ant_right, wm, csf, were kept
model_backward <- lm(VO2max_rel ~ Age + Mass + Total_Lean_Mass +
                       WBTPF + GM_Ant_Left + WM_Ant_Left +
                       CSF_Ant_Left + GM_Ant_Right + WM_Ant_Right +
                       CSF_Ant_Right + GM_WM_Left + WM_WM_Left +
                       CSF_WM_Left + GM_WM_Right + WM_WM_Right +
                       CSF_WM_Right, data = df)
summary(model_backward)
```

```
##
## Call:
## lm(formula = VO2max_rel ~ Age + Mass + Total_Lean_Mass + WBTPF +
##     GM_Ant_Left + WM_Ant_Left + CSF_Ant_Left + GM_Ant_Right +
##     WM_Ant_Right + CSF_Ant_Right + GM_WM_Left + WM_WM_Left +
##     CSF_WM_Left + GM_WM_Right + WM_WM_Right + CSF_WM_Right, data = df)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.0792  -2.9775   0.2452   2.8483  17.5881
##
## Coefficients: (3 not defined because of singularities)
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       99.24224  187.25980   0.530   0.5964
## Age               -0.28432    0.05148  -5.523 5.91e-08 ***
## Mass              -0.31898    0.15781  -2.021   0.0439 *
## Total_Lean_Mass    0.46857    0.22470   2.085   0.0377 *
## WBTPF             -0.53781    0.17532  -3.068   0.0023 **
## GM_Ant_Left     -135.12238  182.27983  -0.741   0.4589
## WM_Ant_Left     -120.76312  180.95386  -0.667   0.5049
## CSF_Ant_Left    -193.63296  168.17280  -1.151   0.2502
## GM_Ant_Right     -83.02012   59.46468  -1.396   0.1634
## WM_Ant_Right    -100.24632   58.98009  -1.700   0.0899 .
## CSF_Ant_Right           NA         NA      NA       NA
## GM_WM_Left        60.92778  151.54992   0.402   0.6879
## WM_WM_Left        67.73545  149.58251   0.453   0.6509
## CSF_WM_Left             NA         NA      NA       NA
## GM_WM_Right      114.93072  124.05404   0.926   0.3547
## WM_WM_Right      116.54302  119.46202   0.976   0.3299
## CSF_WM_Right            NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.992 on 413 degrees of freedom
## Multiple R-squared:  0.6814, Adjusted R-squared:  0.6714
## F-statistic: 67.95 on 13 and 413 DF,  p-value: < 2.2e-16

# R^2=0.6714

# Stepwise selection
ols_step_both_p(model_all, penter=0.05, prem=0.1)

##
##                                 Stepwise Selection Summary

## ----------------------------------------------------------------------------
----------------
##                          Added/                   Adj.

## Step      Variable      Removed     R-Square    R-Square     C(p)
AIC        RMSE
## ----------------------------------------------------------------------------
----------------
##    1       WBTPF        addition       0.647       0.646    35.3170    262
0.4509    5.1801
##    2         Age        addition       0.672       0.670     4.9060    259
1.1336    4.9995
```

```
##    3    CSF_WM_Right    addition       0.674       0.672     4.1300    259
0.3412    4.9890
##    4     WM_Pstr_Left    addition       0.677       0.674     2.2270    258
8.3840    4.9719
## ------------------------------------------------------------------------
---------------
```

```
# Variables kept: WBTPF, age, csf_whitematter_right, and wm_pstr_left
model_stepwise <- lm(VO2max_rel ~ WBTPF + Age + CSF_WM_Right + WM_Pstr_Left,
data = df)
summary(model_stepwise)
```

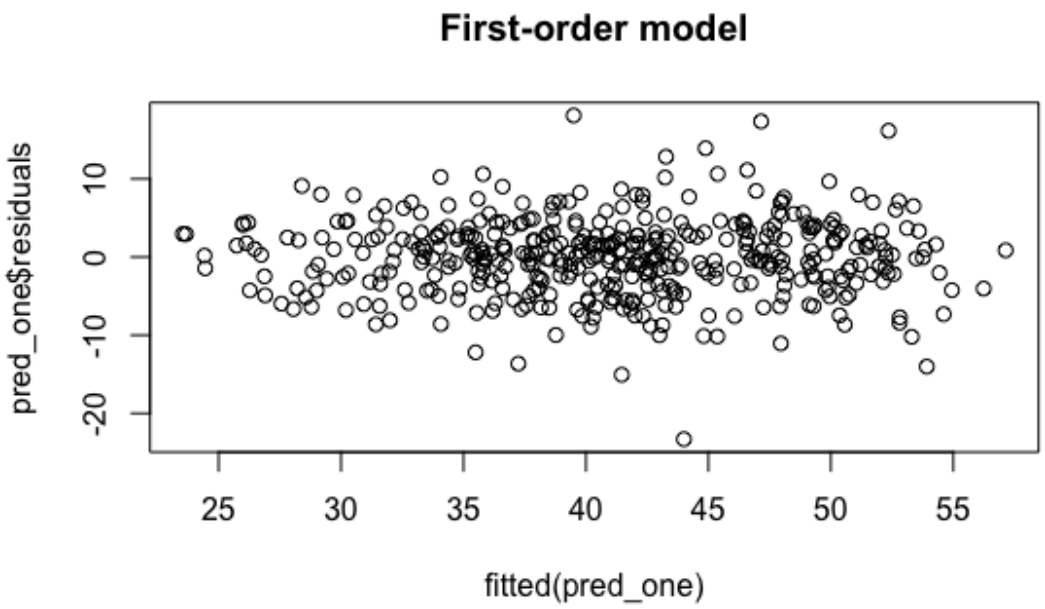
```
##
## Call:
## lm(formula = VO2max_rel ~ WBTPF + Age + CSF_WM_Right + WM_Pstr_Left,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.3008  -2.9998   0.1775   2.9288  18.1028
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   79.50440    2.81769  28.216  < 2e-16 ***
## WBTPF         -0.89357    0.03190 -28.014  < 2e-16 ***
## Age           -0.27031    0.04885  -5.533 5.53e-08 ***
## CSF_WM_Right -98.89377   45.08441  -2.194   0.0288 *
## WM_Pstr_Left -20.75242   10.46954  -1.982   0.0481 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.972 on 422 degrees of freedom
## Multiple R-squared:  0.6772, Adjusted R-squared:  0.6741
## F-statistic: 221.3 on 4 and 422 DF,  p-value: < 2.2e-16
```

```
# R^2=0.6741
```

```
# Interestingly, all three stepwise approahces yielded different results
# I'm curious about sex—let me see how VO2peak varies across sex
vo2_sex <- data.frame('male' = mean(df$VO2max_rel[df$Sex == 'M'], na.rm = TRU
E),
           'female' = mean(df$VO2max_rel[df$Sex == 'F'], na.rm = TRUE)
           )
# Is this a significant difference?
t.test(df$VO2max_rel[df$Sex == 'M'], df$VO2max_rel[df$Sex == 'F'])
```

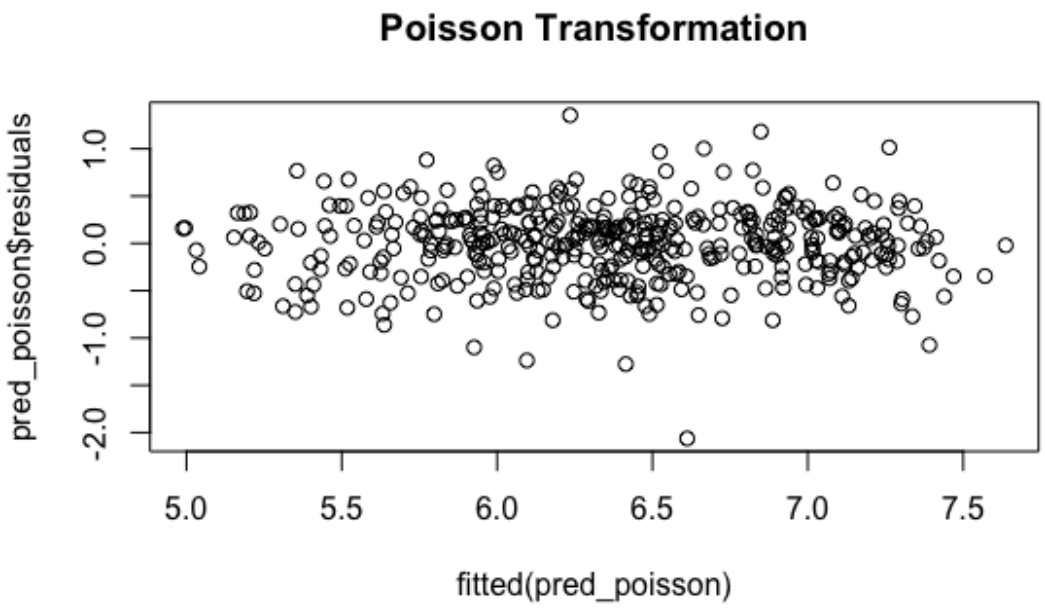
```
##
##  Welch Two Sample t-test
##
## data:  df$VO2max_rel[df$Sex == "M"] and df$VO2max_rel[df$Sex == "F"]
## t = 13.257, df = 416.07, p-value < 2.2e-16
```

```
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   7.988459 10.769949
## sample estimates:
## mean of x mean of y
##  45.61142  36.23221

# Yes, with p < 10^-16
# So I'll include sex in the analysis

model_two <- lm(VO2max_rel ~ WBTPF + Age + CSF_WM_Right + WM_Pstr_Left,
                data = df)
summary(model_two)
# R^2=0.6781

# What if you add sex
model_three <- lm(VO2max_rel ~ Height + Total_Lean_Mass + tNAA_Cr_WM_Left +
                  Age + BMI + WBTPF + Sex, data = df)
summary(model_three)
# R^2=0.6781
# A significant model, but R squared isn't improved

# Let's include this in the model and test for multicollinearity
# by calculating variance inflation factor (VIF)
vif(model_two)

##        WBTPF          Age CSF_WM_Right WM_Pstr_Left
##     1.023762     1.024152     1.105998     1.104093

# Seems like there's no multicollinearity
pred_one <- model_two

# Improving prediction model ---------------------------------------------

# Let's try to improve our prediction model with residual analysis
plot(fitted(pred_one), pred_one$residuals, main = "First-order model")
```




```
# The errors seem pretty normally distributed;
# perhaps greater at higher values

# Perhaps we can try a poisson transformation
df$poisson <- sqrt(df$VO2max_rel)
pred_poisson <- lm(poisson ~ WBTPF + Age + CSF_WM_Right + WM_Pstr_Left + Sex,
                   data = df)
summary(pred_poisson)

##
## Call:
## lm(formula = poisson ~ WBTPF + Age + CSF_WM_Right + WM_Pstr_Left +
##     Sex, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.06200 -0.23902  0.02223  0.23896  1.35441
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.380603   0.238569  39.320  < 2e-16 ***
## WBTPF        -0.068899   0.003331 -20.686  < 2e-16 ***
## Age          -0.022871   0.003886  -5.886 8.08e-09 ***
## CSF_WM_Right -7.752217   3.591063  -2.159   0.0314 *
## WM_Pstr_Left -1.675737   0.831975  -2.014   0.0446 *
## SexM          0.039341   0.050634   0.777   0.4376
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

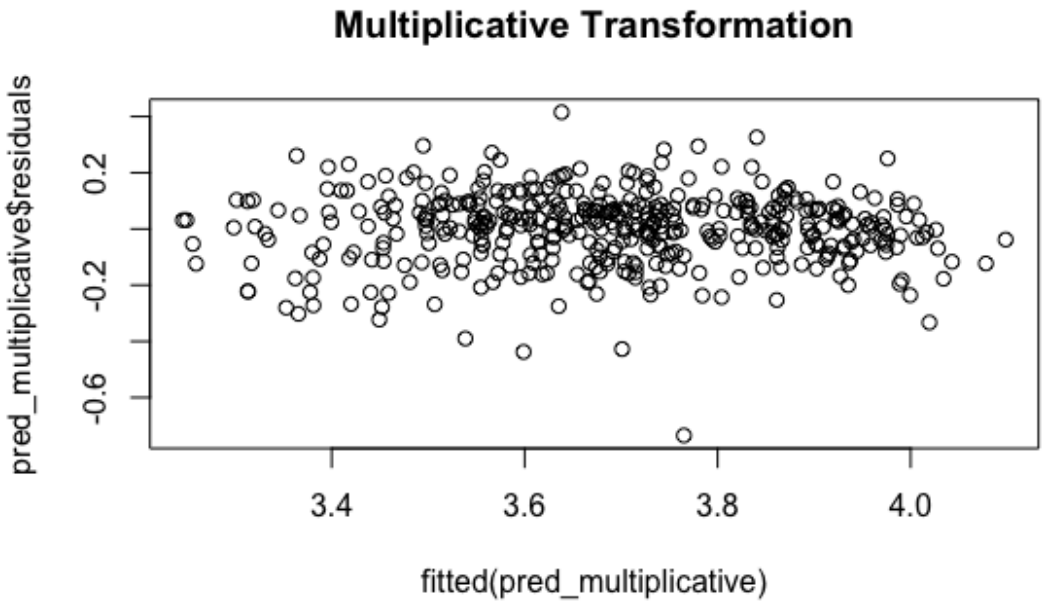
```
## Residual standard error: 0.3949 on 421 degrees of freedom
## Multiple R-squared:  0.6772, Adjusted R-squared:  0.6734
## F-statistic: 176.6 on 5 and 421 DF,  p-value: < 2.2e-16

plot(fitted(pred_poisson), pred_poisson$residuals, main = "Poisson Transforma
tion")
```



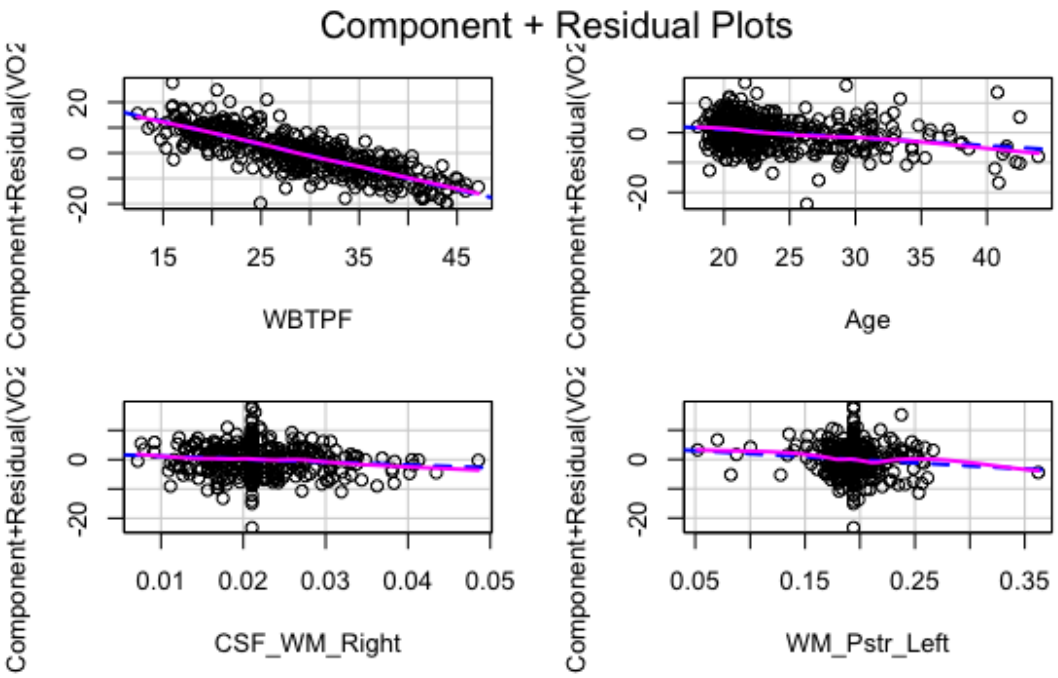

```
# This transformation doesn't improve R^2 (0.67 in both cases), nor does it
# affect p-value of F-test

# How about multiplicative transformation?
df$multiplicative <- log(df$VO2max_rel)
pred_multiplicative <- lm(multiplicative ~ WBTPF + Age + CSF_WM_Right + WM_Ps
tr_Left + Sex, data = df)
summary(pred_multiplicative)

##
## Call:
## lm(formula = multiplicative ~ WBTPF + Age + CSF_WM_Right + WM_Pstr_Left +
##     Sex, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.73495 -0.06889  0.01230  0.07889  0.41522
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.674623   0.078062  59.883  < 2e-16 ***
## WBTPF        -0.022243   0.001090 -20.410  < 2e-16 ***
## Age          -0.007741   0.001271  -6.089 2.56e-09 ***
```

```
## CSF_WM_Right -2.349912   1.175031  -2.000   0.0462 *
## WM_Pstr_Left -0.557276   0.272231  -2.047   0.0413 *
## SexM          0.006419   0.016568   0.387   0.6986
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1292 on 421 degrees of freedom
## Multiple R-squared:  0.668,  Adjusted R-squared:  0.664
## F-statistic: 169.4 on 5 and 421 DF,  p-value: < 2.2e-16
```

```
plot(fitted(pred_multiplicative), pred_multiplicative$residuals, main = "Mult
iplicative Transformation")
```

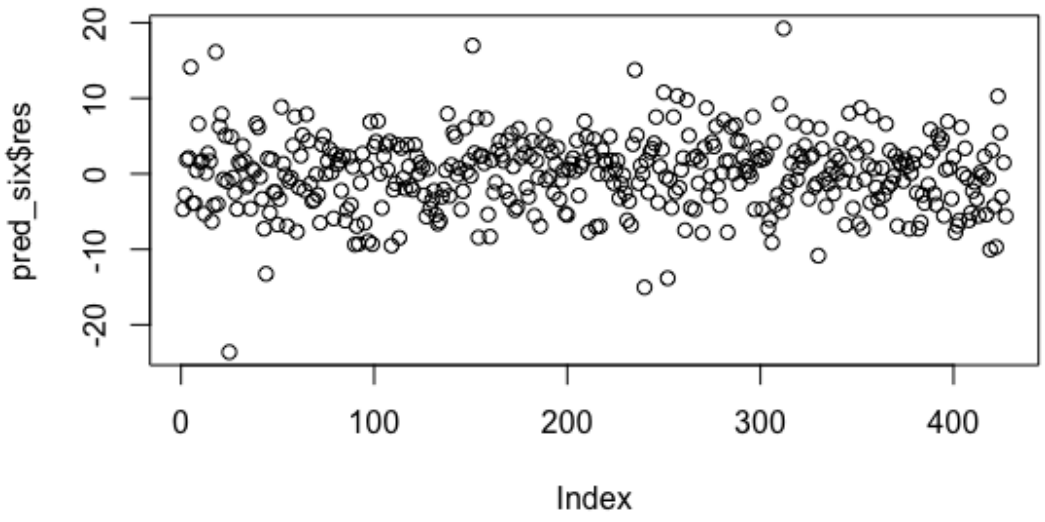



```
# This transformation doesn't improve R^2 (0.68 in both cases), nor does it
# affect p-value of F-test

# How about partial residual plots?
crPlots(pred_one)
```



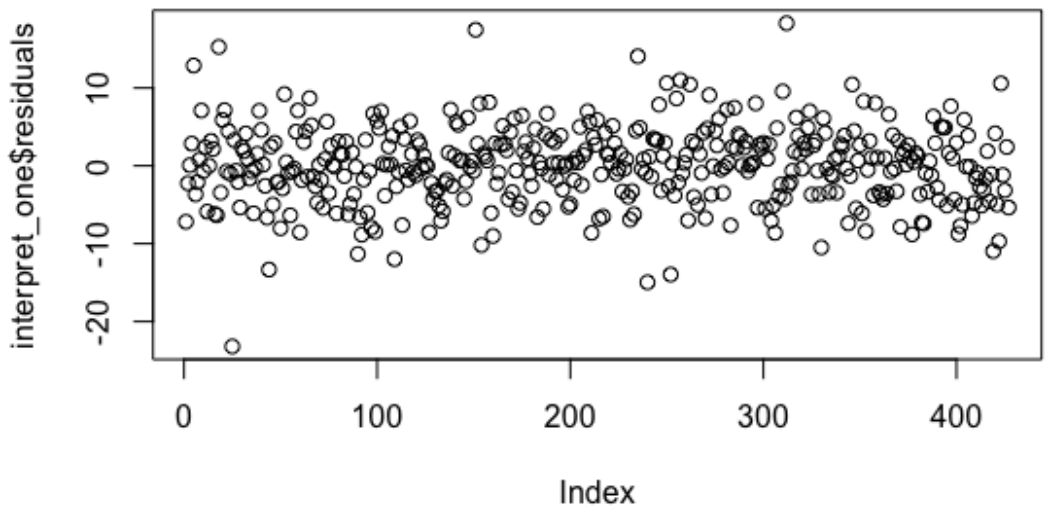

```
# The variables seem to have linear relationships

# I wonder if there is interaction between age and sex
pred_two <- lm(VO2max_rel ~ WBTPF + Age + CSF_WM_Right + WM_Pstr_Left + Sex +
                 Age*Sex, data = df)
summary(pred_two)
anova(pred_one, pred_two, test = 'F') # nested F test says this doesn't help


# WBTPF and age?
pred_three <- lm(VO2max_rel ~ WBTPF + Age + CSF_WM_Right + WM_Pstr_Left +
                   Age*WBTPF, data = df)
summary(pred_three)
anova(pred_one, pred_three, test = 'F') # nested F test
# Also no

# Sex and WBTPF?
pred_four <- lm(VO2max_rel ~ WBTPF + Age + CSF_WM_Right + WM_Pstr_Left + Sex
+
                   Sex*WBTPF, data = df)
summary(pred_four)
anova(pred_one, pred_four, test = 'F') # nested F test

# Age, sex, bmi?
pred_five <- lm(VO2max_rel ~ WBTPF + Age + CSF_WM_Right + WM_Pstr_Left + Sex
+
                   Age*WBTPF + Sex*WBTPF + Age*Sex+WBTPF, data = df)
summary(pred_five)
anova(pred_one, pred_five, test = 'F') # nested F test
```

```
# Also no, but we're closer—p = 0.068

# WBTPF^2?
pred_six <- lm(VO2max_rel ~ WBTPF + Age + CSF_WM_Right + WM_Pstr_Left + Sex +
                   Age*WBTPF + Sex*WBTPF + Age*Sex+WBTPF + I(WBTPF^2), data =
 df)
summary(pred_six)
anova(pred_one, pred_six, test = 'F') # nested F test
# Close; p = 0.0502

# WBTPF^2?
pred_seven <- lm(VO2max_rel ~ WBTPF + Age + CSF_WM_Right + WM_Pstr_Left + Sex
 +
                   Age*WBTPF + Sex*WBTPF + Age*Sex+WBTPF + I(WBTPF^2) +
                   CSF_WM_Right*WBTPF + WM_Pstr_Left*WBTPF +
                   CSF_WM_Right* WM_Pstr_Left, data = df)
summary(pred_seven)
anova(pred_one, pred_seven, test = 'F') # nested F test
# Not better, p=0.14

# How are residuals on pred_six
plot(pred_six$res)
```

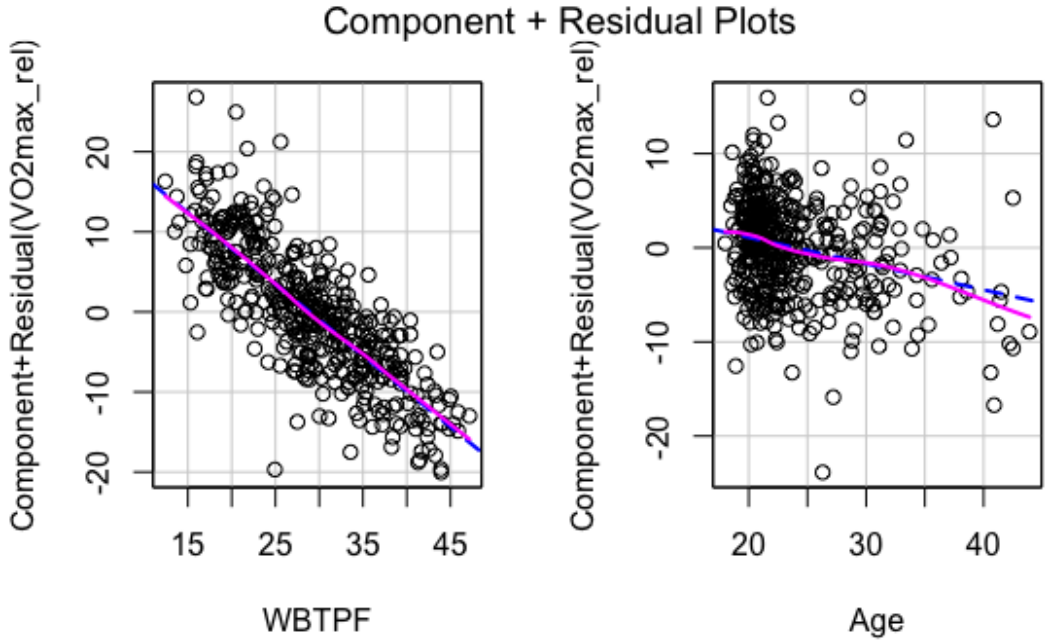



```
# Seem prety good

# Interpretation model ------------------------------------------------
interpret_one <- model_forward
# This forward model was very simple
```

```
# Let's look at its residuals
plot(interpret_one$residuals)
```




```
crPlots(interpret_one)
```




```
# These also seem uniformly distributed

summary(interpret_one)
```

```
##
## Call:
## lm(formula = VO2max_rel ~ WBTPF + Age, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.2057  -3.2876   0.1243   3.0059  18.3147
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 73.45994    1.41009   52.10  < 2e-16 ***
## WBTPF       -0.89142    0.03201  -27.84  < 2e-16 ***
## Age         -0.27847    0.04902   -5.68  2.5e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.999 on 424 degrees of freedom
## Multiple R-squared:  0.672,  Adjusted R-squared:  0.6705
## F-statistic: 434.4 on 2 and 424 DF,  p-value: < 2.2e-16
```

```
# This model is simple with a high R^2. I'll see if adding any interaction/
# quadratic terms improves it but it is good as is

interpret_two <- lm(VO2max_rel ~ WBTPF + Age + Sex, data = df)
summary(interpret_two)
```

```
##
## Call:
## lm(formula = VO2max_rel ~ WBTPF + Age + Sex, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.4086  -3.1835   0.1202   2.9459  18.1316
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 72.36730    1.76303  41.047  < 2e-16 ***
## WBTPF       -0.86310    0.04216 -20.474  < 2e-16 ***
## Age         -0.28111    0.04909  -5.727 1.94e-08 ***
## SexM         0.65890    0.63826   1.032    0.303
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.999 on 423 degrees of freedom
## Multiple R-squared:  0.6728, Adjusted R-squared:  0.6705
## F-statistic:   290 on 3 and 423 DF,  p-value: < 2.2e-16
```

```
anova(interpret_one, interpret_two, test = 'F')
```

```
## Analysis of Variance Table
##
```

```
## Model 1: VO2max_rel ~ WBTPF + Age
## Model 2: VO2max_rel ~ WBTPF + Age + Sex
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1    424 10598
## 2    423 10571  1    26.633 1.0657 0.3025

# Sex isn't worth the addition

# Body fat and age interaction
interpret_three <- lm(VO2max_rel ~ WBTPF + Age + WBTPF*Age, data = df)
summary(interpret_three)
anova(interpret_one, interpret_three, test = 'F')

# Body fat^2
interpret_four <- lm(VO2max_rel ~ WBTPF + Age + I(WBTPF^2), data = df)
summary(interpret_four)
anova(interpret_one, interpret_four, test = 'F')

# Body fat^2
interpret_five <- lm(VO2max_rel ~ WBTPF + Age + I(Age^2), data = df)
summary(interpret_five)
anova(interpret_one, interpret_five, test = 'F')

# None of these transformations were worth it. VO2 ~ WBTPF + Age is the best
model

confint(interpret_one)

##                 2.5 %     97.5 %
## (Intercept) 70.688292 76.2315798
## WBTPF       -0.954349 -0.8284986
## Age         -0.374827 -0.1821090
```