BodyFat

Jillian Maw

3/5/2022

HOMEWORK ANALYSIS #4 - BODY FAT

Measuring body fat is not simple. One method requires submerging the body underwater in a tank and measuring the increase in water level. A simpler method for estimating body fat would be preferred. In order to develop such a method, researchers recorded age, weight, height, and 10 body circumference measurements for 252 men. Each man's percentage of body fat was accurately estimated by an under-water weighing technique. The data can be found in the `BodyFat` dataset (the variable `brozek` is the percentage of body fat).

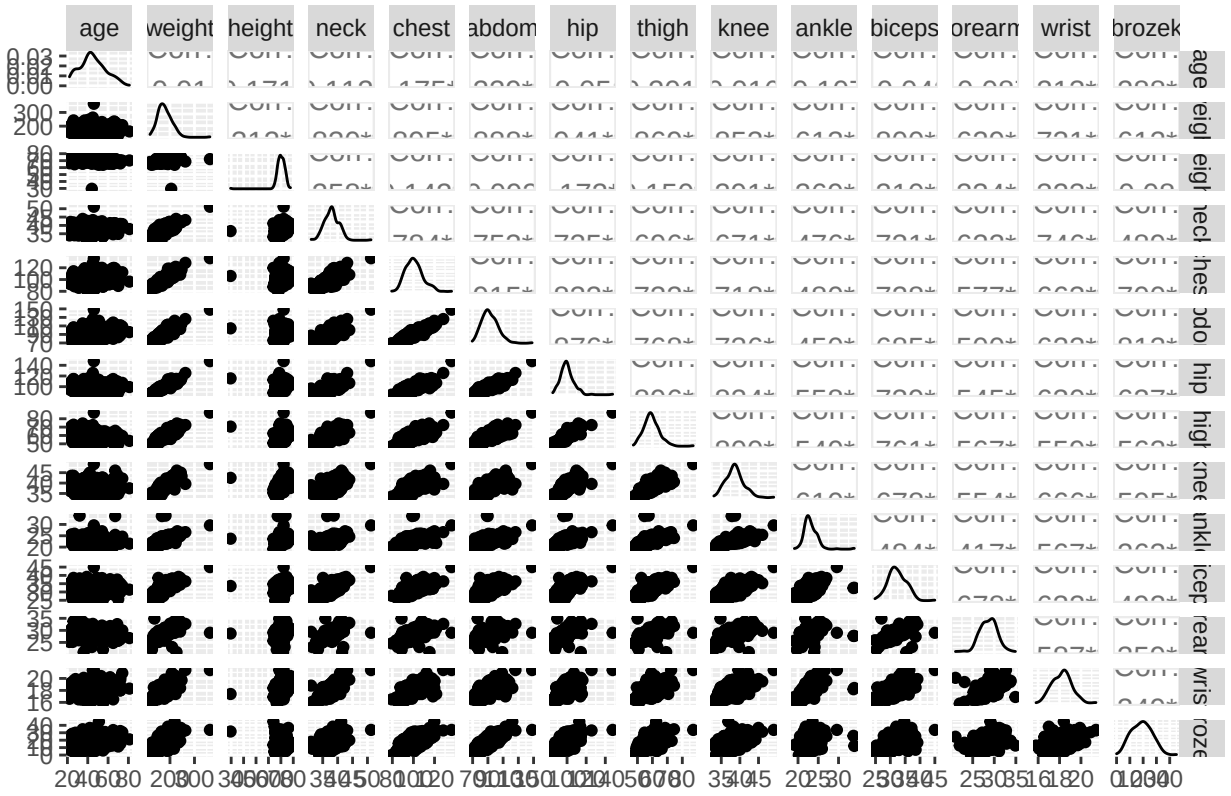
For each of the following questions, assume that your audience are nurses with moderate statistical training. Please attach your clearly commented code (R or Python) to the back of your answers as an appendix.

1. In your own words, summarize the overarching problem and any specific questions that need to be answered using the `BodyFat` data. Discuss how statistical modeling will be able to answer the posed questions.

Measuring body fat usually requires the unwieldy process of submerging oneself underwater in a large tank of water, and measuring the water displacement. To simplify the process of measuring body fat, researchers recorded age, weight, height, and 10 body circumference measurements to determine if a multiple linear regression model can accurately predict a man's percentage of body fat. We will be using statistical modeling to create this multiple linear regression model to determine the linear relationship that could accurately capture this prediction. If this prediction can be proven, it can assist health care providers by providing estimates for a man's percentage of body fat, to help determine health care needs.

2. Explore the data using basic exploratory graphics and summary statistics. Comment on any potential relationships you see through this exploratory analysis. Calculate variance inflation factors and discuss what variables, if any, are collinear. Comment on what effect collinearity can have on a regression analysis.

Below are graphs of the measurement variables in comparison to each other and two tables, one of covariance and one of correlations between each variable. The scatterplot matrix shows that several measurements have an outlier or two that could skew the data. The covariance matrix shows several measurements have positive or negative relationships. The correlation matrix show that there are several measurements that have high correlations with each other, and others that do not. Surprisingly, the percentage of body fat, our response variable, only has a high correlation with one measurement, abdomen, and 3 medium-high correlations with weight, chest, and hip measurements; the rest are low correlations. With the number of measurements we have, which are our explanatory variables, and with the number that have high correlations with each other, is is likely our explanatory variables are collinear with each other. Collinearity is when two (or more) explanatory variables are highly correlated with each other, which inflates the standard errors in statistical modeling and makes significance harder to determine. With inflated standard errors and difficult to interpret significance we would be hard pressed to create an accurate prediction model for body fat percentage.




```
## [1] "Covariance Matrix for Percentage of Body Fat"
```

```
##                age     weight    height      neck      chest      abdom
## age     159.341578  -5.234126 -7.890056  3.445724  18.534120  31.115987
## weight   -5.234126 864.849835 33.622835 59.372359 221.083680 281.033291
## height   -7.890056  33.622835 13.430361  2.296011   4.358799   3.678768
## neck      3.445724  59.372359  2.296011  5.913399  16.024631  19.707660
## chest    18.534120 221.083680  4.358799 16.024631  70.572805  82.715251
## abdom    31.115987 281.033291  3.678768 19.707660  82.715251 115.769524
## hip      -4.615559 198.640520  4.522004 12.827626  50.148704  67.630207
## thigh   -13.330217 134.385186  2.889758  8.897333  32.326150  43.454430
## knee      0.482945  60.472538  2.568863  3.935752  14.546943  19.091603
## ankle    -2.285742  30.536394  1.670345  1.962693   6.839234   8.216576
## biceps   -1.611006  71.178769  2.332086  5.375151  18.512010  22.291037
## forearm  -2.222577  37.355199  1.728658  3.055037   9.791885  10.864927
## wrist     2.520810  20.096337  1.106624  1.696495   5.212692   6.260377
## brozek   28.104537 139.035028 -2.383557  9.187899  45.418820  67.508375
##                hip      thigh      knee      ankle    biceps   forearm
## age      -4.615559 -13.330217  0.482945 -2.2857418 -1.611006 -2.222577
## weight  198.640520 134.385186 60.472538 30.5363939 71.178769 37.355199
## height    4.522004   2.889758  2.568863  1.6703454  2.332086  1.728658
## neck     12.827626   8.897333  3.935752  1.9626931  5.375151  3.055037
## chest    50.148704  32.326150 14.546943  6.8392341 18.512010  9.791885
## abdom    67.630207  43.454430 19.091603  8.2165761 22.291037 10.864927
## hip      51.501969  33.829951 14.258971  6.7905777 16.046210  7.897026
```

```
## thigh    33.829951  27.657837 10.140369  4.8102336 12.112626  6.019647
## knee     14.258971  10.140369  5.814777  2.4940019  4.946936  2.696470
## ankle     6.790578   4.810234  2.494002  2.8739041  2.481025  1.425938
## biceps   16.046210  12.112626  4.946936  2.4810245  9.151209  4.140216
## forearm   7.897026   6.019647  2.696470  1.4259383  4.140216  4.077609
## wrist     4.230325   2.748687  1.501498  0.8990379  1.789602  1.108464
## brozek   34.753740  22.836903  9.405709  3.4393970 11.501034  5.596469
##              wrist     brozek
## age      2.5208096  28.104537
## weight  20.0963368 139.035028
## height   1.1066239  -2.383557
## neck     1.6964953   9.187899
## chest    5.2126916  45.418820
## abdom    6.2603769  67.508375
## hip      4.2303251  34.753740
## thigh    2.7486872  22.836903
## knee     1.5014978   9.405709
## ankle    0.8990379   3.439397
## biceps   1.7896024  11.501034
## forearm  1.1084644   5.596469
## wrist    0.8750473   2.521638
## brozek   2.5216384  59.702404
```

```
## [1] "Correlation Matrix for Percentage of Body Fat"
```

```
##                 age      weight      height      neck     chest      abdom
## age      1.00000000 -0.01409968 -0.17055784 0.1122529 0.1747788 0.22909832
## weight  -0.01409968  1.00000000  0.31197525 0.8302241 0.8948858 0.88815812
## height  -0.17055784  0.31197525  1.00000000 0.2576389 0.1415808 0.09329566
## neck     0.11225285  0.83022415  0.25763893 1.0000000 0.7844238 0.75321566
## chest    0.17477878  0.89488585  0.14158078 0.7844238 1.0000000 0.91510258
## abdom    0.22909832  0.88815812  0.09329566 0.7532157 0.9151026 1.00000000
## hip     -0.05095042  0.94120792  0.17193930 0.7350482 0.8318195 0.87585391
## thigh   -0.20079992  0.86890427  0.14993691 0.6957163 0.7316884 0.76794095
## knee     0.01586597  0.85274916  0.29069026 0.6711855 0.7181026 0.73583209
## ankle   -0.10681358  0.61250729  0.26886008 0.4760994 0.4802337 0.45046144
## biceps  -0.04218842  0.80009377  0.21035920 0.7306895 0.7284431 0.68484757
## forearm -0.08719457  0.62903950  0.23359439 0.6221500 0.5772246 0.50006553
## wrist    0.21348124  0.73051775  0.32280514 0.7457930 0.6633268 0.62199599
## brozek   0.28814824  0.61186832 -0.08417553 0.4889920 0.6997151 0.81201607
##                hip      thigh       knee      ankle      biceps     forearm
## age     -0.05095042 -0.2007999 0.01586597 -0.1068136 -0.04218842 -0.08719457
## weight   0.94120792  0.8689043 0.85274916  0.6125073  0.80009377  0.62903950
## height   0.17193930  0.1499369 0.29069026  0.2688601  0.21035920  0.23359439
## neck     0.73504816  0.6957163 0.67118555  0.4760994  0.73068949  0.62214999
## chest    0.83181952  0.7316884 0.71810261  0.4802337  0.72844313  0.57722458
## abdom    0.87585391  0.7679410 0.73583209  0.4504614  0.68484757  0.50006553
## hip      1.00000000  0.8963556 0.82396680  0.5581600  0.73913086  0.54494020
## thigh    0.89635562  1.0000000 0.79960977  0.5395366  0.76135984  0.56683826
## knee     0.82396680  0.7996098 1.00000000  0.6100903  0.67815746  0.55376593
## ankle    0.55815995  0.5395366 0.61009028  1.0000000  0.48378868  0.41654490
## biceps   0.73913086  0.7613598 0.67815746  0.4837887  1.00000000  0.67776785
## forearm  0.54494020  0.5668383 0.55376593  0.4165449  0.67776785  1.00000000
## wrist    0.63015307  0.5587280 0.66564457  0.5669257  0.63241411  0.58681767
```

```
## brozek   0.62674910  0.5619947 0.50481133  0.2625731  0.49204153  0.35868684
##              wrist      brozek
## age      0.2134812  0.28814824
## weight   0.7305178  0.61186832
## height   0.3228051 -0.08417553
## neck     0.7457930  0.48899197
## chest    0.6633268  0.69971510
## abdom    0.6219960  0.81201607
## hip      0.6301531  0.62674910
## thigh    0.5587280  0.56199465
## knee     0.6656446  0.50481133
## ankle    0.5669257  0.26257312
## biceps   0.6324141  0.49204153
## forearm  0.5868177  0.35868684
## wrist    1.0000000  0.34887603
## brozek   0.3488760  1.00000000
```

3. Use a variable selection technique to determine a MLR model that will answer the questions posed in #1. Justify your choice of your selected variable selection procedure (e.g., state why you chose to use backward instead of forward selection). Justify your choice of a model comparison criterion (e.g. state why you chose to base your variable selection procedure on AIC vs. BIC).

We decided to use the "best subset selection" variable selection procedure because it is the best method for minimizing Aikake Information Criteria, Bayesian Information Criteria, or Predictive Error, and maximizes the Adjusted R^2 for our multiple linear regression model. In addition, the "best subset selection" procedure tests all possible variables with each other, allowing us to know with certainty that we have the best model. We decided to use the Aikake Information Criteria, or AIC, model comparison criterion, because we are looking to make predictions, and the AIC model comparison criterion is optimized for making predictions, compared to the Bayesian Information Criteria, or BIC, which is optimized for making inferences.

4. Write out (in mathematical form with Greek letters) a MLR model that would help answer the questions you stated in #1 using the variables you identified in #3. Provide an interpretation of the intercept and at least 1 slope coefficient included in your model.

Below is the multiple linear regression model written in mathematical form with Greek letters:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7} + \beta_8 x_{i8},\ \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

Above, β_0 represents the intercept, or when all the other measurements are zero, then y is β_0, on average; this is a next-to-useless interpretation, however, and it is better to think of β_0 as the average base percentage of body fat for men. Next, we clarify the representations of the coefficients of this model. β_1 represents the age measurement, β_2 represents the weight measurement, β_3 represents the neck measurement, β_4 represents the abdomen measurement, β_5 represents the hip measurement, β_6 represents the thigh measurement, β_7 represents the forearm measurement, and β_8 represents the wrist measurement. To interpret one of the coefficients of this model, we state that "Holding all other body circumference measurements constant, then as β_8, the wrist measurement, goes up by 1 unit of measurement, then the y_i, or percentage of body fat, goes up by β_8 units of measurement, on average. β_8 is the effect of that wrist measurement on percentage of body fat.

5. Fit your model in #3 to the `BodyFat` data and summarize the results by displaying estimated coefficients in a table (do NOT just provide a screen shot of the R or Python output). Interpret at least 1 of the coefficients (not the intercept) in the context of the problem.

Below is our fitted model, mathematically written in Greek letters, for the `Bodyfat` data, adjusting for the fact this model contains our best estimates for predicting percentage of body fat:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \hat{\beta}_4 x_{i4} + \hat{\beta}_5 x_{i5} + \hat{\beta}_6 x_{i6} + \hat{\beta}_7 x_{i7} + \hat{\beta}_8 x_{i8},\ \epsilon_i \overset{iid}{\sim} N(0, \hat{\sigma}^2)$$

Written with numeric coefficients, the fitted model above becomes: $\hat{y}_i$ = -20.111 + 0.0593 x_{i1}+ -0.0842 x_{i2}+ -0.432 x_{i3}+ 0.8766 x_{i4}+ -0.1855 x_{i5}+ 0.2867 x_{i6}+ 0.4817 x_{i7}+ -1.4026 x_{i8}, $\epsilon_i \overset{iid}{\sim} N(0, \hat{\sigma}^2)$, where $\hat{\sigma}^2$ = 15.7846.

Below is a table of the estimated coefficients in the multiple linear regression model, for easy reading:

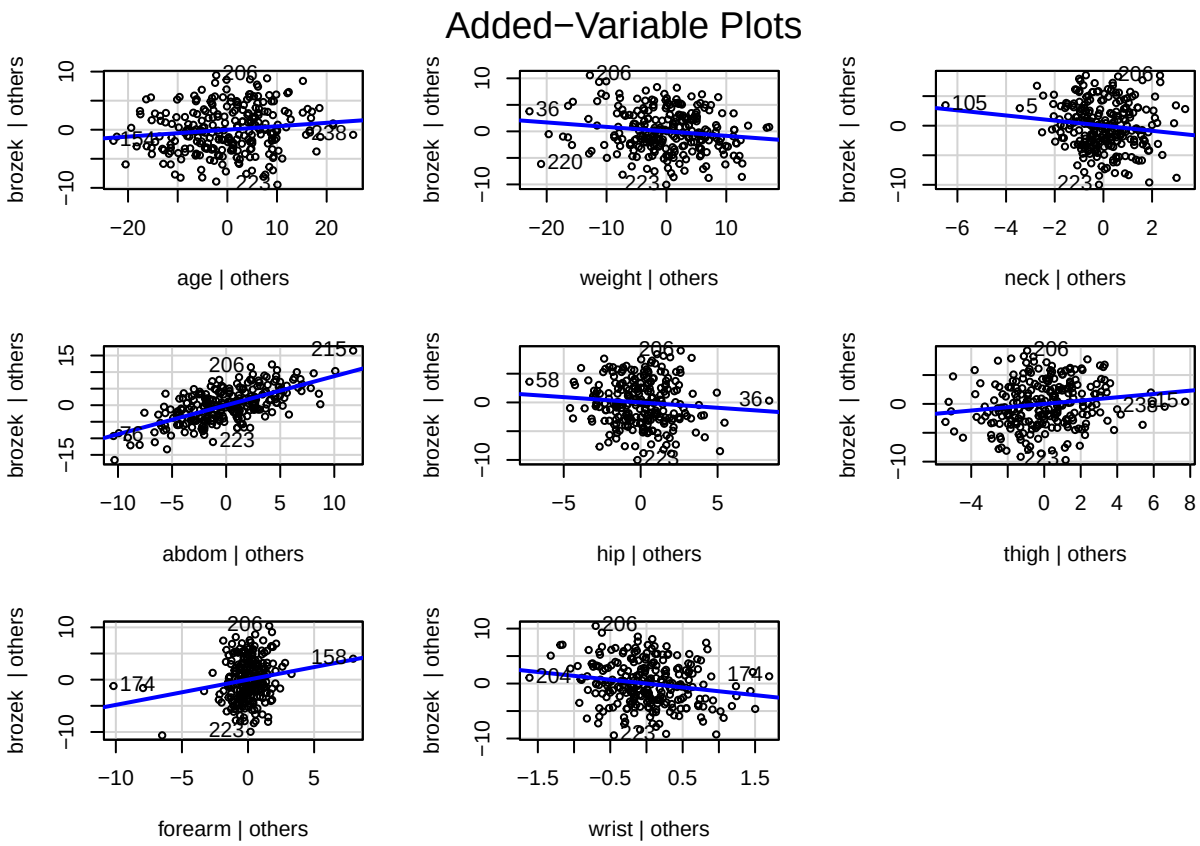
Coefficient Model's Greek Letter Representation	Coefficient Name	Estimated Value
$\hat{\beta}_0$	Intercept	-20.111
$\hat{\beta}_1$	Age	0.0593
$\hat{\beta}_2$	Weight	-0.0842
$\hat{\beta}_3$	Neck	-0.432
$\hat{\beta}_4$	Abdomen	0.8766
$\hat{\beta}_5$	Hip	-0.1855
$\hat{\beta}_6$	Thigh	0.2867
$\hat{\beta}_7$	Forearm	0.4817
$\hat{\beta}_8$	Wrist	-1.4026
With random errors of $\epsilon_i \overset{iid}{\sim} N(0, \hat{\sigma}^2)$, where $\hat{\sigma}^2$ = 15.7846		

To interpret an estimated coefficient, we can say something like "holding all else constant, if a hip measurement increases by 1 unit in the model, the percentage of bodyfat goes up by -0.1855, on average." We were not asked to do a confidence interval on the coefficients, so I will not be doing it for this homework/report. I will add that $\hat{\sigma}^2$ represents the unbiased, estimated variance of the model, or the average distance of the percentage of body fat from the multiple linear regression model's "line" of estimation.

6. List your model assumptions, then justify them using appropriate graphics or summary statistics.

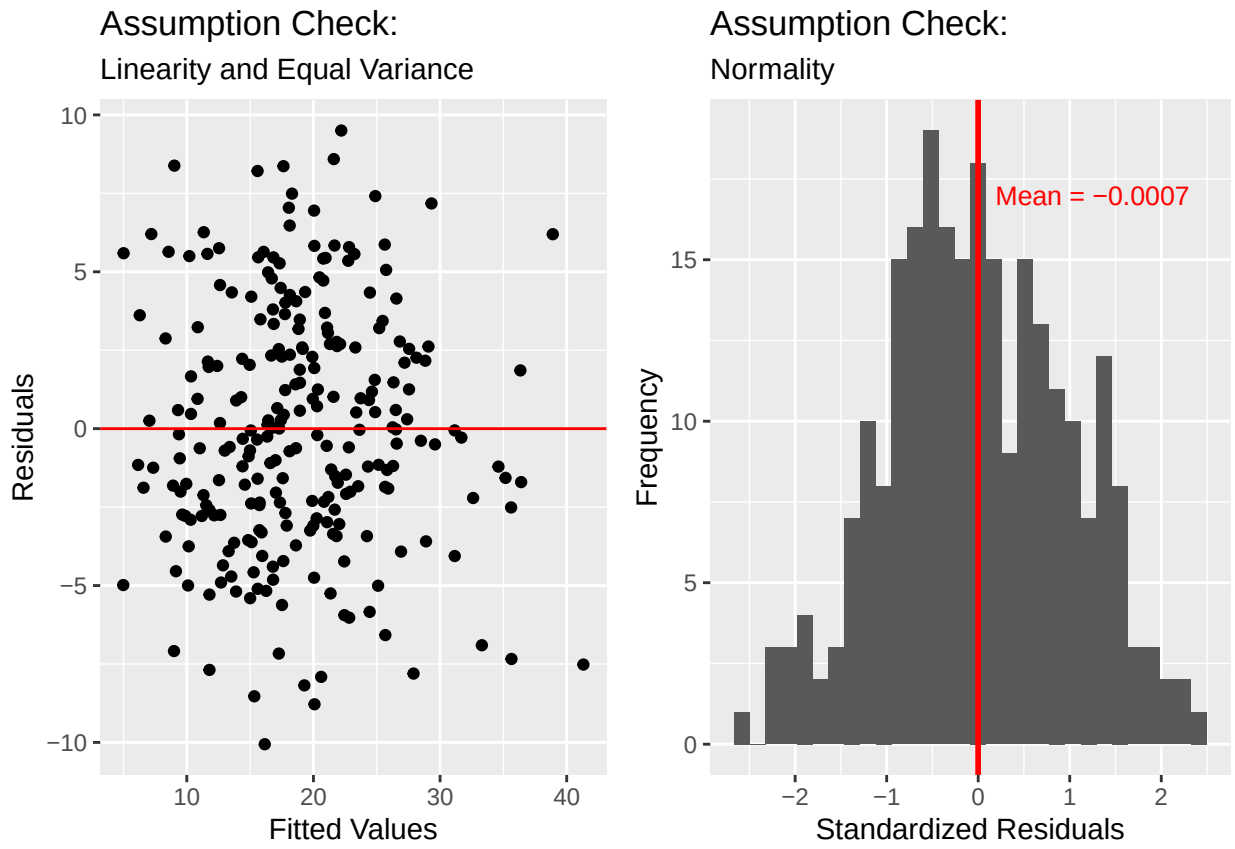
For this multiple linear regression model, we assumed that the regressions are linear, independent, all follow a Normal distribution, and that the data has equal variance about the regression lines. We will prove this with the graphics and summary statistics below.

First, we will look at the Added-Variable Plots below. The graphs for neck, hip, and forearm seem to show they have outliers affecting the slope, in that they pull the slope through a cluster of points in one direction over another. However, there are no patterns in the plots showing that other graphical line would be appropriate (e.g. exponential), so we can assume the linearity assumption is met.



We can assume the data is independent because one man's measurements should not affect another man's measurements. While a man may have measurements for himself that are similar in ratios, due to the average size restraints of any human being, that does not affect another man's measurements. We can safely assume that even if there are men who are related to each other in the study, the mother's genetics may influence a difference in sizes to allow for a reasonable assumption of independence between all mens' sizes.

Finally, for our last graphics, we check the normality and equal variance assumptions. We will prove the equal variance assumption by checking a plot comparing fitted values (predicted value for body fat percentage in the dataset) to the residuals (the difference between the observed and a predicted body fat percentages) in a scatterplot. Since we can see a constant variance with a lack of patterns, the data has equal variance and that assumption seems to be proven correct. Next, we will prove the data follows a Normal distribution by plotting the standardized residuals (residuals transformed to show their difference from the data's mean, if the mean was 0) on a histogram. Below, we see the histogram shows a generally Normal distribution, so we assume the normality assumption is also met.



To further prove we meet the Normality assumption for the model, we conducted a One-sample Kolmogorov-Smirnov test, also called a KS-test, and a Jarque-Bera test for normality, also called a JB-test. These tests conduct hypothesis tests on whether or not a data set follows a Normal distribution or not. For the KS-test, the null hypothesis is that the data comes from a Normal distribution, while the alternative hypothesis is that the data does *not* come from a Normal distribution. For the JB-test, the null hypothesis is that the data's distribution is not skewed, whereas the alternative hypothesis is that the data's distribution *is* skewed. We set the p-value to be 0.05 for both tests, to prove significance. The KS-test produced a p-value of 0.872, so we failed to reject the null hypothesis. The JB-test for normality produced a p-value of 0.253, so we failed to reject the null hypothesis for both tests. We accept that we have the normality assumption met for our data.

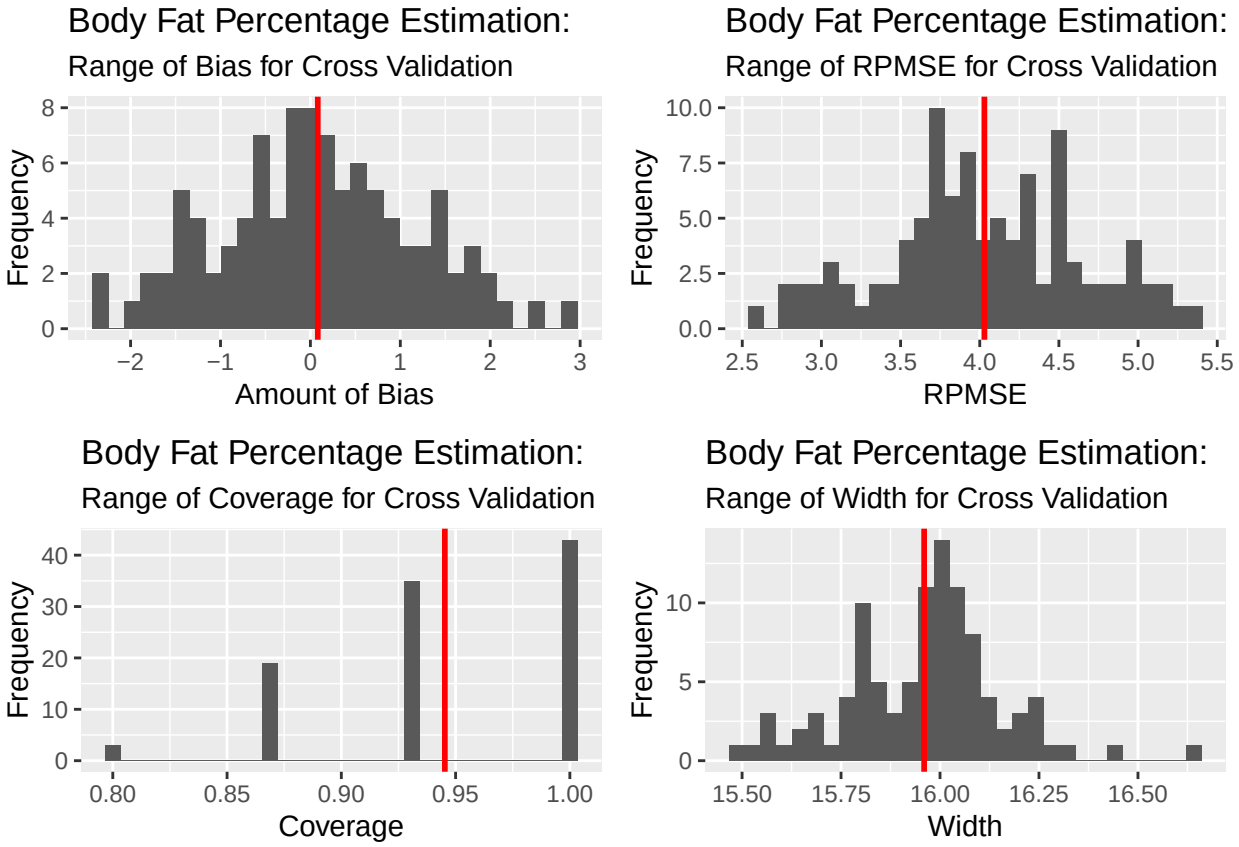
To further prove the Equal Variance assumption for our model, we conducted a Breusch-Pagan test, or BP-test. Checking the scatterplot of fitted values versus residuals above, it appears that we have a model with linearity and mostly equal variance, so we can proceed with the BP-test. The BP-test conducts hypothesis tests on whether or not a data set has homoskedasticity, or equal variance. The null hypothesis is that the data has homoskedasticity, while the alternative hypothesis is that the data has *heteroskedasticity*. We set the p-value to be 0.05, to prove significance. The test produced a p-value of 0.1903, so we failed to reject the null hypothesis. We accept that we have the Equal Variance assumption met for our data.

7. Because prediction is so important, nurses want to know how accurate your predictions are. Carry out an appropriate study that will evaluate the ability of your model to perform predictions. Display appropriate numerical summaries and interpret these summaries on the level of your target audience. Draw conclusions about how "good" your predictions are relative to the original spread of the response variable. Separately, please discuss model fit by reporting and interpreting R^2.

We conducted a cross validation procedure 100 times on 15 randomly selected observations of the data, newly selected each time, and calculated an average bias of 0.0821. This means our predictions are, on average,

slightly higher than the true average percentage of body fat. We also calculated an average Root Predictive Mean Square Error of 4.0293, which means our predictions are off, on average, 4.0293 percentage of body fat. Considering the range of percentage of body fat is between 0% and 45.1%, this seems a reasonable amount of error. To see how far our predictions ranged, we calculated the width to be 15.9599 percentage of body fat, on average. In addition, the coverage, or the percentage of prediction intervals that contain the true average percentage of body fat, to be 0.9453. Below the following paragraph are graphs showing in greater detail the results of the cross validation procedures, with the red lines representing the mean values relative to the results shown.

In addition to the above cross validation results (and below graphical representation of the results), the percent of variability in percentage of body fat explained by the covariate variables in the adjusted multiple linear regression model listed above (i.e. age, weight, and 6 selected body circumference measurements) is R^2, 74.41. With all of these results and summary statistics, we feel it safe to say that our predictions are rather good, relative to the original spread of the response variable percentage of body fat.



8. Nurses wish to make a prediction of percentage of body fat for the following person: `age`= 50, `weight`= 203, `height`= 67, `neck`= 40.2, `chest`=114.8, `abdom`=108.1, `hip`=102.5, `thigh`=61.3, `knee`= 41.1, `ankle`= 24.7, `biceps`= 34.1, `forearm`= 31, `wrist`= 18.3. Describe how you would use your fitted model in #5 to make a prediction for this person. What is the predicted percentage body fat for this patient?

To determine the predicted percentage body fat for this patient (with the measurements stated above), we input the measurements we use in the multiple linear regression model and determine we are 95% confident that if the measurements above are as stated, the associated percentage of body fat would be between 22.997 percent and 38.9485 percent, on average. To be more specific, 22.997 percent is our estimated answer for this patient, but could easily be in the previously stated range.

Appendix of Code

```
knitr::opts_chunk$set(echo = FALSE, include = FALSE)
library(ggplot2)
library(GGally)
library(MASS)
library(normtest)
library(lmtest)
library(SciViews)
library(car)
library(bestglm)
library(knitr)
library(kableExtra)
library(gridExtra)
options(scipen = 999)
bodyfat = read.table("~/R programming/STAT_330/BodyFatData.txt",
                     sep = ' ', header = TRUE)
#reorder dataset to use bestglm
bodyfat <- bodyfat[,c(2:ncol(bodyfat),1)]
bodyfat.mlr <- lm(brozek~age+weight+height+neck+chest+abdom+hip+thigh+knee+ankle+
                    biceps+forearm+wrist, data=bodyfat)
vif(bodyfat.mlr) #weight, abdom, and hip all above 10, chest *very* close to 10
ggpairs(bodyfat, progress=FALSE, title = 'Scatterplot Matrix for Bodyfat Data')
#should I print tables with names?
# kable(cov(bodyfat), "html", digits = 4, booktabs = TRUE, row.names = TRUE,
#       caption = 'Covariance Matrix') %>%
#   kable_styling(latex_options = "striped") %>%
#   landscape()
# kable(cor(bodyfat), "html", digits = 4, booktabs = TRUE, row.names = TRUE,
#       caption = 'Correlation Matrix') %>%
#   kable_styling(latex_options = "striped") %>%
#   landscape()
print("Covariance Matrix for Percentage of Body Fat")
cov(bodyfat)
print("Correlation Matrix for Percentage of Body Fat")
cor(bodyfat)
#do I need 87,178,291,200 instances to find a match in criterion values?
vs.res1 <- bestglm(bodyfat, IC="AIC", method="exhaustive", TopModels=15)
#vs.res12 <- bestglm(bodyfat, IC="AIC", method="forward", TopModels=30)
#vs.res13 <- bestglm(bodyfat, IC="AIC", method="backward", TopModels=30)

vs.res2 <- bestglm(bodyfat, IC="BIC", method="exhaustive", TopModels=15)
#vs.res21 <- bestglm(bodyfat, IC="BIC", method="exhaustive", TopModels=30)
#vs.res23 <- bestglm(bodyfat, IC="BIC", method="backward", TopModels=30)

#vs.res3 <- bestglm(bodyfat, IC="CV", method="backward", t=100)
vs.res3 <- bestglm(bodyfat, IC="CV", method="exhaustive", t=100)
#vs.res32 <- bestglm(bodyfat, IC="CV", method="forward", t=100)

#vs.res1$BestModels[1,]
#vs.res2$BestModels[1,]
vs.res1$BestModels #wrist forearm thigh hip abdom neck weight age for 699.3530
vs.res2$BestModels #wrist forearm abdom weight for 716.6614
```

```r
#use exhaustive method because it is best and we can
#use AIC for prediction of variables and BIC if we WERE doing inference
#new mlr model:
bodyfat.adj.mlr <- lm(brozek~age+weight+neck+abdom+hip+thigh+forearm+wrist, data=bodyfat)
#vif(bodyfat.adj.mlr) #says weight and hip are both collinear, don't include in homework
#create table of coefficient estimates with kable
bodyfat.beta.0 <- round(as.numeric(coef(bodyfat.adj.mlr)["(Intercept)"]), digits = 4)
bodyfat.beta.1 <- round(as.numeric(coef(bodyfat.adj.mlr)["age"]), digits = 4)
bodyfat.beta.2 <- round(as.numeric(coef(bodyfat.adj.mlr)["weight"]), digits = 4)
bodyfat.beta.3 <- round(as.numeric(coef(bodyfat.adj.mlr)["neck"]), digits = 4)
bodyfat.beta.4 <- round(as.numeric(coef(bodyfat.adj.mlr)["abdom"]), digits = 4)
bodyfat.beta.5 <- round(as.numeric(coef(bodyfat.adj.mlr)["hip"]), digits = 4)
bodyfat.beta.6 <- round(as.numeric(coef(bodyfat.adj.mlr)["thigh"]), digits = 4)
bodyfat.beta.7 <- round(as.numeric(coef(bodyfat.adj.mlr)["forearm"]), digits = 4)
bodyfat.beta.8 <- round(as.numeric(coef(bodyfat.adj.mlr)["wrist"]), digits = 4)
bodyfat.var <- round(sigma(bodyfat.adj.mlr)^2, digits = 4)
#Linearity
avPlots(bodyfat.adj.mlr, ask=FALSE) #only for linearity
#Independence (idea of) & Equal Variance
#just think about it
fit.vs.resids.1 <- ggplot(bodyfat, aes(x=bodyfat.adj.mlr$fitted.values,
                                       y=bodyfat.adj.mlr$residuals)) +
  geom_point() +
  xlab('Fitted Values') +
  ylab('Residuals') +
  ggtitle('Assumption Check:',
          subtitle ='Linearity and Equal Variance') +
  geom_hline(yintercept = 0, col = "red", lwd = 0.5)

#Normality
std.resids <- stdres(bodyfat.adj.mlr)
hline <- round(mean(std.resids), digits = 4)
bodyfat.freq <- ggplot() +
  geom_histogram(mapping=aes(x=std.resids)) +
  xlab('Standardized Residuals') +
  ylab('Frequency') +
  ggtitle('Assumption Check:', subtitle = 'Normality') +
  geom_vline(xintercept = mean(std.resids), col = "red", lwd = 1) +
  annotate("text", x = hline + 1.25, y = 17,
           label = paste("Mean =", hline), col = "red", size = 3.5)
#suppressMessages(print(bodyfat.freq))
#suppressMessages(print(fit.vs.resids.1))
suppressMessages(grid.arrange(fit.vs.resids.1, bodyfat.freq, nrow=1))
ks.test(std.resids, "pnorm")
jb.norm.test(std.resids)
bptest(bodyfat.adj.mlr)
#cross validation
set.seed(87) #set seed for reproducibility
n.cv <- 100 #Number of CV studies we'll run
bias <- rep(NA, n.cv) #n.cv empty biases (one for each CV)
RPMSE <- rep(NA, n.cv) #n.cv empty RPMSE (one for each CV)
coverage <- rep(NA, n.cv) #n.cv empty coverage (one for each CV)
width <- rep(NA, n.cv) #n.cv empty width (one for each CV)
```

```
n.test <- 15 #How big my test set is
for(i in 1:n.cv){
  test.obs <- sample(1:nrow(bodyfat), n.test)
  test.set <- bodyfat[test.obs,]
  train.set <- bodyfat[-test.obs,]

  train.lm <- lm(brozek~age+weight+neck+abdom+hip+thigh+forearm+wrist, data=train.set)
  test.preds <- predict.lm(train.lm, newdata=test.set, interval="prediction")

  bias[i] <- mean(test.preds[,1]- test.set$brozek)
  RPMSE[i] <- sqrt(mean((test.preds[,1] - test.set$brozek)^2))
  coverage[i] <- mean((test.preds[,2] < test.set$brozek) &
                        (test.preds[,3] > test.set$brozek))
  width[i] <- mean(test.preds[,3] - test.preds[,2])
}

bodyfat.stddev <- round(sigma(bodyfat.adj.mlr), digits = 4)
mean.bias <- round(mean(bias), digits = 4)
mean.RPMSE <- round(mean(RPMSE), digits = 4)
mean.coverage <- round(mean(coverage), digits = 4)
mean.width <- round(mean(width), digits = 4)
bodyfat.r2 <- round(summary(bodyfat.adj.mlr)$r.squared, digits = 4)
CV.bias <- ggplot() +
  geom_histogram(mapping=aes(x=bias)) +
  xlab('Amount of Bias') +
  ylab('Frequency') +
  ggtitle('Body Fat Percentage Estimation:',
          subtitle = 'Range of Bias for Cross Validation') +
  geom_vline(xintercept = mean.bias, col = "red", lwd = 1) #+
  # annotate("text", x = 1.1, y = 7,
  #          label = paste("Mean=", mean.bias), col = "red", size = 3.5)
#suppressMessages(print(CV.bias))

CV.RPMSE <- ggplot() +
  geom_histogram(mapping=aes(x=RPMSE)) +
  xlab('RPMSE') +
  ylab('Frequency') +
  ggtitle('Body Fat Percentage Estimation:',
          subtitle = 'Range of RPMSE for Cross Validation') +
  geom_vline(xintercept = mean.RPMSE, col = "red", lwd = 1) #+
  # annotate("text", x = 5.05, y = 7.5,
  #          label = paste("Mean=", mean.RPMSE), col = "red", size = 3.5)
#suppressMessages(print(CV.RPMSE))

CV.coverage <- ggplot() +
  geom_histogram(mapping=aes(x=coverage)) +
  xlab('Coverage') +
  ylab('Frequency') +
  ggtitle('Body Fat Percentage Estimation:',
          subtitle = 'Range of Coverage for Cross Validation') +
  geom_vline(xintercept = mean.coverage, col = "red", lwd = 1)# +
  # annotate("text", x = mean.coverage - 0.05, y = 40,
  #          label = paste("Mean=", mean.coverage), col = "red", size = 3.5)
```

```
#suppressMessages(print(CV.coverage))

CV.width <- ggplot() +
  geom_histogram(mapping=aes(x=width)) +
  xlab('Width') +
  ylab('Frequency') +
  ggtitle('Body Fat Percentage Estimation:',
          subtitle = 'Range of Width for Cross Validation') +
  geom_vline(xintercept = mean.width, col = "red", lwd = 1)# +
  # annotate("text", x = 16.3, y = 10,
  #          label = paste("Mean=", mean.width), col = "red", size = 3.5)
#suppressMessages(print(CV.width))
suppressMessages(grid.arrange(CV.bias, CV.RPMSE, CV.coverage, CV.width, nrow=2))
#prediction
bodyfat.predict <- data.frame(age=50, weight=203, neck=40.2, abdom=108.1, hip=102.5,
                              thigh=61.3, forearm=31, wrist= 18.3)
predict.range <- predict.lm(bodyfat.adj.mlr, newdata=bodyfat.predict,
                            interval="prediction", level=0.95)
#End of homework's code
```