Problem Motivation:

A large amount of health professionals recommend that people assess their health, in part, by estimating their percentage of body fat. However, precise measurements of one’s body fat are not easy to come by; such measurements usually require inconvenient and costly techniques. There are simpler techniques for measurement, though they require a good deal of estimation and are not as accurate. It would therefore be of interest to see if percentage of body fat can be predicted using some type of linear regression model.

Main Assumptions:

* It has been shown that percentage of body fat for an individual can be estimated once body density is determined, using the equation D = 1/[(A/a) + (B/b)] where D = Body Density (gm/cm^3), A = proportion of lean body tissue, B = proportion of fat tissue, a = density of lean body tissue (gm/cm^3), b = density of fat tissue (gm/cm^3).
* Taking a=1.10 gm/cm^3 and b=0.90 gm/cm^3 (see Katch and McArdle (1977), p. 111 or Wilmore (1976), p. 123) we derive "Siri's equation": 100B = 495/D – 450.
* It has been shown that the method of underwater weighing accurately measures body density, enabling one to accurately compute percent body fat.
* The underwater weighing technique states Body Density = WA/[(WA-WW)/c.f. - LV], where WA = Weight in air (kg), WW = Weight in water (kg), c.f. = Water correction factor (=1 at 39.2 deg F as one-gram of water occupies exactly one cm^3 at this temperature, =.997 at 76-78 deg F), LV = Residual Lung Volume (liters).
* Finding body density through this technique and using Siri’s equation to solve for percentage body fat enables us to generate data that we can analyze to identify some linear relationship between body fat percentage and other variables.
* We assume that we can use linear regression to derive a model to predict percentage of body fat
* We assume that the relationships between the explanatory variables and the response variable are (roughly) linear, or can be made linear via the use of transformations or additional terms

Summary of Findings:

Through this analysis, I have found that it is possible to accurately measure percentage of body fat using the linear regression model(s) derived through my statistical analysis. Furthermore, the optimal model(s) have been shown to fit new data with similar accuracy.

Outliers in the data set used did not seem to be overly influential, as many were removed through pre-processing and the residuals vs leverage plots for the models indicated that no outliers were overly influential. Therefore, I did not feel the need to apply any further analysis. If this statistical analysis were to be continued, I would conduct a more in-depth analysis of potential outliers.

Finally, though the models predict percentage of body with significant accuracy, they failed the Shapiro-Wilks test at the 5% significance level, indicating that the residuals are not normally distributed. This indicates that, for future work, a transformation of variables may be useful to make these models more robust.

Development of Statistical Analysis:

* I removed the column for Density from the dataset, as this variable is used to compute the values of the response variable using Siri’s equation. It is not a variable that should be included in the regression model.
* I identified two problematic outliers in the dataset, as the measured values differed significantly when compared to the rest of the data points. I removed these two points as they would have influenced the model negatively.
* I removed those data points in which a given explanatory variable fell more than 5 standard deviations away from the mean of the given explanatory variable. Assuming the explanatory variables are normally distributed, this ensures that all points fall within the parametrized normal distribution.
* To get an idea of the relationships between the explanatory variables and the response variable for the data, I ran the data set through pairs.R, which lists the relationships between variables using plots and the correlations between variables (SEE APPENDIX I). The output from pairs.R showed that most of the relationships between the explanatory variables and the response variable were linear, though some slightly deviated and can be said to be somewhat curvilinear. I concluded that this may need to be remedied via the transformation of variables and / or the addition of terms (interaction or polynomial). The output also showed that there were high levels of correlations between variables. I concluded that this would inevitably cause the regression model to be uneasily interpreted, which would need to be remedied.
* To conduct cross validation, I split the data set into two separate data sets: the dataset labeled “train” consists of 80% of the data points from the original data set, while the dataset labeled “test” consists of the remaining 20%. All data points were randomly chosen.
* I constructed a naïve model using the train data set, where the response variable is the calculated percentage of body fat and the rest of the variables in the data set are included as explanatory variables.
* I plotted the diagnostic plots for the naïve model, namely the Residuals vs. Fitted plot, the Normal Q-Q plot, the Scale-Location plot and the Residuals vs. Leverage plot. The purpose of the residuals vs. fitted plot and the scale-location plot are to get a visual idea as to whether the residuals are homoscedastic (have constant variance). The purpose of the normal Q-Q plot is to get a visual idea of whether the residuals are normally distributed. Finally, the purpose of the residuals vs. leverage is to visually identify those points with high leverage that are also influential**.** By looking at these plots, I deemed it safe to assume that the naïve model has normally distributed residuals with constant variance and that there are no highly influential data points present.
* To support the assumption that the model’s residuals have constant variance and normally distributed residuals, I conducted the Breusch-Pagan test and the Shapiro-Wilks test on the model. The Breusch-Pagan test has a null hypothesis which states that the residuals have constant variance, while the alternative hypothesis states they do not. The Shaprio-Wilks test has a null hypothesis that states the residuals are normally distributed, while the alternative hypothesis states they are not. The calculated p-values for the test were higher than our levels of significance for each test, therefore we failed to reject the null hypotheses for both tests.
* Looking at the summary of the naïve model, the majority of variables failed to be significant in the model. However, the global F test for the overall significance of the model concluded that the model with the chosen variables significantly fits the data. This seemed to be a contradiction, as using many insignificant variables to produce a model should produce an insignificant model, though this was not the case. This is due to high levels of correlation between the variables in the data set, which was also indicated by the many negative coefficients for the variables in the model. This can be remedied via the two-way anova test, wherein the reduced model (the model with less explanatory variables) is compared to the full model (the model with more explanatory variables). By removing insignificant variables from the model and analyzing the effect using the two-way anova test, we can derive a simpler and more easily interpretable model.
* Due to the variables in the dataset being measured quantities, all of the variables are numeric. Therefore, generating box plots and contemplating whether to include interaction effects in the model was an unncecessary step for this statistical analysis.
* Before, we noted that we may need to transform the response variable in order to attain a model that fits the data best. A good way to check the best transformation for the y variable is by using the box cox transformation. Using the function boxcox() in R, we can plot the optimal range for the term lambda, where lambda represents the power to put your response variable to in your model (when lambda = 0, you take the logarithm of your response variable in your model). When generating this plot using boxcox(), it was seen that the optimal value of lambda for the response variable was very close to 1. This indicated that we should keep our response variable the way it is. Furthermore, one of the goals of the box cox transformation is to achieve normality and constant variance of residuals in a model, though we have already achieved this. All evidence indicated we should not change the response variable.
* When we had previously analyzed the output of pairs.R on our data set, we saw that some of the explanatory variables had a somewhat curvilinear relationship with our response variable. Using the training data set, I picked out those two variables which were the most curvilinear to analyze more finely, those two variables being Neck Circ. And Forearm Circ (SEE APPENDIX III). When analyzing these two variables vs. percentage of body fat, it is seen that they have what seems to be a cubic relationship with the data, that cubic relationship being stretched across the x-axis. Therefore, I chose to construct a new model wherein I added cubic polynomial terms multiplied by 0.01 (to account to for the stretching). Upon plotting diagnostics for this model and running similar tests as was done for the naïve model, I concluded that this model did not significantly improve model performance. Therefore, I chose not to include the polynomial terms in the model.
* The process of model selection was then conducted using the train data set, while the test data set was used to see how the derived optimal model(s) performed on unseen data. Cross validation was done to prevent the model from overfitting the data.
* To derive an optimal, I enlisted the use of stepwise elimination algorithms. The first stepwise algorithm I enlisted was the backwards elimination algorithm, while the second was a hybrid of the forward and backward stepwise algorithms.
* For the process of cross validation, four optimal models were chosen based upon the metrics of optimal Mallow’s CP, optimal Bayes Information Criterion (BIC), optimal Akaike Information Criterion (AIC), and optimal adjusted R-squared respectively. These models were then tested using the test dataset.

Variable Descriptions:

* Fat\_Percent: The percentage of body fat as determined through Siri’s equation presented above
* Density: Body density computed through the underwater weighing technique
* Age: Subject’s age (in years)
* Weight: Subject’s weight (in pounds)
* Height: Subject’s height (in inches)
* Neck Circ.: Subject’s measured neck circumference (in centimeters)
* Chest Circ.: Subject’s measured chest circumference (in centimeters)
* Abdomen 2 Circ.: Subject’s measured abdomen circumference (in centimeters)
* Hip Circ.: Subject’s measured hip circumference (in centimeters)
* Thigh Circ.: Subject’s measured thigh circumference (in centimeters)
* Knee Circ.: Subject’s measured knee circumference (in centimeters)
* Ankle Circ.: Subject’s measured ankle circumference (in centimeters)
* Bicep Circ.: Subject’s measured bicep circumference (in centimeters)
* Forearm Circ.: Subject’s measured forearm circumference (in centimeters)
* Wrist Circ.: Subject’s measured wrist circumference (in centimeters)

Assessment of Models:

**Naïve Model Metrics:**

Call:

lm(formula = Fat\_Percent ~ ., data = train)

Residuals:

Min 1Q Median 3Q Max

-0.020529 -0.006811 0.000115 0.006552 0.034360

Coefficients:

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

(Intercept) 1.114e+00 5.979e-02 18.624 < 2e-16 \*\*\*

Age -1.285e-04 8.642e-05 -1.487 0.13875

Weight 1.228e-04 1.697e-04 0.724 0.47006

Height 5.506e-04 5.003e-04 1.100 0.27259

`Neck Circ.` 4.459e-04 6.101e-04 0.731 0.46579

`Chest Circ.` 3.468e-04 2.832e-04 1.224 0.22236

`Abdomen 2 Circ.` -2.237e-03 2.482e-04 -9.013 2.58e-16 \*\*\*

`Hip Circ.` 4.437e-04 3.743e-04 1.185 0.23744

`Thigh Circ.` -3.090e-04 3.954e-04 -0.781 0.43555

`Knee Circ.` -1.323e-04 6.559e-04 -0.202 0.84042

`Ankle Circ.` -7.412e-04 9.478e-04 -0.782 0.43522

`Bicep Circ.` -6.607e-04 4.491e-04 -1.471 0.14296

`Forearm Circ.` -6.291e-04 5.117e-04 -1.230 0.22044

`Wrist Circ.` 4.276e-03 1.448e-03 2.953 0.00356 \*\*

---

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

Residual standard error: 0.01002 on 184 degrees of freedom

Multiple R-squared: 0.736, Adjusted R-squared: 0.7173

F-statistic: 39.45 on 13 and 184 DF, p-value: < 2.2e-16

* We can see that we have negative coefficients for 7 / 13 of the explanatory variables

The residual standard error (the amount of error that can be explained by the regression model) is relatively low

* The t-test for the significance of each of the variables in the model states that most of the variables are insignificant in the model
* The global F-test for the significance of the model states that the model fits the data well
* The multiple R-squared and adjusted R-squared are relatively high, indicating that there is a relatively strong positive correlation present between the explanatory variables and the response variable

**Naïve Model Diagnostic Plots:**

* The residuals vs. Fitted plot is used to determine if the residuals exhibit a non-linear pattern; the red line across the center of the plot is nearly horizontal, therefore it is reasonable to assume that the residuals follow a linear pattern.
* The Normal Q-Q plot is used to determine if the residuals of the regression model are normally distributed; if the points fall roughly on the diagonal line, then we can assume they are normally distributed. Some of the points at both ends of the line start to deviate a bit from the diagonal, though most lie nearly on it so it's safe to assume they are normally distributed.
* Scale-Location plot: Used to check the assumption of homoscedasticity / equal variance among the residuals of the regression model. If the red line is roughly horizontal across the plot, then the assumption is likely met. The red line in our plot is nearly horizontal, therefore it is safe to assume constant variance.
* Residuals vs. Leverage plot: Used to identify influential observations. If any points in this plot fall outside of Cook's distance (the dashed line(s)), then the point(s) are influential. In our plot, none of the points cross and therefore there aren't any overly influential points in the data set.

Diagram

Description automatically generated

**Naïve Model Tests:**

Shapiro-Wilk normality test

data: naive\_model$residuals

W = 0.98885, p-value = 0.1249

studentized Breusch-Pagan test

data: naive\_model

BP = 13.025, df = 13, p-value = 0.4459

* We can see here that we fail to reject the null hypothesis for each of these tests, indicating that the residuals for the naïve model have homoscedasticity (constant variance) and are normally distributed

**Naïve Model BoxCox Transformation Plot:**

**Chart, line chart, scatter chart

Description automatically generated**

* Here we can see that the optimal choice of lambda is nearly 1, indicating that we should transform our response variable by raising it to the power of 1. This is the default power that our response variable is raised it, so we should simply leave it as it is

**Naïve Model Training Dataset Predictions:**

**Chart, line chart, scatter chart

Description automatically generated**

**Naïve Model Testing Dataset Predictions:**

**Chart, scatter chart

Description automatically generated**

* Above are two plots of the naïve model’s predictive performance on the train and test datasets respectively. In addition, the red bands indicate the upper and lower bounds of the 95% prediction intervals. We can see that, for both the train and test datasets, the model seems to predict the percentage of body fat with significant accuracy, though some points fall outside of the confidence interval.

**Polynomial Model Metrics:**

Call:

lm(formula = Fat\_Percent ~ . - `Neck Circ.` + poly(0.01 \* `Neck Circ.`,

3) - `Forearm Circ.` + poly(0.01 \* `Forearm Circ.`, 3), data = train)

Residuals:

Min 1Q Median 3Q Max

-0.019871 -0.006882 -0.000080 0.006348 0.034536

Coefficients:

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

(Intercept) 1.101e+00 6.051e-02 18.200 < 2e-16 \*\*\*

Age -1.247e-04 8.780e-05 -1.421 0.15717

Weight 9.204e-05 1.851e-04 0.497 0.61969

Height 6.288e-04 5.387e-04 1.167 0.24460

`Chest Circ.` 3.776e-04 2.966e-04 1.273 0.20452

`Abdomen 2 Circ.` -2.209e-03 2.566e-04 -8.608 3.63e-15 \*\*\*

`Hip Circ.` 4.425e-04 3.815e-04 1.160 0.24760

`Thigh Circ.` -2.793e-04 4.087e-04 -0.683 0.49519

`Knee Circ.` -1.049e-04 6.648e-04 -0.158 0.87482

`Ankle Circ.` -6.680e-04 9.647e-04 -0.692 0.48955

`Bicep Circ.` -6.314e-04 4.674e-04 -1.351 0.17844

`Wrist Circ.` 4.279e-03 1.490e-03 2.873 0.00456 \*\*

poly(0.01 \* `Neck Circ.`, 3)1 1.384e-02 2.106e-02 0.657 0.51192

poly(0.01 \* `Neck Circ.`, 3)2 1.295e-03 1.268e-02 0.102 0.91875

poly(0.01 \* `Neck Circ.`, 3)3 3.801e-03 1.147e-02 0.332 0.74064

poly(0.01 \* `Forearm Circ.`, 3)1 -1.805e-02 1.772e-02 -1.019 0.30975

poly(0.01 \* `Forearm Circ.`, 3)2 5.622e-03 1.215e-02 0.463 0.64404

poly(0.01 \* `Forearm Circ.`, 3)3 3.233e-03 1.362e-02 0.237 0.81258

---

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

Residual standard error: 0.01011 on 180 degrees of freedom

Multiple R-squared: 0.7369, Adjusted R-squared: 0.7121

F-statistic: 29.66 on 17 and 180 DF, p-value: < 2.2e-16

* We can see that we have negative coefficients for 7 / 17 of the explanatory variables

The residual standard error (the amount of error that can be explained by the regression model) is relatively low

* The t-test for the significance of each of the variables in the model states that most of the variables are insignificant in the model
* The global F-test for the significance of the model as a whole states that the model fits the data well
* The multiple R-squared and adjusted R-squared are relatively high, indicating that there is a relatively strong positive correlation present between the explanatory variables and the response variable

**Polynomial Model Diagnostic Plots:**

* The residuals vs. Fitted plot: The red line across the center of the plot is nearly horizontal, therefore it is reasonable to assume that the residuals follow a linear pattern.
* The Normal Q-Q plot : Some of the points at both ends of the line start to deviate a bit from the diagonal, though most lie nearly on it so it's safe to assume they are normally distributed.
* Scale-Location plot: The red line in our plot is nearly horizontal, therefore it is safe to assume constant variance.
* Residuals vs. Leverage plot: In our plot, none of the points cross and therefore there aren't any overly influential points in the data set.

Chart, diagram

Description automatically generated

**Polynomial Model Tests:**

Shapiro-Wilk normality test

data: poly\_model$residuals

W = 0.98853, p-value = 0.1121

studentized Breusch-Pagan test

data: poly\_model

BP = 14.881, df = 17, p-value = 0.6041

* We can see here that we fail to reject the null hypothesis for each of these tests, indicating that the residuals for the naïve model have homoscedasticity (constant variance) and are normally distributed.

**Polynomial Model Training Dataset Predictions:**

**Chart, line chart

Description automatically generated**

* Above is the plot of the polynomial’s prediction on the train dataset. The red bands indicate the upper and lower bounds of the 95% prediction intervals. We can see that the model seems to predict the percentage of body fat with significant accuracy, though a few points fall outside of the prediction interval.

**Back Selection Model Metrics:**

Call:

lm(formula = Fat\_Percent ~ Age + Height + `Chest Circ.` + `Abdomen 2 Circ.` +

`Bicep Circ.` + `Wrist Circ.`, data = train)

Residuals:

Min 1Q Median 3Q Max

-0.019746 -0.007174 0.000517 0.006511 0.036110

Coefficients:

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

(Intercept) 1.0773047 0.0219167 49.154 < 2e-16 \*\*\*

Age -0.0001407 0.0000658 -2.137 0.033836 \*

Height 0.0009092 0.0003270 2.780 0.005974 \*\*

`Chest Circ.` 0.0004976 0.0002302 2.162 0.031901 \*

`Abdomen 2 Circ.` -0.0019895 0.0001802 -11.038 < 2e-16 \*\*\*

`Bicep Circ.` -0.0006826 0.0003771 -1.810 0.071854 .

`Wrist Circ.` 0.0042127 0.0011437 3.684 0.000299 \*\*\*

---

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

Residual standard error: 0.009974 on 191 degrees of freedom

Multiple R-squared: 0.7282, Adjusted R-squared: 0.7197

F-statistic: 85.29 on 6 and 191 DF, p-value: < 2.2e-16

* We can see that we have gone from 13 explanatory variables down to 6. Furthermore, all of the variables are significant at the 5% level, with 3 / 6 of the coefficients still being negative
* The t-test for the significance of each of the variables is significant
* The global F-test for the significance of the model states that the model fits the data well
* The multiple R-squared and adjusted R-squared are relatively high, indicating that there is a relatively strong positive correlation present between the explanatory variables and the response variable

**Back Selection Model Diagnostic Plots:**

* The residuals vs. Fitted plot: The red line across the center of the plot is nearly horizontal, therefore it is reasonable to assume that the residuals follow a linear pattern.
* The Normal Q-Q plot : Some of the points at both ends of the line start to deviate a bit from the diagonal, though most lie nearly on it. That being said, it may be the case that the residuals are not normally distributed.
* Scale-Location plot: The red line in our plot is nearly horizontal, therefore it is safe to assume constant variance.
* Residuals vs. Leverage plot: In our plot, none of the points cross and therefore there aren't any overly influential points in the data set.

Chart, diagram

Description automatically generated

**Back Selection Model Tests:**

Shapiro-Wilk normality test

data: backsel$residuals

W = 0.98469, p-value = 0.03011

studentized Breusch-Pagan test

data: backsel

BP = 9.17, df = 6, p-value = 0.1642

* We can see here that we reject the null hypothesis for the Shapiro-Wilks test and fail to reject the null hypothesis for the Breusch-Pagan test. Therefore, we conclude that the residuals are not normally distributed and have constant variance.

**Back Selection Model Testing Dataset Predictions:**

Chart, scatter chart

Description automatically generated

* Above is the plot of the back selection model’s prediction on the test dataset. The red bands indicate the upper and lower bounds of the 95% prediction intervals. We can see that the model seems to predict the percentage of body fat with significant accuracy, with no points being outside the prediction interval.

**BIC Model Metrics:**

Call:

lm(formula = as.formula(form), data = train)

Residuals:

Min 1Q Median 3Q Max

-0.020331 -0.007433 0.000518 0.006422 0.037560

Coefficients:

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

(Intercept) 1.181e+00 1.784e-02 66.226 < 2e-16 \*\*\*

Weight 3.201e-04 7.054e-05 4.538 9.99e-06 \*\*\*

`Abdomen 2 Circ.` -2.272e-03 1.434e-04 -15.845 < 2e-16 \*\*\*

`Bicep Circ.` -9.009e-04 3.938e-04 -2.288 0.02324 \*

`Wrist Circ.` 3.102e-03 1.104e-03 2.809 0.00548 \*\*

---

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

Residual standard error: 0.01 on 193 degrees of freedom

Multiple R-squared: 0.7239, Adjusted R-squared: 0.7182

F-statistic: 126.5 on 4 and 193 DF, p-value: < 2.2e-16

* We can see that we have gone from 13 explanatory variables down to 4. Furthermore, all of the variables are significant at the 5% level, with 2 / 4 of the coefficients still being negative
* The t-test for the significance of each of the variables is significant
* The global F-test for the significance of the model states that the model fits the data well
* The multiple R-squared and adjusted R-squared are relatively high, indicating that there is a relatively strong positive correlation present between the explanatory variables and the response variable

**BIC Model Diagnostic Plots:**

* The residuals vs. Fitted plot: The red line across the center of the plot is nearly horizontal, therefore it is reasonable to assume that the residuals follow a linear pattern.
* The Normal Q-Q plot : Some of the points at both ends of the line start to deviate a bit from the diagonal, though most lie nearly on it. That being said, it may be the case that the residuals are not normally distributed.
* Scale-Location plot: The red line in our plot is nearly horizontal, therefore it is safe to assume constant variance.
* Residuals vs. Leverage plot: In our plot, none of the points cross and therefore there aren't any overly influential points in the data set.

Chart, diagram

Description automatically generated

**BIC Model Tests:**

Shapiro-Wilk normality test

data: bicmod$residuals

W = 0.98543, p-value = 0.03872

studentized Breusch-Pagan test

data: bicmod

BP = 4.8075, df = 4, p-value = 0.3076

* We can see here that we reject the null hypothesis for the Shapiro-Wilks test and fail to reject the null hypothesis for the Breusch-Pagan test. Therefore, we conclude that the residuals are not normally distributed and have constant variance.

**BIC Model Testing Dataset Predictions:**

Chart, scatter chart

Description automatically generated

* Above is the plot of the BIC model’s prediction on the test dataset. The red bands indicate the upper and lower bounds of the 95% prediction intervals. We can see that the model seems to predict the percentage of body fat with significant accuracy, with no points being outside the prediction interval.

**AIC Model Metrics:**

Call:

lm(formula = as.formula(form), data = train)

Residuals:

Min 1Q Median 3Q Max

-0.020331 -0.007433 0.000518 0.006422 0.037560

Coefficients:

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

(Intercept) 1.181e+00 1.784e-02 66.226 < 2e-16 \*\*\*

Weight 3.201e-04 7.054e-05 4.538 9.99e-06 \*\*\*

`Abdomen 2 Circ.` -2.272e-03 1.434e-04 -15.845 < 2e-16 \*\*\*

`Bicep Circ.` -9.009e-04 3.938e-04 -2.288 0.02324 \*

`Wrist Circ.` 3.102e-03 1.104e-03 2.809 0.00548 \*\*

---

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

Residual standard error: 0.01 on 193 degrees of freedom

Multiple R-squared: 0.7239, Adjusted R-squared: 0.7182

F-statistic: 126.5 on 4 and 193 DF, p-value: < 2.2e-16

* We can see that we have gone from 13 explanatory variables down to 4. Furthermore, all of the variables are significant at the 1% level, with 2 / 4 of the coefficients still being negative
* The t-test for the significance of each of the variables is significant
* The global F-test for the significance of the model states that the model fits the data well
* The multiple R-squared and adjusted R-squared are relatively high, indicating that there is a relatively strong positive correlation present between the explanatory variables and the response variable

**AIC Model Diagnostic Plots:**

* The residuals vs. Fitted plot: The red line across the center of the plot is nearly horizontal, therefore it is reasonable to assume that the residuals follow a linear pattern.
* The Normal Q-Q plot : A greater amount of residuals seem to deviate at the left and right ends of the line than for previous models. Depending on the level of significance chosen, we may conclude that the residuals are not normally distributed.
* Scale-Location plot: The red line in our plot is nearly horizontal, therefore it is safe to assume constant variance.
* Residuals vs. Leverage plot: In our plot, none of the points cross and therefore there aren't any overly influential points in the data set.

Chart, diagram

Description automatically generated

**AIC Model Tests:**

Shapiro-Wilk normality test

data: aicmod$residuals

W = 0.98543, p-value = 0.03872

studentized Breusch-Pagan test

data: aicmod

BP = 4.8075, df = 4, p-value = 0.3076

* We can see here that we reject the null hypothesis for the Shapiro-Wilks test and fail to reject the null hypothesis for the Breusch-Pagan test. Therefore, we conclude that the residuals are not normally distributed and have constant variance.

**AIC Model Testing Dataset Predictions:**

Chart, scatter chart

Description automatically generated

* Above is the plot of the AIC model’s prediction on the test dataset. The red bands indicate the upper and lower bounds of the 95% prediction intervals. We can see that the model seems to predict the percentage of body fat with significant accuracy, with no points being outside the prediction interval.

**Adjusted R-Squared Model Metrics:**

Call:

lm(formula = as.formula(form), data = train)

Residuals:

Min 1Q Median 3Q Max

-0.019626 -0.006664 0.000608 0.006087 0.035813

Coefficients:

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

(Intercept) 1.075e+00 2.244e-02 47.904 < 2e-16 \*\*\*

Age -1.116e-04 7.549e-05 -1.478 0.14115

Height 7.893e-04 3.418e-04 2.309 0.02203 \*

`Chest Circ.` 5.142e-04 2.317e-04 2.220 0.02762 \*

`Abdomen 2 Circ.` -2.169e-03 2.291e-04 -9.466 < 2e-16 \*\*\*

`Hip Circ.` 3.544e-04 2.925e-04 1.212 0.22715

`Bicep Circ.` -6.245e-04 4.078e-04 -1.531 0.12735

`Forearm Circ.` -5.364e-04 4.932e-04 -1.088 0.27816

`Wrist Circ.` 4.360e-03 1.189e-03 3.666 0.00032 \*\*\*

---

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

Residual standard error: 0.009957 on 189 degrees of freedom

Multiple R-squared: 0.7319, Adjusted R-squared: 0.7206

F-statistic: 64.5 on 8 and 189 DF, p-value: < 2.2e-16

* We can see that we have gone from 13 explanatory variables down to 8. Furthermore, all of the variables are significant at the 1% level, with 4 / 8 of the coefficients still being negative
* The t-test for the significance of each of the variables is significant
* The global F-test for the significance of the model states that the model fits the data well
* The multiple R-squared and adjusted R-squared are relatively high, indicating that there is a relatively strong positive correlation present between the explanatory variables and the response variable

**Adjusted-R-Squared Model Diagnostic Plots:**

* The residuals vs. Fitted plot: The red line across the center of the plot is nearly horizontal, therefore it is reasonable to assume that the residuals follow a linear pattern.
* The Normal Q-Q plot : A greater amount of residuals seem to deviate at the left and right ends of the line than for previous models. Depending on the level of significance chosen, we may conclude that the residuals are not normally distributed.
* Scale-Location plot: The red line in our plot is nearly horizontal, therefore it is safe to assume constant variance.
* Residuals vs. Leverage plot: In our plot, none of the points cross and therefore there aren't any overly influential points in the data set.

Chart, diagram

Description automatically generated

**Adjusted R-Squared Model Tests:**

Shapiro-Wilk normality test

data: adjr2mod$residuals

W = 0.98571, p-value = 0.04261

studentized Breusch-Pagan test

data: adjr2mod

BP = 9.9807, df = 8, p-value = 0.2664

* We can see here that we reject the null hypothesis for the Shapiro-Wilks test and fail to reject the null hypothesis for the Breusch-Pagan test. Therefore, we conclude that the residuals are not normally distributed and have constant variance.

**Adjusted R-Squared Model Testing Dataset Predictions:**

Chart, scatter chart

Description automatically generated

* Above is the plot of the adjusted R-Squared model’s prediction on the test dataset. The red bands indicate the upper and lower bounds of the 95% prediction intervals. We can see that the model seems to predict the percentage of body fat with significant accuracy, with no points being outside the prediction interval.

**Cp Model Metrics:**

Call:

lm(formula = as.formula(form), data = train)

Residuals:

Min 1Q Median 3Q Max

-0.020331 -0.007433 0.000518 0.006422 0.037560

Coefficients:

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

(Intercept) 1.181e+00 1.784e-02 66.226 < 2e-16 \*\*\*

Weight 3.201e-04 7.054e-05 4.538 9.99e-06 \*\*\*

`Abdomen 2 Circ.` -2.272e-03 1.434e-04 -15.845 < 2e-16 \*\*\*

`Bicep Circ.` -9.009e-04 3.938e-04 -2.288 0.02324 \*

`Wrist Circ.` 3.102e-03 1.104e-03 2.809 0.00548 \*\*

---

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

Residual standard error: 0.01 on 193 degrees of freedom

Multiple R-squared: 0.7239, Adjusted R-squared: 0.7182

F-statistic: 126.5 on 4 and 193 DF, p-value: < 2.2e-16

* We can see that we have gone from 13 explanatory variables down to 4. Furthermore, all of the variables are significant at the 1% level, with 2 / 4 of the coefficients still being negative
* The t-test for the significance of each of the variables is significant
* The global F-test for the significance of the model states that the model fits the data well
* The multiple R-squared and adjusted R-squared are relatively high, indicating that there is a relatively strong positive correlation present between the explanatory variables and the response variable

**Cp Model Diagnostic Plots:**

* The residuals vs. Fitted plot: The red line across the center of the plot is nearly horizontal, therefore it is reasonable to assume that the residuals follow a linear pattern.
* The Normal Q-Q plot : A greater amount of residuals seem to deviate at the left and right ends of the line than for previous models. Depending on the level of significance chosen, we may conclude that the residuals are not normally distributed.
* Scale-Location plot: The red line in our plot is nearly horizontal, therefore it is safe to assume constant variance.
* Residuals vs. Leverage plot: In our plot, none of the points cross and therefore there aren't any overly influential points in the data set.

Chart, diagram

Description automatically generated

**Cp Model Tests:**

Shapiro-Wilk normality test

data: cpmod$residuals

W = 0.98543, p-value = 0.03872

studentized Breusch-Pagan test

data: cpmod

BP = 4.8075, df = 4, p-value = 0.3076

* We can see here that we reject the null hypothesis for the Shapiro-Wilks test and fail to reject the null hypothesis for the Breusch-Pagan test. Therefore, we conclude that the residuals are not normally distributed and have constant variance.

**Cp Model Testing Dataset Predictions:**

Chart, scatter chart

Description automatically generated

* Above is the plot of the Cp model’s prediction on the test dataset. The red bands indicate the upper and lower bounds of the 95% prediction intervals. We can see that the model seems to predict the percentage of body fat with significant accuracy, with no points being outside the prediction interval.

**Conclusion (summary of findings)**

The statistical analysis of this data set has indicated that it Is advantageous to use linear regression when estimating one’s percentage of body fat, given that person is a male. Furthermore, it has been shown that such a linear regression model can have different combinations of explanatory variables and still accurately predict percentage of body fat. Take the number of explanatory variables in the five optimal models generated: the models determined via back selection, Bayes Information Criterion (AIC), Akaike Information Criterion (AIC), adjusted R-squared and Mallow’s Cp had different combinations and / or numbers of variables respectively.

For future work, I would recommend that gender specific data be collected. To elaborate, I believe that this model could benefit from adding a categorical variable for women. The variable could simply be additive, and another regression line could be generated in order to explain the trend in women’s body fat. This would enable not only men to have a much easier and more cost effective way to estimate body fat, but women as well.

I believe that this model could benefit from more diverse data, such as additional data gathered from men and women in different countries. This would likely force the model to become more complex, though it would undoubtedly increase the model’s ability to predict the percentages of body fat of men and women with more diverse backgrounds.

I believe that, if the number of data points were significantly increased through more data collection, other methods of statistical learning could be applied to this problem. To elaborate, it has been shown that the predictive power of deep learning has a direct relationship with the number of data points in a dataset. If there were enough data, deep learning could be applied to predict percentage of body fat with much more certainty.

Finally, I believe that with more time this model could be improved by analyzing more finely transformations of variables and outliers in the model. The problem of curvilinearity failed to be completely handled in the naïve model, and thus this model could stand to improve if that were accurately remedied. In addition, a deeper analysis of the significance of outliers could also stand to improve this model; having the ability to remove more problematic outliers or collect data akin to outliers the model failed to predict would be useful.

**Appendix**

1. **Pairs.R Output:**

* The below images list the plotted relationships and correlations from the variance-covariance matrix for the full data set. Most of the plotted relationships are linear, while some have a slight deviation towards being curvilinear. In addition, there are high levels of correlation between variables.

**Company name

Description automatically generated with medium confidence**

**Arrow

Description automatically generated with medium confidence**

**Chart, histogram

Description automatically generatedA picture containing company name

Description automatically generated**

1. **Training Dataset Relationship Plots**

**Chart, diagram, scatter chart

Description automatically generated**Chart, scatter chart

Description automatically generated

**Graphical user interface

Description automatically generated with medium confidence**Diagram, schematic, scatter chart

Description automatically generated

1. **Prolematic Relationships Transformed**

**Chart, scatter chart

Description automatically generated with medium confidence**

1. **Naïve Model Residuals Plot**

**Chart, scatter chart

Description automatically generated**

1. **Polynomial Model Residual Plots**

**Chart, scatter chart

Description automatically generated**

1. **Optimal Models Residual Plots**

**Diagram, qr code

Description automatically generated**

**Citations:**

Bailey, Covert (1994). \_Smart Exercise: Burning Fat, Getting Fit\_,

Houghton-Mifflin Co., Boston, pp. 179-186.

Behnke, A.R. and Wilmore, J.H. (1974). \_Evaluation and Regulation of Body

Build and Composition\_, Prentice-Hall, Englewood Cliffs, N.J.

Siri, W.E. (1956), "Gross composition of the body", in \_Advances in

Biological and Medical Physics\_, vol. IV, edited by J.H. Lawrence and C.A.

Tobias, Academic Press, Inc., New York.

Katch, Frank and McArdle, William (1977). \_Nutrition, Weight Control, and

Exercise\_, Houghton Mifflin Co., Boston.

Wilmore, Jack (1976). \_Athletic Training and Physical Fitness: Physiological

Principles of the Conditioning Process\_, Allyn and Bacon, Inc., Boston.

Library lmtest

Library car

Library leaps