This report uses the abdomen predictor(measured in cm) to best estimate the body fat percentage of the given data sample of 252 men using a single regression (SLR) model. Analysis began by finding the correlation coefficients and R^2 values of each predictor with respect to body fat percentage (our dependent variable) using R (Figure X). We narrowed down the predictors to continue analyzing to the abdomen, adiposity, and chest because they had the best R^2 values and were over the 0.4 R^2 benchmark we set for ourselves. Eliminating leverage points was the next order of importance. We used R to plot Cook's Distance influence values for abdomen, adiposity, and chest (Figure X) to find these leverage points. Row 39 contained leverage points for all three predictors and six more predictors upon further analysis (make sure to include the entire matrix of 14 Cook's distance plots in slideshow) leading us to remove row 39 from analysis because of its high influence. To discover outliers we used an outlier function in R finding rows 41 and 216 to contain abdomen and adiposity outliers. We removed rows 41 and 216 from the data as well after this analysis. To finalize our SLR model we plotted residual and QQ plots to measure linearity, homoscedasticity, and normality for abdomen and adiposity. We eliminated chest as a potential predictor because of its excessive variance in the Cook's Distance plot.

Shreya

Next step is to check whether our predictor values satisfy the three key assumptions to the SLR Model: Linearity, Homoscedasticity, and Normality.

Figure 1

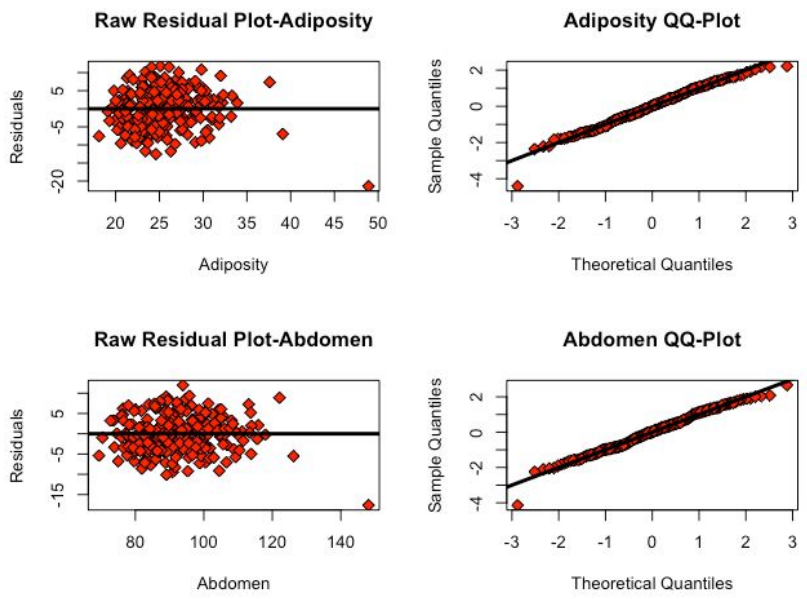

Figure 1 clearly suggests that there are 3 main outliers on both the adiposity and abdomen residual plots. The QQ plots further confirm the existence of these outliers with one point sticking out on the bottom left of the QQ plots. Upon inspection, it was figured that those three points corresponded to our previously found outliers: row 39, 41, 216.

Figure 1 weakly violates homoscedasticity, as there is no strong spreading of data points. Since clustering takes place, the figure suggests a violation of linearity. Normality seems to be justified because the QQ plots are linear and data points aren't sticking out too much, except for the outliers except for a slight curve on the top right of the abdomen QQ plot.

To get a better perspective, we decided to view these residual and QQ plots after removing those leverage points and outliers, which is shown in Figure 2.

Figure 2

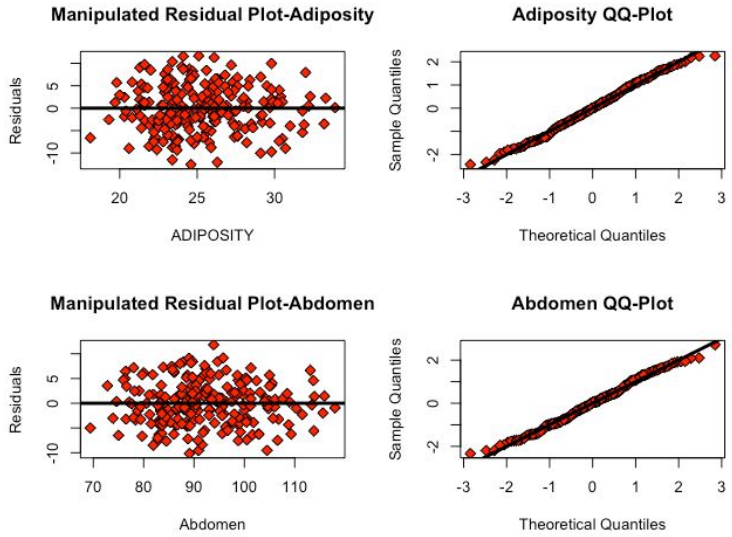

Figure 2 shows that without the leverage points and outliers, the clustering effect has faded, suggesting a more satisfactory linearity. However, the slight curve on the top right of the QQ plots that hasn't changed much with the removal of outliers, potentially suggesting that for the most part the data is normal, however one should be careful around the end points.

Figure 2 also suggests that both the plots are free of the "fanning" or the "funneling" effect since values are spread out in an even manner indicating they don't have nonconstant error variance.

Overall, all group member contributed evenly, but if we had to specify parts, Shaelyn did the intro and leverage points, Shreya did the assumptions regarding the SLR model such as homoscedasticity and normality and Xiangyu did the SLR model and confidence intervals. The presentation follows the same lineup as well.

We decided to use the SLR. Next we compared the coefficient values before and after removing the outliers and leverage points. The R^2 changes from 0.813 to 0.817 after we removed the outliers and leverage points in rows 39,41, and 216, for instance. (GRAPH: not too much difference, so no graph is better?)
It is noted that row 42 in the dataset contains a leverage point that is a below normal height and that row 182 has 0% body fat which is an outlier (researchers might fill this in by mistake) but they fail to affect our potential SLR model body predictors. Moreover, our correlation value increases by keeping these rows in our data. Based on those facts, we decided not to remove these rows of data. On the next stage, we decided to choose the predictors for our SLR model

based on the values of correlation coefficients, deciding on abdomen. Considering the interpretability of those data, we also decide that abdomen is relatively easy to measure compared with other measurements like adiposity. By applying the lm() function we get the predicted slope and intercept values of our model. To make sure the accuracy of our model, we compare the graphs with and without those variables model (Figure X). After using the predict() function in R and randomly sampling rows from the BodyFat data, we found we are 95 percent confident that our prediction will correctly predict body fat percentage 11% of the time using the predict function.

Citations:
Komsta, Lukasz. "CRAN - Package Outliers." *CRAN - Package Outliers*. R Project, 24 Jan. 2011. Web. 18 Feb. 2017.

We found the abdomen predictor to be the best estimate of the body fat percentage of the data sample using a single linear regression (SLR) model out of a potential 17 total predictors. Analysis began by finding the correlation coefficients and R^2 values of each predictor with respect to body fat percentage (our dependent variable) using R. We narrowed down the predictors to continue analyzing to the abdomen, adiposity, and chest because they had the best R^2 values and were over the 0.4 R^2 benchmark we set for ourselves (Table 1). Eliminating leverage points was the next order of importance. We used R to plot Cook's Distance influence values for abdomen, adiposity, and chest (Figure 1) to find these leverage points. Row 39 contained leverage points for all three predictors and six more predictors upon further analysis leading us to remove row 39 from analysis because of its high influence. To discover outliers we used an outlier function in R finding rows 41 and 216 to contain abdomen and adiposity outliers. We removed rows 41 and 216 from the data as well after this analysis.

To continue searching for the predictor we would use in our SLR model we plotted residual and QQ plots to measure linearity, homoscedasticity, and normality for abdomen and adiposity. We eliminated chest as a potential predictor because of its excessive variance in the Cook's Distance plot. Figure 2a clearly suggests that there are 3 main outliers on both the adiposity and abdomen residual plots. The QQ plots further confirm the existence of these outliers with one point sticking out on the bottom left of the QQ plots. Upon inspection, it was figured that those three points corresponded to our previously found outliers: row 39, 41, 216. Figure 2a weakly violates homoscedasticity, as there is no strong spreading of data points. Since the data points aren't very spread out, and clustering takes place, the figure suggests a violation of linearity. Normality seems to be justified because the QQ plots are linear and data points aren't sticking out too

much, except for the outliers. However, there is a slight curve on the top right of the abdomen QQ plot. To get a better perspective, we decided to view these residual and QQ plots after removing those leverage points and outliers, which is shown in Figure 2b. It is clear from the graph, that without the leverage points and outliers, the clustering effect has faded, suggesting a more satisfactory linearity and normality. Figure 2b also suggests that both the plots are free of the "fanning" or the "funneling" effect since values are spread out in an even manner rather than clustered indicating they satisfy the assumptions of homoscedasticity or in other words don't have non-constant error variance. Figure 2b residuals display a much better spread than residuals in figure 2a.

To finalize our SLR model we compared the coefficient values before and after removing the outliers and leverage points. The R^2 changes from 0.813 to 0.817 after we removed the outliers and leverage points in rows 39,41, and 216, for instance. The high value of the abdomen R^2 value further confirmed our belief an SLR model would fit the data just as well as an MLR. It is noted that row 42 in the dataset contains a leverage point that is a below normal height and that row 182 has 0% body fat which is an outlier (researchers might fill this in by mistake) but they fail to affect our potential SLR model body predictors. Moreover, our correlation value increases by keeping these rows in our data. Based on those facts, we decided not to remove these rows of data. On the next stage, we decided to choose the predictors for our SLR model based on the values of correlation coefficients, deciding on abdomen. Considering the interpretability of those data, we also decide that abdomen is relatively easy to measure compared with other measurements like adiposity. By applying the lm function we find our SLR model shown in the graph, along with the 95% confidence intervals that indicate the significance of our statistical analysis (Figure 3).

After using the summary function in R, we found our p-value for abdomen to be 2.2e-16. Since abdomen as a predictor has a low p-value, it is likely to be a meaningful addition to a model because changes in the predictor's value are related to changes in the response variable. This is one way our analysis could be improved by switching to an MLR model that uses abdomen as a predictor as part of its analysis.