**Introduction:** For generations, obesity and fat were considered to be the problem of the ruling classes, mainly royalty and their officials, while everyone else had the problem of malnutrition. Today, the opposite seems to be true as more food has become available to more people. The downside to this is that unhealthy food tends to be cheaper than healthier foods, leading to a crisis of obesity in developed countries. Over the past half a century, there have been many strides to study the average body fat percentage and what biological factors can alter this. As a case study, we will use the data collected by William E. Siri in 1956.

**Background Info/Data Cleaning:** The mean body fat percentage (Y) is 18.94 rounded up. We will also remove the 39th data point as this individual contains several maximums (including weight) and will have a pronounced effect on our model; this individual's weight is 363.15 lbs and the next smallest weight is 262.75 lbs. With the individual included, the mean weight is 178.92 lbs, and without him it is 178.19 lbs. We will also remove the 42nd data point.

**Motivation for Model:** Our proposed MLR Model is **Body Fat %=0.122Age+1.66Adiposity-28.675.** As an example, a man whose age is 23 and Adiposity is 31.2 is expected to have a body fat % of 25.92265 based on our model. His 95% prediction interval is between [24.19345, 27.65185]. Our estimated coefficients are Age and Adiposity, which are in the units of years and bmi. This means that for every age increases in one year(keep the same adiposity), the model predicts that body fat % will increase, on average, by 0.122%; and for every adiposity increases in one bmi(keep the same age), the model predicts that body fat % will increase, on average, by 1.66%. We chose this model because of the following reasons. First, we hope to select the data that are easy to measure and not prone to produce errors, so we deleted the DENSITY that was difficult to measure, and predictors like NECK, KNEE, CHEST, ABDOMEN, HIP, THIGH, ANKLE, BICEPS, FOREARM, WRIST, which were less likely to measure accurately. Secondly, we tested every remaining predictor, and the p-values of WEIGHT and HEIGHT were 0.3672499 and 0.3916344 and larger than 0.05, so we failed to reject the null hypothesis that the two individuals were not useful.The p-values of AGE and ADIPOSITY were both smaller than 0.05, so we rejected the null hypothesis. Third, considering the adjusted R^2, the Adjusted R^2 of our final model is 0.5930276, which was larger than the Adjusted R^2 (0.5911449) of our former model with 4 predictors. Therefore, we conclude our final model that Body Fat Pct=0.122Age+1.66Adiposity-28.675 is better.

**Statistical Analysis/Hypothesis Testing/Inference**

We conducted some F tests.

Firstly, we use case 1 F test to test if any of the age, weight, height and adiposity are useful for prediction.

H0:They are all not useful. H1:At least one of them are useful.P value is $0 < 0.05$,

We fail to reject H0. Thus, at least one of the four predictors is useful.

Then, we test age after accounting weight, height,and adiposity by using case 2 F test.

H0: age is not useful. H1: age is useful. P value is $4.179128e\text{-}06 < 0.05$, we reject the null in favor of the alternative. So, we keep age as a predictor. Similarly, we test weight after accounting age, height,and adiposity by using case 2 F test. H0: weight is not useful. H1: weight is useful. P value is $0.3556107 > 0.05$, we fail to reject H0. So, weight is not useful. We test height after accounting age, weight,and adiposity by using case 2 F test. H0: height is not useful. H1: height is useful. P value is $0.3571696 > 0.05$, we fail to reject H0. So, height is not useful.

We test adiposity after accounting age, weight,and height by using case 2 F test.

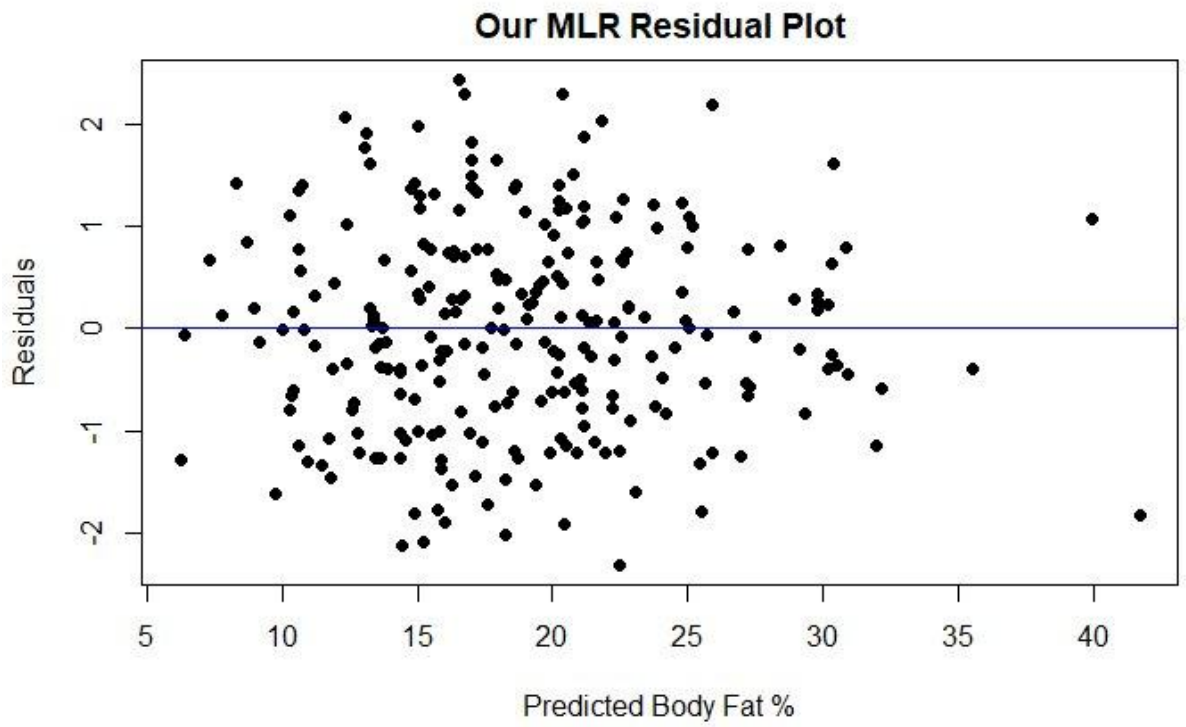
H0: adiposity is not useful. H1: adiposity is useful. P value is $0.003958038 < 0.05$, we reject the null in favor of the alternative. So, we keep the adiposity.

Then, we test if age and adiposity are as good as all of the variables by T test case 3.
H0: The subset of age and adiposity are not important for prediction, H1: The subset of age and adiposity are important for prediction. Since p value is close to $0 < 0.05$ we reject the null in favor of the alternative. So the subset is as good as the model based on all the variables.
We test whether having just one of the variables age is sufficient compared to age and adiposity. This also falls under case 3. H0: The subset of age is not good as two predictors. H1: The subset of age is as good as two predictors. Since p value is close to $0 < 0.05$ we reject the null in favor of the alternative, so the subset is as good as the model based on age and adiposity. We test whether having just one of the variables adiposity is sufficient compared to age and adiposity. This also falls under case 3. H0: The subset of age is not good as two predictors.
H1: The subset of age is as good as two predictors. Since p value is $1.655e-06 < 0.05$ we reject the null in favor of the alternative. So the subset is as good as the model based on age and adiposity. Thus, both age and adiposity are good predictors. Since we want to use the MLR model, we choose age and adiposity as our predictors.
2. Adjusted $R^2 = 0.5930276 > 0.5911449$. Because $R^2$ takes into account useless predictors and our observed adjusted $R^2$ is larger than the normal $R^2$, we can safely conclude that our model explains ~59.3% of the variability in body fat %
3.The estimated intercept is -28.675 and the estimated slopes are 0.122 for age and 1.66 for adiposity. We used Frank's data, age = 23, adiposity=31.2. The estimated body fat rate is 25.92265.
Confidence interval for mean response: We are 95% confident that average body fat rates of people who are like Frank are between 24.19345 and 27.65185. Prediction interval: There is a 95% probability that Frank's body fat rate is in between 16.11717 and 35.72812.
The truth is that Frank's body fat rate is 26.8, which is very close to the estimated value 25.92265.
**Model Diagnostics:** No model can assuredly be used without first checking its corresponding residual and QQ plots, as well as checking for outliers. Because we did the latter in Step 1, our subsequent diagnostics will confirm if more outliers exist that were not noticed at first. Diagram (I) is the residual plot for our created MLR model and diagram (II) is the Cook's distance graph for our model. As seen, these models show no violations of linearity, homoskedasticity or normality (the three assumptions) so our model passes diagnostics. In the residual plot, the three points with a body fat % of over 35% may look alarming but they aren't a significant distance away from the mean body fat % of 18.83% with an SD of 7.68%. One might note the higher Cook's distance in this graph, but it is no larger than 0.1 and far less than 1, so it does not hold much influence over the model. Individually checking the diagnostics of each predictor and the outcome variable (with respect to our created MLR model) is important as well, and these can be seen in graphs (III), (IV) & (V), which show the QQ plots of age, adiposity and body fat % respectively. Overall, the three assumptions are satisfied.
**Model Strengths/Weaknesses:** Perhaps the best strength our model has is that its predictors are easy for anyone to calculate; anyone can test themselves using this model. A person's age is already known to them and adiposity is simply that person's weight divided by the square of their height, which themselves are simple to acquire. Also, our model's adjusted $R^2$ is relatively strong at almost 60%; when compared to other combinations of predictors (and the fact that data rarely surpasses 20%) this number becomes much more important. However, having only two predictors can lead to more uncertainties, such as during diagnostics and when creating models. Other statisticians or data scientists might not be satisfied with just two predictors given their

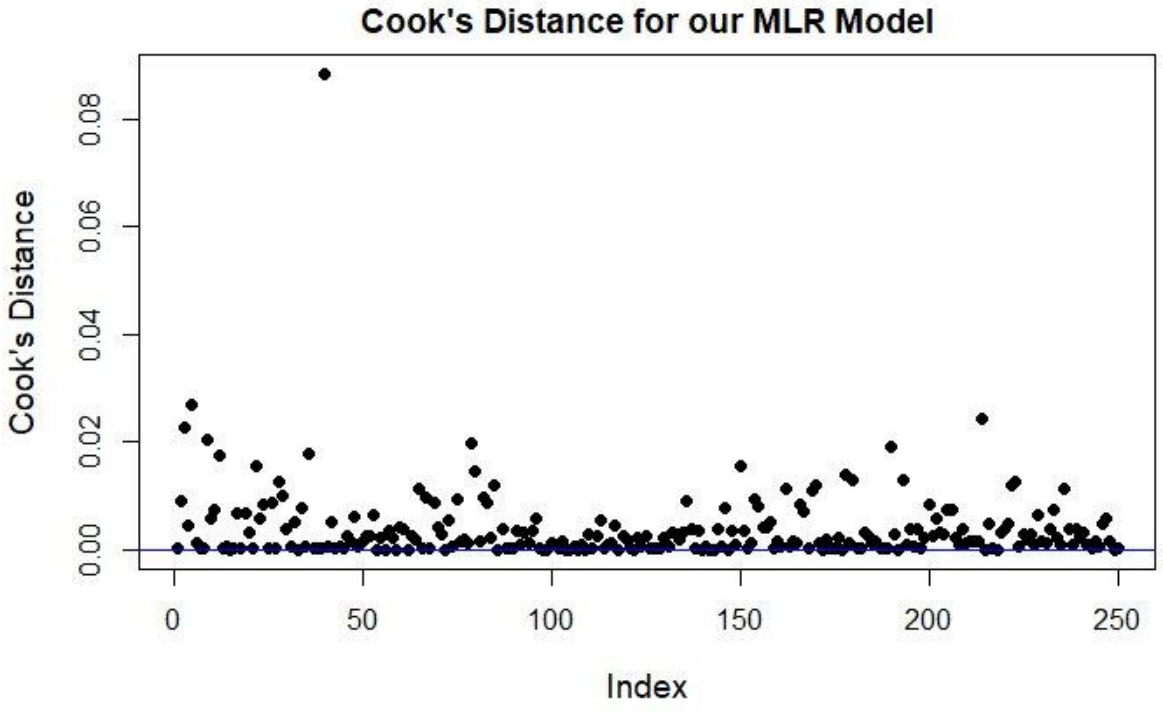
respective R^2 and F tests, which would lead to a scenario where more predictors would be ideal; additionally, this model only works for males as no females were included in the data set.
**Conclusion/Discussion**: Our proposed model is Body Fat %=0.122Age+1.66Adiposity-28.675. We discussed that the body fat rate of young people is often higher than the average, and that of the elderly is often lower than the average. Age is one of the important predictors of the model. It is possible that when young people use the model to measure, the actual situation is often higher than the predicted value, while when the elderly use the model to measure, the actual situation is prone to lower than the predicted value.
**Contribution:**
Zhuoyu: Introduction; Motivation for Model; Conclusion/Discussion; Contributions
Zhijiang: Statistical Analysis/Hypothesis Testing/Inference
Andrew: Data cleaning; Model Diagnostics; Model Strengths/Weaknesses

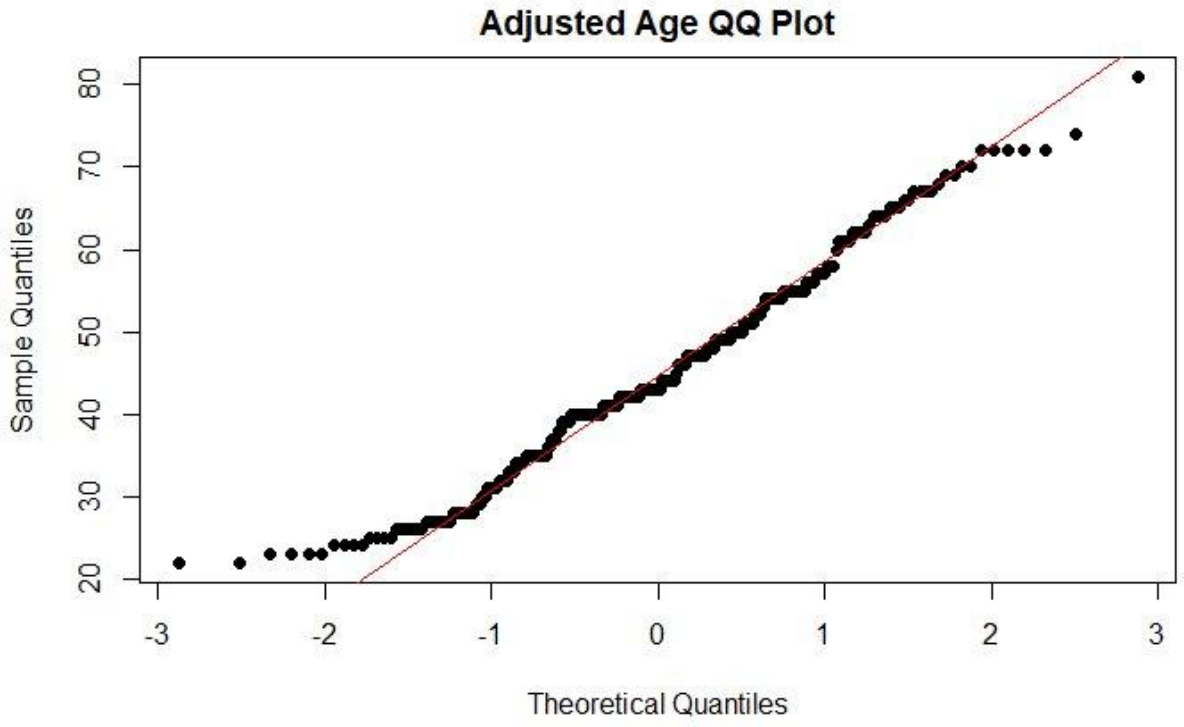
## APPENDIX AND GRAPHS

(I)



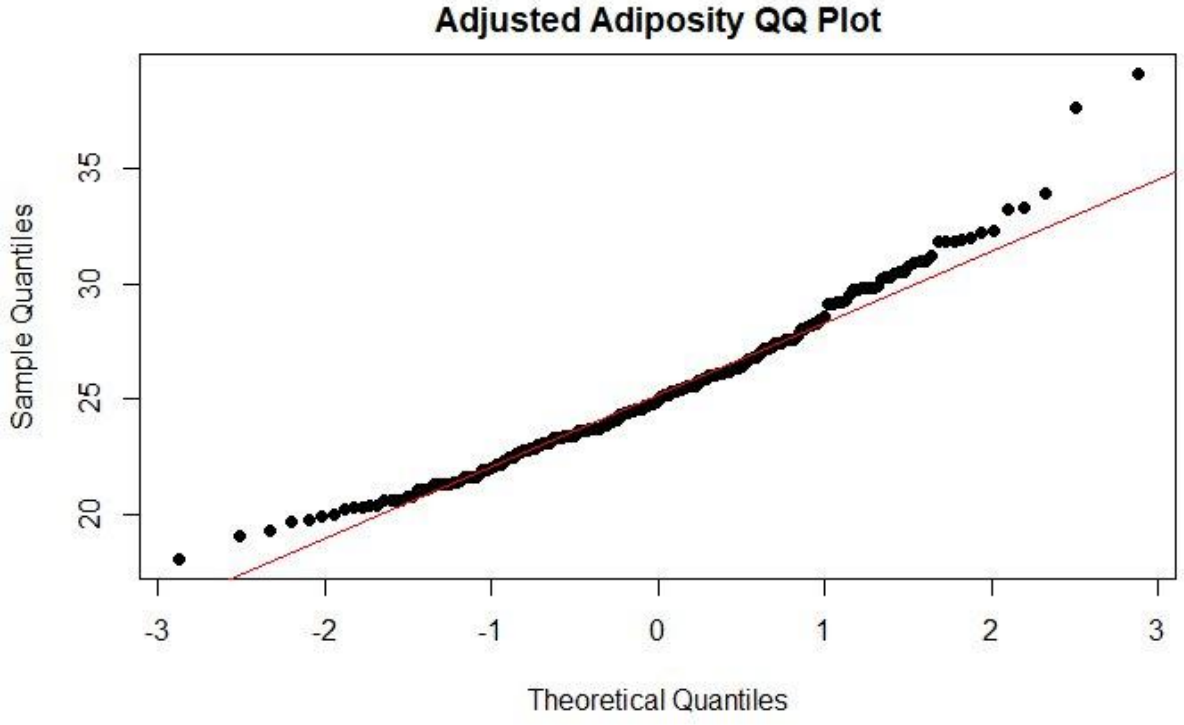

(II)

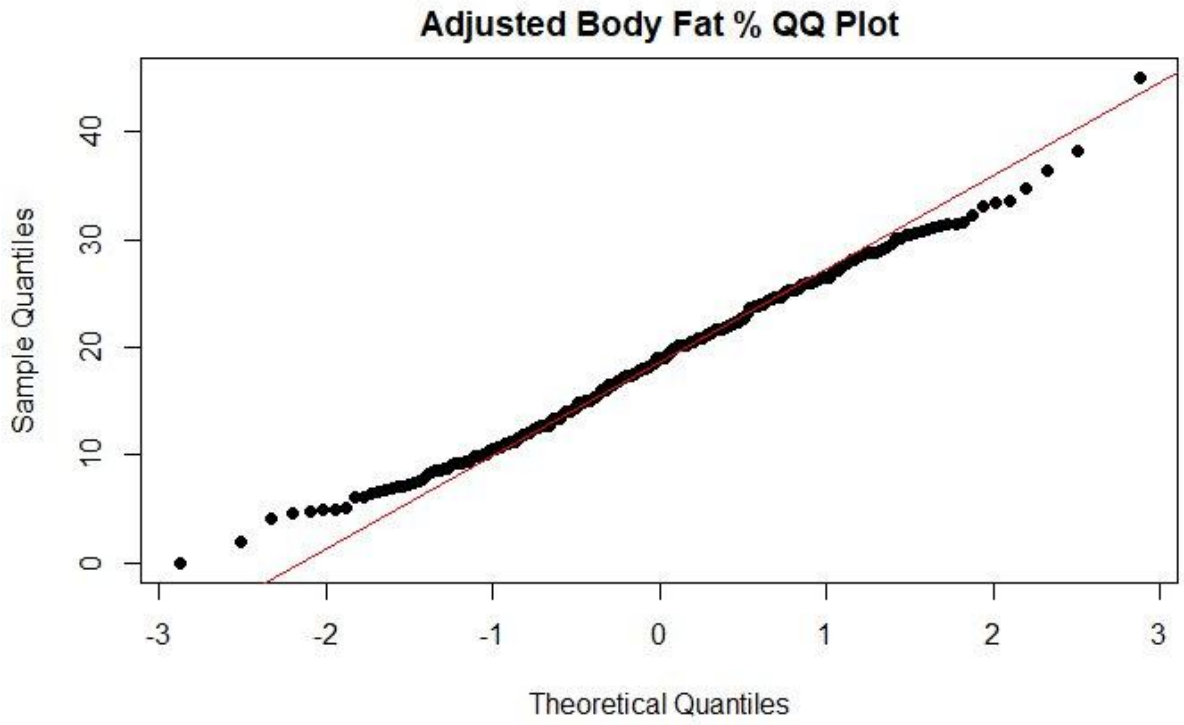



Adjusted Age QQ Plot
Sample Quantiles
Theoretical Quantiles
20
30
40
50
60
70
80
-3
-2
-1
0
1
2
3

(III)

(IV)


Adjusted Adiposity QQ Plot
Sample Quantiles
Theoretical Quantiles
20
25
30
35
-3
-2
-1
0
1
2
3


Adjusted Body Fat % QQ Plot
Sample Quantiles
0
10
20
30
40
-3
-2
-1
0
1
2
3
Theoretical Quantiles

(V)