Module 2 Executive Summary

Introduction:

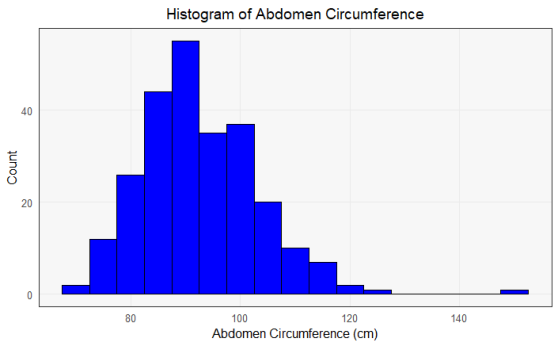
With a new found focus on personal health in the post Covid world it is no shock that most people know the common calculation for BMI using exclusively height and weight is incorrect more often than not. Using this as motivation our group took to looking at multiple data points ranging from weight to appendage circumferences in attempts to create a better predictive model for BMI. And after statistical analysis looking at both multi and singular regression we found the best combination of simplicity and accuracy in a model came from using the circumference of the abdomen to measure BMI in the following way:

$$BMI = -37.54 + 0.61\,(Abdomen\ Circumference)$$

Data Processing:

Besides body fat percentage calculated by measured density of 252 males, each individual's age and body measurements such as height, weight, and circumferences around certain limbs and core body parts were taken. For data processing, we chose to re-scale the height measurements into centimeters and weight into kilograms in order to match the metric units for the rest of the body circumference measurements. Next, we attempted to recalculate the body fat percentages of some extreme observations (namely the individuals with 0% and <45% body fat), but the Siri equation does not give any values that differed from what data was given, so we decided to eliminate these rows (Durnin, 1965).

Lastly, we created histogram plots of each feature, such as the one for abdomen circumference shown here. We used these graphs to determine if there are extreme values, then identified which observations and removed them. Most of the extreme values for each feature were from the same 1-2 rows.



Methodology:

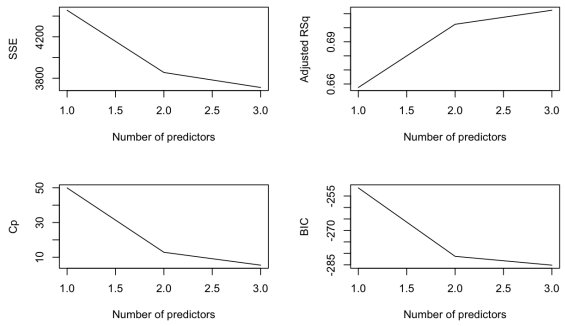
Since the goal of this project was to make both a simple and accurate prediction model, we decided on linear regression, first, fitting a linear model with every feature. This initial model had only two significant predictors (abdomen and wrist circumference) with p-values smaller than the significance level of 0.05, and many obvious multicollinearity issues among various predictors.

Next, we decided to use some interpretable measures such as SSE, adjusted R-squared, Mallows Cp, and BIC to select the best subset of predictors for our model. The line plots for these metrics are shown below, with values increasing or decreasing as predictors are added.

To decrease the sum of squared errors, increase adjusted R-squared, and decrease both Mallows Cp and BIC, we can pick between 2 to 4 variables for our model. The group decided to choose a

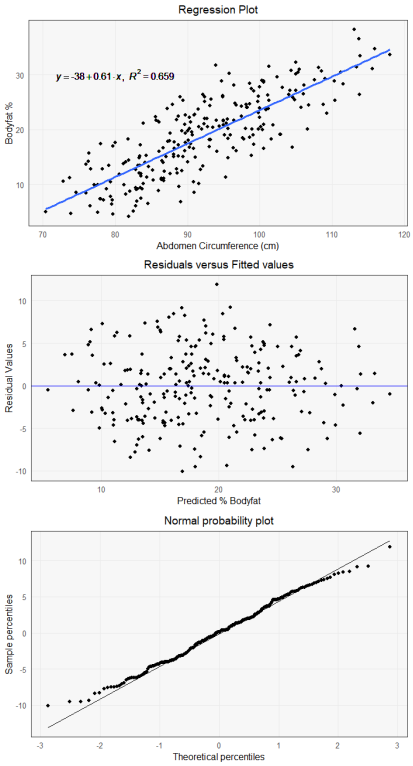
Module 2 Executive Summary

maximum of three predictors in order to ensure simplicity. Our top 3 predictors using best subsets regression are abdomen circumference, weight, and wrist circumference. However, looking at the Variance Inflation Factor (VIF) table for these 3 predictors indicated weight was highly correlated and both abdomen and wrist circumference both evidence of collinearity. Because of this we chose to fit a model without weight, and singular regression models for wrist and abdomen circumference. In which our best R-squared fit (.66) belonged to the abdomen circumference model which followed the equation you were shown in our introduction.



Diagnostics and Assumption Checking:

For assumption checking we first want to verify a linear relationship exists, looking at our regression plot we do see a clear linear grouping between abdomen circumference and BMI. Next we want to check both the independence of our errors vs. BMI along with equal variance of our residuals. Both of these can be done by looking at our residuals vs. fitted plot in which we see no clear pattern or correlation across the fitted value range which indicates both assumptions hold. Finally we check to ensure our errors are normally distributed, this is done by looking at the normal probability plot (Q-Q) in which we see a strong linear trend with very minimal skewing towards either end. This is expected with real data and we can assume normality.



Conclusions:

Our model walks the line between statistical significance and simplicity in application, taking a singular easily taken body measurement of abdomen circumference to give a more accurate BMI reading. With the strength of this model being in simplicity it does have some weaknesses, mainly in handling extreme cases. In average cases BMI and abdomen circumference will be linearly related. However, as BMI increases to extreme levels (>50%) this linearity begins to falter, thus in application, users should take a more complex approach when handling extreme BMI cases. But in general cases, our simple model using only abdomen circumference can be used without a significant decrease in BMI predictive power.

Module 2 Executive Summary

Contribution

1) Amy Qin wrote the data processing and methodology parts of the executive summary, wrote the code related to data cleaning and model fitting, produced each plot shown in the summary, and worked on slides .
2) Chao-sheng designed and implemented the shiny app, structured the Github repo, wrote the README, and suggested using a single variable model
3) Max wrote and edited the Executive Summary, helped with final predictor choice analysis, and created the presentation

References

Durnin, J. V., & Rahaman, M. M. (1967). The assessment of the amount of fat in the human body from measurements of skinfold thickness. *British Journal of Nutrition*, *21*(3), 681–689.