Executive Summary

Our group aims to predict body fat percentage based on relative body measurements. We want to design a prediction model such that people without Statistics expertise can use it well. Supposedly, a user can input his or her body measurements, and the model will accurately predict this user's body fat percentage.

We used the Bodyfat Dataset in the 1970s provided by STAT 628 as our only dataset. It includes 252 people with measurements such as Age, Bodyfat Percentage, etc. After examining the dataset, we decided to delete Index 172, 182, and change the Height of Index 42. For Index 172, it has an abnormal Bodyfat Percentage of 1.9%. For Index 42, we used the formula between Adiposity, Weight, and Height[1] to find that its measurements did not align with each other. We used the formula to recover its Height 69.5. For Index 182, it has a Bodyfat Percentage of 0, and the formula[2] gives a negative percentage, so Index 182 was deleted. For Index 39 and index 219, their Weight and Bodyfat Percentage were higher than any other data entries, and we considered them harmful to our prediction model. They were both deleted.

Our final model is a weighted linear regression model:

Bodyfat Percentage = -40 + Abdomen Circumference * 0.635 (+/-7.25)

For example, a man with Abdomen Circumference 100 (cm) will have a bodyfat percentage of 23.5%, and his 95% prediction interval is between 16.25% and 30.75%. Our estimated coefficients are -40 and 0.635, which are in the units of 'Percentage' and 'Percentage per cm'. This means that for every centimeter increase in Abdomen Circumference, the model predicts that body fat Percentage will increase, on average, by 0.635%; the (+/-7.25) part is from 95% prediction interval, meaning that we are 95% confident that a user's true Bodyfat percentage will fall in range of our prediction +/- 7.25%.

We chose the weighted linear regression model to account for the time difference between the dataset (1970s) and today (2023). The age distribution of the American male population shifted over the decades, which potentially impacted the robustness of a normal linear regression model. So, we found a new dataset regarding the 2022 American male population by age.[3] For each age group, its weight was based on the percentage of that group in both 1970s dataset and 2022 dataset. The weighted 1970s percentages should be the same as 2022 percentages. Regarding the weighted linear regression, we believed that for each sample in each 1970s age group, its impact to the model was the same as if the age group were to have a 2022 percentage.

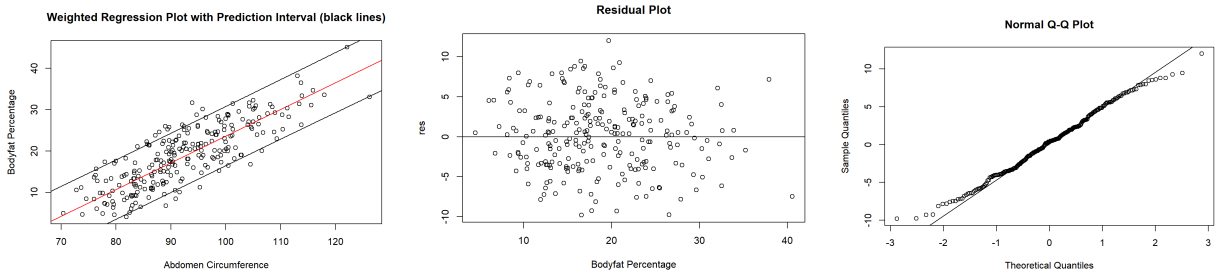
Our model only uses one independent variable: Abdomen Circumference. The R-squared value is 0.72, suggesting that 72% of variability in the dataset can be explained by the regression model. We produced the following 3 plots to analyze the model:

---

[1] BMI Calculation Formula Explained(https://www.registerednursern.com/bmi-calculation-formula-explained/)

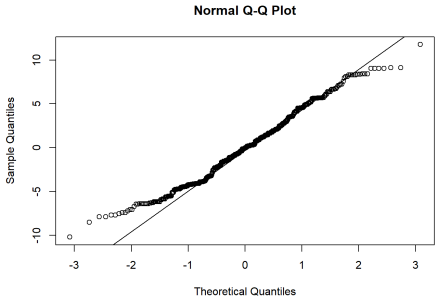
[2] Siri, W.E. (1956), "Gross composition of the body", in _Advances in Biological and Medical Physics_, vol. IV.

[3] Resident population of the United States by sex and age as of July 1, 2022*(in millions)*



The first plot shows the red regression line and black prediction interval. We can see that most data points fall inside the prediction interval. The second plot shows the residuals, the black line being 0 residual for reference. The residuals appear random, which suggests that our model does not miss any trend in the dataset. The third plot is QQ-plot with standard normal distribution. The points are mostly on the diagonal line, suggesting that residuals follow a standard normal distribution. This further proves the robustness of our model.

Our main alternative idea was to use "restricted" bootstrapping to simulate a 2022 age distribution dataset. We first tried to randomly select samples from the 1970s dataset, until the age distribution became similar to 2022 age distribution. But it was computationally impossible (1 million tries, sample size 20, 5 age groups yielded no result). Then we tried to sample each age group separately. It guaranteed 2022 age distribution at the cost of being less random. We found a maximum sample size 97 that satisfied the 2022 age distribution. We explored different models with this bootstrapping dataset.



If we use Abdomen Circumference as the only independent variable, R-squared value is 0.72; if we use all measurements from the dataset, R-squared value is 0.79, not significant enough to guarantee its complexity. We also tried to sample 5 times and combine the datasets of sample size 97, so we can get more information out of the 1970s dataset. The Abdomen Circumference model produced R-squared value 0.7, but the QQ-plot showed that the model did not work well with edge cases.

The main strength of our weighted linear regression model is simplicity. Other than the highest R-squared value, the model asks the user for one measurement only and gives easy to interpret outputs: a Bodyfat Percentage and a prediction range. Also, the weights account for age distribution differences in a simple way. The main problem with weighted regression was that we could not check the model's performance against a 2022 age distribution. Although our restricted bootstrapping method can simulate a 2022 age distribution, its performance was not good enough to be our final choice.

All things considered, we believe that our weighted linear regression model is the best model to predict Bodyfat Percentage. During our exploration, we thought about many ideas, but the small dataset became a major limitation. With only 252 samples, many statistical methods became impractical. If given a larger sample size, we can produce a more accurate model.

References:

1.BMI Calculation Formula Explained
(https://www.registerednursern.com/bmi-calculation-formula-explained/)

2.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.

3.Resident population of the United States by sex and age as of July 1, 2022*(in millions)*
(https://www.statista.com/statistics/241488/population-of-the-us-by-sex-and-age/)

Contributions	Ruofeng Tang	Xiaoyang Wang	Xingyu Tang
Presentation	Reviewed/edited slide 5-8 (bootstrap model).	Reviewed/edited and provided feedback on all slides.	Made all slides
Summary	Wrote the whole document.	Reviewed/edited and provided feedback on the whole document.	Reviewed/edited and provided feedback on the whole document.
Code	Explored data cleaning Responsible for Rmd file and bootstrap model	Explored data cleaning Reviewed/edited and provided feedback on all codes	Explored data cleaning Responsible for weighted linear regression model
Shiny App	Reviewed/edited and provided feedback on Shiny app	Responsible for Shiny app	Reviewed/edited and provided feedback on Shiny app