# Body Fat Project Executive Summary

## Introduction and Motivation

The goal of this analysis is to develop a predictive model for body fat percentage using body measurable variables. Accurately predicting body fat percentage is crucial for assessing health risks and determining fitness levels. By using simple-to-measure variables, we aim to create a reliable model that can be applied in health and fitness assessments.

## Data Cleaning Process

The raw data contains a few potential issues, such as outliers and measurement inaccuracies. To ensure the model's reliability, we performed exploratory data analysis on the BodyFat dataset, identified 3 suspect variables from 17, and used box plots for a more intuitive analysis of these variables, several data cleaning steps were performed:

1. A subject with a BMI greater than 40 (45.1) was identified and removed as an outlier. Extremely high BMI values can skew the regression results, leading to distorted predictions. By eliminating this data point, the model's generalizability to more typical observations is improved.
2. A subject had an abnormally low height measurement (29.5''), which was likely due to a data entry error. To correct this, the subject's height was recalculated using their known BMI and weight.
3. Body fat percentages below 7% were considered unrealistic for a general population. To address this, we fitted a linear regression model using other variables with body fat percentages greater than or equal to 7%. The fitted model was then used to re-estimate the body fat percentages.

## Final Model Selection and Rationale

After examining several models, we finalized the following predictive model for body fat percentage:

$$BODYFAT = -6.94 + 0.676*ABDOMEN - 2.135*WRIST + 0.06*AGE$$

This model was chosen based on a consideration of statistical significance, model simplicity, and predictive power. By including ABDOMEN and WRIST, two easily measurable variables with strong correlations to body fat, and adding AGE to improve model performance, we achieve a balance between accuracy and practical applicability.

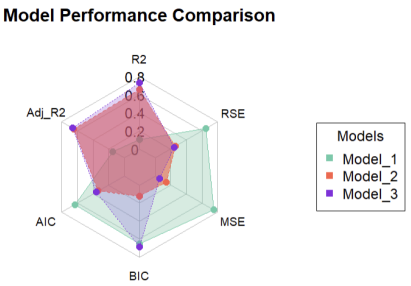
## Statistical Analysis and Key Metrics

Multiple Linear Regression was used to model the relationship between body fat percentage and the predictor variables. We evaluated different models, including combinations of circumference measurement variables but ultimately found the three variables mentioned above to be the most significant contributors. Below are some key metrics of our model:

1. $R^2 = 0.730$: This indicates that approximately 73% of the variation in body fat percentage is explained by the final model.
2. Adjusted $R^2 = 0.727$: After adjusting for the number of predictors, the model still retains strong explanatory power.
3. Residual Standard Error (RSE) = 3.76: The typical difference between the observed and predicted values of body fat percentage is about 3.76%.
4. P-values for ABDOMEN and WRIST were both highly significant ($p < 0.001$), indicating that these variables are strong predictors. AGE had a smaller, but still meaningful contribution ($p < 0.01$).

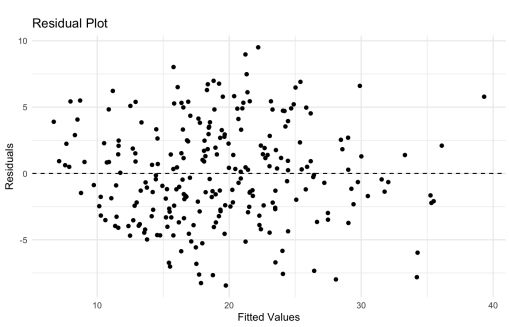
## Model Comparison

We compared the final model with a simpler model using only ABDOMEN and WRIST, and a more complex model using additional variables (CHEST, HIP, ABDOMEN, and WRIST). The final model performed better based on adjusted $R^2$ and residual error, striking a balance between simplicity and accuracy. As the radar chart shown below:




From the radar chart, we see that Model_2 performs best across several key metrics. It has the lower AIC, BIC, RSE, and MSE, indicating it offers the best balance between accuracy and simplicity. Model_3, which includes four circumference measures, slightly improves $R^2$ and Adjusted $R^2$, but the gain is minimal compared to the added complexity. The higher AIC and BIC values further show that the added complexity of Model_3 is not justified by the small improvement in explained variance. Model_1, which excludes AGE, performs worse overall with higher error metrics, validating AGE's importance in improving the model. In summary, Model_2 is the most suitable choice, offering near-optimal explanatory power with fewer predictors.

**Model Diagnostics and Assumptions**





We conducted diagnostic checks to ensure the validity of our linear regression assumptions. The residuals were roughly normally distributed, as confirmed by the residual plot. The residual vs. fitted values plot showed no clear pattern, suggesting that the variance of the residuals is constant (i.e., homoscedasticity). The QQ plot shows the model residuals appear to be approximately normally distributed in the middle of the data, but there may be some non-normality or outliers in the extremes. This could suggest that while the model is good for most predictions, it might struggle with extreme values.

**Strengths and Weaknesses of the Model**

The final model is simple and interpretable, using only three variables that are easy to measure and significantly associated with body fat percentage. However, the model may not fully capture more complex, nonlinear relationships between the predictors and body fat percentage, which could lead to underfitting in some cases. In summary, This model provides a strong balance between simplicity and predictive power, and its performance is supported by sound statistical evidence and diagnostic checks. Future iterations could explore adding non-linear terms or interaction effects to further refine predictions.

**Conclusion**

The final model effectively balances simplicity and accuracy, using three key variables — abdominal, wrist circumference, and age — to explain 73% of the variation in body fat percentage. While it performs well for most predictions, it may have limitations with extreme values. Overall, this model is practical for real-world use and could be improved by exploring more complex relationships.

**Contributions**

| Parts | Hengyu Yang | Yi Ma | Leyan Sun | Tianle Qiu |
|---|---|---|---|---|
| Presentation | Review data clean and give ideas | Edit format and give ideas | Make ppt of our thinking and modeling, and tabulate the results in the code. | Some advice |
| Summary | Edit format, add lost parts | Add some plots and edit format | Check and add some parts | Responsible |
| Code | Maintainer of GitHub repo, rewrite model selection, plot generation | Initial version of analysis code | Data cleansing | Model diagnostic |
| Shiny App | Responsible | First edition of shiny app | Format adjustment | Format adjustment |