# Body Fat Regression Model by L10G01

**510428929, 530675262, 520050392, 530802983, and Abdullah Fashola**

This version was compiled on November 3, 2024

**Our aim was to find a regression model to predict body fat percentage. Candidate models were produced via Forward and Backwards stepwise searches using Akaike information criterion (AIC), and exhaustive searches using both AIC and the Bayes information criteria (BIC). The Backward AIC model exhibited the lowest Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and error range. It explained 70.4% of the variance in body fat ($R^2$ = 0.704), with an AIC of 1339.243 indicating a strong fit. Abdomen circumference emerged as a key predictor, with a 1 cm increase leading to a 0.86% increase in body fat.**

**Introduction.** The dataset includes male body fat data among other measurements. Body fat percentage is an important indicator of health, with higher levels increasing the risk of diseases such as coronary artery disease to cancer (Dix, 2019). Body fat also impacts physical appearance, with research suggesting lower levels are more attractive (Brierley *et al.*, 2016), and, as a result, are likely to experience greater workplace and personal success (Henson, 2022).

Thus, knowing one's body fat percentage is pivotal to both health and success. However, conventional measuring methods, including X-rays, and Hydrostatic Weighing (Tinsley, 2023) are expensive and time consuming. Therefore, to empower the average man to obtain this vital measurement more conveniently, we present a regression model that offers a convenient, accurate estimation using simple, easily-obtained body measurements.

**Data set.** Our analysis used the SOCR Body Composition dataset, which includes body measurements of 252 male individuals. The data were provided by Dr. A. Garth Fisher for non-commercial purposes. The dataset contains 14 continuous numerical variables: Density (mean=1.06, sd=0.0190), Age (mean=44.9, sd=12.6), Weight (mean=81.2, sd=13.3), Height (mean=178, sd=9.30), circumferences of the neck, chest, abdomen, hip, thigh, knee, ankle, extended biceps, forearm, wrist,and one ratio variable Percent Body Fat (mean=19.2, sd=8.37), calculated based on Siri's equation (Siri, 1956).

Volumes and weights were measured using underwater weighing, a method based on Archimedes' principle that obtains precise data through water displacement. The volumes and weights measured underwater can be used to calculate body fat percentage using the Siri equation (Siri, 1956). Various body circumference measurements were also conducted according to the standard procedures described in Behnke and Wilmore (Behnke and Wilmore, 1974). In the dataset, lengths are uniformly measured in centimeters, weight in kilograms, density in g/cm$^3$, and body fat percentage is expressed as a percentage.These data are ideal for studying how to estimate body fat by body circumference, can help us predict health risks, and help doctors design wellness programs .Additionally, this dataset only includes male participants,which presents some limitations in gender diversity.

**Analysis.**

***Data cleaning.*** Data cleaning was performed on the dataset through several steps to ensure accuracy and reliability. Impossible values, such as 0% body fat, were manually removed to eliminate invalid entries. The body density variable was dropped as it did not align with the goal of developing a cost-effective model with easy-to-measure predictors. Outliers were identified and removed using the Interquartile Range method. Any data points lying outside 1.5 times the IQR above the third quartile or below the first quartile for any variable were excluded. This process reduced the dataset from 250 to 232 entries. By eliminating extreme physical values that did not reflect the general population, the performance of the final model was improved.

***Data visualization.*** In order to explore the dataset and understand the relationships between variables, we conducted comprehensive data visualization analysis using the `ggplot2` package. First, we performed univariate analysis by plotting histograms for each variable and overlaying density curves to examine the distribution of the data. This helped us identify central tendencies, variability, and any skewness present. To our surprise, we found that, except for the Age variable and the multimodal distributions of Neck Circumference and Ankle Circumference, most variables exhibited approximately normal distributions.

Next, we carried out bivariate analysis by plotting scatter plots between body fat percentage and each independent variable to identify potential linear relationships. We discovered that Abdomen Circumference and Chest Circumference showed strong positive linear relationships with body fat percentage and are key predictor variables. Weight was also positively correlated with body fat percentage but exhibited some dispersion.

Finally, we calculated the Pearson correlation coefficients between variables and visualized them using a heatmap. Abdomen Circumference and Chest Circumference appeared in deep red, indicating a strong positive correlation. The Age variable showed negative or weak correlations with other variables.

***Full model assumption checking.*** To prepare for regression, we evaluated the linearity between predictors and the dependent variable body fat percentage Appendix A. Most predictors followed approximate linear relationships with body fat, but height and ankle circumference appeared randomly distributed without a linear trend. Logarithm transformations did not improve this, hence these variables were excluded for better model feasibility.

***Candidate models selection.*** In this research, we conducted four model selection procedures to identify the best predictive model for the cleaned dataset: forward and backward stepwise searches using the Akaike Information Criterion (AIC) and exhaustive searches using both AIC and Bayesian Information Criterion (BIC).

We began with forward and backward stepwise searches on the null and full models, respectively, using AIC as the penalty. These searches produced two distinct candidate models. To further refine our selection, we conducted exhaustive searches using AIC and BIC. The exhaustive AIC search resulted in the same model as the backward stepwise method.

The three final models were: the forward AIC model (predictors: Abdomen2Circumf, Weight, WristCircumf, ExtendBicepsCircumf); the backward and exhaustive AIC model (Age, Weight, Abdomen2Circumf, ThighCircumf, WristCircumf); and the exhaustive BIC model (Weight, Abdomen2Circumf, WristCircumf).

***Final model selection.*** In the final model selection process, we applied 10-fold cross-validation to evaluate the predictive accuracy of the three candidate models. This technique involves dividing the dataset into ten parts, training the model on nine parts, and testing it on the remaining part, repeated across all ten folds. We assessed the models by comparing their Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), where lower MAE and RMSE values indicate better accuracy, and a smaller range suggests greater stability. A plot of MAE and RMSE values from the cross-validation process illustrated the range of errors, showing from the highest to the lowest mean values.

The Backward AIC model was chosen as the final model due to its lowest mean MAE and RMSE values and minimal error range Appendix B, demonstrating better accuracy and consistency compared to other candidates.

***Final model assumption checking.***

- Linearity: In the residual plot (the left plot in Fig.1), the residuals are symmetrically distributed above and below 0, and no curved pattern is shown.
- Independence: Due to the nature of the data collection in this experiment, the errors were independent.
- Homoskedasticity: The spread of errors is fairly constant over the range of fitted values in the residual plot in Fig.1.
- Normality: In the QQ plot (the right plot in Fig.1), the majority of the scatters are clustered reasonably tight to the diagonal line. In addition, with a sufficiently large data size, the Central Limit Theorem ensures that the distribution of the estimated coefficients tends to be normal.

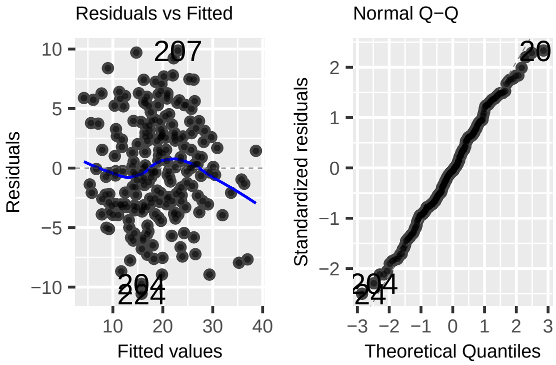

**Fig. 1.** Residual plot and QQ plot

**Results.** Through our analysis, we developed a linear regression model to predict body fat percentage using the following five predictors:

$$\begin{aligned}\text{Body Fat} = & -30.07\\ & +0.86 \times \text{Abdomen Circumference}\\ & -0.21 \times \text{Weight}\\ & -1.74 \times \text{Wrist Circumference}\\ & +0.06 \times \text{Age}\\ & +0.27 \times \text{Thigh Circumference}\end{aligned} \tag{1}$$

The intercept, although not meaningful by itself, anchors the model, while the individual coefficients provide insights into how each variable contributes to body fat levels. Abdomen circumference is a key predictor, with a 1 cm increase leading to a 0.86% increase in body fat. This highlights the central role of abdominal fat in overall body fat percentage. Weight has a negative coefficient, suggesting that for a given abdomen circumference, an increase in weight may be associated with more lean mass rather than fat. This counterintuitive result indicates that weight alone, without any other physical contexts, is not a straightforward indicator of body fat percentage. Age has a small effect on the model with each additional year increasing body fat by 0.06%. This provides insight into the gradual increase of body fat that tends to occur with age due to metabolic changes. Wrist has an extremely strong negative association, potentially reflecting the impact of body frame size on fat distribution. It likely serves as a proxy for overall body frame and bone structure, with larger wrists being associated with leaner body compositions, holding the other variables constant. The positive thigh circumference coefficient may capture the fat stored in the lower body.

The final model demonstrates strong performance, explaining 70.4% of the variance in body fat ($R^2 = 0.704$) with low out of sample prediction errors (RMSE = 4.272, MAE = 3.524). AIC (1339.243) suggests a good fit, and similar out-of-sample $R^2$ (0.708) indicates reliable predictive power on new data.

**Discussion and conclusion.** A key limitation of our analysis is the potential multicollinearity among predictors. This is indicated in the model by the unexpected negative coefficient for weight that may distort the coefficient estimates, making it more difficult to interpret the true impact of each variable on body fat percentage. Additionally, the model's assumption of linearity may not fully capture complex, non-linear relationships between variables. The moderate prediction errors (RMSE = 4.272, MAE = 3.524) suggest that while the model is adequate for general predictions, it may lack the precision required for clinical or medical applications.

To address these limitations in future research, we plan to reduce the effect of multicollinearity by removing or combining highly correlated variables and exploring new uncorrelated variables. Exploring non-linear models or including interaction terms could better represent the underlying relationships. Further, we intend to enhance the dataset with a larger and more diverse sample to improve the models' generalisability. These can both be implemented by incorporating additional relevant body fat predictors variables like physical activity, dietary habits and genetic analysis.

Despite its limitations, the model effectively identifies key predictors of body fat percentage, with abdomen circumference emerging as a significant positive contributor. The model explains a substantial portion of the variance in body fat percentage ($R^2 = 0.704$) and demonstrates consistent predictive performance on new data (out-of-sample $R^2 = 0.708$). Refining the model by addressing multicollinearity and expanding the scope of variables will enhance its precision and utility, particularly for applications requiring high accuracy in body fat assessment.

## Appendix A

Relationship of predictors against the dependent variable

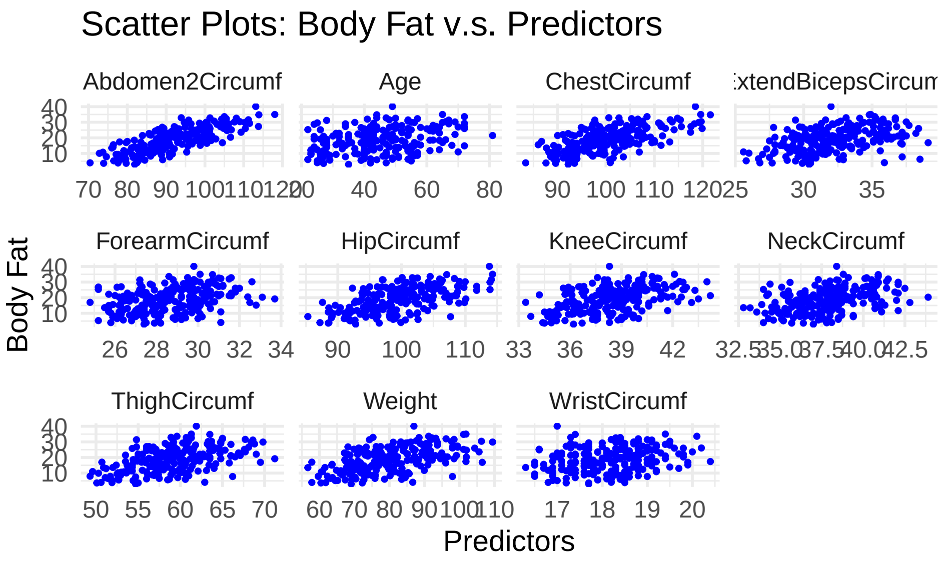

**Fig. 2.** Scatter plot: Body Fat v.s. Predictors

## Appendix B

MAE and RMSE comparison of the candidate models

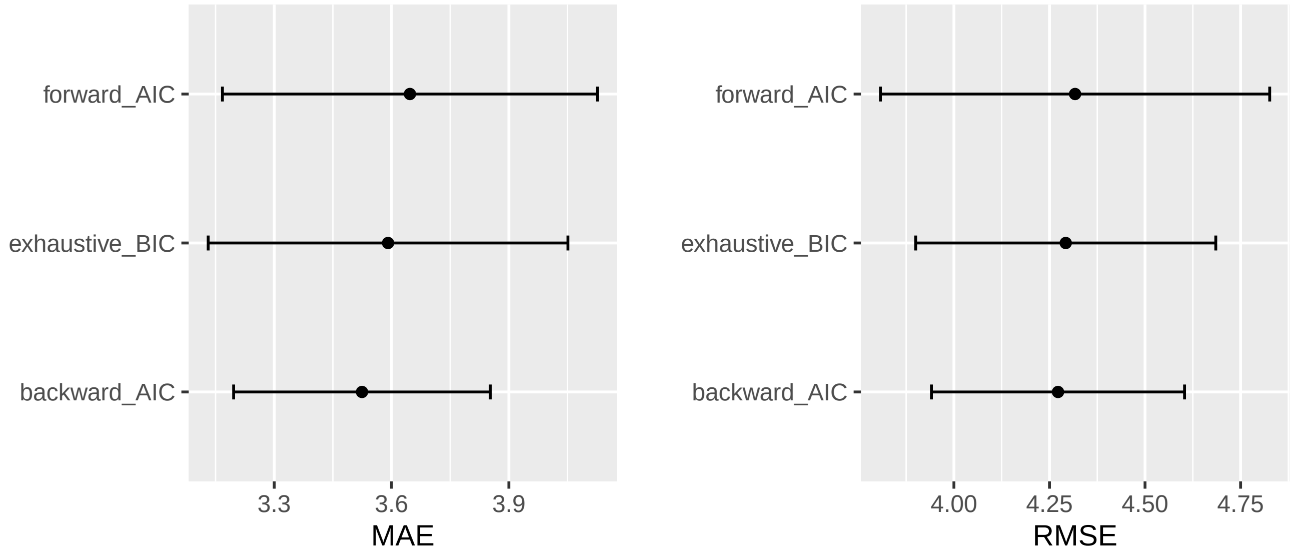

**Fig. 3.** MAE Comparisson

## References

Behnke A, Wilmore J (1974). *Evaluation and Regulation of Body Build and Composition.* Prentice-Hall, Englewood Cliffs, N.J.

Brierley ME, Brooks KR, Mond J, Stevenson RJ, Stephen ID (2016). "The Body and the Beautiful: Health, Attractiveness and Body Composition in Men's and Women's Bodies." *PLOS ONE*, **11**(6), e0156722. `doi:10.1371/journal.pone.0156722`. Accessed: 2024-11-03, URL https://pmc.ncbi.nlm.nih.gov/articles/PMC4892674/.

Dix M (2019). "Types of Body Fat: Benefits, Dangers, and More." Accessed: 2024-11-03, URL https://www.healthline.com/health/types-of-body-fat.

Henson D (2022). "What's the Price of Pretty Privilege?" Accessed: 2024-11-03, URL https://bond.edu.au/news/whats-price-of-pretty-privilege.

Siri W (1956). *Gross Composition of the Body*, volume IV. Academic Press, Inc., New York.

Tinsley G (2023). "The 10 Best Ways to Measure Your Body Fat Percentage." Accessed: 2024-11-03, URL https://www.healthline.com/nutrition/ways-to-measure-body-fat#TOC_TITLE_HDR_5.