Final Project

Yixuan Jiao, Landi Guo, Fengdi Zhang

2022-12-13

Abstract

In this study, a multiple linear regression model is developed to predict body fat in men using scale and tape measurements. Candidate models are created using a dataset of 252 male subjects and are generated through the application of automatic procedures, criterion based procedures, and LASSO. Interactions between variables are also considered through two-way ANOVA. The final model chosen, based on its root mean square error, included age, height, neck circumference, abdomen circumference, forearm circumference, wrist circumference, and the interaction between neck and abdomen as predictors. These measurements are simple and widely available, making the model a useful tool for predicting body fat in a convenient manner.

Introduction

Although fat is an important source of stored energy, excessive amounts of body fat have been linked to various health issues, including type 2 diabetes, heart disease, and stroke[1]. According to widely accepted standards, an adult male with more than 25% body fat and an adult female with more than 32% body fat is considered obese[2]. The National Health and Nutrition Examination Survey from 2017-2020 found that 41.9% of adults in the United States have obesity, with rates over 35% in 19 states[3]. Given the prevalence of obesity, it is important to have accurate and convenient methods for measuring body fat. However, measuring body fat can be challenging. This project aims to develop a multiple linear regression model using scale and tape measurements to predict body fat in men, using a dataset of 252 male subjects with data on body fat percentage, age, height, weight, and other body circumference measurements.

Methods

Exploratory analysis is conducted to check for patterns, distributions, and anomalies in the dataset. This dataset contains 252 observations which are all male, and 16 variables of interest. The variables include three measures of body fat (Brozek's, Siri's, and body density) as well as age, weight, height, and various body circumference measurements including neck, chest, abdomen, hip, thigh, knee, ankle, bicep, forearm, and wrist. In the following analysis, percent body fat using Brozek's equation is chosen as the outcome. Firstly, one observation with 0% body fat is removed from the dataset. The mean and range are summarized for the remaining observations(table2). Then marginal distributions for each variable and pairwise relationship between each pair of variables are plotted(fig1). The distributions for all variables are symmetric, so no transformation is necessary. To confirm the normality of `bodyfat_brozek`, formal Shapiro's test is conducted, which supports the conclusion based on the histogram (fig2)that the body fat percentage is normally distributed. The pairwise scatterplot shows that all variables are linearly correlated with body fat percentage. Additionally, there are many variables that are highly correlated with other variables, which warrants further investigation.

Variance inflation factor (VIF) is calculated for each variable to assess collinearity. Variables with VIF greater than 5 may have misleading coefficients due to high collinearity. Upon calculation, the variables `weight`, `hip`, `abdomen`, `chest`, and `thigh` have VIF greater than 5[(fig3a)]. However, according to the p-values of the complete linear model, only `abdomen` is significant. Therefore, all of these variables except `abdomen` are excluded. After recalculating the VIFs, no further collinearity is identified[(fig3b)]. Various model selection procedures are then applied to the remaining variables to generate candidate models, including automatic procedure (stepwise)[(table3)], criterion based procedure[(table4)] (Cp[(fig4a)] value and Adjusted R square[(fig4b)]), and LASSO ($\lambda = 0.1$)[(fig5)(table5)]. The interaction between main effects are also considered through a two-way ANOVA test. A 10-fold cross validation is used to compare the candidate models based on predictive ability and select a final "best" model. Diagnostic plots are also generated for comparison and final model selection.

Result

The candidate model from stepwise regression includes `age`, `height`, `neck`, `abdomen`, `forearm`, and `wrist` as predictors to predict `bodyfat_brozek`. The candidate models from LASSO and criterion approach include the same predictors as the model from stepwise regression, but also include an additional predictor, `bicep`. Among the predictors in the candidate models, two-way ANOVA test reveals interaction between neck and abdomen has a significant p-value of 0.000477 for model 1[(table7)] and a p-value of 0.000379 for model 2[(table8)], indicating that it should be included in the model at a significance level of 5%. As a result, the final candidate models are as follows:

- Candidate model 1: `bodyfat_brozek` ~ `age` + `height` + `neck` + `abdomen` + `forearm` + `wrist` + `neck:abdomen`
- Candidate model 2: `bodyfat_brozek` ~ `age` + `height` + `neck` + `abdomen` + `forearm` + `wrist` + `bicep` + `neck:abdomen`

The root mean squared error (RMSE) for 10-fold cross validation is 3.95 for model 1 and 3.98 for model 2[(table6)], indicating that model 1 has better predictivity than model 2. The adjusted R-squared for both models is 0.7385, which means that `bicep` does not add value to the model. Diagnosis plots also suggest that both models fit the underlying assumptions of linear regression. Based on these results, the final model chosen is candidate model 1[(table9)].

Discussion

Summary of the final model(Table8) for predicting body fat percentage:

Table 1: Model Summary

Parameter	Coefficient	Interpretation (Holding all other variable constant)
Intercept	-61.05	-
Age	0.05	As age goes up by 1 year, the predicted body fat percentage will increase by approximately 0.05%.

Parameter	Coefficient	Interpretation (Holding all other variable constant)
Height	-0.32	As height goes up by 1 inch, the predicted body fat percentage will decrease by approximately 0.32%.
Neck	1.39	As neck circumference goes up by 1 cm, the predicted body fat percentage will increase by approximately 1.39%.
Abdomen	1.48	As abdomen circumference goes up by 1 cm, the predicted body fat percentage will increase by approximately 1.48%.
Forearm	0.28	As forearm circumference goes up by 1 cm, the predicted body fat percentage will increase by approximately 1.28%.
Wrist	-1.61	As wrist circumference goes up by 1 cm, the predicted body fat percentage will decrease by approximately 1.61%.
Neck:abdomen	-0.02	The effect of neck circumference on body fat percentage is slightly weaker when the abdomen circumference is also high.

There are several limitations to the study that should be considered when interpreting these results. First, the model might not be as accurate as ideal because it only considers the interactions between two variables, as determined through an ANOVA test. This means that other potential predictor variables or interactions may have been overlooked and could potentially influence the final model. Additionally, our dataset consisted of only 252 men, which may not be representative of the larger population. As a result, the final model has limited usage and accuracy if applied to a different population of men. Another limitation of our study is the noticeable error on body fat percentage, where one man was recorded as having a body fat percentage of 0%. This raises concerns about the reliability of the other data points and could potentially impact the accuracy of the model. Given these limitations, there are several avenues for future research that could help to improve upon the findings of our study. One possibility would be to replicate the study in a larger and more diverse sample, including both male and female subjects. Doing so could help to confirm the generalizability of our findings and identify any gender-specific differences in the predictors of body fat percentage. Additionally, future research could consider a wider range of predictor variables and interactions in order to further improve the model.

Reference

1: Body fat. The Nutrition Source. (2022, August 2). Retrieved December 16, 2022, from https://www.hsph.harvard.edu/nutritionsource/healthy-weight/measuring-fat/

2: Neumann, K. D. (2022, December 9). How to reduce body fat. Forbes. Retrieved December 16, 2022, from https://www.forbes.com/health/body/how-to-reduce-body-fat/#:~:text=With%20that%20said%2C%20general%20body,boo

3: State of obesity 2022: Better Policies for a healthier america. tfah. (2022, September 27). Retrieved December 16, 2022, from https://www.tfah.org/report-details/state-of-obesity-2022/

Appendix

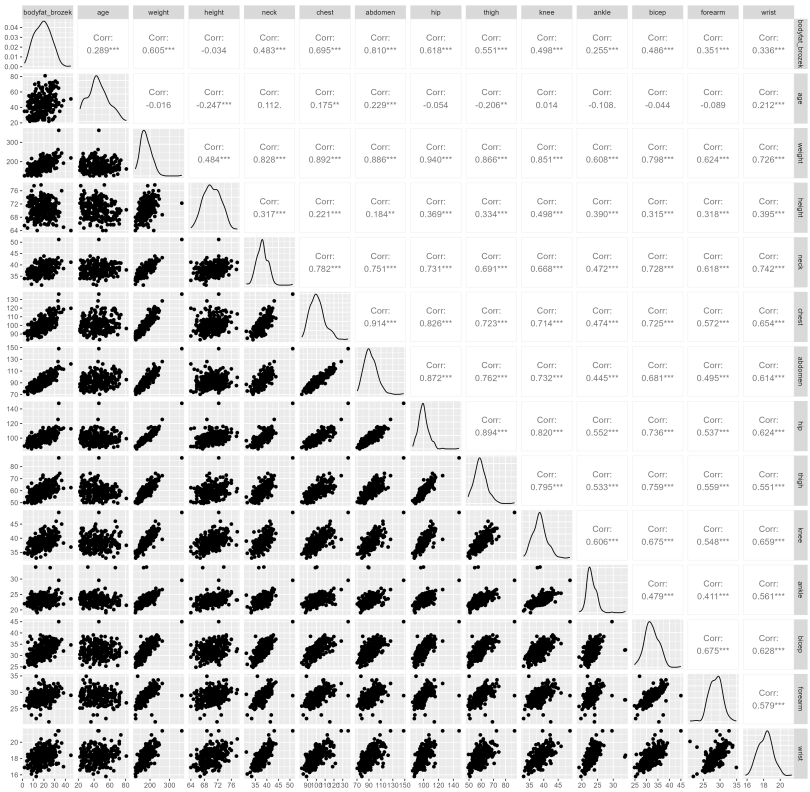


Fig1: Distributions for each Variable and Pairwise Relationship between each pair of Variables

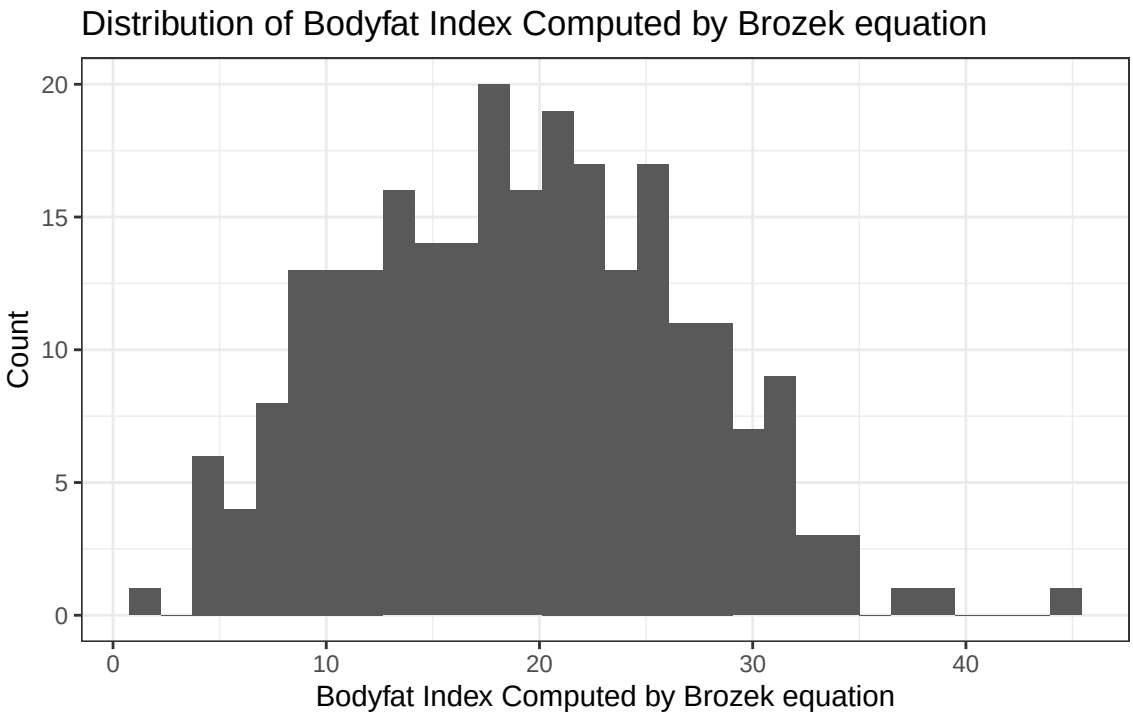


Fig2: Distribution of Bodyfat Index Computed by Brozek equation

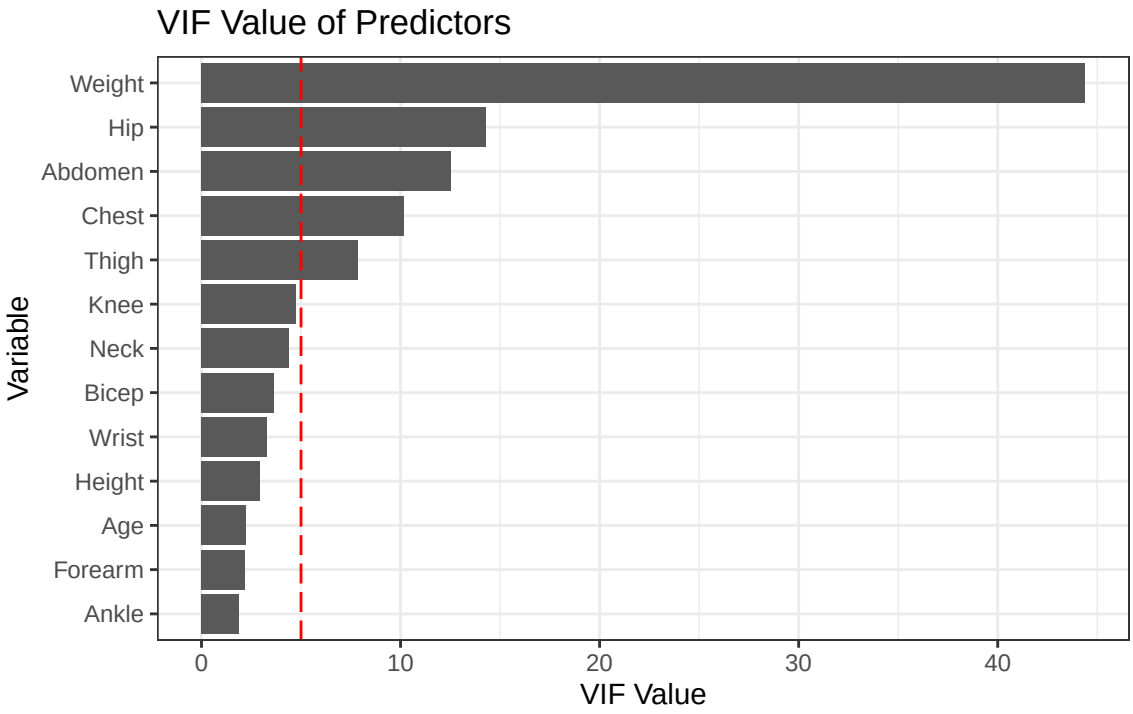


Fig3a: VIF Value of Predictors

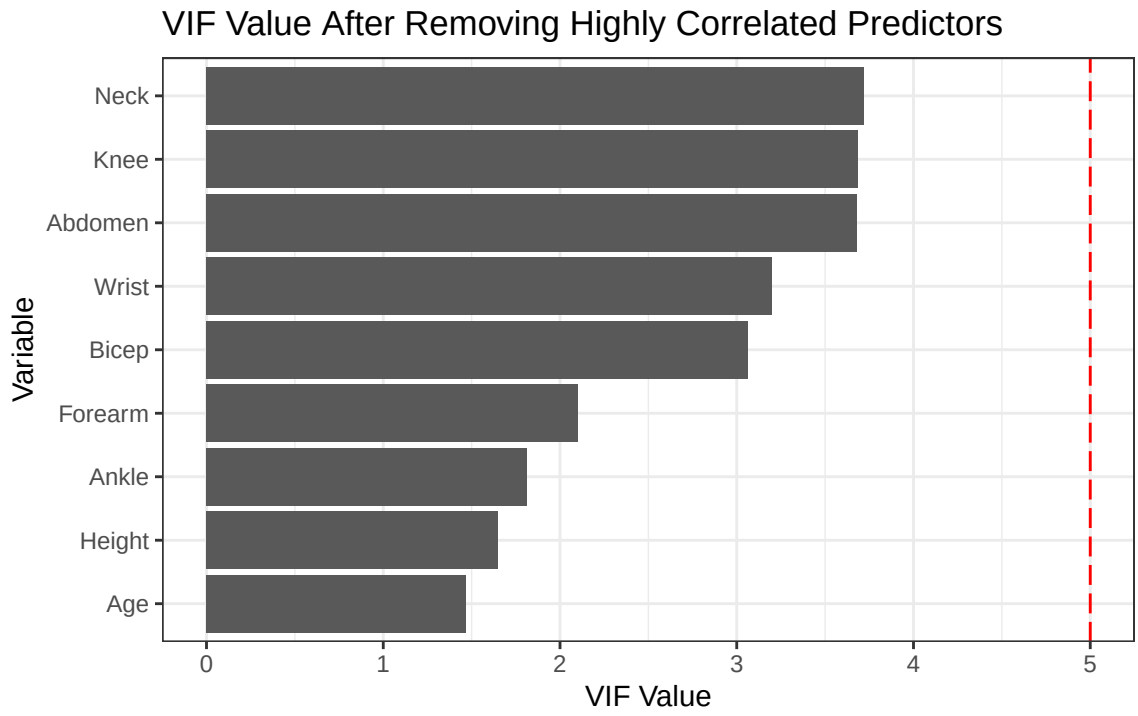


Fig3b: VIF Value After Removing Highly Correlated Predictors

Fig4a Cp Vs. Model Size

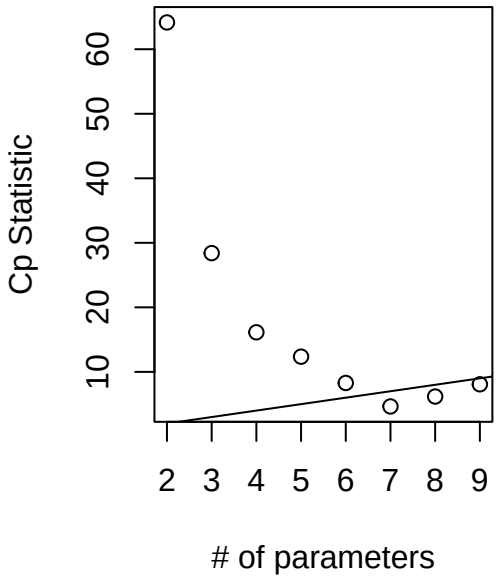
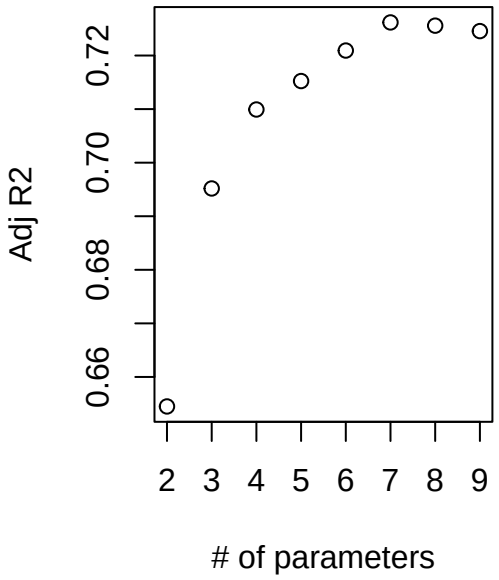


Fig4b Adj R2 Vs. Model Size



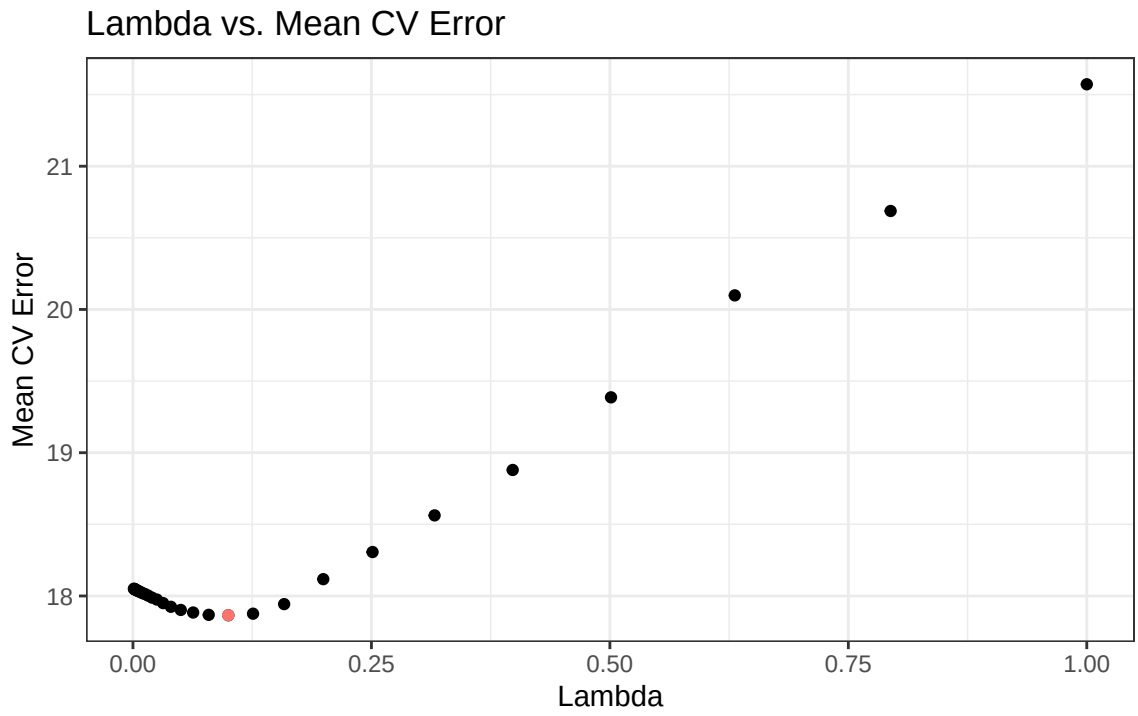


Fig5: Plot for Selecting the Optimal Lambda

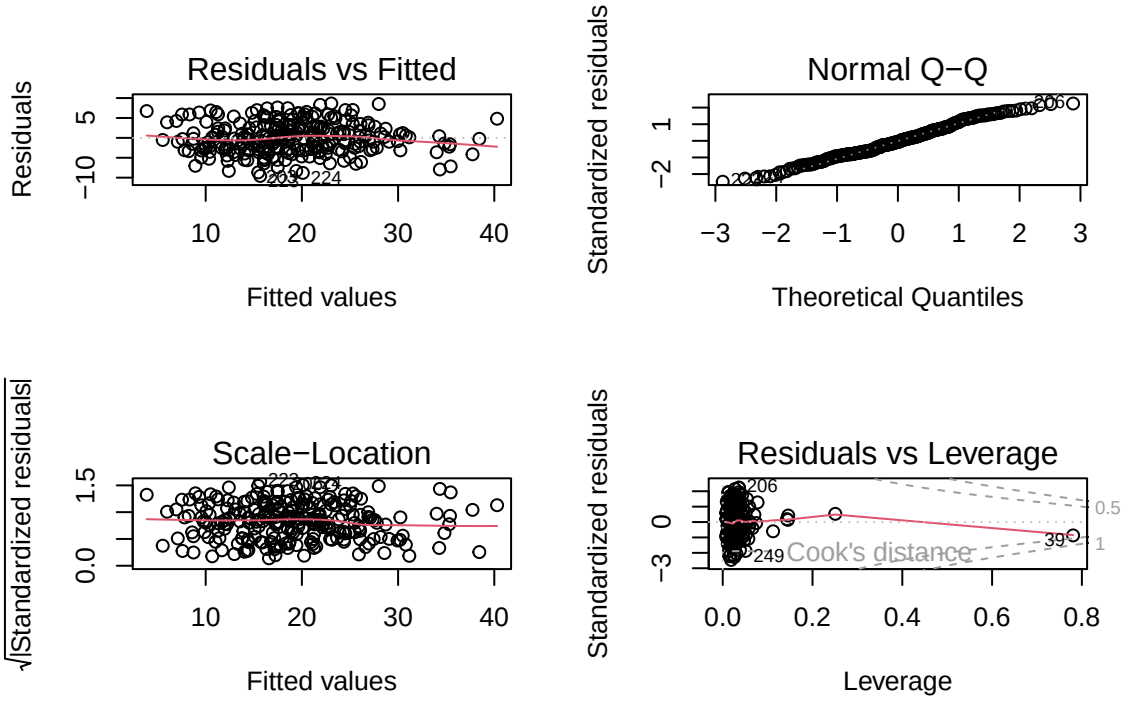


Table 2: Descriptive Statistics

Characteristic	**N = 251**
bodyfat_brozek	19 (13, 25)
age	43 (36, 54)
height	70.00 (68.38, 72.25)
neck	38.00 (36.40, 39.45)
abdomen	91 (85, 99)
knee	38.50 (37.05, 39.95)
ankle	22.80 (22.00, 24.00)
bicep	32.10 (30.25, 34.35)
forearm	28.70 (27.30, 30.00)
wrist	18.30 (17.60, 18.80)

Table 3: Model Parameter for Stepwise Procedure

Term	Estimate	Standard Error	Test Statisitcs	P-value
(Intercept)	7.21	7.70	0.94	0.34950
age	0.07	0.02	2.93	0.00371
height	-0.27	0.11	-2.39	0.01777
neck	-0.53	0.20	-2.69	0.00771
abdomen	0.71	0.04	19.17	0.00000
forearm	0.47	0.17	2.73	0.00685
wrist	-1.72	0.46	-3.77	0.00020

Table 4: Criterion Score For Different Model Sizes

Number of Parameter	Cp Value	Adjusted R Square
2	64.15	0.654
3	28.41	0.695
4	16.14	0.710
5	12.36	0.715
6	8.28	0.721
7	4.64	0.726
8	6.18	0.726
9	8.08	0.725
10	10.00	0.724

Table 5: Model Parameter for Lasso Fit

Term	Esitmate
(Intercept)	4.171
age	0.058
height	-0.266
neck	-0.359
abdomen	0.678
bicep	0.010
forearm	0.301
wrist	-1.488

Table 6: Validation Results for Candidate models

Model	RMSE	R-square	MAE
1	3.95	0.75	3.25
2	3.98	0.74	3.29

Table 7: Two-way ANOVA Test for Model Candidate 1

Term	SSE	MSE	F-stats	P-value
age	1230	1230	79.8	1.07e-16
height	22.3	22.3	1.45	0.229
neck	3270	3270	212	6.03e-35
abdomen	5980	5980	388	3.12e-52
forearm	54.6	54.6	3.55	0.0608
wrist	230	230	14.9	0.000144
bicep	7.52	7.52	0.489	0.485
neck:abdomen	200	200	13	0.000379

Table 8: Two-way ANOVA Test for Model Candidate 2

Term	SSE	MSE	F-stats	P-value
age	1230	1230	79.8	1.04e-16
height	22.3	22.3	1.45	0.229
neck	3270	3270	212	5.48e-35
abdomen	5980	5980	389	2.58e-52
forearm	54.6	54.6	3.55	0.0608
wrist	230	230	14.9	0.000144
neck:abdomen	193	193	12.5	0.000477

Table 9: Final Model Parameter

Term	Estimate	Standard Error	Test Statisitcs	P-value
(Intercept)	-61	20.7	-2.95	0.00348
age	0.0534	0.0231	2.31	0.0215
height	-0.316	0.112	-2.82	0.00518
neck	1.39	0.575	2.42	0.0164
abdomen	1.48	0.22	6.74	1.16e-10
forearm	0.279	0.176	1.58	0.115
wrist	-1.61	0.448	-3.59	0.000403
neck:abdomen	-0.0194	0.00548	-3.54	0.000477