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 essential part of our body and an important source of stored energy, excessive amounts of body fat is associated with type 2 diabetes, heart diseases, and stroke. More than 25 percent body fat is considered obese for adult male, while more than 32 percent body fat is considered obese for adult women. National data from the 2017-2020 National Health and Nutrition Examination Survey revealed that 41.9 percent of adults in the U.S. have obesity, and the adult obesity rate is over 35 percent in nineteen states. As obesity becomes one of the most common medical conditions in the U.S., it is important to find an accurate and easy way to find out one's body fat. Unfortunately, body fat is not always straightforward to measure. We are given a dataset containing percentage of body fat, age, height, weight, and other body circumference measurements for 252 men. How can we possibly estimate body fat for men in a more convenient way? This project aims to build a multiple linear regression model using scale and tape measurements to predict body fat for men.

Methods

Exploratory Analysis

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 first three variables are body fat measured in three different ways (Brozek's, Siri's, body density). The rest of the variables are age(years), weight(lbs), heights(inches), neck(neck circumference in cm), chest(chest circumference in cm), abdomen(abdomen circumference in cm), hip(hip circumference in cm), thigh(thigh circumference in cm), knee(knee circumference in cm), ankle(ankle circumference in cm), bicep(extended biceps circumference in cm), forearm(forearm circumference in cm), and wrist(wrist circumference in cm). These are simple measurements that can potentially be used to predict body fats.

In the rest of the analysis, percent body fat using Brozek's equation is chosen as the outcome. Firstly, one entry with 0 body fat is removed from the dataset. Mean and range are summarized for the remaining observations(Table1). Then marginal distributions for each variable and pairwise relationship between each pair of variables are plotted(Fig1). The distributions for all variables are symmetric, therefore no transformation is needed. To confirm the normality of `bodyfat_brozek`, formal Shapiro's test is conducted. The results of the test align with the histogram that `bodyfat_brozek` is normally distributed (Fig2). Pairwise scatterplot shows that all variables are linearly correlated with `bodyfat_brozek`. Additionally, there are many variables that are highly correlated with other variables, which require further investigation.

Model Building

Variance inflation factor (VIF) for each variable is calculated for checking collinearity. Variables with VIF > 5 suggest that the coefficients might be misleading due to collinearity and high collinearity. `weight`, `hip`, `abdomen`, `chest`, and `thigh` have VIF > 5, but according to the p-values of the complete linear model, only `abdomen` is significant. Therefore, all these variables except `abdomen` are excluded. Re-calculate the VIFs and no more collinearity is found. Different model selection procedures are conducted on the remaining variables to generate candidate models. Procedures include automatic procedure (stepwise), criterion based procedure (Cp value and Adjusted R^2), and LASSO. Interactions between main effects are considered by conducting 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 all the same predictors as the model from stepwise regression, but also include an extra predictor `bicep`. Among the predictors in the candidate models, two-way ANOVA test(table?) reveals interaction between neck and abdomen has significant p-value = 0.000477 for model 1 and p-value = 0.000379 for model 2, suggesting it should be included in the model under 5% significance. Therefore, the final candidate models are as following: 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 RMSE for 10-fold cross validation is 3.95 for model 1, and 3.98 for model 2. Therefore, model 1 has better predictivity than model 2. The adjusted R-squared for is 0.7385 for both model 1 and model 2, which means that `bicep` is not adding value to the model. Diagnosis plots(figure?) are also suggest that model 1 fits the underlying assumptions of linear regression better than model 2, because *add reasons*. In conclusion, the final model chosen is candidate model 1.

Appendix

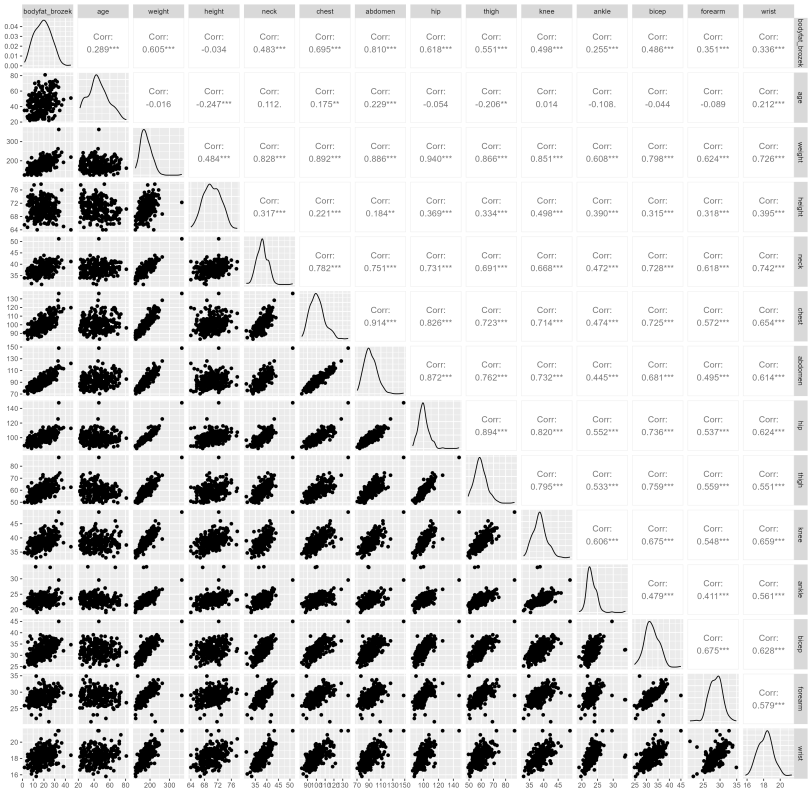


Fig1: Distributions for each Variable and Pairwise Relationship between each pair of Variables

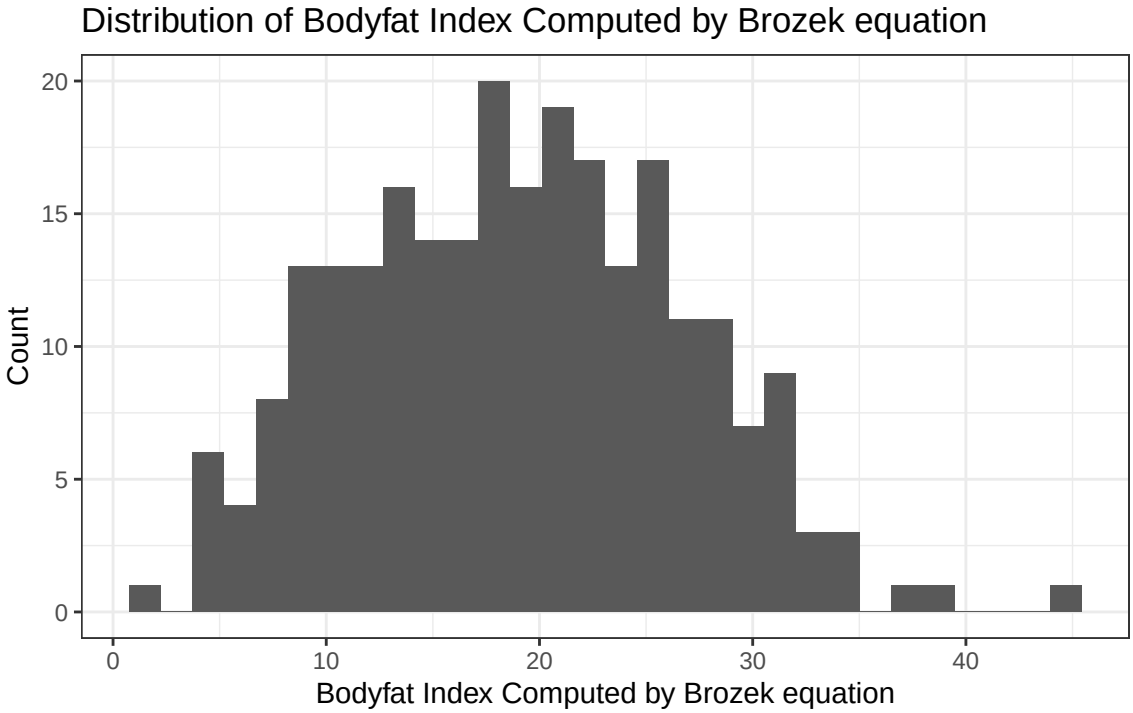


Fig2: Distribution of Bodyfat Index Computed by Brozek equation

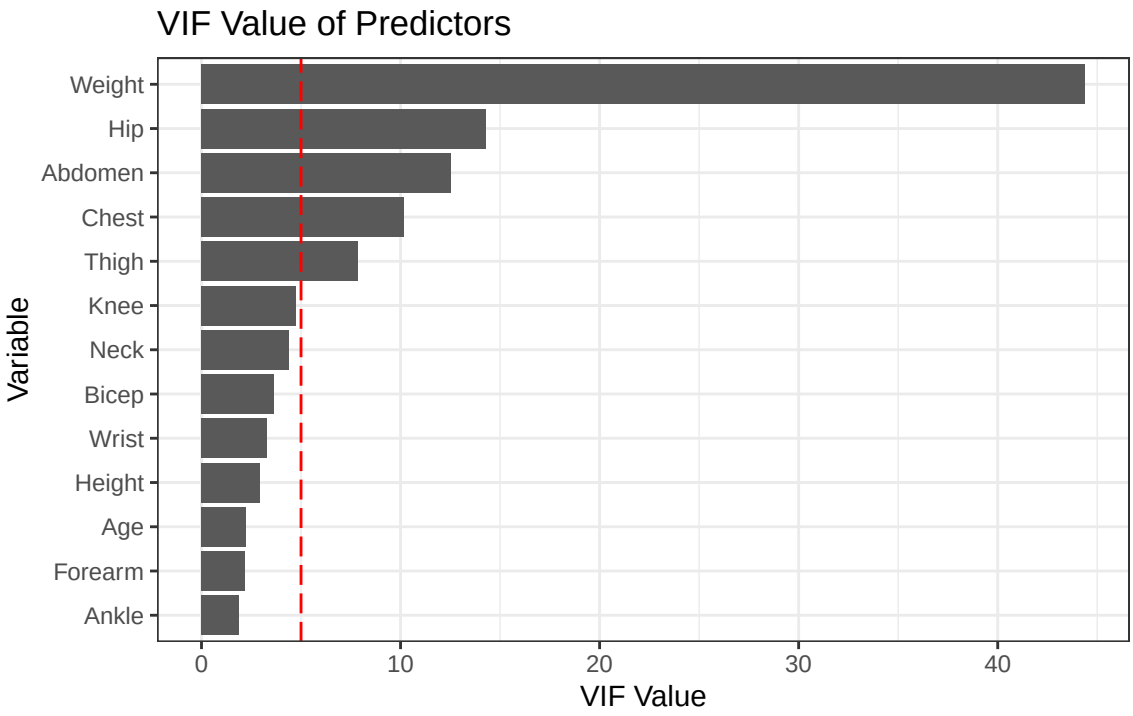


Fig3a: VIF Value of Predictors

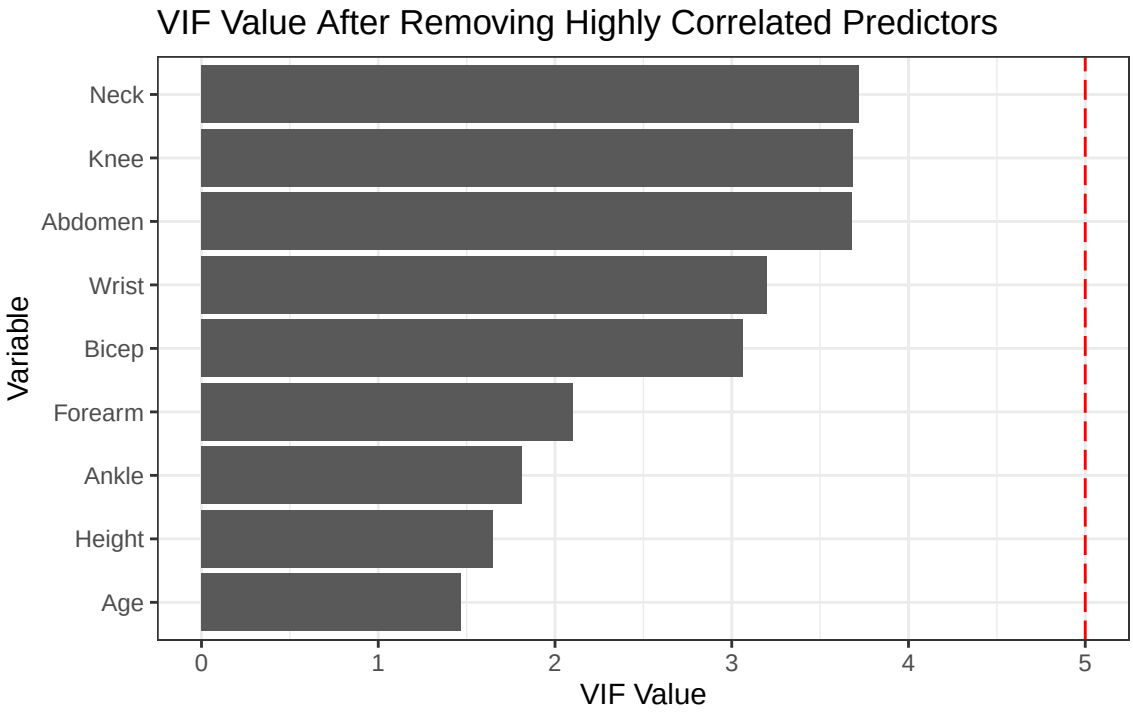


Fig3b: VIF Value After Removing Highly Correlated Predictors

Fig4a Cp Vs. Model Size

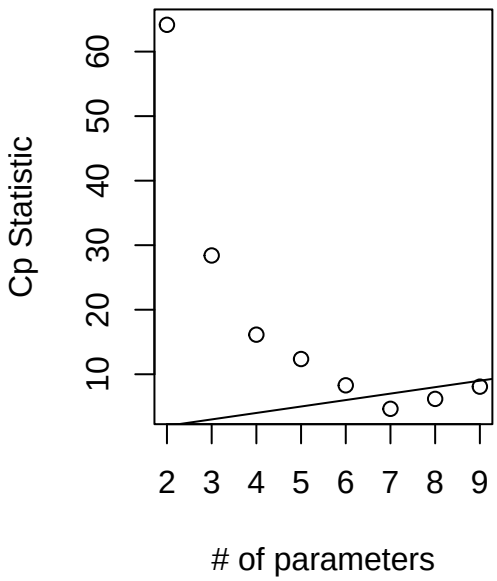
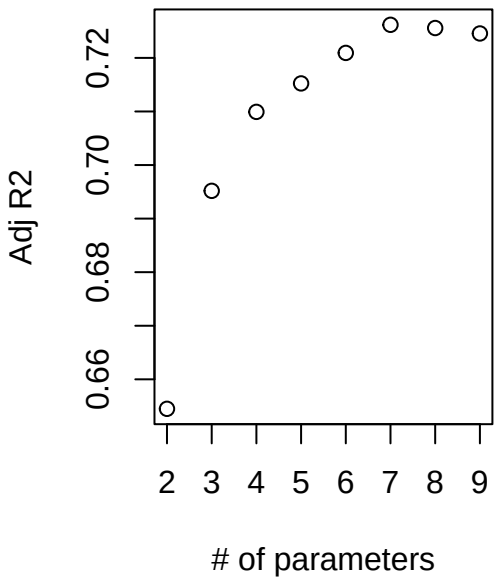


Fig4b Adj R2 Vs. Model Size



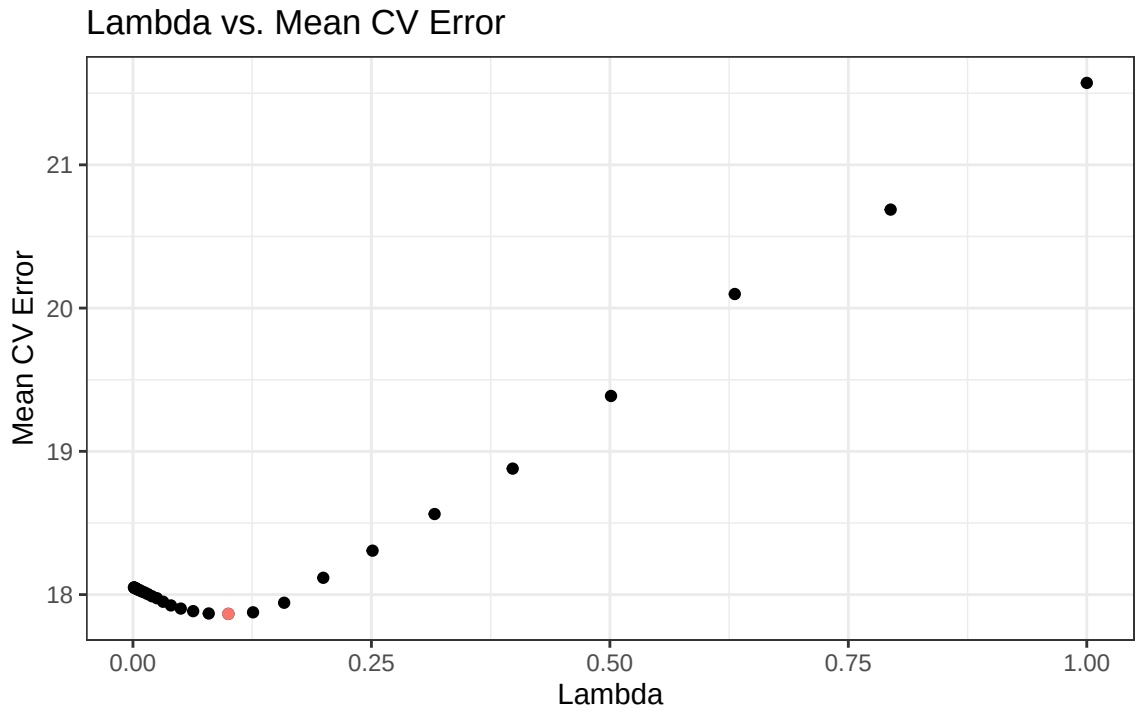


Fig5 Plot for Selecting the Optimal Lambda

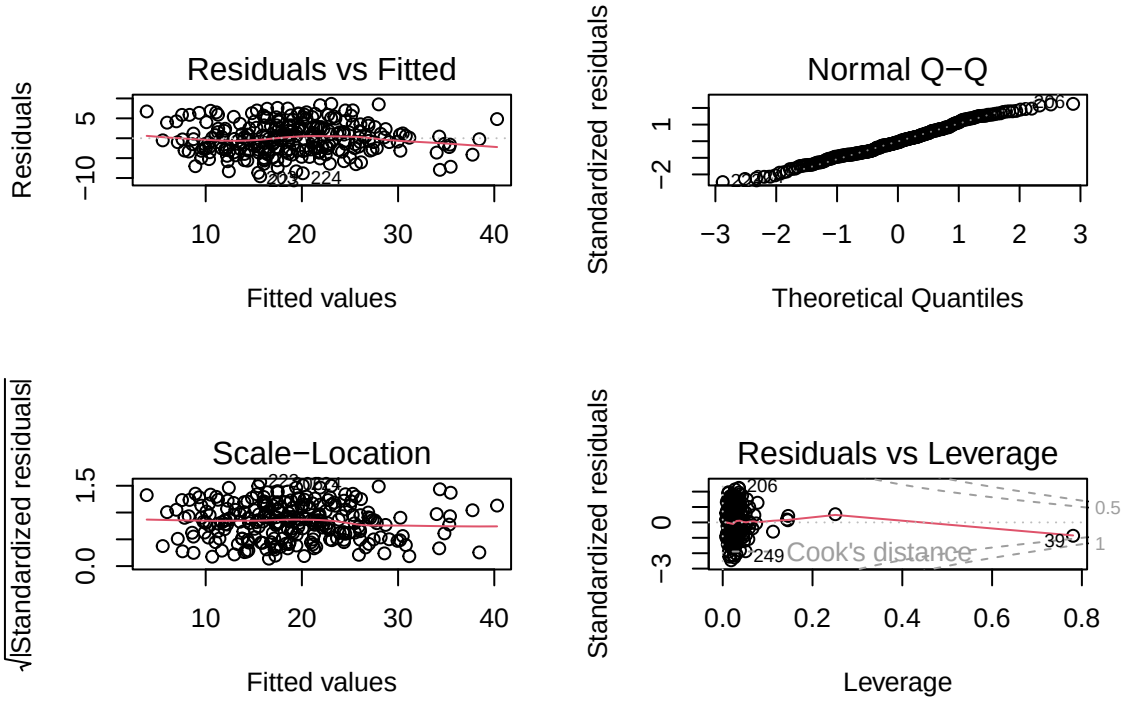

Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

Table 1: 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 2: 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 3: 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 4: 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 5: Validation Results for Candidate models

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

Table 6: 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 7: Two-way ANOVA Test for Model Candidate 2

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: 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