P8130 Final Project

Group Members: Runze Cui, Zhengwei Song, Kristen He

12/15/2022

Abstract

Body fat is an essential indicator of overall health condition, and body density is one of the significant indicators to measure body fat. The goal of this study is dedicated to predicting body density with multiple continuous variables possibly obtained from reports of routine health examination or clinical diagnosis. Specifically, this project mainly focuses on the critical discussions, comparisons, and integrations for multiple linear regression model building, selection (backward elimination, forward selection, stepwise regression, Mallow's $C_{p,}$ Criteria, and LASSO), diagnosing (residual plots analysis), validation and any necessary methodologies behind. Based on the considerations of the Principle of Parsimony and on the premise of ensuring further predictive ability, the interaction term is removed and a 7-parameter model is confirmed as the final model in this study.

Introduction

In a research laboratory setting, the overall density of the body (Db) is calculated through its mass and volume (Db = mass/volume). The mass of the body is found by simply weighing a person on a scale, and the body's volume is most easily and accurately determined by completely immersing a person in water and calculating the volume of water from the weight of water that is displaced (via "underwater weighing"). The proportions of water, protein, and minerals in the body are found by various chemical and radiometric tests [1]. The densities of water, fat, protein, and minerals are either measured or estimated. However, based on the available dataset, we can only choose from 13 potential variables to explain and predict human body density, of which there are three basic human indicators: *age, height,* and *weight*; and 10 circumference indicators: *neck, chest, abdomen, hip, thigh, knee, ankle, bicep, forearm,* and *wrist*. We want to examine exactly how many of these indicators are needed at least, and the weights of each indicator so that we can more comprehensively explain and predict human body density.

Since everyone has a different body fat density distribution. The goal of this project is to build a simple model in which doctors can easily use these variables to measure the body density of patients. Given a dataset consisting of 14 body measurement variables from 252 observations, we used different methods to compare and choose the variables to build the model.

Methods

Data Exploration

The raw data contains 17 columns and 252 observations. However, we decided to get rid of certain columns and set up a cleaner dataset for further model building. Specifically, we removed the *bodyfat_brozek* and *bodyfat_siri* since we decided to choose *body_density* as our outcome. For convenience, we assumed the random effect between sample members is not obvious, so we also removed *id*. Considering the small value of body density, which is two orders of magnitude smaller than the other independent variables, we multiply the values of *body_density* by 1000 to make the preceding estimated coefficients can be easily read for analysis.

After cleaning the dataset, we got a cleaned dataset from 252 observations, including 14 variables: *body_density, age, weight, height, neck, chest, abdomen, hip, thigh, knee, ankle, bicep, forearm, wrist* (Table 1). All variables are numeric, and our primary outcome of interest is *body_density*. None of these variables shows an extremely skewed distribution; therefore, we kept the original data without applying transformations (Figure 1). In the correlation plot (Figure

2), we see that the variables have a high correlation with each other such as *abdomen* and *chest*. The high correlation variables could cause possible poor estimations and inflated standard errors due to multicollinearity. However, further model selection techniques are able to address this concern.

Also, we need to find the potential interaction terms based on related literature and refer to the correlation matrix shown above. Based on the background information and the present literature, there are interactions existing among *chest, abdomen, thigh* [2], so we determine to keep all of them by combining into an interactive term in our model for subsequent model selection and validation.

Model Fitting

We have used automatic selections, Mallow's C_p criteria-based methods, and LASSO to fit the best model. In the backward elimination and stepwise selection, we obtained the same result, which included 11 parameters (10 terms): intercept, *age, height, neck, bicep, wrist, chest, abdomen, thigh,* and *chest:thigh.* However, the forward selection model generated 7 parameters (6 predictors): intercept, *abdomen, weight, wrist, forearm, bicep,* and *neck.* Given two different results, we have also generated the C_p value and adjusted R^2 value graphs to guide a better model (Figure 3). After that, we applied LASSO to perform variable selections (Figure 4). Results generated from LASSO provided a model of 10 parameters (9 predictors) by selecting a bigger lambda rather than the recommended one by cross validation, to have a larger penalty and smaller model. Since we want the model to be as simple as possible (principle of parsimony) and simultaneously maintain the same predictive power. We will choose the model with a relatively low C_p value and a high adjusted R^2 value. Therefore, we have decided to keep 7 parameters (6 predictors) in the linear regression since they provide the simplest model and relatively good C_p and R^2 values.

Results

Model Diagnostics

In order to see how each observation fits the selected regression model by forward selection, we ran model diagnostics, including Residuals vs. Fitted plot, normal Q-Q plot, and scale-location plot (Figure 5). Overall, these plots show that the model is fitting the data well. In the collinearity graph , *weight* has a higher VIF value than other parameters. Although a high VIF value might indicate the possible existence of multicollinearity, we have decided to keep the parameter, since it is one of the important and basic parameters in evaluating *body density*.

Cross Validation

From the cross-validation results, we have obtained a Root Square Mean Error (RMSE) value of 10.081 and a Mean Absolute Error (MAE) of 8.343, indicating that the observed data points are close to the model's predicted values based on the expanded *body density* values. This shows the model is performing well in predicting new values. Additionally, the variance of these measures is relatively low, suggesting that our model has a good predictive ability.

Also, we compared different models selected by automatic, criterion, and LASSO with respect to the cross-validated prediction error. By and large, the models between those methods showed a similar prediction ability by the low predicition errors (RMSE), but the final model with the least parameters is hence potentially the best model, when comparing the medians (the line in the middle of the box) and the overall distribution of the box plots (Figure 6).

Conclusion

Based on the results from automatic selection (Figure 7), stepwise selection, and LASSO, the final regression model that we have selected is

$$Body\_density = 10^{-3} * (1168.022 - 2.321 * abdomen + 0.336 * weight + 3.142 * wrist - 0.975 * forearm - 0.759 * bicep + 0.987 * neck)$$

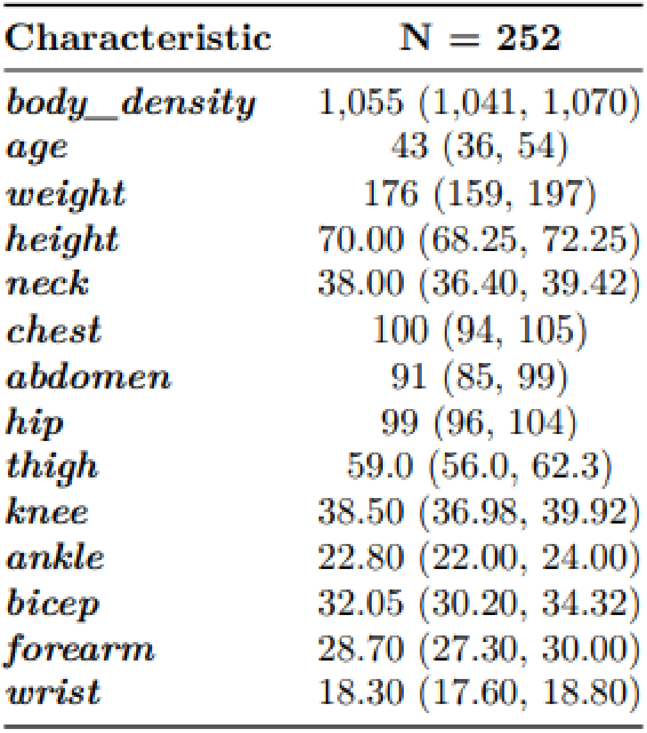
From this model we can see that, controlling for other covariates, as the measurements of the abdomen, forearm, and bicep decreases, the body density increases, while increases in weight and neck measurements will lead to increases in body density. From our 10-fold cross-validation, we can see that the model has a good predictive ability. However, the dataset only reflects body measurements of 252 men from one US state in 1985. Therefore the model will require additional adjustments before being used to predict men's body density from other locations.

References

[1] Siri William E (1956). "The gross composition of the body". Advances in Biological and Medical Physics. 16 4: 239–280. doi:10.1016/B978-1-4832-3110-5.50011-X. ISBN 9781483231105. PMID 13354513:239–278

[2] Jackson, A. S., & Pollock, M. L., 1978,Generalized equations for predicting body density of men, 40,497-504 Brozek, J., Grande, F., Anderson, J. T., & Keys, A., ,Densit
'=][ometric analysis of body composition: Revision of some quantitative assumptions, 110,113-140

Appendix



Characteristic	**N = 252**
body_density	1,055 (1,041, 1,070)
age	43 (36, 54)
weight	176 (159, 197)
height	70.00 (68.25, 72.25)
neck	38.00 (36.40, 39.42)
chest	100 (94, 105)
abdomen	91 (85, 99)
hip	99 (96, 104)
thigh	59.0 (56.0, 62.3)
knee	38.50 (36.98, 39.92)
ankle	22.80 (22.00, 24.00)
bicep	32.05 (30.20, 34.32)
forearm	28.70 (27.30, 30.00)
wrist	18.30 (17.60, 18.80)

Table 1: Descriptive statistics summary of all of the parameters

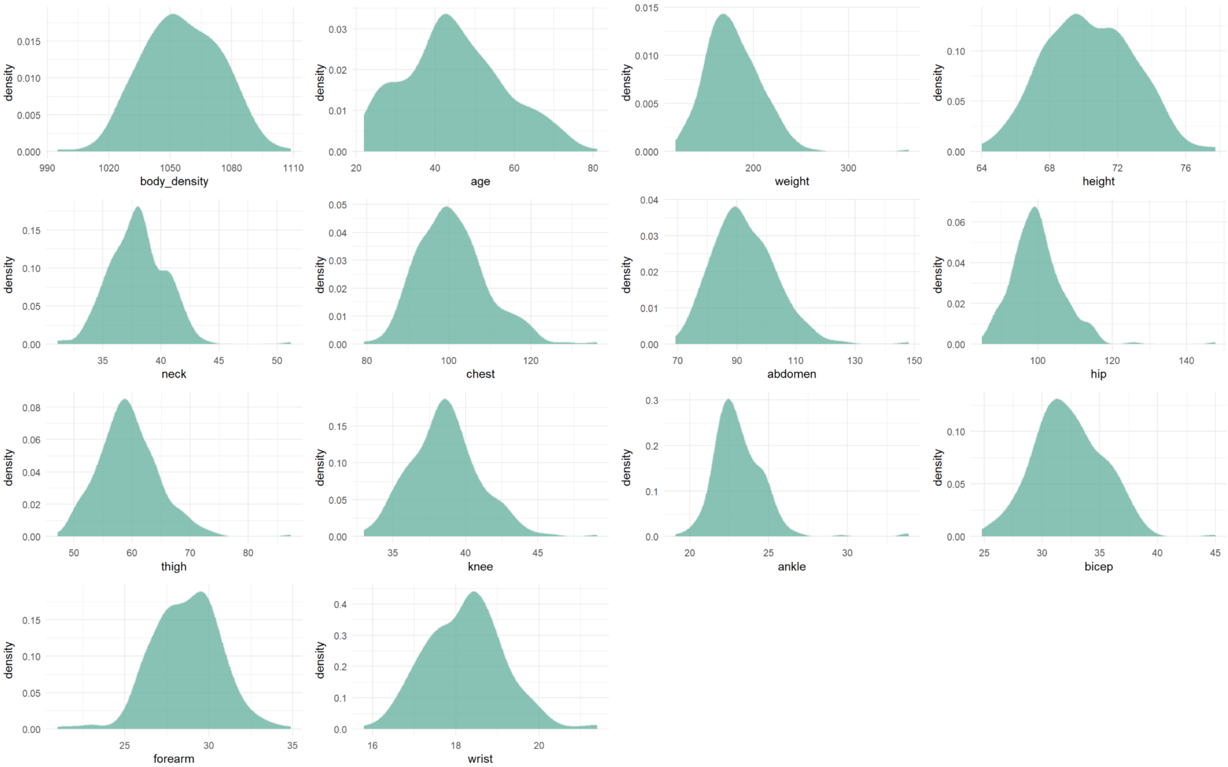


Figure 1: Distribution density plots of all the parameters

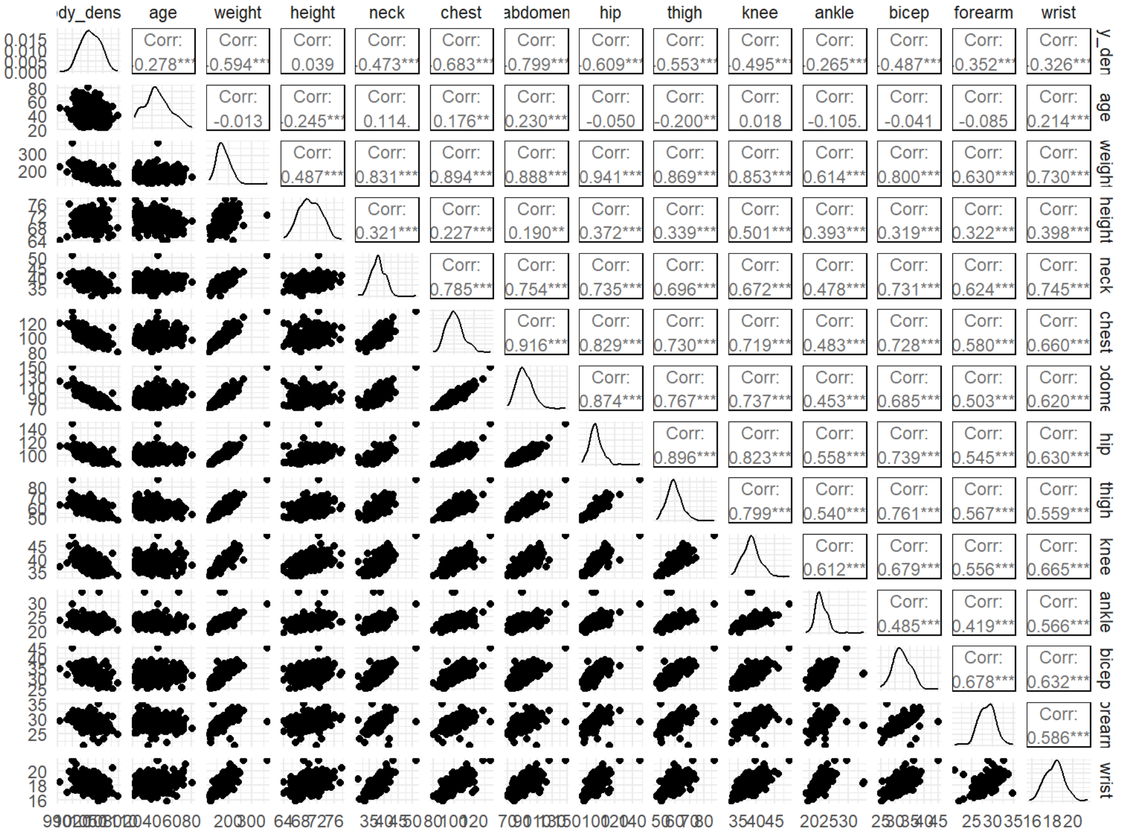


Figure 2: Correlation plot among all the parameters

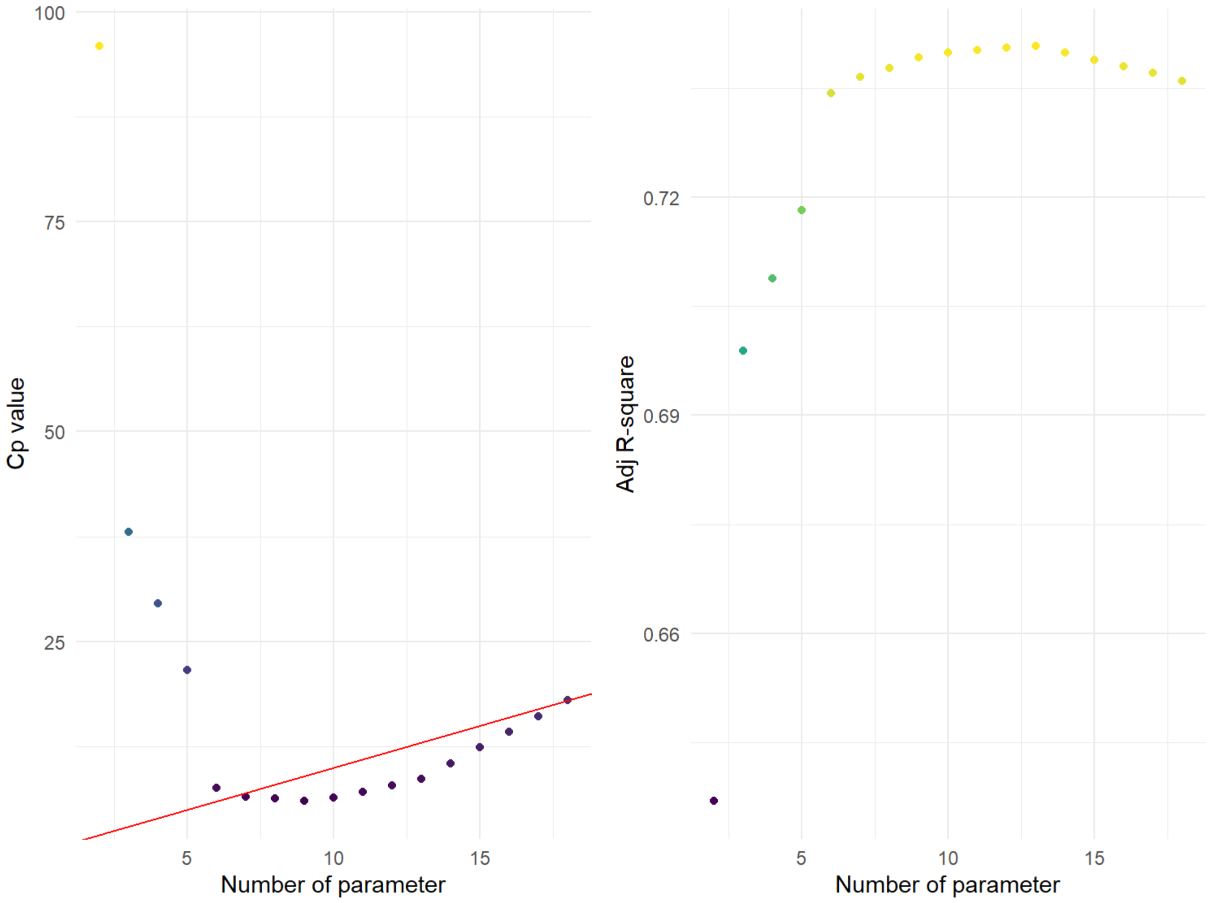
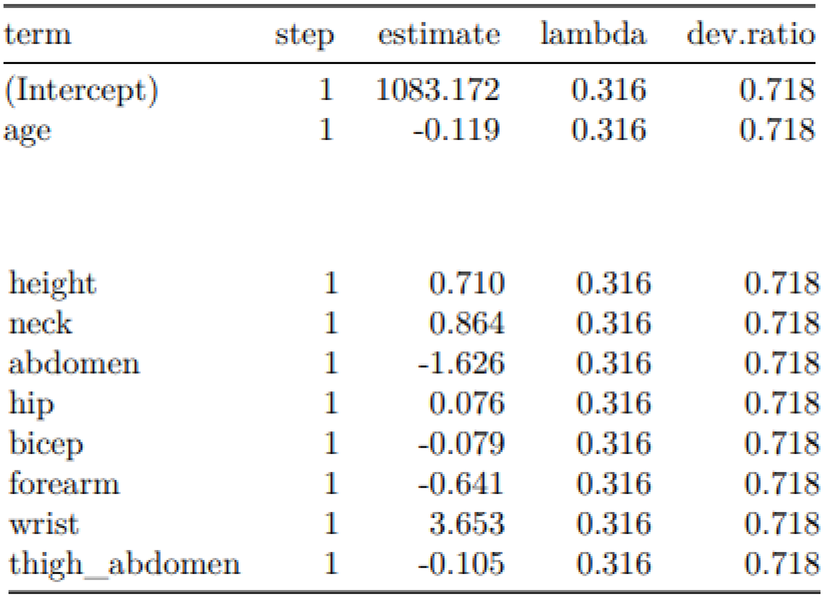


Figure 3: C_p and adjusted R^2 value graphs of parameters



term	step	estimate	lambda	dev.ratio
(Intercept)	1	1083.172	0.316	0.718
age	1	-0.119	0.316	0.718
height	1	0.710	0.316	0.718
neck	1	0.864	0.316	0.718
abdomen	1	-1.626	0.316	0.718
hip	1	0.076	0.316	0.718
bicep	1	-0.079	0.316	0.718
forearm	1	-0.641	0.316	0.718
wrist	1	3.653	0.316	0.718
thigh_abdomen	1	-0.105	0.316	0.718

Figure 4: Parameters selected from LASSO

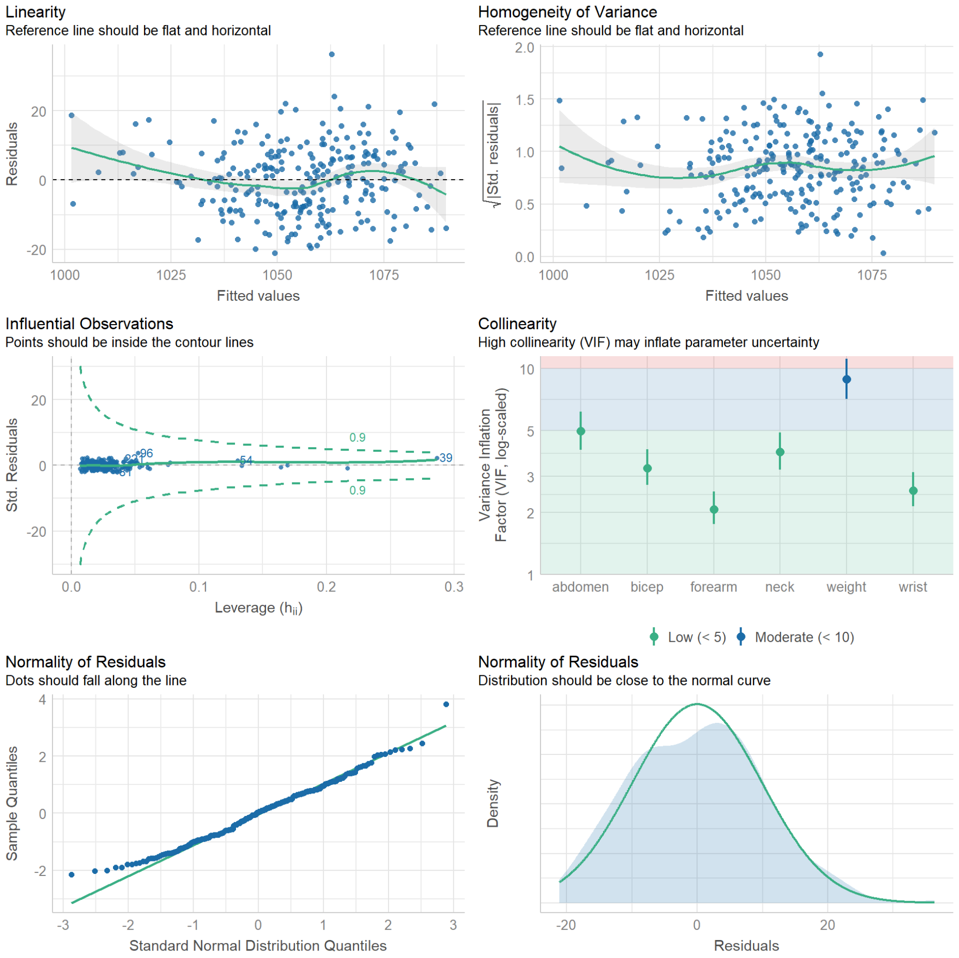


Figure 5. Diagnostic plots of the preliminary model

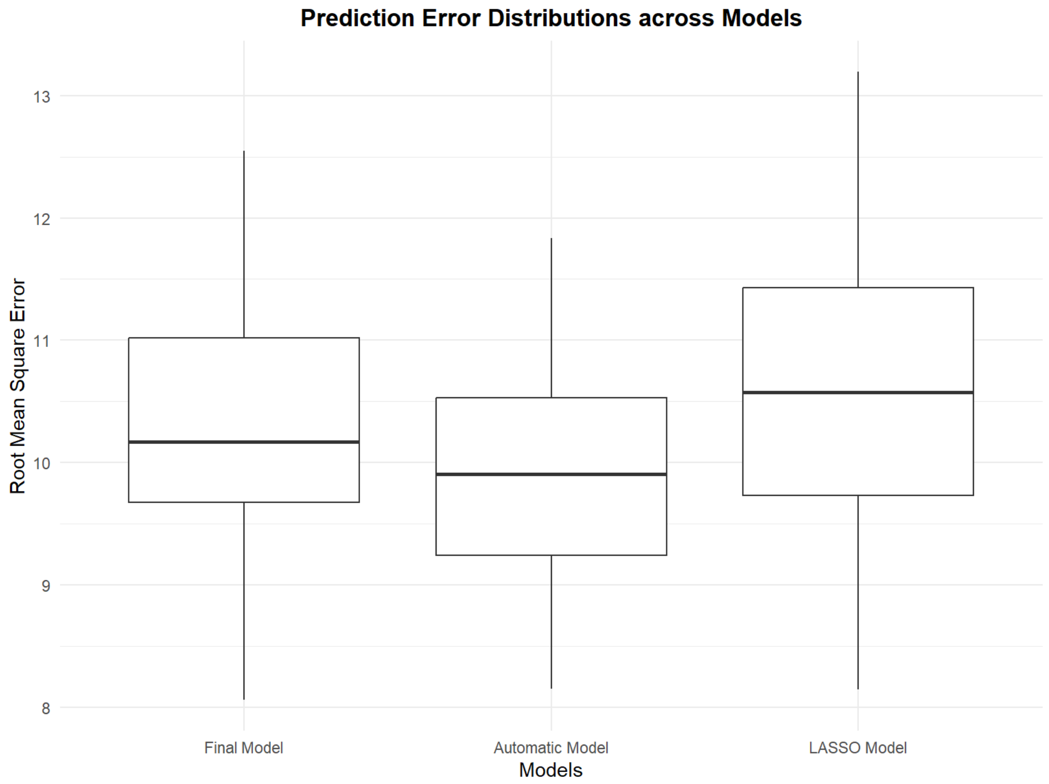
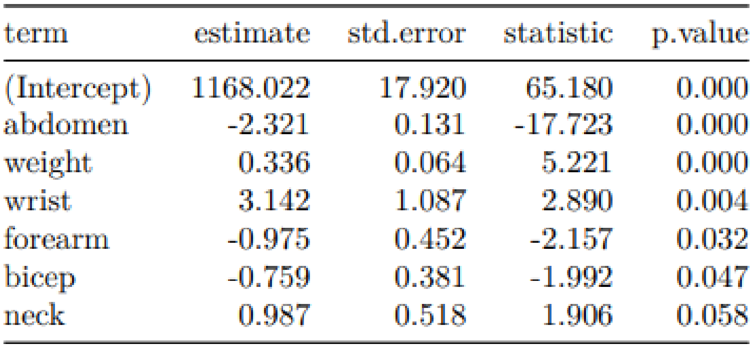


Figure 6: Prediction error distributions of three models



term	estimate	std.error	statistic	p.value
(Intercept)	1168.022	17.920	65.180	0.000
abdomen	-2.321	0.131	-17.723	0.000
weight	0.336	0.064	5.221	0.000
wrist	3.142	1.087	2.890	0.004
forearm	-0.975	0.452	-2.157	0.032
bicep	-0.759	0.381	-1.992	0.047
neck	0.987	0.518	1.906	0.058

Figure 7. Parameters selected from the forward selection