1 Background Information and Data Preprocess

In this report, we present the procedure of analyzing a bodyfat data set and building a reasonable and satisfying model to estimate the bodyfat percentage based on several common body measurement. This bodyfat data set contains 252 men with measurements of their percentage of bodyfat, body density and various body circumference measurements e.g. chest circumference, hip circumference. Our goal is to give a "rule of thumb" to predict the percentage of bodyfat based on these body measurements.

We list the procedure of preprocessing data below. We first check missing value in our data set and there is no missing value. Then we check outliers and unreasonable value in each variable.

According to American Council on Exercise, male bodyfat should be greater than 2% and we find two individuals with bodyfat less than 2%. One is with bodyfat = 0% and another with bodyfat 1.9%. We also check the maximum bodyfat in our data set which is 45.1%. We use online bodyfat calculator toverify their bodyfat values. We consider only to impute the individual with 0% bodyfat and replace it with 7.3% based on the result of online bodyfat calculator. In terms of height variable, we find an individual with height = 29.5 inches which is too short. We verify it with weight and BMI value and the result does not match with 29.5. We use the calculated result 69.5 to replace it.

- id=182: bodyfat=0% to bodyfat=7.3%
- id=42: height=29.5 inches to height=69.5 inches

We find two outliers in terms of weight with 262.75 and 363.15 pounds. We verify them with BMI and height and they are correct. There is no outlier in terms of age but we notice that the the maximum value of age is 81 which has a gap with the second largest value 74.

2 Motivation

2.1 Model Statement

Our final model is:

$$Lasso(\lambda = 1) \quad = -38.873 + 0.002Age + 0.838Abdomen(cm) - 0.245Weight(kg)$$

That is, for males, for every age increase in one unit, the model predicts that bodyfat will increase, on average, by 0.002%. When someone's abdomen increase 1cm or weight increase 1kg, his bodyfat will increase 0.838% or decrease 0.245%, respectively.

For example, a 40 year-old man with 80cm Abdomen circumferences and 70kg weight is expected to have a bodyfat of 11.137% based on our model.

2.2 Model Selection

We mainly came up with the following 5 models: The comparison of three method we use is as following:

Method	Variables	$adjR^2$	RMSE
Full Model	all variables	0.732	3.861
Stepwise	Age, Neck, Abdomen, Hip, Thigh Forearm, Wrist, Weight	0.736	3.883
Split in age groups: Age$\leq$ 40	Abdomen, Hip, Thigh, Wrist	0.733	3.470
40 < Age$\leq$ 60	Neck, Abdomen, Ankle, Forearm, Wrist	0.741	3.916
Age> 60	Abdomen, Ankle	0.695	3.165
Lasso(λ=0.2)	Age, Abdomen, Weight, Abdomen, Hip Thigh, Biceps, Forearm, Wrist, Weight	0.727	3.898
Lasso(λ=1)	Age, Abdomen, Weight	0.698	4.101

We made 3 criteria when choosing models: First, easy measured and less variables are better. Second, model can be easily explained(no machine learning method applied). Last, model has high adjusted R-square and low RMSE values, which are estimates of how well the model fits. And we chose the last model based on the following reasons: First, the variables are easy to measure and commonly available. Second, according to the adjusted R^2 and RMSE in the table, 5 models only have relatively small differences. Therefore, we prefer to use the least number of variables and sacrifices the goodness of fit slightly.

3 Statistical Analysis

First, we found that $AdjustedR^2 = 0.698$ of this Lasso regression, which implies that our model can explain 69.8% variation of response variable around its mean. That can also be illustrate that 69.8% of our samples fit this lasso regression. Then, we conducted the following t-test to see whether the predictors in Lasso regression ($\lambda = 1$) we have chosen are significant in predicting the outcome. For our model,

$$Bodyfat = \beta_0 + \beta_1 Age + \beta_2 Abdomen + \beta_3 Weight + \epsilon \tag{1}$$

The null hypothesis is $\beta_i = 0$ where $i = 1, 2, 3$. We used the t-statistics to test whether the parameter has significant influence on our response variable. For our model, we return the summary includes estimation, 95% confidence interval, t-statistics and p-value, the table includes includes these information is as following:

Variable	Estimation	Confidence Interval	t-statistics	p-value
Intercept	-38.873	(-71.265,-6.480)	-2.3635	0.018868
Age	0.002	(-0.043,0.047)	0.0910	0.927549
Abdomen	0.838	(0.697,0.979)	11.6954	0.000000
Weight	-0.245	(-0.328,-0.161)	-5.7696	0.000000

The estimation and confidence interval imply that when someone grow up for one year, his bodyfat will increase 0.002%, at least -0.043% and at most 0.047%. This confidence interval includes 0, which implies age may not have significant influence on bodyfat. When someone's abdomen increase 1cm, his bodyfat will increase 0.838%, at least 0.697% and at most 0.979%. And his weight increase 1kg, his bodyfat will decrease 0.245%, at least 0.328% and at most 0.161%. Besides, based on the 95% CI, we can reject the null hypothesis of β_2 and β_3. But we don't have enough evidence to reject the null hypothesis of β_1, because the confidence interval includes 0, negative values and positive values. In other words, we have enough evidence to say that abdomen and weight have significant influence on bodyfat, but we don't have enough evidence to say age has significant influence on bodyfat.

Then, we consider the t-statistics and p-values. Under the 0.05 Type I error, we can state that abdomen and weight have significant influence on bodyfat. From another word, we can tolerate that there is at most 5% of variables relations in the sample may be due to chance. Under this circumstance, we are almost 100% sure that abdomen and weight are related to the response variable bodyfat.

4 Model Diagnostics

Linear model assumptions: Basically, the performance of the Lasso is well understood under the assumptions of the standard linear model. Here we mainly focus on homoscedasticity. The residual plots shows that residuals are randomly distributed around zero and there is no evidence for heteroscedasticity. The table below is the summary of residuals. Levene's test also supports our conclusion.

Min	1Q	Median	3Q	Max
-10.1755	-2.9786	-0.0265	2.9991	11.3068

Outlier and high leverage point: We also want to detect high leverage points and outliers. Cook's distance is adequate for detection. From our analysi report, there are two high influence points. One has the largest weight in the whole data set, while the other one doesn't show any obvious difference from other samples. Anyway, we drop these two observations and fit a lasso model again. The adjusted R^2 improves to 0.704 after dropping.

Overfitting and test performance By now, what we show is still the model performance on the train data set. Out-of-sample performance is also necessary for our sample. The coefficient of determination is computed as the performance score on test data and train data. Two results were displayed in the score. First, With the larger lambda, the variables in the model are less, but test score is non-decreasing. Therefore, the models with more strict penalty on variable selection will perform as well as those without penalty.Secondly, the larger lambda results in a drop in the train score. However, it not a bad idea. The improving performance on test data claims that we actually are getting out of the overfitting trap.

5 Model Strengths/Weakness

Pros: 1) Lasso helps to avoid problems like multicollinearity, which are common in data sets with small numbers of sample but large numbers of variables. 2) We use cross-validation to find a believable model. 3) After variable selection, the model is simple enough to interpret and use. Concerns about overfitting are also solved. 4) The model is simple enough to interpret and use.

Cons: 1) Slight inconsistency of results due to the weaknesses of lasso. 2) The variable "age" is not significant in our model though it makes sense in many researches.

6 Conclusion

In this report, we analyze the bodyfat data set. We select our final model from various model like stepwise regression, split in age groups and etc. In our final model, we use age, abdomen and weight to estimate the bodyfat, which $R_2 = 0.698$. Compare with other models, this model has the least variables and the variables are easy to measure and interpret. We found that age variable in our model is not significant, but we still keep this variable in our model. Even though, it is not significant, it still meets our expectation and can be illustrated in our real life.

7 Contributions

Bruce Zheng:
Cleaned the data in python, wrote the ”Background Information/Data Cleaning” part in summary and presentation, finished the Shiny APP.
Suhui Liu:
Wrote the code about full model, stepwise regression model, split in age groups model in R program. Wrote the ”Model Motivation” part in summary and presentation.
Yixuan Wang:
Wrote the code about full model, two lasso regression models, and wrote the code about k-fold to select the best lambda in python. Wrote the ”Model Analysis” in summary and presentation.
Haishuo Chen:
Wrote the code about model diagnostics on linear model assumptions and high leverage points, validate performance on test data in python. Wrote the ”Diagnostics” and ”Strengths and weaknesses” in summary and presentation.

References

[1] A single threshold value of waist girth identifies normal-weight and overweight subjects with excess visceral adipose tissue-S Lemieux, D Prud’homme, C Bouchard, A Tremblay, J P Després The American Journal of Clinical Nutrition, Volume 64, Issue 5, November 1996, Pages 685–693, https://doi.org/10.1093/ajcn/64.5.685

[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6906176/

[3] Behnke, A.R., Wilmore, J.H. (1974). Evaluation and regulation of body build and composition.