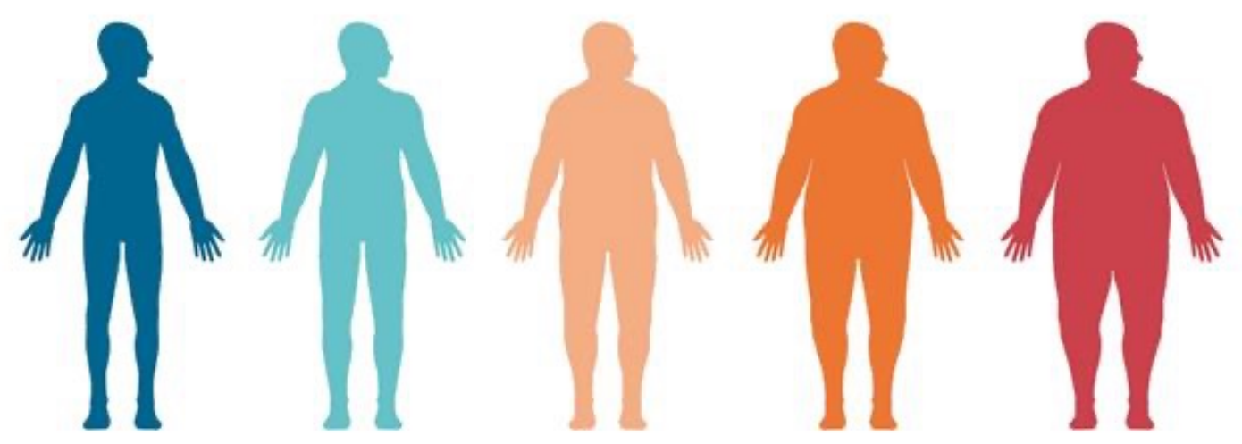
Estimating Bodyfat Percentage Using Practical Physical Measurements

Anusha P Anilkumar, Ben White, Chu Zhang, Jinzhuang Cheng, Juwu Pu



Introduction

The aim of this project is to find a more accurate and convenient way to estimate the percentage of bodyfat by various practical physical measurements.

Data Summary

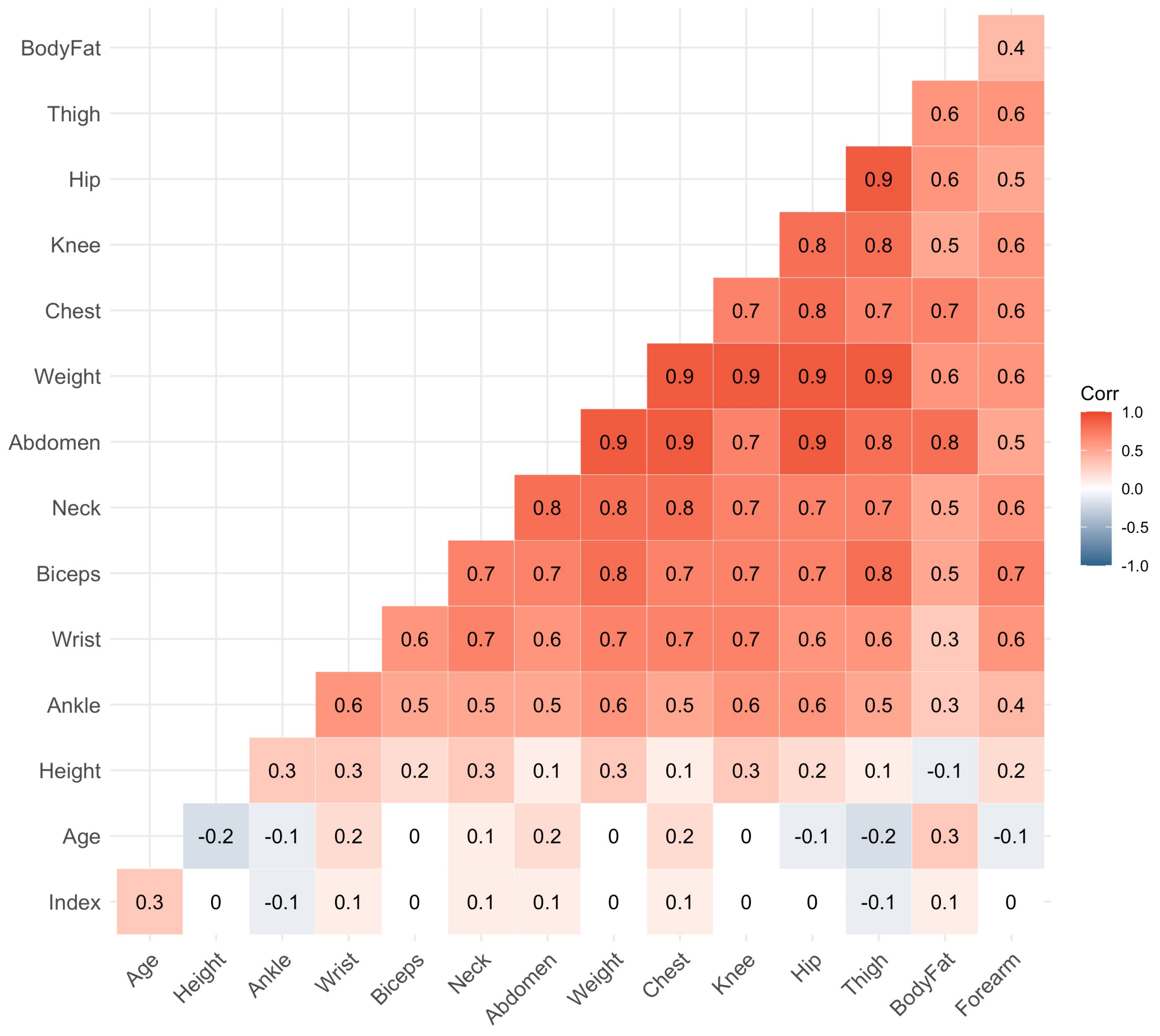
The collected data contains 252 observations (men) for 14 variables. These variables are percent BodyFat from Siri's (1956) equation: Age (years), Weight (lbs), Height (inches), Neck circumference (cm), Chest circumference (cm), Abdomen circumference (cm), Hip circumference (cm), Thigh circumference (cm), Knee circumference (cm), Ankle circumference (cm), Biceps (extended) circumference (cm), Forearm circumference (cm) and Wrist circumference (cm). Summary statistics of these variables are presented in the following table:

Variable	n	Mean	SD	P-25	P-50	P-75
BodyFat	252	19.15	8.37	12.48	19.2	25.3
Age	252	44.88	12.60	35.75	43	54
Weight	252	178.92	29.39	159	176.5	197
Height	252	70.15	3.66	68.25	70	72.25
Neck	252	37.99	2.43	36.4	38	39.43
Chest	252	100.82	8.43	94.35	99.65	105.38
Abdomen	252	92.56	10.78	84.58	90.95	99.325
Hip	252	99.90	7.16	95.5	99.3	103.53
Thigh	252	59.41	5.25	56	59	62.35
Knee	252	38.59	2.41	36.98	38.5	39.93
Ankle	252	23.10	1.69	22	22.8	24
Biceps	252	32.27	3.02	30.2	32.05	34.325
Forearm	252	28.66	2.02	27.3	28.7	30
Wrist	252	18.23	0.93	17.6	18.3	18.8

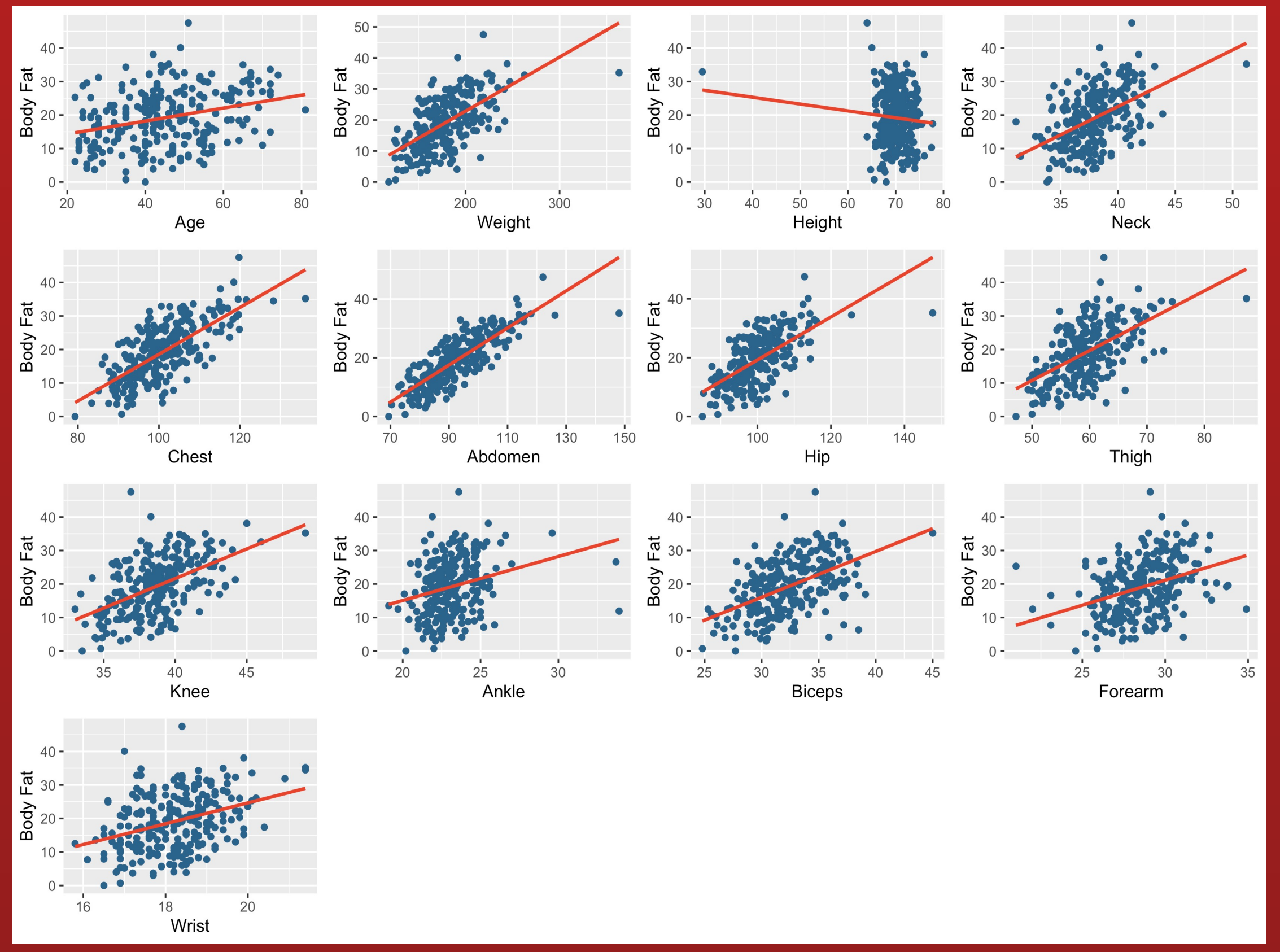
Methodology

In order to estimate and predict bodyfat percentage , it is crucial to identify any potential predictors through plotting correlation diagrams between variables.

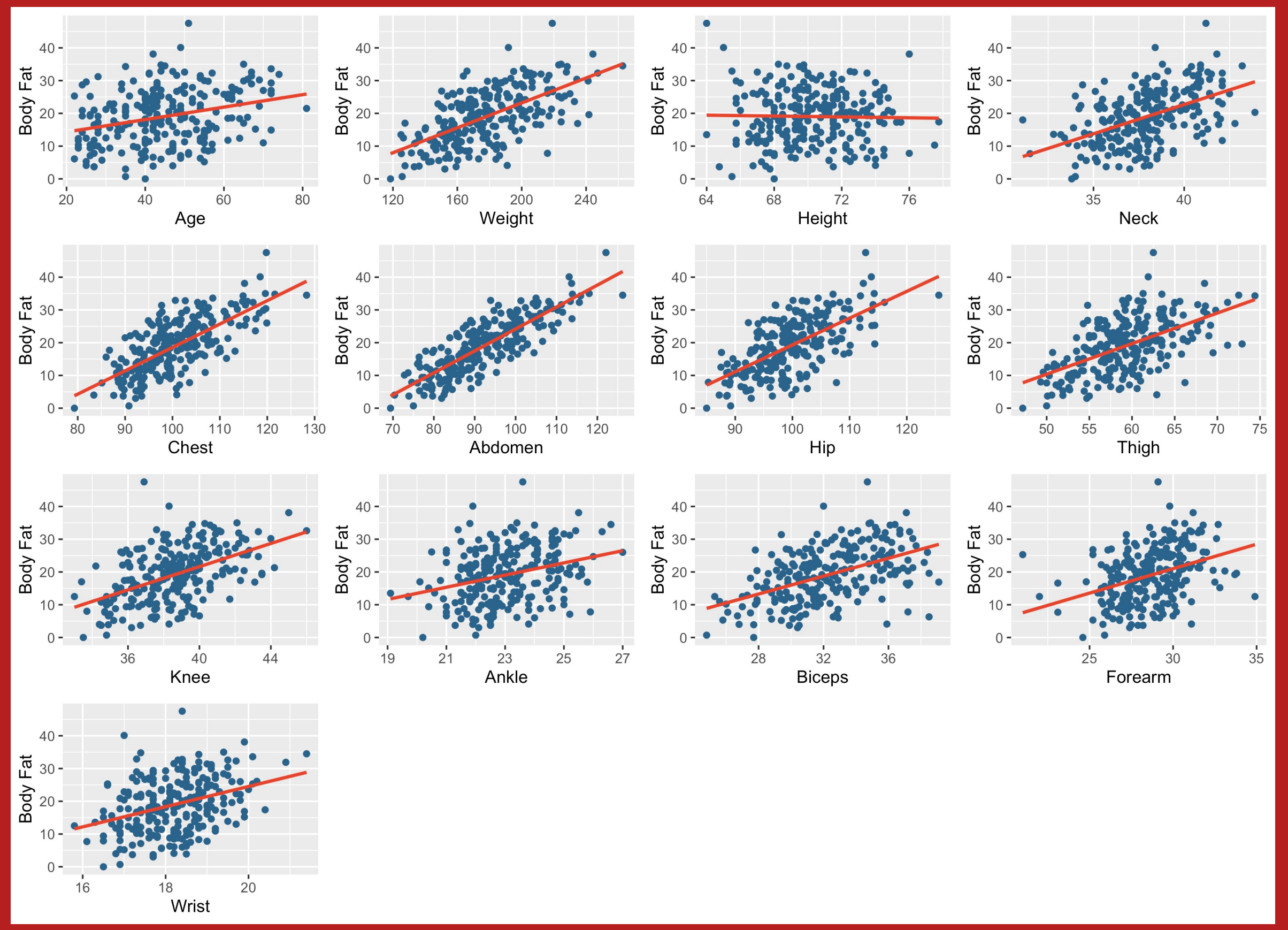
As shown in the figure below, there is a strong positive correlation between all the variables, which implies that there is high multicollinearity. Thus, it is important to choose appropriate variables carefully. Abdomen looks promising - however it looks like there may be some outliers relating to weight.



We can view the plots of all the predictors individually against bodyfat.



After plotting the relationship between the variables, it is obvious that there is indeed a good positive relationship between most of them except one of the explanatory variables. The plot between Height and BodyFat shows that there is a very weak relationship between them. In addition, there are some clear outliers which we can identify and remove, then, plotting again to see if this has had the desired effect of providing a cleaner data set:



This looks much better - however there are still some outliers in the Hip vs Bodyfat plot, but we will keep this in order to avoid removing too much data.

Model Selection

Performing backward and forward stepwise AIC selection and also dropping variables to fit different models, we select one based on the best AIC, R^2 and adj_R^2, values. Finally, Abdomen, Weight and Wrist are chosen as the most effective predictors of variation in BodyFat.

$$\widehat{BodyFat} = \hat{\alpha} + \hat{\beta}_1 \cdot Abdomen + \hat{\beta}_2 \cdot Weight + \hat{\beta}_3 \cdot Wrist$$

Results

On carrying out linear fitting on the cleaned data, the results show that the Multiple R-squared = 0.7371 and the Adjusted R-squared = 0.7339. It can be observed that 73.39% of the body fat information can be explained by the above data.

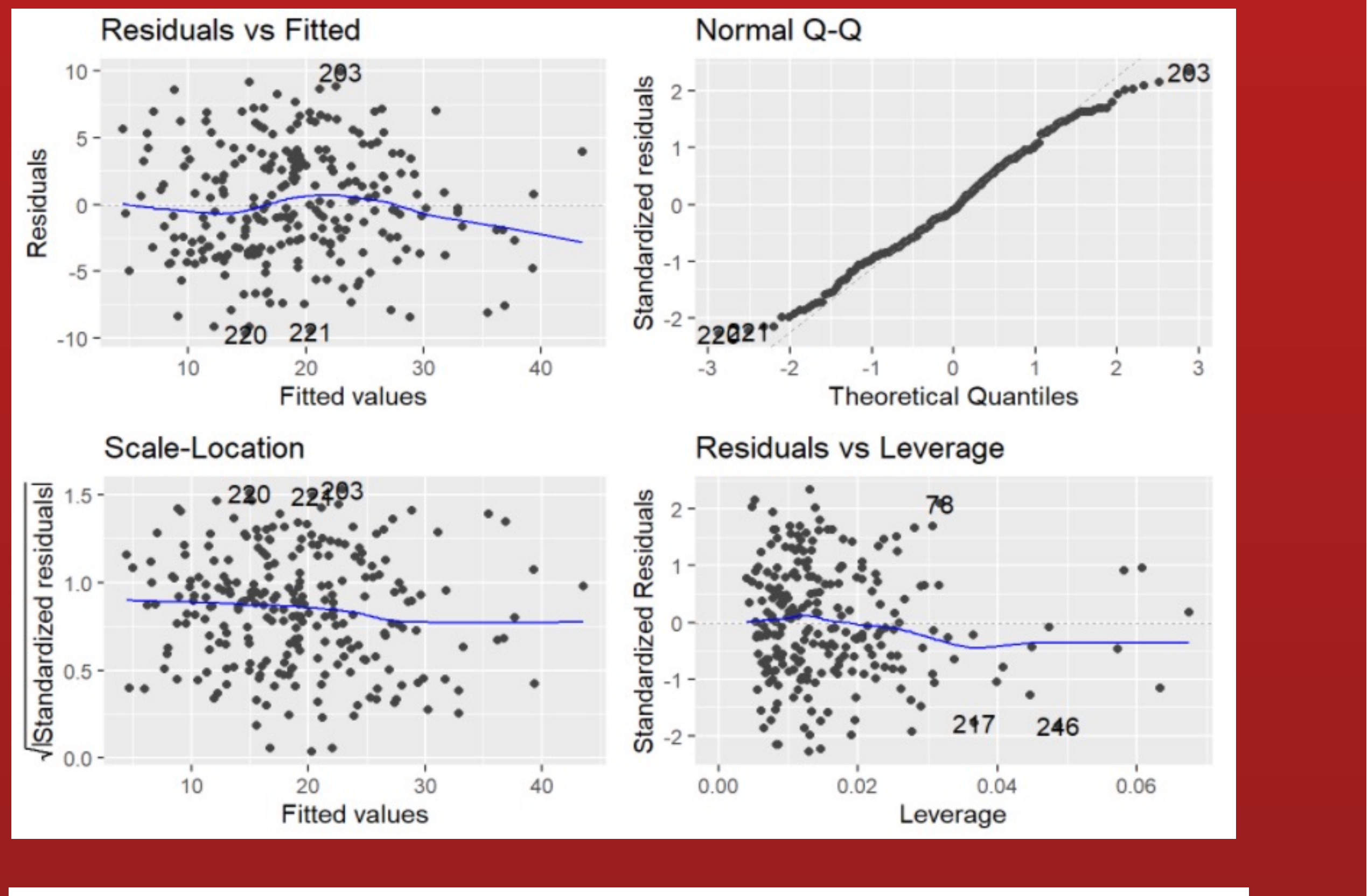
Variable	Estimate	Std Error	P-Value	Lower CI	Upper CI
Intercept	-27.556	6.689	0	-40.731	-14.381
Abdomen	0.956	0.055	0	0.848	1.064
Weight	-0.091	0.024	0	-0.139	-0.044
Wrist	-1.396	0.433	0.001	-2.249	-0.542

$$\widehat{BodyFat} = -\widehat{27.556} + \widehat{0.956} \cdot Abdomen - \widehat{0.091} \cdot Weight - \widehat{1.396} \cdot Wrist$$

Model Assumptions

The model conforms to the basic assumptions of the linear model.

- Residual vs Fitted Values Plot: The residuals are randomly scattered around zero line which means they have with zero mean and constant variance.
- QQ Plot: The residuals are approximately consistent with a normally distribution.
- Scale Location Plot: There is no problem of heteroscedasticity as the error terms are equally scattered.
- Standardized Residuals vs Leverage Plot: There are no pints with a massively high leverage, so looks good.



Conclusion

The model selected 3 variables to predict the body fat percentage of adult men, namely weight, abdominal circumference and wrist circumference. The analysis shows that the use of this model can be more economical and time-effective to get a rough adult men's body fat percentage.

Reference

https://www.kaggle.com/fedesoriano/body-fat-prediction-dataset