DATA2902 Group Project Executive Summary

490428362[a], 470398544[a], 500460173[a], and 490411445[a]

[a]The University of Sydney

This version was compiled on November 20, 2020

A series of multiple linear regression models were developed aiming to robustly predict body fat percentage for males based on the measurements of 250 men of various ages. The selected model using backward selection consisted of eight parameters; age, height, and neck, abdomen, hip, thigh, forearm, wrist circumefences. Although the selected model generated relatively accurate predictions, limitations of the model and direction for future research are examined.

1 Introduction. One way to assess an individual's health and fitness level is to estimate their percentage of body fat. However, traditional measurement techniques to estimate body fat percentage are often inconvenient and costly. Generally, body fat percentages are calculated using Siri's equation (SOCR Data BMI Regression - Socr, 2020) given by

$$PBF = \frac{495}{D} - 450$$

where D is body density in gm/cm^3. Given the number of variables necessary to determine D (Section 2), it is evident that estimating body fat percentage is an inefficient process. Hence, our analysis aims to determine which variables are best predictors for percentage body fat and from these variables determine which ones are cost-efficient and practical.

2 Dataset description. The data was obtained from the BYU Human Performance Research Center (Bodyfat | DASL, 2020) and details measurements for 250 men. There are fifteen variables in total associated with each observation. These variables are as follows:

- Body density given by $BD = \frac{WA}{\frac{WA-WW}{CF} - LV}$ where
  - WA is weight in air (kg)
  - WW is weight in water (kg)
  - CF is the water correction factor (how much space 1 gram of water takes up at particular temperature)
  - LV is residual lung volume
- Percentage body fat calculated using Siri's equation $\left(PBF = \frac{495}{D} - 450\right)$
- Age in years
- Weight in kg (converted from lbs)
- Height in cm (converted from inches)
- Neck, Chest, Abdomen, Waist, Hip, Thigh, Knee, Ankle, Bicep, Forearm and Wrist circumferences in cm

No additional details have been provided about data collection in the study such as how individuals were selected or where they were from.

3 Analysis.

3.1 Model Selection. An interactive correlation matrix was adopted to visualise the correlation coefficients, and the corresponding scatters plot between the variables. It indicated that the waist and abdomen variables are perfectly identical ($r = 1.00$), and density and percentage of body fat have a strong negative relationship ($r = -0.99$). Then we conducted stepwise variable selection using the Akaike Information Criterion (AIC) to determine a reasonable model for multiple linear regression. The backward variable selection begins with a full model that excludes the density, and the final model that contains age, height, neck, abdomen, hip, thigh, forearm and wrist. The forward variable selection begins with a null model, and the final model contains waist, weight, wrist, bicep, age and thigh.

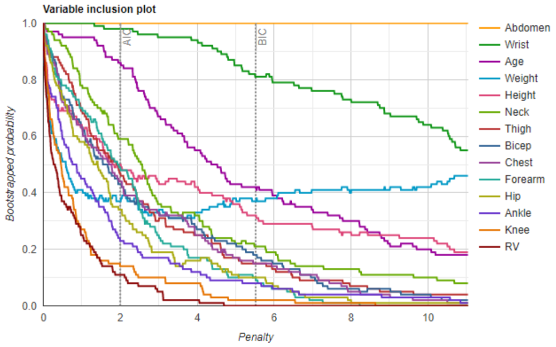
3.2 Assumption Checking. The diagnostic plot, scatter plot and quantile-quantile plot were used to examine assumptions in the backward and forward selection models. The diagnostic plot of residuals versus fitted values, and the scatter plot of each variable and percentage of body fat were used for checking the assumption of linearity and homoskedasticity, and the quantile-quantile plot was used for checking the normality assumption (Appendix 1.1).

The residuals from both models are equally spread without any distinct pattern, and their locally weighted regression fit lines are almost flat and close to horizontal zero residual line . The scatter plot is also utilised to verify that each variable in both models has a linear trend with the percentage of body fat. Therefore, the linearity and homoskedastic assumptions reasonably hold for both models.

Moreover, the standardised residuals from both fitted models follow the straight dashed line in the normal quantile-quantile plot, and the data size is sufficient to rely on the central limit theorem. Hence, there is an indication that the normality assumption for the residuals in both models are reasonably satisfied. However, the assumption of independence between the errors could not be assessed due to limited information available concerning data collection and measurement processes.

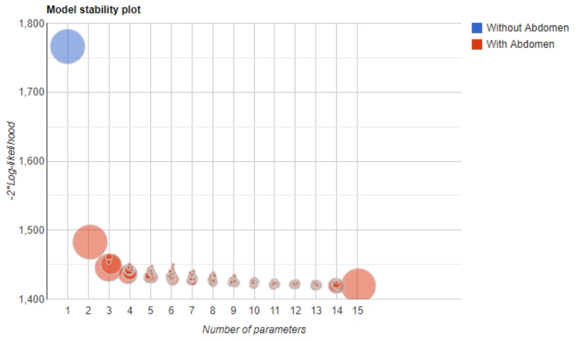
4 Results.

4.1 Variable Inclusion Plot.



The `Abdomen` and `Waist` variables are more likely to remain in the model as the penalty increases, indicating that these variables are crucial to the predictive power of the model.

4.2 Model Stability.



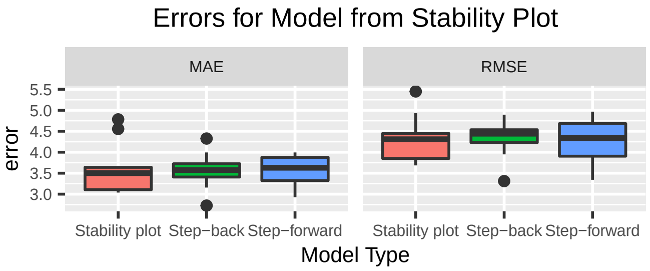
The penalties decrease significantly when there are less than 4 parameters and only decrease slightly after there are more than 4 parameters. Therefore, we also considered a simple model with only 4 parameters to compare its accuracy against more complex models. The selected variables were Weight, Abdomen and Wrist which were chosen with a probability of 0.28.

5 Model performance.

5.1 In-sample performance. We observe that the r^2 values for all models are quite similar, ranging from 0.7 to 0.75. Hence, all three models perform comparably when tested on in-sample data.

	Forward model	Backward model	Model from stability plot
R-squared	0.7538	0.726	0.7193

5.2 Out of sample performance. 10-fold cross validation testing was performed to assess the models' predictive power on previously unseen data.



The means of MAEs and RMSEs were similar across the 3 models, however the spreads of the errors vary quite significantly. Specifically, the RMSEs for the step-back model have a much tighter spread indicating that the model is more consistent. Hence, we chose the step-back optimised model as our final model.

6 Discussion.

6.1 Limitations. One of the biggest areas of concern in analysing our data and developing a robust regression model, were the uncertainties concerning the underlying dataset itself. Limited information was available about how specific measurements were taken, and hence replication of these methods becomes difficult. To develop an easier method of measuring one's body fat percentage, this would require understanding exactly how each measurement was taken, for example where on the waist one measures the circumference. In addition, the independence of each measurement is questionable when information about how the population was sampled is unavailable. It is not clear whether observations were only made in the United States, from a specific state, or even worldwide. Given there exists an extreme range in body types globally, it is difficult to conclude that the regression model can accurately predict an individual's body fat percentage based on only 250 measurements. When considering in sample performance of the different models, it is also apparent that the final model selected is not necessarily a major improvement in AIC, or r^2 value. However, we can see in out of sample performance, generally the spread of residuals is more stable. Upon investigating the coefficient values of the final regression model, it should be noted that some go against what one would naturally expect. For example, most coefficients only have a value less than one; with wrist having a value of -1.73 ($\beta_{\text{wrist}} = -1.73$). Intuitively, one would expect that a larger wrist circumference would lead to a potentially higher percentage body fat however this is not the case. Other coefficients that are also negative raise the same issue, however to a lesser extent. The intercept value is difficult to interpret physically speaking, as an individual with 0 measurements for areas such as neck, height would not exist. In terms of looking to find an accurate method of determining one's state of health, it is also apparent that body fat is not the only, or rather necessarily the best method available.

6.2 Comparison. When examining the MAE and RMSE of the regression model boxplots, it is clear that there is often an error ϵ of 3.5 and above. In some areas this is satisfactory however when attempting to classify one into a specific body fat category, this discrepancy is concerning. Body fat percentages are often split into five categories (2-5%: Essential fat; 6-13%: Athletic; 14-17%: Fit; 18-24%: Acceptable; $>$ 25%: Obese). An error of above three ($\epsilon > 3$) can classify one into various different incorrect categories and hence conclude inaccurate information. When comparing with the BMI rating, it is clear that BMI provides a more general outlook (Appendix Table 1), with body fat percentage attempting to be more specific (Appendix 1.2). Hence, we can interpret this as using BMI as a good early check of an individual level of health, with further investigation into body fat percentage providing greater accuracy.

6.3 Improvement. The most effective resolutions to the aforementioned issues are both an increase in observations as well as ensuring that the data has been taken from a widespread population of various ethnicities, body types as well as genders. Ensuring that the exact methodology of taking measurements is consistent is also of great importance. Conducting further research, we also found that often a nonlinear model worked best. For future research this would ideally be the optimal approach with a greater range of measurements taken, such as skin fold for example.

7 Conclusion. Although relatively accurate in determining an estimate of one's state of health, ambiguities in the data collection processes for this dataset used in the model may not provide the level of specificity needed to accurately aid in one's understanding of their health. The final equation for the multiple linear regression model is given by

$$\text{Pct.BF} = 5.04 - 0.45(\text{Neck}) + 0.82(\text{Abdomen}) - 0.19(\text{Hip}) + 0.22(\text{Thigh}) + 0.3(\text{Forearm}) - 1.73(\text{Wrist}) + 0.07(\text{Age}) - 0.11(\text{Height}) + \epsilon$$

References. Wiki.stat.ucla.edu. 2020. SOCR Data BMI Regression - Socr. [online] Available at: http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_BMI_Regression [Accessed 2 November 2020].

Table 1. Comparing BMI to Body fat classifications

	Underweight	Normal	Overweight	Obese
Essential Fat	1	2	0	0
Athletic	0	54	2	0
Fit	0	47	16	0
Acceptable	0	19	57	4
Obese	0	1	28	19

Dasl.datadescription.com. 2020. Bodyfat | DASL. [online] Available at: https://dasl.datadescription.com/datafile/bodyfat [Accessed 2 November 2020].

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.1. https://CRAN.R-project.org/package=dplyr

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Broman KW (2015) R/qtlcharts: interactive graphics for quantitative trait locus mapping. Genetics 199:359-361 doi:10.1534/genetics.114.172742

C. Sievert. Interactive Web-Based Data Visualization with R, plotly, and shiny. Chapman and Hall/CRC Florida, 2020.

Barret Schloerke, Di Cook, Joseph Larmarange, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg and Jason Crowley (2020). GGally: Extension to 'ggplot2'. R package version 2.0.0. https://CRAN.R-project.org/package=GGally

Thomas Lumley based on Fortran code by Alan Miller (2020). leaps: Regression Subset Selection. R package version 3.1. https://CRAN.R-project.org/package=leaps

Max Kuhn (2020). caret: Classification and Regression Training. R package version 6.0-86. https://CRAN.R-project.org/package=caret

Tarr G, Müller S, Welsh AH (2018). "mplot: An R Package for Graphical Model Stability and Variable Selection Procedures." *Journal of Statistical Software*, *83*(9), 1-28. doi: 10.18637/jss.v083.i09 (URL: https://doi.org/10.18637/jss.v083.i09).

Daniel Anderson and Andrew Heiss (2020). equatiomatic: Transform Models into 'LaTeX' Equations. R package version 0.1.0. https://CRAN.R-project.org/package=equatiomatic

Hadley Wickham (2007). Reshaping Data with the reshape Package. Journal of Statistical Software, 21(12), 1-20. URL http://www.jstatsoft.org/v21/i12/.

Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie and Jonathan McPherson (2020). shiny: Web Application Framework for R. R package version 1.5.0. https://CRAN.R-project.org/package=shiny

Hao Zhu (2020). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.3.1. https://CRAN.R-project.org/package=kableExtra

Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.30.

Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963

Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595

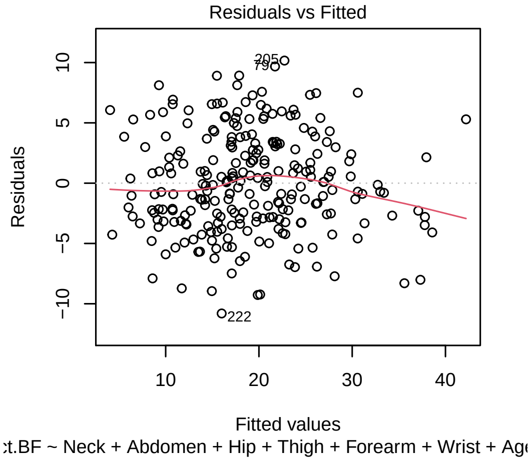


Fig. 1. Appendix 1.1

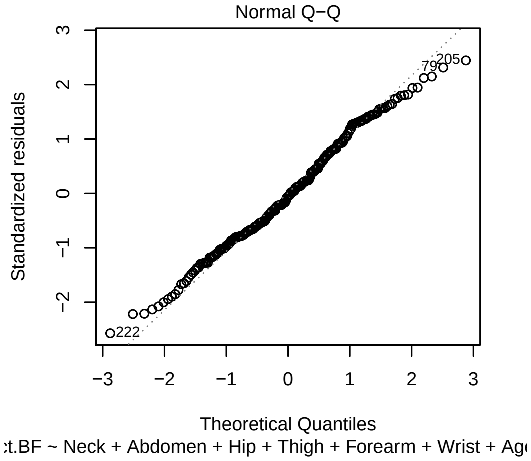


Fig. 2. Appendix 1.1