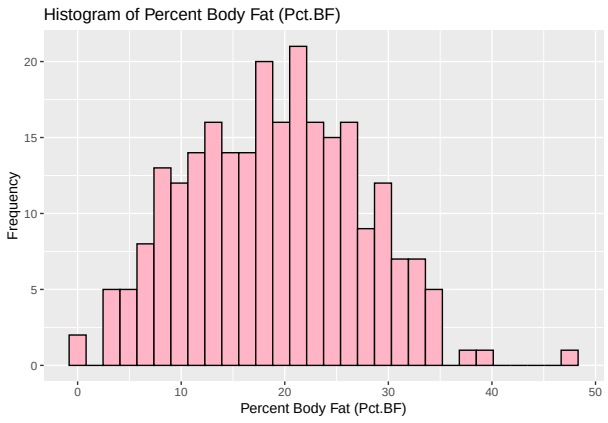
013E01 Executive Summary

John Fu 500212859 Lim Joanne 530233785

Abstract. This report undertakes analysis on the effect of potential risk factors on the percentage body fat. Utilizing a data set of 250 men, multiple linear regression models were constructed and refined using both forward and backward selection methods based on Akaike Information Criterion (AIC). The final model, chosen for its lower AIC and higher adjusted R-squared value, identifies eight key predictive measurements. Cross-validation techniques affirmed the model's predictive power, with a slight improvement in performance over the full model. In our discussion, we highlight the limitations of our analysis, i.e., the potential bias in our sample caused by the homogeneous nature, consisting only men, which might not accurately represent broader populations.

Introduction. In this report, we consider effect of several factors, i.e., Age, Height, Neck, Abdomen, Hip Thigh, Forearm and Wrist. We apply a multiple linear regression framework to understand how these factors will affect percentage body fat.

Data set. The data set we use in our analysis was originally collected by SOCR to estimate the percentage of body fat determined by underwater weighing and various body circumference measurements for 250 men. There are 16 variables included in the data and the key dependent variable for this analysis is `Pct.BF` - a variable measuring percentage body fat of men. The histogram showcases the distribution of individuals across various percent body fat levels,with a noticeable peak between 20 and 25%



Analysis. Our modeling began by first eliminating the 'density' variable from the data set due to its challenging nature of measurement in real-world settings and limited practical utility. Following this, we employed both backward and forward selection model based on the AIC to identify the optimal predictors for body fat percentage.

For backward selection, we firstly run a full regression model which contains all the explanatory variables in our data set. Then we try to drop variables to lower the AIC, and this narrowed down to eight key predictors: Age, Height, Neck, Abdomen, Hip Thigh, Forearm and Wrist.

Conversely, the forward selection began without any predictors, incrementally adding the most statistically significant ones. This process ended up including six predictors: Waist, Weight, Wrist, Bicep, Age and Thigh.

In conclusion, backward selection model appears to be more compelling with its slightly higher adjusted R-squared value and lower AIC value.

Assumptions. Before proceed to analyse our models results, we must first ensure that all our assumptions are satisfied. The key assumptions for our model are: 1. Linearity - the relationship between Y and x is linear. 2. Independence - all the errors are independent of each other. 3. Homoskedasticity - the errors have constant variance. 4. Normality - the error follows a normal distribution.

By looking at appendix 1, we can determine if our assumptions are met. Since there is no obvious pattern (e.g. no smiley or frowny face) in the residual vs fitted values plot, therefore the linearity assumption is met. The residuals don't appear to be fanning out or changing their variability over the range ofthe fitted values so the constant error variance assumption is met, and thus the Homoskedasticity assumption is met. Also in the QQ plot, apart from the top 6 or so points, the majority of points lie quite close to the line in the QQ plot. Hence, the normality assumption for the residuals is reasonably well satisfied. Additionally, we have quite large sample size so we can also rely on the central limit theorem to give us approxi-

mately valid inferences. Lastly, the independence of error terms is crucial and typically addressed during the initial phases of experimental design, i.e. **before data collection**. Each variable is designed to maintain its independence and since each observation doesn't inherently influence another, we can conclude that they are independent of each other.

Therefore given all our assumptions are met, our multiple linear regression model can be reliably analysed.

Results. Our final model is:

$$\begin{aligned}\widehat{Pct.BF} = 5.04 &+ 0.0726 \times Age - 0.268 \times Height \\ &- 0.451 \times Neck + 0.822 \times Abdomen \\ &- 0.195 \times Hip + 0.224 \times Thigh + 0.295 \\ &+ 0.295 \times Forearm - 1.731 \times Wrist\end{aligned}$$

Our full model has 15 variables and an R-squared of 0.737. However, in our new simplified model, we dropped 7 variables and obtained an R-squared of 0.739 which is slightly higher than the full model. The in-sample performance of the final model provided an R-squared that shows that approximately 73.9% of the total variability in Pct.BF is explained by the explanatory variables.

We used 10-fold cross validation to measure out-of-sample performance. From the output we are able to see that in the full model, it has a RMSE of 4.346 and MAE of 3.597. Conversely, the simplified model has a RMSE of 4.263 and MAE of 3.508. Thus we can see that the simplified model outperforms the full model.

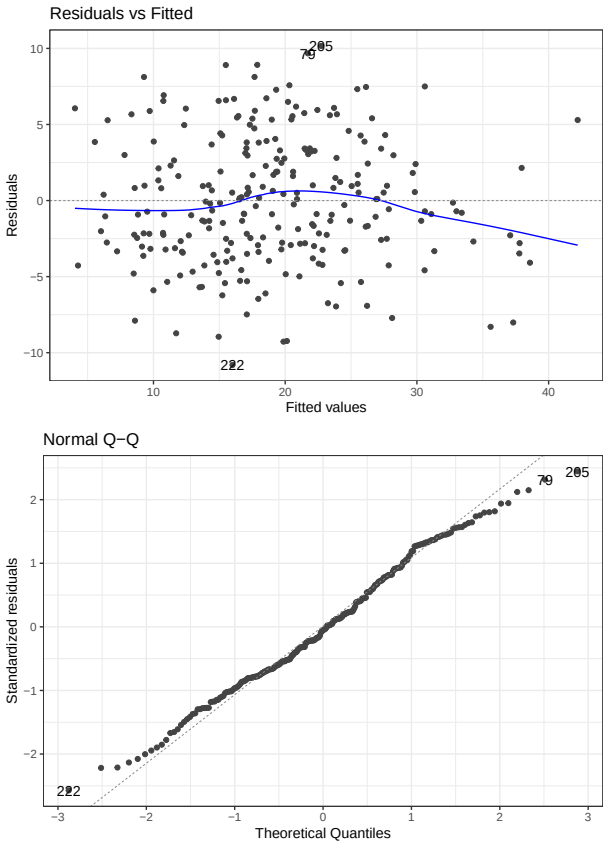
Now we interpret the coefficients: (See Appendix 2)

Making all the explanatory variables equal to zero, on average the predicted percentage body fat is 5.04%. Holding other variables constant, a year increase in age will leads to a 0.07 increase in body fat percentage. Holding all other variables constant, a 1cm increase in Height, on average would have a predicted decrease in body fat percentage by 0.268%. Holding other variables constant, a 1cm increase in Neck circumference, on average body fat percentage will decrease by 0.451%. Holding other variables constant, a 1cm increase in Abdomen circumference will lead to 0.822% increase in body fat percentage. Holding all variables constant, a 1cm increase in Hip circumference, on average would have a predicted decrease in body fat percentage by 0.195%. Holding other variables constant, a 1cm increase in Thigh circumference will lead to 0.224% increase in body fat percentage. Holding other variables constant, a 1cm increase in Forearm circumference will lead to 0.295% increase in body fat percentage. Holding other variables constant, a 1cm increase in Wrist circumference, on average body fat percentage will decrease by 1.731%.

Discussions and conclusions. In conclusion, we examined how certain risk factors effect the percentage body fat of men by applying a multiple linear regression model to outline the effect of each of the factors. We found that each of the risk factors we examined have statistically significant effect on body fat percentage, except for Hip, Thigh and Forearm.

However, there are some potential limitations. Most significantly the data itself being recorded on a homogeneous sample restricted to men, which may not generalize to women or to populations with diverse health profiles and demographics.

Appendix 1



Appendix 2

Coefficients

Full Model

Linear Regression

250 samples 14 predictor

No pre-processing Resampling: Cross-Validated (10 fold) Summary of sample sizes: 224, 225, 224, 226, 225, 224, ... Resampling results:

RMSE Rsquared MAE
4.345675 0.7245596 3.596767

Tuning parameter 'intercept' was held constant at a value of TRUE

Simple Model

Linear Regression

250 samples 8 predictor

No pre-processing Resampling: Cross-Validated (10 fold) Summary of sample sizes: 226, 224, 224, 225, 225, 225, ... Resampling results:

RMSE Rsquared MAE
4.233898 0.742006 3.502025

Tuning parameter 'intercept' was held constant at a value of TRUE

References

[MacFarlane(2017)] MacFarlane J (2017). *Pandoc: A Universal Document Converter*. Version 1.19.2.1, URL http://pandoc.org.

[Xie(2017)] Xie Y (2017). *knitr: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.17, URL https://yihui.name/knitr/.

[SOCR(2023)] *SOCR Data BMI Regression*. Retrieved April 2023, URL http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_BMI_Regression.

[Eddelbuettel(2023)] Eddelbuettel D (2023). *pinp: 'pinp' is not 'PNAS'*. GitHub repository, URL https://github.com/eddelbuettel/pinp.