

013E01 Executive Summary

John Fu

Lim Joanne

0.1 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.

0.2 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.

0.3 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 ___, 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 of the 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 approximately 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.

0.4 Results.

Our final model is:

$$\widehat{Pct.BF} = 5.04 + 0.0726Age - 0.268Height - 0.451Neck + 0.822Abdomen$$

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.

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 lead 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%.