Alternative Methods to Estimate Body Fat Percentage

Aidan Noney[a], **Ashish Dravid**[a], **Enes Gelebek**[a], **and Rajan Shankar**[a]

[a]The University of Sydney, Camperdown, NSW 2006

This version was compiled on November 21, 2020

ABSTRACT: Measuring body fat percentage is a complex procedure. In this report, we search for a multiple linear regression model that accurately and reliably estimates body fat percentage of adult males. Taking into account in-sample performance, out-of-sample performance and model stability, we find that a man's abdomen circumference and weight are the best predictors to achieve this. Our final selected model yields an out-of-sample mean absolute error (MAE) of 3.6%.

body fat | regression | ML | health

1. Introduction

Body fat percentage has long been a metric of importance in both the clinic and the lab. It is suggested that body fat percentage has negative influences such as morbidity and mortality on the body, and its ability to process drugs (Durnin and Womersley, 1974).

Current methods to estimate body fat percentage such as electric calipers, hydrostatic weighing and electrical impedance are rather technical and expensive. The cheapest of these methods — electronic calipers — are still usually off by ±3% (Marcin, 2018).

As such, the aim of this study is to build a multiple linear regression model to estimate body fat percentage within an error of ±3% using common and easily obtainable body measurements.

2. Data Description

The data set contains 250 observations of adult males. It was originally collected by the Human Performance Research Center at Brigham Young University as part of an academic study on modelling body composition (Penrose *et al.*, 1985). The response variable for our purposes is `pct_bf` and the predictor variables are easy-to-measure body characteristics.

3. Analysis

Linearity Assumption. This assumption needs to be checked before undertaking any model selection processes. We do this by looking at each scatter plot of the response variable `pct_bf` against each predictor variable (Figure 4 in the appendix). All of the predictor variables exhibit a fairly linear relationship with `pct_bf`, so we conclude that the linearity assumption is satisfied.

Independence Assumption. As mentioned earlier, our data was originally collected for an academic study on modelling body composition. Since such a study would have required the subjects to be independent of one another (i.e. not deliberately blood-related), we conclude that the observations in the data set of independent of each other.

Step-wise Search. We try two step-wise search methods: forward search and backward search. These methods both use the Akaike Information Criterion (AIC) to determine which variables to add to the null model or drop from the full model. After this is done, we clean up the resulting two models by dropping insignificant variables until all remaining variables are significant. These models are denoted by `M_forw` and `M_back` in future plots.

Stability Analysis. A more reliable method of picking out models is to look at stability plots. This helps us identify variables and models that are consistently favoured despite slight changes in the data set. We can simulate these slight changes by bootstrapping the original data set so that we end up with 100 re-sampled data sets. Then through an exhaustive search of the model space, we can find the best model on each of these data sets.

The `mplot` package (Tarr *et al.*, 2018) allows us to do the above and also produce a Variable Inclusion Plot (VIP) by repeating the whole bootstrapping process for a range of different General Information Criterion (GIC) penalties. The VIP is shown below:

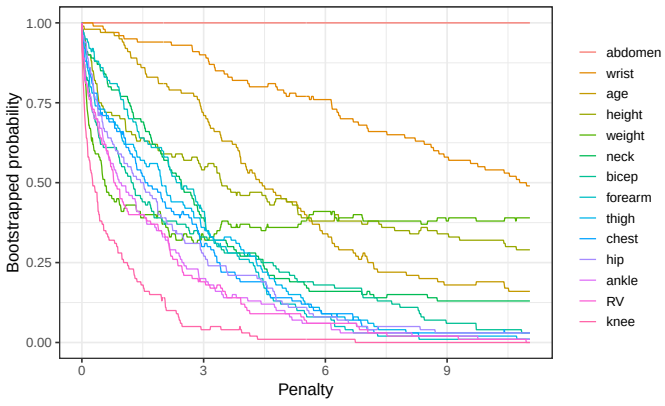

Fig. 1. Variable Inclusion Plot

Interestingly, `abdomen` is selected in *every* bootstrapped model for all values of the GIC penalty, so clearly it is a must-have in our final model. The `weight` variable also shows interesting behaviour; it initially sees a dip in usage, but then rises back up as the GIC penalty increases. This suggests that `weight` includes information from several predictor variables, hence why more parsimonious models prefer it.

Next, we analyse a Model Stability Plot (MSP). This plot is also produced using the `mplot` package and is shown on the next page in Figure 2. Each bubble represents a model, and its size represents the proportion of bootstrapped data sets that favoured that model for a particular number of parameters (intercept + predictor variables).

The 2-parameter case has one large bubble — this means that every bootstrapped data set favoured the *same* 2-parameter model. This model happens to consist of `abdomen` and the intercept, which aligns with what we saw in the VIP in Figure 1. We also see a dominant 3-parameter model that is shaded red. This means it includes `weight`, which reflects the tendency of `weight` to be included in parsimonious models, as seen in the VIP.

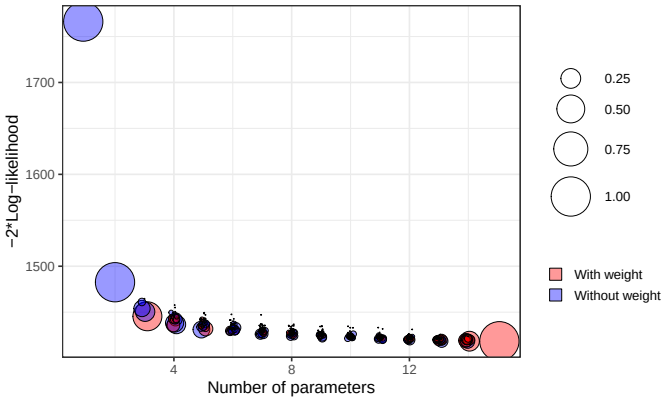

Fig. 2. Model Stability Plot

As a result, two models that we will investigate further are denoted `M_ab` (`abdomen` by itself) and `M_abwt` (`abdomen` + `weight`).

Out-of-sample Performance. We can evaluate the out-of-sample performance of each of our four candidate models by using 10-fold cross-validation (CV). This is done by partitioning the data into 10 folds and training the models on all 9-fold combinations. Then, we calculate the MAE on the remaining fold in each case and average the results.

This is all done using the `caret` package, and the CV results are shown below:

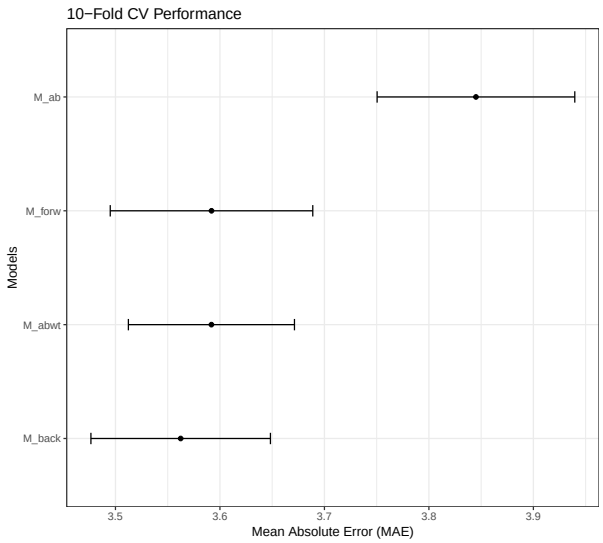

Fig. 3. Cross-validation results

We identify three best-performing models according to CV MAE: `M_forw`, `M_abwt` and `M_back`. Taking this information into account along with what we saw in the stability plots earlier, we select `M_abwt` as the best-performing model. It is very parsimonious, very stable and has a low CV MAE score. The model output for `M_abwt` can be found in Table 1 in the appendix.

Homoskedasticity Assumption. With our final selected model, we need to check that its residuals all have equal variance. After plotting the residuals against the fitted values of the model (Figure 5 in the appendix), we see that the residuals are distributed quite evenly about the zero line and do not exhibit any clear patterns. As such, we conclude that the homoskedasticity assumption is satisfied.

Normality Assumption. We require that the errors follow a normal distribution; $\varepsilon_i \sim N(0, \sigma^2)$. A good way to check for this is by comparing our residual quantiles with the corresponding theoretical normal quantiles through a Q–Q plot (Figure 5 in the appendix). In this plot, we see that residuals mostly adhere to the Q–Q line, but there is some departure at the extremes. However, our data set contains a lot of observations, so any inferences that rely on the residuals being normally distributed are still valid due to the central limit theorem.

4. Results

The equation for our final model `M_abwt` is:

pct_bf = -47.45 + 0.98 abdomen - 0.29 weight + ε

This means that a 1cm increase in abdomen circumference results in a 0.98% increase in body fat on average, holding weight constant. Alternatively, a 1kg increase in weight results in a 0.29% decrease in body fat on average, holding abdomen circumference constant. The negative coefficient for weight seems counter-intuitive. This is most likely a consequence of the high linear correlation of 0.87 between the `weight` and `abdomen` variables. However, the negative could still be interpreted to mean that if someone gains weight while keeping abdomen circumference constant, they are probably building muscle so their body fat percentage goes down.

The model output in Table 1 shows that our coefficient p-values are highly significant (<0.01). Even though the linear regression assumptions were all fine, these p-values may not be reliable due to the high correlation issue between the variables.

5. Discussion and Conclusion

Our final model had a CV MAE score of ~3.6; this means that our out-of-sample predictions were on average 3.6% off the true body fat percentage. This is very close to our goal of being about 3% off. Importantly, our model uses features that are much easier to obtain than using electronic calipers.

The multiple linear regression assumptions were all fine. However, the two variables in our final model were quite highly correlated, and this resulted in a strange negative coefficient for `weight`, which degraded the quality of our coefficient inferences. Another limitation is that while body fat percentage can only range between 0 and 100, our model can predict outside this range. Therefore, we need to be careful not to over-extrapolate.

Future research can involve looking at other statistical learning algorithms such as kNN and LASSO. Collecting more data would make these algorithms perform better, and would also allow us to implement a more rigorous train-validation-test workflow into our model selection and evaluation.

6. Bonus: Shiny App

We have developed a Shiny app that allows users to interact with our linear regression models for themselves. You can view scatter plots between variables, look at summary outputs for the 4 models discussed in this paper, and even use our final model `M_abwt` to make predictions! The app is available online at:

https://adra8108.shinyapps.io/Bodyfat/

Appendix

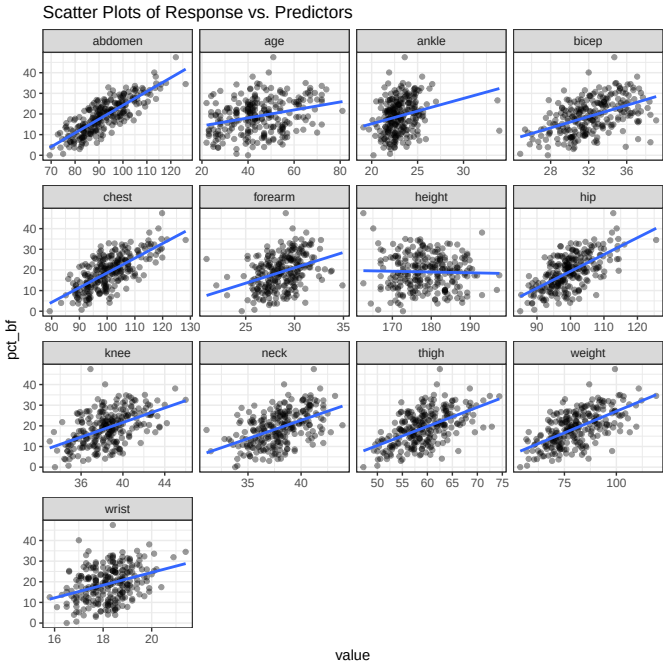

Fig. 4. pct_bf against each predictor variable

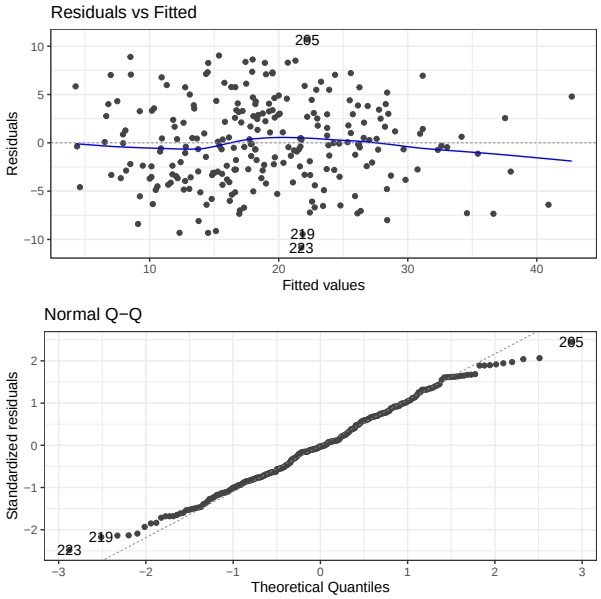

Fig. 5. Residual vs. Fitted plot and Normal Q–Q plot

Table 1. Results summary of abdomen + weight model

	Dependent variable:
	pct_bf
abdomen	0.977***
	(0.056)
weight	−0.293***
	(0.047)
Constant	−47.445***
	(2.636)
Observations	250
R^2	0.723
Adjusted R^2	0.721
Residual Std. Error	4.385 (df = 247)
F Statistic	321.966*** (df = 2; 247)
Note:	*p<0.1; **p<0.05; ***p<0.01

References

Durnin JVGA, Womersley J (1974). "Body fat assessed from total body density and its estimation from skinfold thickness: measurements on 481 men and women aged from 16 to 72 Years." *British Journal of Nutrition*, **32**(1), 77–97. doi:10.1079/BJN19740060.

Hlavac M (2018). *stargazer: Well-Formatted Regression and Summary Statistics Tables*. Central European Labour Studies Institute (CELSI), Bratislava, Slovakia. R package version 5.2.2, URL https://CRAN.R-project.org/package=stargazer.

Horikoshi M, Tang Y (2018). *ggfortify: Data Visualization Tools for Statistical Analysis Results*. URL https://CRAN.R-project.org/package=ggfortify.

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

Lüdecke D (2020). *sjPlot: Data Visualization for Statistics in Social Science*. R package version 2.8.6, URL https://CRAN.R-project.org/package=sjPlot.

Marcin A (2018). "How to Measure Body Fat: 6 Methods to Try." URL http://www.healthline.com/health/how-to-measure-body-fat.

Penrose K, Nelson A, Fisher A (1985). "FACSM." *Human Performance Research Center, Brigham Young University. Provo, Utah*, **84602**.

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.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." *Journal of Open Source Software*, **4**(43), 1686. doi:10.21105/joss.01686.