STA 206 Project

Samuel Van Gorden

11/14/2022

There are several observations that seem to be outliers: an observation with 0.0 percent body fat, an observation with 29.5-inch height . These should probably be removed. Many of the variables seem to be roughly normally distributed, with some having slight right skew.

Make some histograms more granular?

Possibly make some variables categorical (height, age). Maybe convert height to cm?

Forearm, ankle, age, density, wrist, and height seem to be least correlated with other predictors. (Try manually removing a few obviously correlated ones like hip and thigh and re-check). The potential predictors that do seem to be correlated with pbf appear to have a mostly linear relationship so higher order models are probably not needed. Age and (especially) height appear to have minimal correlation with pbf. Because of the aforementioned factors, it may be interesting to look at a first-order model with just forearm and ankle as predictors, though this model probably utilizes too few predictors.

First though, let's remove any remaining influential points.

Try model with relatively high correlation with pbf and low correlation with other variables.

Looks like a pretty bad model (ankle only significant at 0.1 level, very low R^2).

Perhaps we try a first order model with everything except hip and thigh (and density since we will be using this in its own model)?

Better, but we can probably do even better better.

Density, weight, height, and thigh are the only variables significant at < 0.05. Try fitting a model with just those.

All variables are significant at $< .001$. What does it look like with original data?

The results are quite different (only density and weight are significant at < 0.05). Let's continue using the model with the influential points removed.

It is becoming apparent that density is by far the best single predictor for a simple linear model, with abdomen being a distant second best. However, from the explanation of the data it appears that density is difficult and costly to measure, and being able to estimate pbf using the other variables may be beneficial. Let's start with all other variables and the data with influential points removed.

Why not just let R do the work for us and try using MSEE-based model selection techniques?

All BIC procedures produce the same model (pbf~weight+abdomen+forearm+wrist) with R_a^2=0.6984. The foreward AIC procedures produce the model (pbf~abdomen+weight+wrist+forearm+neck+bicep) with R_a^2=0.7026 and backward AIC procedures produce the model (pbf~age+weight+neck+abdomen+hip+thigh+forearm+wrist) with R_a^2=0.7064. Based on R_a^2, the best model of these appears to be the one selected by backward stepwise/selection.

What if we converted some of these variables to categorical variables? Lets try it with age (young, middle-aged, old) and height (short, tall).

Retry the model selection procedures with this new dataset.

Now let's compare the density-only model with the other models we have come up with.

Check assumptions for all models being used

The non-density models actually seem to meet the assumptions of linearity, constant variance, and normality of errors better than the density-only model. What if we perform a transformation of pbf before fitting the density-only model? Let's use the Box-Cox procedure to see which transformation would be best.

Looks like raising pbf to the 0.93 power greatly resolved the assumption violations!