Multiple Regression For the Cost-Effective Measurement of Body Fat.

Sydney[1], **Dhruvin**[1], **Steve**[1], **and Bie**[1]

[a]University of sydney

This version was compiled on November 20, 2020

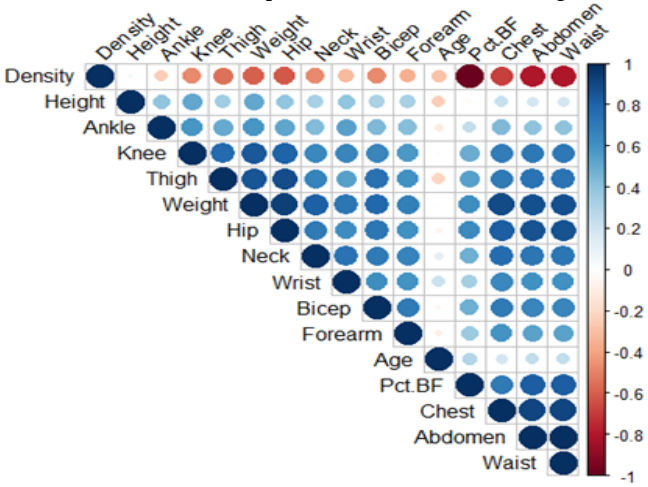
The premise for our multiple regression was to evaluate which easily performed measurements can be best used to predict a male's bodyfat percentage - aside from underwater density, since it is a major input into BMI. To achieve our goal, we have conducted an exhaustive search to find a suite of variables that minimises AIC and then have extensively adapted our model to the fit the assumptions of multiple linear regression.

Our Model: $ln(PercentageBodyFat) = 0.25ln(age) - 1.59ln(neck) + 15.05ln(abdomen) - 0.11(abdomen)^2 - 0.02(hip) + 0.02(thigh) + 0.67ln(bicep) - 2.01ln(wrist) - 46.02$

Introduction. We aimed to address the premise of our regression – To find the cheapest and most easily performed measurements to estimate a male's body fat percentage. From this we sought to address:

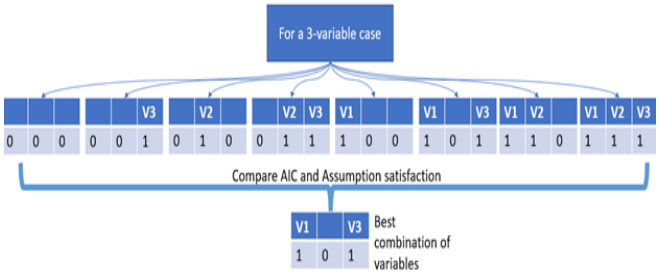
1. Does the suite of variables with the lowest AIC fit all the assumptions, if not, why?
2. What is the relationship between a unit increase in each of the selected variables (holding others constant) and a male's body fat percentage? (not including underwater density.)

Dataset:. The dataset – **SOCR Body Density Data** – originates from BYU and observes body fat percentages, and various body circumference estimates for 252 men. Although we only used 251 samples due to an error in the data. The following variables are in the dataset where percentage body fat is our dependent variable, whilst the others are independent (except density): *Density, Percentage Bodyfat, Age, Weight, Height, Neck, Chest, Abdomen, Waist, Hip, Thigh, Knee, Ankle, Bicep, Forearm, Wrist*. Included is a correlation graph that indicates basic relationships between the variables (Figure 1).



Model selection. The Primary comparison parameter we used to decide which model was better was the **Akaike information criterion (AIC)**. So, between two competing models, the one with the lower AIC wins. This was of course constrained by if that model sufficiently passed the assumptions of *Linearity, Homoskedasticity, Normality of errors, No Multicollinearity*. Our model selection processes consisted of the following two main steps iterated may times over.

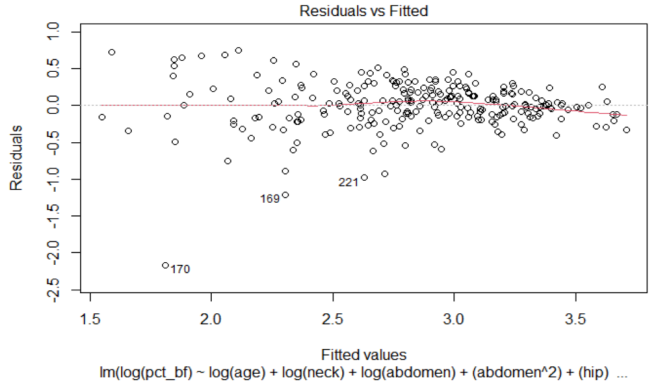
The first one is, **selecting the variables** that are used in the model and which ones are not. In order to do this, we used an exhaustive search algorithm (Figure 2). we coded up a program that would, for each set of variables go through all the different combinations of variables and output a data frame with the AIC scores for the different models (Figure 3). Then based on this we selected the model.



The second component is **Applying transformations** on the variables to better comply with the assumptions of the model. For this we looked at all the different variables and the considered how they can better fit the assumption, then applied the appropriate transformation.Then we ran it through the all the combinations and looked at the affect it had on our overall model. Depending on if the effect was positive or negative, we either kept the transformation or reverted it.

Assumption checking. After getting the final Model for body fat, we need to do a series of assumption check.

The first thing we need to check is the **linearity** between body fat and those transformed variables to make sure the good performance of prediction of the model. This can be best tested by the scatter plot (Figure 4).

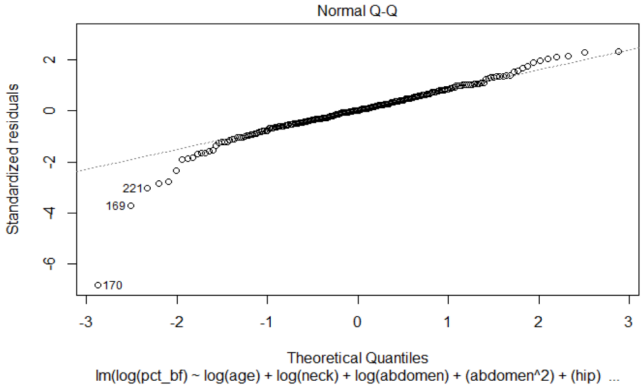


This plot indicates that plots are symmetrically distributed around the horizontal line on which residuals are zero, which confirms the assumption. However when the fitted values go beyond 2.5 the line shows a slight "bowed pattern", which is an evidence that the model may make systematic error when doing large predictions.

Besides linearity, Figure.4 also checks the **homoskedasticity** which indicates the spread looks plausibly constant whereas the points on the left side look wider than the right side. To confirm

if there is heteroskedasticity we also did the Goldfeld-Quandt test and got the p-value 0.02, therefore we have to reject the null hypothesis and stick to the alternative hypothesis that the model does not have constant error variance, which may result in the confidence interval to be too wide or narrow. So even after a few non-linear corrections of our model, this assumption still fails.

Third assumption is the **normality of the error distribution**. From the (Figure 6)



which is the normal quantile plot of the residuals, most of the points are lying on or close to the diagonal line except a few points on the lower and upper tail. Overall, the points follow a normal distribution and we can hold this assumption for our model.

Another important thing we need to check is that our model has **no or very little multicollinearity** between different variables. We use Variance Inflation Factor table (Figure 7)

```
   log(age)    log(neck) log(abdomen)      abdomen          hip        thigh
   1.849840     3.256886   220.294566   222.178833     8.983704     5.936652
log(bicep)   log(wrist)
   2.802291     2.464316
```

to decide if there's strong correlation among variables and remove the variable having high VIF values(VIF value larger than 10). After dropping the variables forearm and waist and introduce a new variable the square of abdomen, the VIF decreases significantly. We noticed the VIF of log(abdomen) and the square of abdomen are still large, we assume that is due to the high correlation with the variable abdomen but not with the other variables.

Furthermore, the model needs to satisfiy the assumption of **independence**. Because independence is always dealt with before the data collection, and we know the sample comes from random data. Therefore, we do not need to worry about it too much when claiming that our model fits this assumption.

Results. The final model we were able to select was:

$ln(PercentageBodyFat) = 0.25 \ln age - 1.59 \ln neck + 15.05 \ln abdomen - 0.11(abdomen)^2 - 0.02(hip) + 0.02(thigh) + 0.67 \ln bicep - 2.01 \ln wrist - 46.02$

This was the best model we were able to produce that satisfied most of the assumptions and had the best predictive power from the options available to us. Even so, this model does not perform particularly well when compared to the percentage body fat calculation using underwater density. If **AIC and adjusted R-squared** are used to measure the performance of our model, our selected model has an **AIC of 161.12** and an **adjusted R-squared of 0.65**. On the other hand, A model based purely upon the underwater density has an AIC of -55.03 and an adjusted R-squared of 0.85.

Now onto whether all our selected variables are statistically significant. As discussed before, our model did not sufficiently pass the assumption of Homoskedasticity. Therefore, in order to accurately calculate the significance, we've used **white's standard errors** (Figure 8). What we found was that: ln(abdomen), ln(neck) and ln(wrist) have a P value of <0.001. The variables ln(age) and abdomen2 have a P value of < 0.01 and the variables, hip and thigh have a P value of < 0.05. The only non-statistically significant variable in our model was the ln(bicep) with a P value of 0.16.

From our model we can **infer**, how much change in a given variable would constitute a 1% rise in the body fat percentage. This would of course only be true if all the other variables were kept constant.

Variable	Amount of movement
Age	0.25% rise
Neck	1.59% drop
Abdomen	$15.05(\ln(\text{units})) - 0.11(units^2)$
Hip	0.02 units drop
Thigh	0.02 units rise
Bicep	0.67% rise
Wrist	2.01% drop

The only anomaly here is the abdomen variable. Where, it is not so clear how the changes in this variable would affect the percentage body fat. The only way to figure that out is to do the calculation.

Limitation.

1. About limitations for our dataset, In terms of estimating for a whole population our sample **size is considered small**. The context of our sample is unknown, for example, we do not know how participants were chosen and if the sample is random. We do not know when this data was obtained, whether it is old or new.
2. Also, in regard to the context of our sample, some **external variables** which may affect an individual's body fat have not been recorded in the data. some improvements or variables to include in the future could include variables such as an individual's muscle percentage.
3. Moreover, our data **only included male participants**, therefore we were not able to test and look into similar questions relating to females and see if there were any differences between female and male body measurements in relation to body fat.

Conclusion.

In conclusion, the central focus of this study is selecting the appropriate model from a potentially large class of candidate models. Obviously, linear regression models with the lowest AIC are capable of estimating body fat, during our research, we determined the main variables that affect an individual's body fat, those being age, neck, abdomen, hip, thigh, bicep, wrist, which have an acceptably strong correlation with body fat. if future research was to be conducted, it would be crucial to have a muscle variable included to improve the accuracy and practicality of the prediction..

Appendix.

Figure 3

	Variables	Adj_R-Squared	AIC	BGtest
6008	pct_bf, age, neck, abdomen, waist, hip, thigh, bicep, forearm...	0.644388070249559	162.792469859367	0.049924123924305
6007	pct_bf, age, neck, abdomen, waist, hip, thigh, forearm, wrist	0.64268154440649	163.024191306085	0.049442462675831
6974	pct_bf, age, height, neck, abdomen, waist, hip, thigh, forear...	0.643244579485404	163.591859232614	0.049575847528208
7407	pct_bf, age, weight, abdomen, waist, bicep, wrist	0.638857085293249	163.741534101764	0.043719620365080
2391	pct_bf, height, neck, abdomen, waist, hip, bicep, wrist	0.640084609100123	163.862684751016	0.033423169244561
6460	pct_bf, age, height, abdomen, waist, bicep, wrist	0.638566304677034	163.94194019391	0.037433207391361
2358	pct_bf, height, neck, abdomen, waist, bicep, wrist	0.638357081257711	164.086037325118	0.045301889346659
6456	pct_bf, age, height, abdomen, waist, wrist	0.636314161727874	164.515484955906	0.037397734201128
6488	pct_bf, age, height, abdomen, waist, thigh, wrist	0.637731138672862	164.51664156934	0.046772436900451
2356	pct_bf, height, neck, abdomen, waist, wrist	0.636174800021296	164.610881626302	0.044464753535395
995	pct_bf, neck, abdomen, waist, hip, bicep, wrist	0.637590660891183	164.613178123234	0.027192088822125
3694	pct_bf, weight, neck, abdomen, waist, hip, bicep, wrist	0.638902665398615	164.679048276622	0.040063980674736

Figure 8

Regression output with White's standard errors

```
                Estimate  Std. Error t value  Pr(>|t|)
(Intercept)  -46.0218125  12.8218867 -3.5893 0.0004019 ***
log(age)       0.2457388   0.0860792  2.8548 0.0046831 **
log(neck)     -1.5862298   0.4754205 -3.3365 0.0009831 ***
log(abdomen)  15.0462230   3.5186357  4.2762 2.746e-05 ***
abdomen       -0.1105790   0.0374263 -2.9546 0.0034425 **
hip           -0.0164758   0.0076136 -2.1640 0.0314522 *
thigh          0.0175526   0.0079597  2.2052 0.0283906 *
log(bicep)     0.6732097   0.4824920  1.3953 0.1642222
log(wrist)    -2.0109696   0.5121158 -3.9268 0.0001125 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```