Introduction

Previous research suggests that body fat percentage can be estimated once body density has been determined. Though accurate, methods used to acquire density (e.g. underwater submersion) is quite complicated. Thus, proposing a simple, accurate and robust rule-of-thumb to estimate body fat percentage using clinically available measurements is necessary.

Data Cleaning

Summary statistics of body fat percentage and some other important predictors are shown below.

Table 1 Summary Statistics

	BODYFAT	WEIGHT	HEIGHT	ADIPOSITY	CHEST	ABDOMEN	HIP
Min.	0.00	118.50	29.50	18.10	79.30	69.40	85.00
1st Qu.	12.80	159.00	68.25	23.10	94.35	84.58	95.50
Median	19.00	176.50	70.00	25.05	99.65	90.95	99.30
Mean	18.94	178.90	70.15	25.44	100.82	92.56	99.90
3rd Qu.	24.60	197.00	72.25	27.32	105.38	99.33	103.50
Max.	45.10	363.10	77.75	48.90	136.20	148.10	147.70

WEIGHT and HEIGHT are measured in pounds and inch respectively, with step length 0.25 unit. Relationship between these two measurements and ADIPOSITYd is

$$ADIPOSITY = \frac{WEIGHT\ (lbs)}{HEIGHT(inch)^2} \cdot 703$$

Other circumference measurements are measured in centimeters.

Record 182 and 172 are two abnormal data points, whose body fat percentages are 0 and 1.9. Since the size of the dataset is relatively small and other predictors for records with 0 body fat percentage are almost minimum, it's quite hard to fix these records and we removed them from the dataset. Another abnormal point is record 42, whose weight, height and adiposity are 205, 29.5 and 29.9. After comparing record 42 with other records, we fixed its height using adiposity formula mentioned above. We also dropped IDNO and DENSITY because they are useless in prediction. There are 250 observations, 1 response variable and 14 predictors left after data cleaning.

Motivation for Model

Due to serious multicollinearity risen from multiple linear regression, we want to handle it using principal component regression. Since transformation coefficients for neck, knee, ankle, biceps, forearm and wrist circumferences are insignificant compared to the others, we drop these predictors.

Principal component regression is implemented as follows:

First, calculate covariance matrix of the predictors and find eigen values $\lambda's$ and eigen vectors $e's$ of covariance matrix.

Second, keep those principal components whose cumulated variance reaches 95% of total variance, i.e. proportion of sum of eigen values for selected transformation greater than 0.95.

Finally, do multiple linear regression on these selected principal components.

The final model is

$$\begin{aligned} BODYFAT = -41.16 + 0.02 \cdot AGE - 0.13 \cdot WEIGHT - 0.26 \cdot HEIGHT + 0.18 \cdot ADIPOSITY + 0.30 \cdot CHEST \\ + 0.54 \cdot ABDOMEN + 0.11 \cdot HIP + 0.09 \cdot THIGH \end{aligned}$$

The model indicates positive linear relationship between body fat % and age, adiposity and circumference of chest, abdomen, hip and thigh, while negative effects of weight and height. The coefficients can be interpreted as increase in body fat % when predictor increase 1 unit.

Statistical Analysis

R^2 is 0.6936, which implies about 69.36% of total variation in body fat percentage can be explained by this model. Estimated standard deviation for error term in linear regression is 4.215.

We conducted t-test on each coefficient. Regression estimates with respect to three selected principal components, standard error, t-statistic and p-values are shown below.

$$BODYFAT = -41.1569 \quad + \quad \underset{(sd = 0.0089, -18.058, < 0.01)}{0.1611 \cdot pc1} \quad + \quad \underset{(sd = 0.0204, 8.984, < 0.01)}{0.1838 \cdot pc2} \quad + \quad \underset{(sd = 0.0579, 11.800, < 0.01)}{0.6827 \cdot pc3}$$

$$pc1\ (weight\ component) = 0.88 \cdot WEIGHT + 0.04 \cdot HEIGHT + 0.10 \cdot ADIPOSITY + 0.24 \cdot CHEST$$
$$+0.30 \cdot ABDOMEN + 0.20 \cdot HIP + 0.14 \cdot THIGH$$

$$pc2(age\ component) = 0.96 \cdot AGE - 0.08 \cdot WEIGHT - 0.06 \cdot HEIGHT + 0.04 \cdot ADIPOSITY + 0.12 \cdot CHEST$$
$$+0.20 \cdot ABDOMEN - 0.04 \cdot HIP - 0.09 \cdot THIGH$$

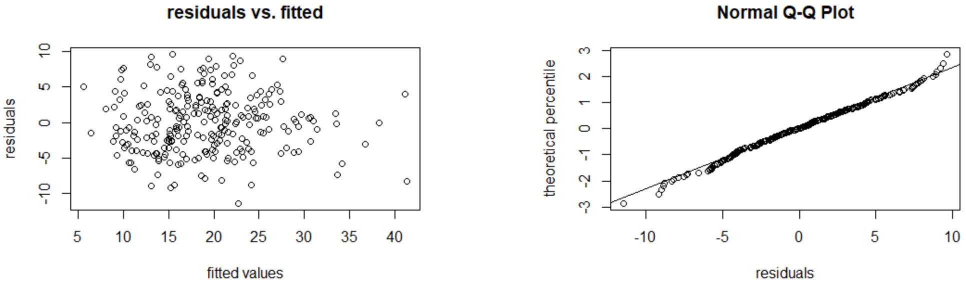
$$pc3(circumference\ component) = -0.23 \cdot AGE - 0.38 \cdot WEIGHT - 0.38 \cdot HEIGHT + 0.23 \cdot ADIPOSITY$$
$$+0.35 \cdot CHEST + 0.67 \cdot ABDOMEN + 0.13 \cdot HIP + 0.12 \cdot THIGH$$

Under 95% significance level, all three principal components are highly significant in predicting body fat percentage. Small standard error indicate the estimates are quite accurate.

Model Diagnostics

Basic assumptions are no outliers, no interference between different observations and body fat percentage is normally distributed. Diagnostics plots are shown below.

Figure 2 Diagnostic plots without potential outliers



After dropping potential outlier and influential point obs.39, estimates change not too much. Scatter plot for residuals versus fitted values indicates independency and equal spread. QQ plot shows the data does not violate normality assumption. No other influential points emerged.

Model Strengths/Weaknesses & Discussion

Strengths: The method can avoid multicollinearity between predictors and produce pretty accurate estimate. We abandon some predictors which are not quite important and redo principal component analysis, which can help make final model more concise.

Weaknesses: The estimated variance is quite large and R^2 is not close to 1, which implies the prediction is not precise enough. When choosing principal components, we just use empirical 95% of cumulative total variance, rather than using variable selection method in linear regression. For further improvement, we may try to use variable selection methods and include nonlinear term in the model or use other technique to better deal with multicollinearity.

Contributions

YF initial git repo, report and shiny-app. Data cleaning and modelling is done by MC. YS write slide for presentation.