**INTRODUCTION**

SUMMARY: Accurate measurement of body fat is inconvenient/costly and it is desirable to have easy methods of estimating body fat that are not inconvenient/costly. Using regression techniques, it is possible to come up with the best results for the measurement of body fat. A variety of popular health books suggest that the readers assess their health, at least in part, by estimating their percentage of body fat. In Bailey (1994), for instance, the reader can estimate body fat from tables

using their age and various skin-fold measurements obtained by using a caliper.

**DATA SET:**

The dataset was obtained from CMU StatLib website. It consists of the percentage of body fat determined by underwater weighing and various body circumference measurements for 252 men. There are 14 regressors in this data set which are as follows

Density determined from underwater weighing X1

Age (years) X2

Weight (lbs) X3

Height (inches) X4

Neck circumference (cm) X5

Chest circumference (cm) X6

Abdomen 2 circumference (cm) X7

Hip circumference (cm) X8

Thigh circumference (cm) X9

Knee circumference (cm) X10

Ankle circumference (cm) X11

Biceps (extended) circumference (cm) X12

Forearm circumference (cm) X13

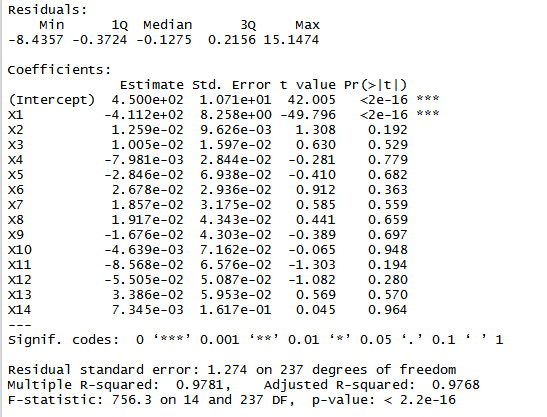
Wrist circumference (cm) X14

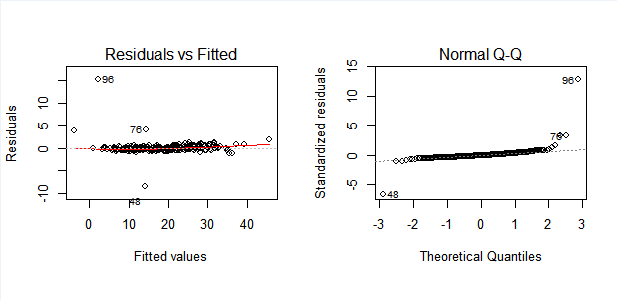
The response variable which was estimated here using these regressors was Percent body fat. (Y)

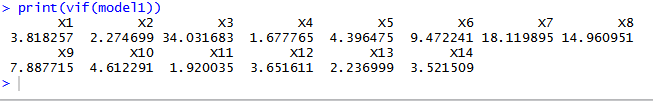
OBJECTIVE: The objective of the project is to provide significant parameters which influences the response. The analysis includes fitting a linear model followed by checking for the model assumptions, multicollinearity, steps followed to eliminate it (principal component regression), significance test of the model.

**DATA ANALYSIS:**

ORIGINAL MODEL: Regression analysis was performed on the full model with all regressors and the results obtained from R are displayed below.



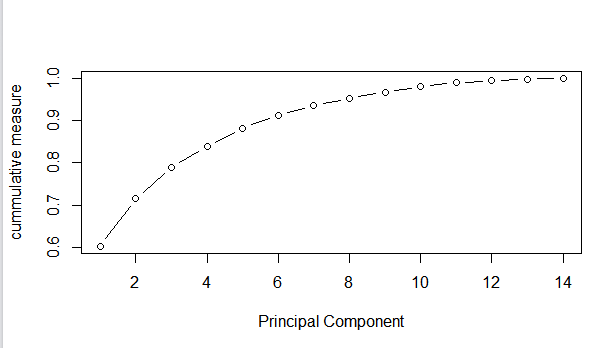




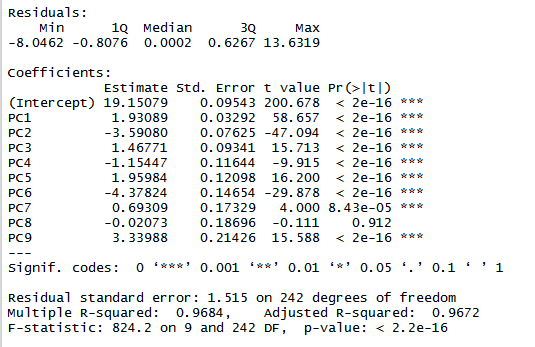
* OBSERVATIONS

From the inspection of residual plots, we see that there is nothing wrong in the normal probability plot as all residuals falls along the straight line as shown in the figure. Also since the residuals vs fitted plot depicts no pattern, we can conclude that the there is no need of any data transformation. R squared value and Adjusted R squared value are 97.81% and 97.68% respectively suggests that the model answers the variability in the response. Also from the table we see that there is severe multicollinearity in the model since VIF of X3,X7,X8 are greater than 10 and VIF of X6 and X9 are above 5.

* MULTICOLLINEARITY

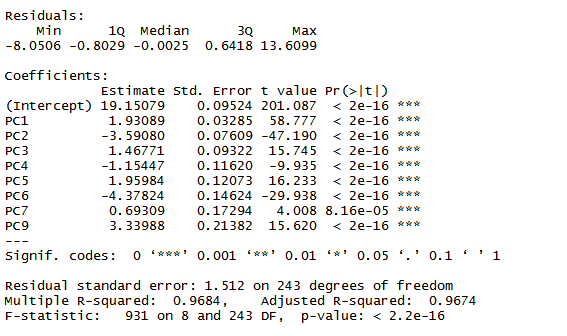


In order to get rid of multicollinearity, I used ridge regression but the output graph which I got was series of parallel lines which could not be interpreted. Hence, I used principal – component regression technique. The principal – component estimator reduces the effects of multicollinearity by using a subset of the principal components in the model. The plot is as shown in the figure. At cumulative measure of 0.95 we can see that the number of significant principal components are 9. In other words, we can conclude that only 9 principal components explain most of the variability of the data. Hence these 9 PC’S are retained for further analysis.

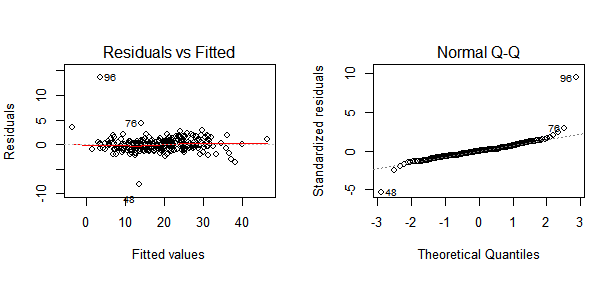


The coefficients of these 9 principal components are as shown in the table below.

From the table, we see that **‘PC8’** is further not significant and hence can be eliminated from the model. The final coefficient table is as shown below.



The coefficients of the principal components are as shown in the table. All the components are significant. Although there is no change in the value of R squared but the adjusted R squared value is increased very slightly.



**CONCLUSIONS & RECOMMENDATIONS**

* The data was fit with simple linear equation, but the data showed showed multicollinearity between regressors. It was eliminated through principal component regression. Then the significance test was done to eliminate **PC8** to obtain the final model. The final R-squared of 96.84% and Adjusted R-squared of 96.745% explains the variability in the model. Thus, this model can be used for further in-depth analysis and recommended to user.
* Although principal component regression is used here for to remove multicollinearity in the model but the suitable method is by collecting additional data. Since in this data set there are 252 observations, the model might be a good fit having more data. But also, if the new data points are unusual and atypical, their presence could be highly influential in the fitted model.
* For future scope, the model can be used to calculate the Body Mass Index (BMI) of a person and fit the regression line and used for further analysis which can serve as an indicator of a person’s obesity and underweight.

**REFERENCES**

* Montgomery, D. C., Peck Elizabeth, A., & Geoffrey Vining, G. Introduction to Linear Regression Analysis, Fifth Edition
* http://lib.stat.cmu.edu/datasets/bodyfat
* Bailey, Covert (1994). \_Smart Exercise: Burning Fat, Getting Fit\_,Houghton-Mifflin Co
* <https://people.duke.edu/~rnau/rsquared.htm>