How Much Fat Does Your Body Contain?

Name: Negin Bolkhanian
Student's ID: 301261754
Instructor's Name: Jiguo Cao

Abstract

Obesity is one of the more talked about health problems in both the medical world and the media today. BMI has been the go to equation when determining whether obesity exists in a person; however, the most concerning factor in obesity is arguably the body fat percent. This makes it important to be able to calculate this percentage number, in order to take the next steps in eliminating obesity in a given person. This paper aims to achieve a multiple linear regression model using a collected measurable – such as density, abdomen, neck, hip and other measurements– data set from 250 men. AIC and BIC value of forward and backward selection was used to choose the model. As a result, forward selection model based on its AIC value was selected as the final model with linear associated factors density, age and abdomen.

Keywords: *Multiple Linear Regression, Model Selection, Body Fat Percent, BMI, Obesity*

Table of Contents

Introduction

In the past decade, both the medical professionals and the media have been concerned by the problem of obesity. Obesity has a direct correlation with percent body fat, making it easier to keep track of weight problems by monitoring body fat percentage. One of the more convenient ways of tracking body fat percent has been the calculation of Body Mass Index (BMI). With all the data available with regards to this subject including BMI tables for different age groups, this method has been established as a frontrunner in this field.

However, this paper focuses on how all of this can be achieved without the use of complicated equations such as BMI, and with a linear equation which contains multiple regressions instead. By doing so, it will be possible to achieve the goal of this paper which is determining an estimation for body fat percent using very simple tools such as a measuring tape and a scale.

Data Description

The data used is on percent body fat measurements for a sample of 252 men, along with various measurements of body size. Weighing the person underwater is usually the way to measure percent body fat, which is a cumbersome procedure. The goal –as mentioned before– is acquiring a regression model that will allow an accurate estimation

of percent body fat, given easily obtainable body measurements. The data were taken from the *Journal of Statistics Education* website.

Variables

Response variable

- BODYFAT: Percent body fat using Brozek's equation

Regressor variables

- DENSITY: Density (gm/cm^3) weight/volume

  (Volume: lean tissue volume + fat tissue volume)
- AGE: Years
- WEIGHT: lbs
- HEIGHT: Inches
- ADIPOSITY: Adiposity index = Weight/Height^2 (kg/m^2)
- NECK: neck circumference (cm)
- CHEST: Chest circumference (cm)
- ABDOMEN: Abdomen circumference (cm) "at the umbilicus and level with the iliac crest"
- HIP: Hip circumference (cm)
- THIGH: Thigh circumference (cm)
- KNEE: Knee circumference (cm)
- ANKLE: Ankle circumference (cm)

- BICEPS: Extended biceps circumference (cm)
- WRIST: Wrist circumference (cm) "distal to the styloid processes"
- FORARM: Forearm circumference (cm)

Special Notes

There are a few errors in the available dataset. For instance, the body densities for cases 96, 76, and 48 each appear to have one digit in error as is visible from the two body fat percentage values. Also, the presence of a man less than 3 feet tall who is over 200 pounds in weight (case 42) is noteworthy. (Presumably there has been a mistake in recording the height. 69.5 is more feasible instead of 29.5). The percent body fat approximations are shortened to zero when negative (case 182). These data can be used to show the utility of multiple regression and to provide practice in model building.

Method and Models

Using the multiple linear regression models usually leads to the response variables being considered as a linear function of the regressor variables with an error variables term ε. Five assumptions are usually in order if the data are suitable for using the linear regression models:

1. The relationship between the response y and the regressors should be at least approximately linear.

2. The error term, ε has zero mean.

3. The error term, ε has constant variance σ_2.

4. The errors between different individuals are uncorrelated.

5. The errors are normally distributed.

Consequently, we would initially create the scatterplot matrix for the dataset so as to check whether linear relationships exist between the response variable and some explanatory variables (the first assumption).

Full model:

BODYFAT = β0 + β1 DENSITY + β2 AGE + β3 WEIGHT + β4 HEIGHT + β5 ADIPOSITY + β6 NECK + β7 CHEST + β8 ABDOMEN + β9 HIP + β10 THIGH+ β11 KNEE + β12 ANKLE + β13 BICEPS + β14 FOREARM + β15 WRIST + ε

For illustration purposes, we are going to go through a part of the scatter plot (the first ten variables). Due to there being 17 variables, it would not be as appropriate to print them all for demonstration. The linear relation between response variable and the independent variables are seen with the help of the following figure. To check for the made assumptions not being violated, and also in order to see the linearity relation with more clarity, the full model will be analyzed.

The two methods that are used in order to create a multiple linear regression model with appropriate independent variables are Akaike's Information Criterion (AIC) and Bayesian's Information Criterion (BIC).

Stepwise model selection will be used to choose the final model from the candidate models. We would set two base models for the model selection step. One is the full model (M_1) with the other one being M_0=1. Four models were selected here using the two talked about methods (AIC and BIC) in forward, backward selection. The aim is to choose the model with the smallest AIC and BIC.

Results

Scatter Plot

Figure 1 is demonstrative of the circumstance that DENSITY is in fact directly correlated with body fat percent. This is to be expected, as density directly correlates with body mass, and the more the body mass within a specific body frame (Volume), the more body fat percent is anticipated. The remaining variables other than DENSITY also appear to have this relationship with body fat percent, as it can be seen in ADIPOSITY, NECK, CHEST and ABDOMEN below in the figure.

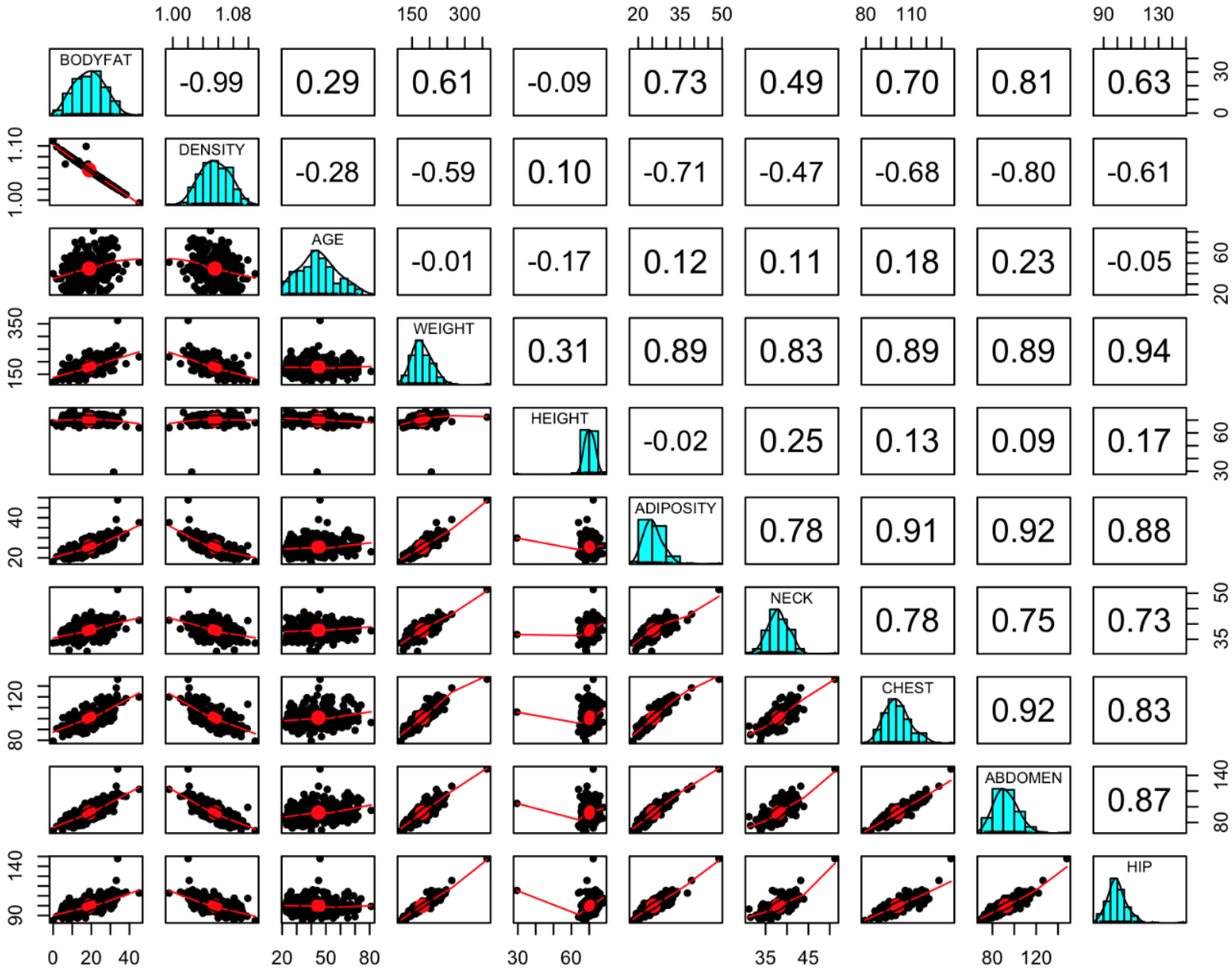

Figure 1 - Scatter plots metrics of the response variables and 9 independent variables.

Full Model Analysis

After creating the full model with sixteen independent variables, the following summary table was acquired.

Table 1: Summary of full model M_1					
Estimated σ	1.66		Significant Predictors	Estimated value	P-value
R^2	0.9787				
F-statistic	723.9		(Intercept)	4.190e+02	< 2e-16
P-value	< 2.2e-16		DENSITY	-3.816e+02	< 2e-16

As the table above shows, even though the model did follow the first assumption of the linear regression model, DENSITY was the only variable among all sixteen to be significant with P-value of less than 5 percent. The high number of variables may have been the reason which resulted in R^2 to be close to one, which concludes that the model might not be a very good one.

Model Selection

The AIC and BIC model selection steps were used in order to find the final model with appropriate variables. The following table illustrates the summary of all the four candidate models with their AIC and BIC values.

Table 2: Summary of Model Selection			
	AIC/BIC	Value	Fitted Model
Forwards Selection	AIC	75.39	BODYFAT = β0 + β1 DENSITY + β2 ABDOMEN + β3 AGE
	BIC	86.51	BODYFAT = β0 + β1 DENSITY + β2 ABDOMEN
Backwards Selection	AIC	76.97	BODYFAT = β_0 + β_1 DENSITY + β_2 AGE + β_3 WEIGHT + β_4 CHEST + β_5 ANKLE + β_6 BICEPS
	BIC	92.21	BODYFAT = β_0 + β_1 DENSITY + β_2 AGE + β_3 WEIGHT

Table 2 shows how we studied both models based on their AIC values as the better one, because of the fact that the AIC values as seen above are not very significant. Finally, two models were selected from the table above (table 2). The final selected models are as stated below. The two models were called Forward.Selection.AIC and Backward.Selection.AIC.

Forward.Selection.AIC = β0 + β1 DENSITY + β2 ABDOMEN + β3 AGE

Backward.selection.AIC = β_0 + β_1 DENSITY + β_2 AGE + β_3 WEIGHT + β_4 CHEST + β_5 ANKLE + β_6 BICEPS

Table 3: Summary of Model Forward.Selection.AIC					
Estimated σ	1.152		Significant Predictors	Estimated value	P-value
R^2	0.98		(Intercept)	4.14e+02	< 2e-16 ***
Adj R^2	0.98		DENSITY	-3.79e+02	< 2e-16 ***
F-statistic	3703		ABDOMEN	4.79e-02	2.71e-05 ***
P-value	< 2.2e-16		AGE	9.51e-03	0.115

Table 4: Summary of Model Backward.selection.AIC					
Estimated σ	1.16		Significant Predictors	Estimated value	P-value
R^2	0.9779		(Intercept)	4.23e+02	< 2e-16 ***
Adj R^2	0.9777		DENSITY	-3.85e+02	< 2e-16 ***
F-statistic	3663		AGE	1.115e-02	0.088 .
P-value	< 2.2e-16		WEIGHT	1.32e-02	0.067 .
			CHEST	3.29e-02	0.145
			ANKLE	-8.29e-02	0.136
			BICEPS	-6.22e-02	0.123

Tables 3 and 4, as shown above, demonstrate the summary of the two fitted models in order to give a better understanding of the two models.

Forward.Selection.AIC appears to be the better model as shown in the tables above, due to the fact that all its independent variables have linear relationships with P-value of less than 5 percent. It becomes clear that the second model is also acceptable when we take a look at the residual plot and the QQ plot of the model. However, the reason that the first model was chosen was its simplicity, plus the second model includes some factors with no strong linear relationships with body fat percent.

Final Model Analysis

As shown in table 3, R^2 and adjusted R^2 values show that the final model can explain 98 percent variability. The following four diagnosis plots constructed for the final model (AIC) are shown below to make sure that our five assumptions are held intact.

In the diagnosis plots below, there are a number of outliers, which are cases 48, 76 and 96 as talked about beforehand for having one digit in errors for body fat percent. These outliers are influential, which is the reason that we keep them as they are in order to achieve a linear regression model. Getting rid of them or changing their values to more realistic values would result in non-linearity in our residual, which violates our pre-set assumption.

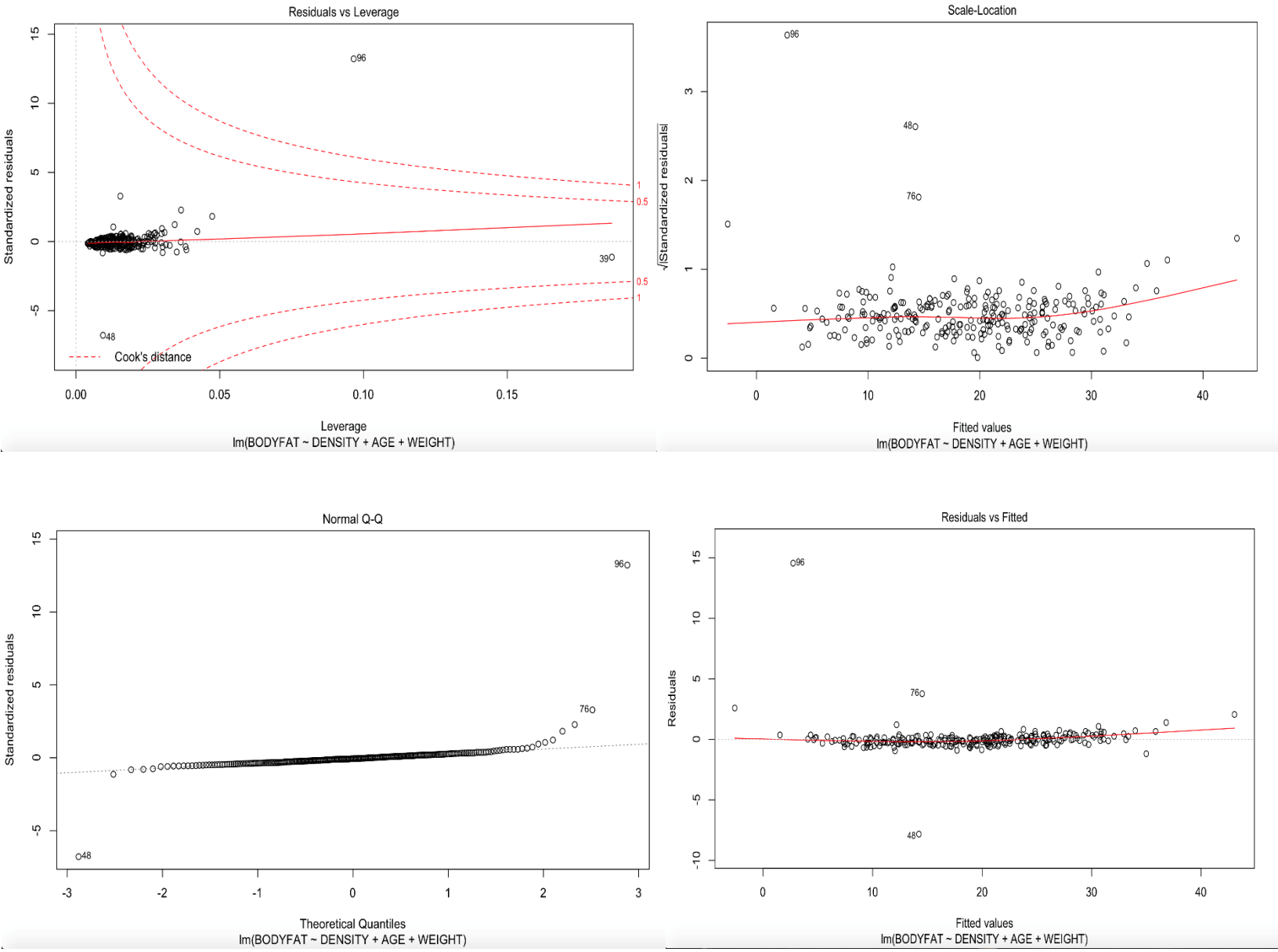

Figure 2 – Linear regression model assumption diagnosis plots for the final Forward.Selection.AIC model

The Residual vs. Fitted plot shows that residuals are spread with a minimal distance from the fitted line, proving them to be random, as they are not non-linear. The Normal Q-Q plot shows whether residual are normally distributed or not. Since the residuals follow

the straight line well, the normality assumption is satisfied. The outliers have caused the pulling up that happens at the end of the plot.

The Scale-Location plot displays whether residuals are spread equally along the ranges of predictors, helping us in checking the satisfaction of the assumption of constant variance. The residuals appear to be spread randomly which holds up the assumption. The leverage plot helps us in finding influential cases, proving our special cases (e.g. case 96) are located outside of the Cook's distance line, meaning that they have high Cook's distance scores, causing them to be significant to the regression results.

Conclusion

Our final multiple linear regression model is

BODYFAT= 4.14e + 02 - 3.79e + 02 DENSITY + 4.79e - 02 ABDOMEN + 9.51e - 03 AGE

According to the model, we can conclude that body fat percent is linearly associated with the following factors

- DENSITY: Density (gm/cm^3) weight/volume – Negative relationhsip
- AGE: Years- Positive relationship
- ABDOMEN: Abdomen circumference (cm) – Positive relationship

Other factors –as analyzed– were involved; however, these three factors mentioned here are significantly linear related to body fat percent, while keeping the model simple enough. Based on the constructed model, the following conclusion statements were made.

1. Density has a strong negative correlation with body fat percent. If density increases by 1unit, body fat percent will be dropped by $3.79x10^{2}$.
2. Body fat percent is more in older people. If age raises 1 unit, the point estimation of body fat percent will also increase by $9.51x10^{-03}$.
3. If Abdomen circumstance change 1 unit, the point estimation of body fat percent also change $4.79x10^{-02}$ units.

Discussion

One of the limitations of this project is the fact that the data set is collected from over 252 men, which indicates that the accuracy of the model when used for women may not be ideal. In order to have a stronger model, a bigger data set is needed, which includes other important factors such as gender. Due to having influential outliers, we need to treat them in a way that does not involve removing them. One suggestion is applying a non-parametric regression model for this data set and change the values of the outliers in a way that makes them more realistic.

Body fat percent in our dataset is calculated using Brozek's equation. However another way to calculate it is using Siri's equation which requires more time and other estimations. Since this model is based on linearity, many of factors that might be significant are not considered here. This is another good reason to apply a non-parametric regression model.

References

Johnson, R.W. (1996), "Fitting percentage of body fat to simple body measurements," Journal of Statistics Education [Online], 4(1). ww2.amstat.org/publications/jse/v4n1/datasets.johnson.html

Seber, G.A.F. and Lee, A.J. (2003), Linear Regression Analysis, 2nd Edition, New York: John Wiley & Son

Behnke, A., and Wilmore, J. (1974), Evaluation and Regulation of Body Build and Composition, Englewood Cliffs, N.J.: Prentice Hall.

Brown, P.J. (1993), Measurement, Regression, and Calibration, Oxford, U.K.: Oxford University Press.