Prediction of Body Fat Percentage using Multiple Linear Regression

Anh Phuong Dinh BS21DON037

Acknowledge

I have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. I would like to extend my sincere thanks to all of them.

I am highly indebted to Professor Suchismita Das for her guidance and constant supervision as well as for providing necessary information regarding the project & also for her support and encouragement in shaping my overall understanding of statistical data analysis. My thanks and appreciations also go to my colleagues who have willingly helped me out with their abilities. Finally, we would also like to thank SP Jain School of Global Management for giving us the opportunity to do this wonderful project.

1. Introduction

1.1. Introduction of Body Fat data.

This is a large dataset that includes estimations of body fat percentage calculated by underwater weighing and various body circumference measures for 252 males. Dr. A. Garth Fisher generously provided this data and granted permission to freely share and utilize the data for non-commercial purposes. These data are used to produce the predictive equations for lean body weight given in the abstract "Generalized body composition prediction equation for men using simple measurement techniques",

(The predictive equation was obtained from the first 176 of the 252 cases that are listed below).

1.2. Theories about BMI and Body Fat Percentage

According to some experts, BMI (body mass index) is the most precise and easiest technique to measure the impact of weight on health. In reality, the majority of modern medical research uses BMI to predict someone's health state and disease risk. There is some disagreement regarding where on the BMI scale the thresholds for 'underweight,' 'overweight,' and 'obesity' should be put. However, the following criteria are employed. (http://healthiack.com/body-fat-percentage-calculator)

- 18.5 < BMI < 25 : optimal weight,
- 25 < BMI < 30 : overweight,
- BMI > 30: obese.

Meanwhile, in September 2000, the American Journal of Clinical Nutrition published a study showing that body-fat percentage may be a better measure of your risk of weight-related diseases than BMI.(http://www.webmd.com/diet/body-fat-measurement). The percentage of body fat for an individual can be estimated by Siri's formula(1956) once body density has been determined. The American Council on Exercise provides the following ranges for men's body-fat percentage:

- Essential fat : 2-5%
- Athletes: 6-13%
- Fitness: 14-17%
- Normal: 18-24%

- Obese: more than 24%

Accurate measurement of body fat is inconvenient/costly and it is desirable to have easy methods of estimating body fat that is convenient/costly.

This project aims to build a prediction model to predict the body fat percentage of a person using multiple linear regression by including various factors from 252 men.

1.3. Working with dataset

The dataset has been split into 2 subparts. One for training the model and the other for testing the model. The training sub dataset consists of 176 men and the testing dataset has data from 77 men. It is divided into a 70/30 ratio.

The model's accuracy is measured by implementing the model on the training dataset and then comparing the predicted body fat percentage with the testing dataset's actual percentage.

Volume, or body density, can be accurately measured in a variety of ways. The technique of underwater weighing "computes body volume as the difference between body weight measured in air and weight measured during water submersion. In other words, body volume is equal to the loss of weight in water with the appropriate temperature correction for the water's density"

Tools like Excel and Python are used for analysis and prediction.

2. Descriptive Statistics:

2.1. Data description

Training Data consists of 176 rows and 15 columns (attributes).

Below is a short description of each feature (attributes) of the data:

1. Density: Individual's body density as determined by underwater weighing (float)
2. BodyFat: The individual's determined body fat percentage (float). This is what we want to predict
3. Age: Age of the individual (Numeric)
4. Weight: Weight of the individual in pounds (Numeric)
5. Height: Height of the individual in inches (Numeric)
6. Neck: Circumference of the individual's neck in cm (Numeric)
7. Chest: Circumference of the individual's chest in cm (Numeric)
8. Abdomen: Circumference of the individual's abdomen in cm (Numeric)
9. Hip: Circumference of the individual's hips in cm (Numeric)
10. Thigh: Circumference of the individual's thigh in cm (Numeric)
11. Knee: Circumference of the individual's knee in cm (Numeric)
12. Ankle: Circumference of the individual's ankle in cm (Numeric)
13. Biceps: Circumference of the individual's biceps in cm (Numeric)
14. Forearm: Circumference of the individual's forearm in cm (Numeric)
15. Wrist: Circumference of the individual's wrist in cm (Numeric)

Dependent variable (Y): *BodyFat*
Independent variable (X_i): Density, Age, Weight, Height, Neck, Chest, Abdomen, Hip, Thigh, Knee, Ankle, Biceps, Forearm, Wrist.

First 5 rows of the dataset :

	Density	BodyFat	Age	Weight	Height	Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle	Biceps	Forearm	Wrist
0	1.0708	12.3	23	154.25	67.75	36.2	93.1	85.2	94.5	59.0	37.3	21.9	32.0	27.4	17.1
1	1.0853	6.1	22	173.25	72.25	38.5	93.6	83.0	98.7	58.7	37.3	23.4	30.5	28.9	18.2
2	1.0414	25.3	22	154.00	66.25	34.0	95.8	87.9	99.2	59.6	38.9	24.0	28.8	25.2	16.6
3	1.0751	10.4	26	184.75	72.25	37.4	101.8	86.4	101.2	60.1	37.3	22.8	32.4	29.4	18.2
4	1.0340	28.7	24	184.25	71.25	34.4	97.3	100.0	101.9	63.2	42.2	24.0	32.2	27.7	17.7

Table 2.1.1

2.2 Descriptive Statistics for Dependent and Independent variables

Table 2.2.1 depicts the descriptive stats like count (number of observations), mean, standard deviation, minimum value, maximum value, and the quartiles:

	Density	Age	Weight	Height	Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle	Biceps	Forearm	Wrist	BodyFat
count	252	252	252	252	252	252	252	252	252	252	252	252	252	252	252
mean	1.055574	44.88492	178.9244	70.14881	37.99206	100.8242	92.55595	99.90476	59.40595	38.59048	23.10238	32.27341	28.66389	18.22976	19.15079
std	0.019031	12.60204	29.38916	3.662856	2.430913	8.430476	10.78308	7.164058	5.249952	2.411805	1.694893	3.021274	2.020691	0.933585	8.36874
min	0.995	22	118.5	29.5	31.1	79.3	69.4	85	47.2	33	19.1	24.8	21	15.8	0
25%	1.0414	35.75	159	68.25	36.4	94.35	84.575	95.5	56	36.975	22	30.2	27.3	17.6	12.475
50%	1.0549	43	176.5	70	38	99.65	90.95	99.3	59	38.5	22.8	32.05	28.7	18.3	19.2
75%	1.0704	54	197	72.25	39.425	105.375	99.325	103.525	62.35	39.925	24	34.325	30	18.8	25.3
max	1.1089	81	363.15	77.75	51.2	136.2	148.1	147.7	87.3	49.1	33.9	45	34.9	21.4	47.5

Table 2.2.1

Mean = $\mu = (\Sigma X_i) / N$

Standard deviation = $\sigma = \text{sqrt} [\Sigma (X_i - \mu)^2 / N]$

Where,

N is the count of each variable.

X_i is are the variables.

Some key observations can be made:

1. Density has a low standard deviation indicating that almost men get the same density.

2. All variables have mean and median roughly equal, so the data set doesn't have many outliers and is a balanced data

2.3 Graphs and fitted lines

Scatter plots between independent and dependent variables:

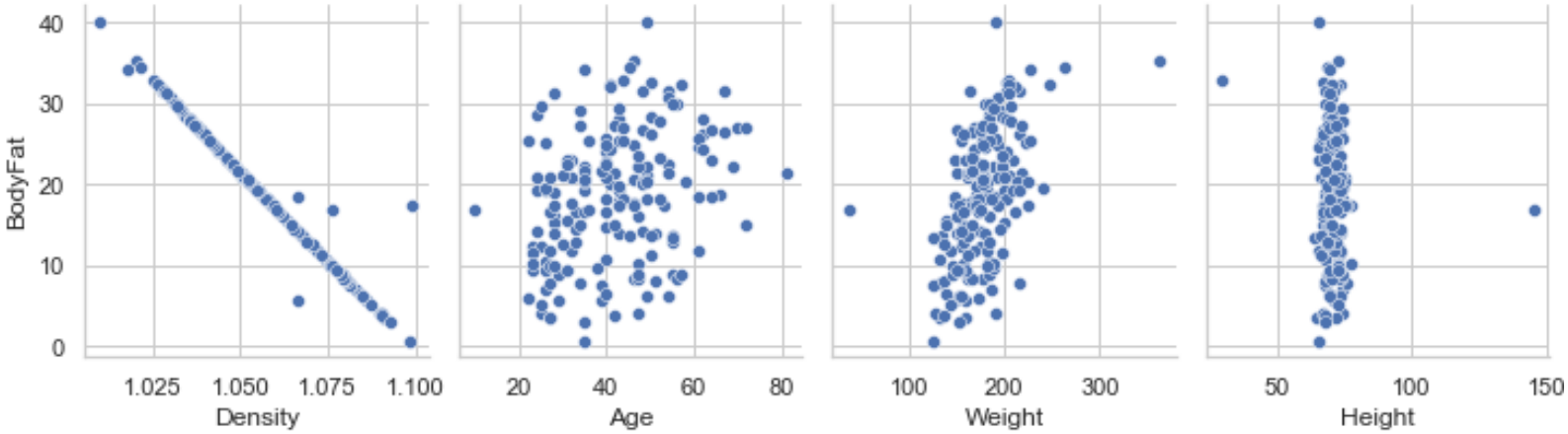

Figure 2.3.1

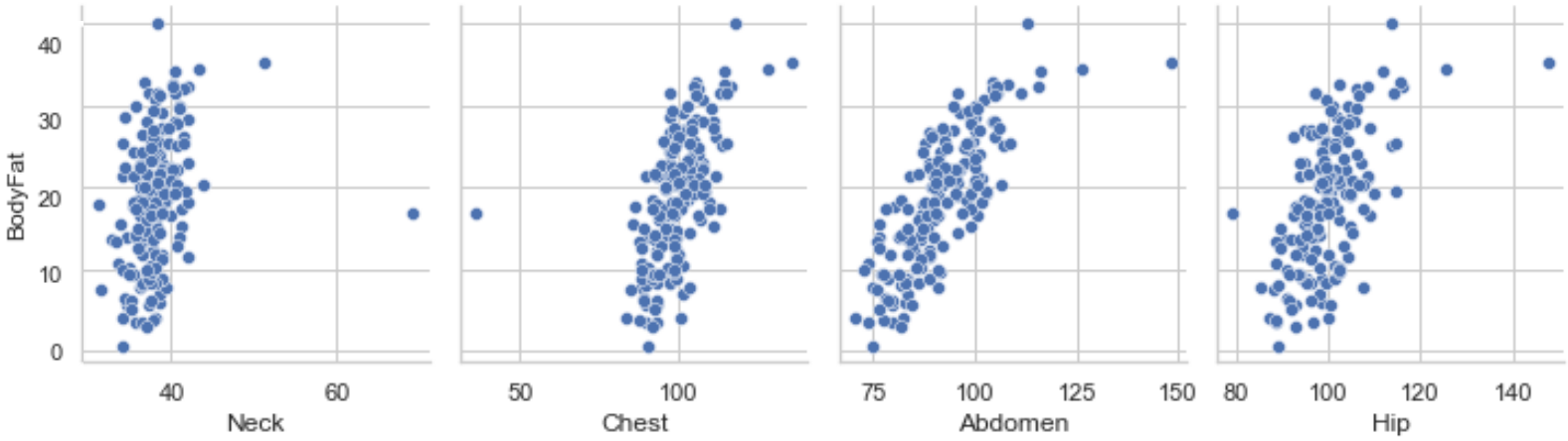

Figure 2.3.2

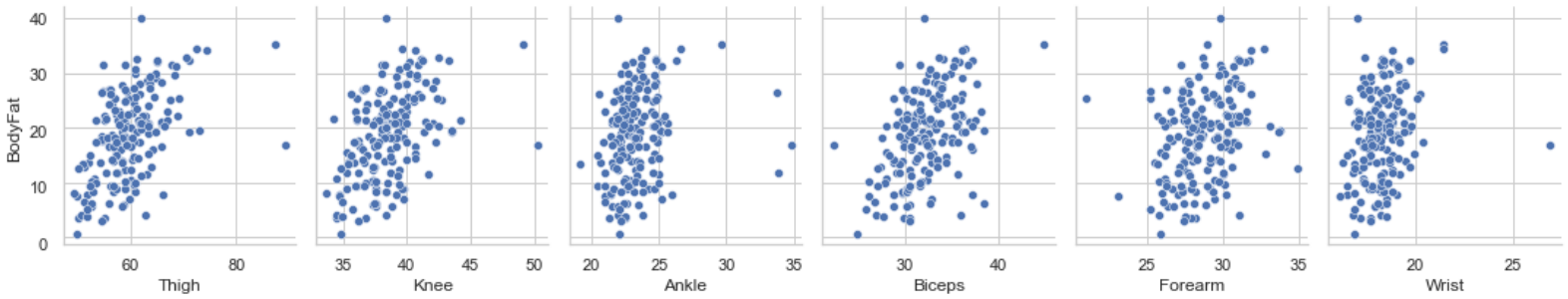

Figure 2.3.3

The models appear to have a strong possibility of a linear relationship (Figure 2.3.1):

- *BodyFat* with *Density, Abdomen*

The model appears to have a moderate possibility of a linear relationship (Figure 2.3.2):

- *BodyFat* with *Weight, Chest, Hip, Thigh, Knee, Biceps*

The models appear to have a weak(or almost no) possibility of a linear relationship(Figure 2.3.3):

- *BodyFat* with *Age, Height, Neck, Ankle, Forearm, Wrist*

2.4 Correlation chart

Correlation coefficient (r) matrix of 14 numeric variables:

$$r = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}} \qquad (2.4.1)$$

Where, $\overline{X}$ = mean of X variable

$\overline{Y}$ = mean of Y variable

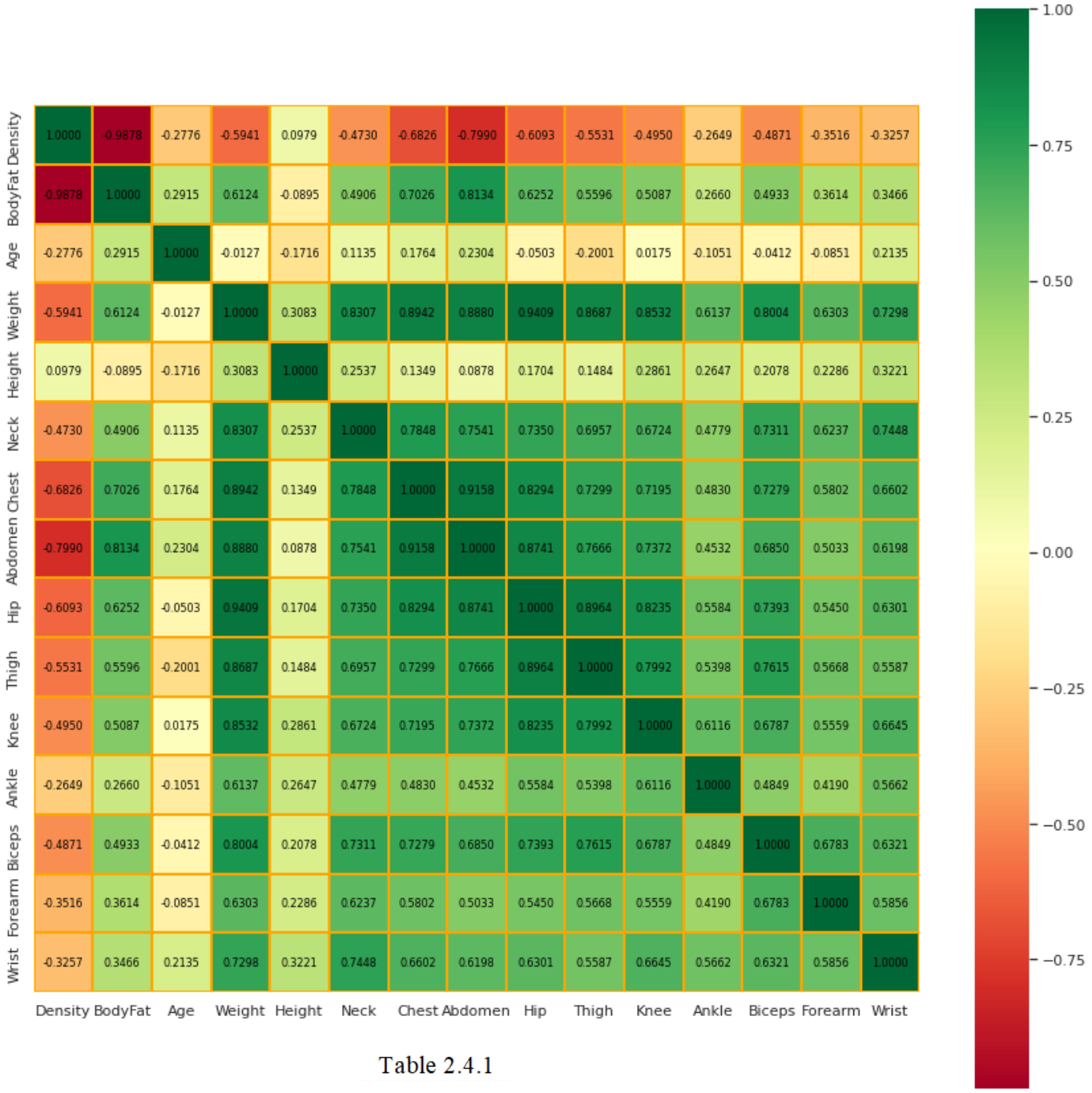

Table 2.4.1

Correlation coefficient r provides a measure of a linear relationship between X and Y.

- Strongly linearly correlated variables ($|r| > 0.7$):

  BodyFat with Density, Chest, and Abdomen

- Moderately linearly correlated variables($0.4<|r|<0.7$)

  BodyFat with Weight, Neck, Hip, Thigh, Knee, and Biceps

- Weakly linearly correlated variables ($|r|<0.4$):

  Bodyfat with Age, Height, Ankle, Forearm, and Wrist

3. Multiple Linear Regression Prediction Model

3.1 Model

Multiple linear regression prediction model for Percentage(Y) and X_i ():

$$\hat{Y}=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\beta_5X_5+\beta_6X_6+\beta_7X_7+\beta_8X_8+\beta_9X_9+\beta_{10}X_{10}+\beta_{11}X_{11}+\beta_{12}X_{12}+\beta_{13}X_{13}+\beta_{14}X_{14} \quad (3.1.1)$$

Where,

$\hat{Y}$ is the predicted value of dependent variable Y.

X_i is the actual value of the independent/explanatory variable:

$X_{1:}$ *Density*

X_2: *Age*

X_3: *Weight*

X_4: *Height*

X_5: *Neck*

$X_{6:}$ Chest

X_7: *Abdomen*

X_8: *Hip*

X_9: *Thigh*

X_{10}: *Knee*

X_{11}: *Ankle*

X_{12}: *Biceps*

X_{13}: *Forearm*

X_{14}: *Wrist*

β_i is the regression coefficient of respective $X_{i.}$

This section is divided into 3 parts dedicated to

Analyze the regression statistic table

Hypothesis test to check model utility

Regression coefficient table and final model

3.2 Regression statistics table

The following table is the regression statistics table. R^2(coefficient of determination) is the most important factor in it.

Regression Statistics	
Multiple R	0.98486747
R Square	0.969963933
Adjusted R Square	0.967352101
Standard Error	1.449155205
Observations	176

Table 3.2.1

Explanation of the terms in the table –

1. Multiple R - square root of R^2

2. R square – Coefficient of determination given by the formula:

$$R^2 = 1 - \frac{\text{SSResid}}{\text{SSTo}}$$

Where,

$$\text{SSResid} = \Sigma(Y_i - \hat{Y})^2$$

$$\text{SST}_o = \Sigma(Y_i - \bar{Y})^2$$

R-squared is a statistical measure of how close the data are to the fitted regression line. An R^2 value of 0.97 means that our model predicts with an accuracy of 97 percent or 97% of variation of Y_i around $\hat{Y}$ is explained by the regressors

3. Adjusted R square - Adjusted R-squared adjusts the statistic based on the number of independent variables in the model.

$$\text{Adjusted } R^2 = R^2 - \frac{(1-R2)*(k-1)}{(n-k)}$$

4. Standard error – Standard deviation of the error/residual

5. Observations – Total number of observations

The table gives the overall goodness of fit measures.

3.3 Hypothesis Testing

To check model utility.

Hypothesis -

H_0 : $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \ldots = \beta_{14} = 0$

There is no linear relationship between the dependent and independent variables.

H_A : $\beta_j \neq 0$ where j = 1,2,3,..,14

There is at least one independent variable that has a linear relationship with dependent variable.

ANOVA table

Formulas -

Source	Sum of squares	Degree of Freedom	Mean squares	F
Treatment	SS_T	k-1	$MS_T = \frac{SS_T}{k-1}$	$F = \frac{MS_T}{MS_E}$
Error	SS_E	N-k	$MS_E = \frac{SS_E}{N-k}$	
Total	TotalSS	N-1		

Here k is the number of coefficients (15) and N is the total number of observations (176).

Table 3.3.1

	Degrees of freedom	*Sum of Square*	*Mean Square*	*F*	*Significance F(p-value)*
Regression	14	10918.63131	779.9022363	371.3730323	1.1471E-114
Residual	161	338.10818	2.100050808		
Total	175	11256.73949			

Table 3.3.2

With significance $\alpha = 0.05$ (from Table 3.3.1)

F – critical (df1 = 14, df2 = 161) = 1.748; F – statistic = 371.373

Since F – statistic > F – critical,

We reject the null hypothesis. Hence, there is at least one independent variable that has a linear relationship with the dependent variable

3.4 Regression coefficient table and final model

The following table gives the value, standard error (SE), t statistic, p-value, and confidence interval of regression coefficients –

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	421.94	14.29321463	29.52065	7.3547E-67	393.7186242	450.17134	393.718624	450.171336
Density	-397.82	11.92559493	-33.3589	3.2754E-74	-421.3750181	-374.2735	-421.375018	-374.273495
Age	0.02	0.012909708	1.818765	0.0708055	-0.002014468	0.0489739	-0.00201447	0.04897392
Weight	-0.02	0.016430622	-1.27863	0.2028675	-0.053456008	0.0114386	-0.05345601	0.01143864
Height	0.04	0.030839393	1.15779	0.2486643	-0.025196332	0.0966074	-0.02519633	0.09660743
Neck	0.02	0.079440782	0.226073	0.82143152	-0.138920896	0.1748397	-0.1389209	0.1748397
Chest	0.06	0.039730033	1.509575	0.13311182	-0.018483709	0.1384347	-0.01848371	0.13843467
Abdomen	0.04	0.041622539	0.973661	0.33168548	-0.041670283	0.1227228	-0.04167028	0.12272277
Hip	0.01	0.060360454	0.243275	0.8081019	-0.104516111	0.1338845	-0.10451611	0.13388451
Thigh	0.06	0.058571284	0.93945	0.34890706	-0.060642255	0.1706918	-0.06064225	0.17069184
Knee	0.09	0.102259514	0.916028	0.36102228	-0.108270348	0.2956155	-0.10827035	0.29561548
Ankle	-0.06	0.080360636	-0.71808	0.47374811	-0.216402273	0.1009914	-0.21640227	0.10099139
Biceps	-0.11	0.067909559	-1.58794	0.11426094	-0.241944923	0.0262718	-0.24194492	0.02627177
Forearm	0.08	0.081723266	0.939283	0.34899252	-0.084626496	0.238149	-0.0846265	0.23814904
Wrist	0.07	0.221900569	0.297558	0.76642435	-0.372182812	0.5042393	-0.37218281	0.50423927

Table 3.4.1

- Column "**Coefficient**" gives the least-squares estimates of β_j.
- Column "**Standard error**" gives the standard errors (i.e.the estimated standard deviation) of the least-squares estimates b_j of β_j.
- Column "**t Stat**" gives the computed t-statistic for H0: $\beta_j = 0$ against Ha: $\beta_j \neq 0$. This is the coefficient divided by the standard error. It is compared to at with (n-k) degrees of freedom where here n = 176 and k = 15.
- Column "**P-value**" gives the p-value for test of H0: $\beta_j = 0$ against Ha: $\beta_j \neq 0$. This equals the P{|t| > t-Stat}where t is an at-distributed random variable with n-k degrees of freedom and t-Stat is the computed value of the t-statistic.
  Note that this p-value is for a two-sided test. For a one-sided test divide this p-value by 2 (also checking the sign of the t-Stat).
- Columns "Lower 95%" and "Upper 95%" values define a 95% confidence interval for β_j.

A simple summary of the above output is that the fitted line is (By substituting coefficients in equation 3.1.1):

$$\hat{Y} = 421.94 - 397.82\,X_1 + 0.02X_2 - 0.02\,X_3 + 0.04\,X_4 + 0.02\,X_5 + 0.06X_6 + 0.04\,X_7 + 0.01\,X_8 + 0.06\,X_9 + 0.09\,X_{10} - 0.06\,X_{11} - 0.11X_{12} + 0.08\,X_{13} + 0.07\,X_{14}$$

(3.4.1)

Although the coefficient value of density has a much negatively value, it is follow by the larger value of intercept.

Coefficients of *Density, Weight, Ankle,* and *Biceps* are negative indicate that the smaller these values, the fatter a person can be.

4. Testing

The prediction model (equation 3.4.1) is tested on the data of case numbers from 177 to 252

First 10 rows of prediction data:

Case Number	Std. Residual	BodyFat	Predicted Value	Residual
177	0.095	13.1	12.979	0.1206
178	-0.136	29.9	30.073	-0.1735
179	0.259	22.5	22.169	0.3305
180	-0.482	16.9	17.514	-0.6137
181	0.052	26.6	26.533	0.0665
182	3.159	0.0	-4.025	4.0252
183	-0.126	11.5	11.661	-0.1609
184	-0.073	12.1	12.193	-0.0927
185	-0.193	17.5	17.746	-0.2462
186	-0.065	8.6	8.683	-0.0831
187	-0.480	23.6	24.212	-0.6116

Table 4.1

Residual Plot:

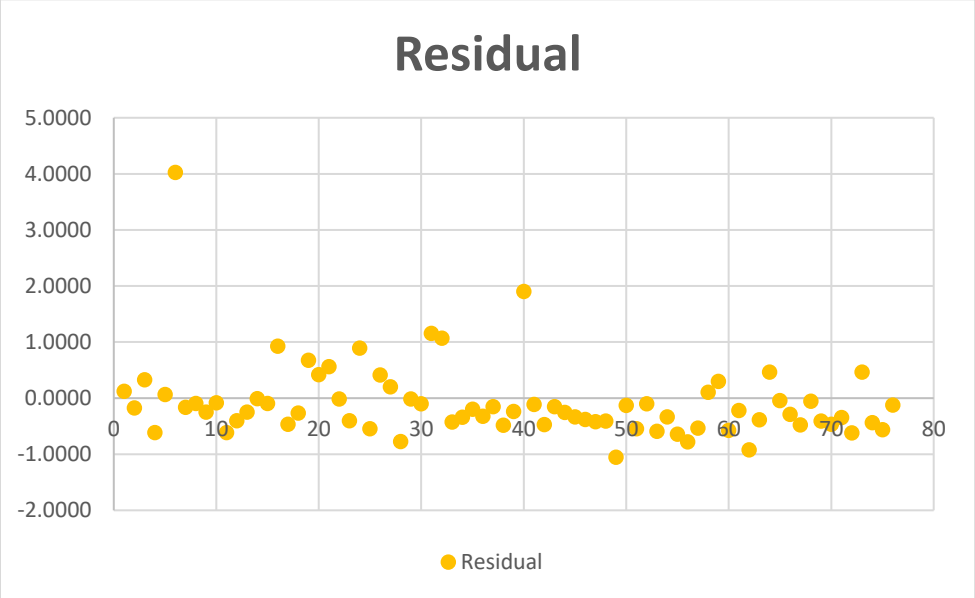

Table 4.2

The plot shows a random pattern, indicating a good fit for a linear model.

5. Conclusion

Body fat prediction accuracy is critical for assessing obesity and related disorders. However, researchers are finding it difficult to analyze the massive amounts of medical data generated. The primary goal of this study is to analyze and assess the predictive power of a model using multiple linear regression.

From the graphs and correlation coefficient, we can see that Density and Abdomen have strongly correlated with the Bodyfat percentage and they could be used as factors for the future prediction models.

We get an R square approximately one indicates that the model predicts with high accuracy

As you can see, a prediction has been generated for each row in the test table. Additionally, our predicted bodyfat percentages align closely with the original values which indicate that the model is good model for predicting body fat percentage.

6. References

6.1 Link to Datasets

https://www.kaggle.com/datasets/fedesoriano/body-fat-prediction-dataset?resource=download

6.2 Data sources

The data were generously supplied by Dr. A. Garth Fisher who gave permission to freely distribute the data and use for non-commercial purposes.

Roger W. Johnson
Department of Mathematics & Computer Science
South Dakota School of Mines & Technology
501 East St. Joseph Street
Rapid City, SD 57701

email address: rwjohnso@silver.sdsmt.edu
web address: http://silver.sdsmt.edu/~rwjohnso

6.3 Textbooks

Introduction to statistics and Data Analysis by Roxy Peck, Chris Olsen and Jay L. Devore

6.4 Appendix

Python scripts and Excel file (submitted with project).