Exploration of the Relationship between Multiple Anthropometric Data and Obesity and Prediction of Obesity using a Linear Regression Model

Xinyi Han[1, †], Ziheng Zhang[2, *, †], Zhengkai Zhuang[3, †]

[1]Department of Mathematics, University of California, Santa Barbara, Santa Barbara, CA 93106, USA;

[2]Department of Applied Mathematics, Xi'an Jiaotong-Liverpool University, 111 Ren'ai Road, Suzhou Industrial Park, Suzhou, China;

[3]Department of Biotechnology, Huazhong University of Science & Technology, 1037 Luoyu Road, Hongshan District, Wuhan, China

* Corresponding Author Email: Ziheng.Zhang19@student.xjtlu.edu.cn

†These authors contributed equally.

Abstract. Obesity prevalence has risen sharply in the United States from 1987 to the present, and there is merely a sign that obesity rates are declining. Obesity is associated with numerous medical and psychological abnormalities and the economic expenditures of individuals. What factors affect obesity has been widely discussed, while still has no conclusion. This article provides a study using the linear regression model to analyze the relationship between some anthropometric data and obesity and predict obesity. A linear regression model including 252 observations and 15 variables is constructed for discovering the relationship and predicting obesity. Some improvements are also pointed out to increase the accuracy of the model. A primary linear regression model is built by R-studio. Moreover, LASSO is used to remove outliers to improve the model. The accuracy is verified to prove the practicality of the model. The study can facilitate the study of forecasting obesity by using anthropometric measurement data and provide a reference for other researchers who are interested in this topic. Moreover, the regression model enables the public to become familiar with their obesity status simply by measuring the circumference of their body parts, contributing to a lower risk of obesity-related diseases.

Keywords: obesity; linear Regression; R-studio; LASSO.

1. Introduction

The growing problem of obesity brings serious health and social difficulties to society around the world. According to the World Health Organization statistics, by 2015, approximately more than 700 million were obese [1]. Obesity is a condition of the accumulation of excess body fat to the extent that it negatively impacts health. Obese people can be defined as having a BMI over 30 kg/m^2; the range 25–30 kg/m^2 is defined as overweight, where BMI refers to the Body mass index derived from the mass (weight) and height of a person [1, 2]. BMI is defined as the body mass divided by the square of the body height, resulting from mass in kilograms and height in meters.

Obesity has been placed under the spotlight driven by many pertinent factors. It is one of the leading risk factors for early death because it may give rise to diabetes, heart disease, stroke, and various types of cancer. Obese youth are more likely to have cardiovascular risk factors than people with normal weight resulting in cardiac structural and hemodynamic alterations [3], which further leads to diseases like hypertension [4], increases in ventricular mass [5], endothelial dysfunction, early coronary, aortic fatty streaks, fibrous plaque [6], and atherosclerosis [5]. In addition, obesity may also cause other severe diseases like type 2 diabetes, gastroesophageal reflux disease, obesity-associated sleep apnea [7], and non-alcoholic fatty liver disease with resultant cirrhosis. Furthermore, according to the latest research, obesity can also trigger some psychological and psychiatric morbidity [1].

In addition to causing disease, obesity also has a negative impact on the economy. People who are obese develop comorbidities that can lead to health care expenditures. According to statistics, there was a $74 billion expenditure related to obesity in the United States in 1998, and this figure nearly doubled to $147 billion in 2008 [8]. In 2011, medical costs which were relevant to obesity in the United States arrived at $209.7 billion, approximately 20% of annual fiscal expenditures related to health care [9]. Although the growth of population is the primary trigger for the increased costs during this time, the rise in obesity rates was also not negligible. The prevalence rate of obesity witnessed a noticeable increase of 27% in the general population from 1987 to 2001. In 1987, obese people spent on average 15% more on health care per capita than healthy-weight individuals. By 2001, that number had more than doubled, far outpacing the growth rate of the total per capita population over the same period. Therefore, without an effective countermeasure, the prevalence of obesity will be bound to give rise to the increase in obesity-related costs. Moreover, higher incidence and prevalence of chronic diseases will simultaneously flourish in society, further increasing the costs associated with obesity. Therefore, it is necessary to find a reasonable way to predict obesity. Once we find out what factors are associated with obesity, the problem can be prevented to some extent. Thus, our goal is to predict obesity using a predictive modeling technique and test its accuracy. This paper is a study of the use of linear regression serving as a method to predict obesity.

Regression analysis is one of the most common statistical procedures used for predicting and forecasting. The purpose of linear regression is to predict the trend of data by fitting a straight line to the data or to predict the value of a dependent variable from the values of other independent variables. Linear regression represents the connection link between the independent variable-carrier and dependent variable-response. If we plot on the X and Y coordinates, it should be a straight line. Linear regression shows a straight line that thoroughly represents or predicts the value of the response variable, given the labeled value of the carrier variable [10]. Therefore, we applied linear regression to our dataset to determine if there was any relationship between obesity and other variables.

Due to the practical meaning of data, the discussion of causality is still a problem for obesity prediction. By observing our data, it is easy to find some existence of causality among our variables. Technically, the causality should not have an impact on the linear regression model. However, considering the fact, that the variables should have causal relationships to make the research meaningful. Another limitation arises from the outliers. It was not determined whether outliers in our data would affect the predictive model to a great extent. Most studies show that the outliers should conclude into the model since the extreme values are also parts of the data. And eventually, we choose to pay much attention to outliers in our predictive model.

2. Methods

2.1. Data Collection

The data we used throughout this paper were produced by Dr. A. Garth Fisher generously. He gave permission to distribute the data freely, but it should be used only for non-commercial purposes. The percentage of obesity was listed, and it used the method of underwater weighing to judge. Meanwhile, other various body circumference measurements were also included. And there were 252 men samples in this data set. Therefore, there were 252 observations and 15 variables in this data set. and the 15 variables were density determined from underwater weight, percent obesity from Siri's equation [11], age (years), weight (lbs), height (inches), neck circumference (cm), check circumference (cm), abdomen circumference (cm), hip circumference (cm), thick circumference (cm), knee circumference (cm), ankle circumference (cm), biceps (extended) circumference (cm), forearm circumference (cm), waist circumference (cm). Obesity of Siri’s equation [11] calculated by body density was:

$$Obesity = \frac{495}{Body\ density\ (gm/cm^3)} - 450$$

The body density in the data set was obtained in the following ways [12]:

$$Body\ density = \frac{WA}{(WA - WW)/CF - LV}$$

The rest values of variables about the circumference of each part of the body were measured with a meter ruler.

2.2. Research Protocol

Firstly, we divided the data into a training set and a test set with a ratio of 7:3, built the model on the training set, and validated it on the test set. Then we tried to use LASSO linear regression, aiming to choose a suitable series of variables. By comparing with models by judging their R-square, we would be able to get our best model. We also made some further improvements to our model: we attempted to remove outlier points and fitted a better model. Finally, we tested our model and got the results.

3. Results

3.1. Establish Model

3.1.1. Primary Linear Regression

Putting all x variables excluding body density into the model, we got our primarily simple linear regression.

Table 1. Statistical Analysis of Model 1

	Coefficient	Std.Error	t-value	P(>|t|)
Age	0.05	0.04	1.19	0.23
Weight	-0.13	0.06	-2.20	0.03*
Height	-0.07	0.10	-0.66	0.51
Neck Circumference	-0.33	0.28	-1.17	0.25
Chest Circumference	0.07	0.12	0.59	0.56
Abdomen Circumference	0.89	0.10	8.75	<0.01*
Hip Circumference	-0.07	0.18	-0.40	0.69
Thigh Circumference	0.17	0.16	1.02	0.31
Knee Circumference	0.27	0.29	0.91	0.36
Ankle Circumference	0.24	0.29	0.83	0.41
Biceps Circumference	0.08	0.20	0.39	0.70
Forearm Circumference	0.53	0.23	2.29	0.02*
Wrist Circumference	-1.70	0.63	-2.69	<0.01
Inspection standards	Multiple R-squared		0.7479	
	Adjusted R-squared		0.7285	
	p-value		<0.01	

Note:'*' means significant.

3.1.2. LASSO Linear Regression

LASSO regression is a compressed estimation. A penalty function is used, aiming to build a more accurate model. So that it compresses some regression coefficients, forcing the sum of the absolute values of the coefficients to be less than a value we set to restrict. Meanwhile, some regression coefficients would be set to zero. Therefore, the advantage of subset shrinkage is preserved, which is a biased estimation for dealing with complex collinear data.

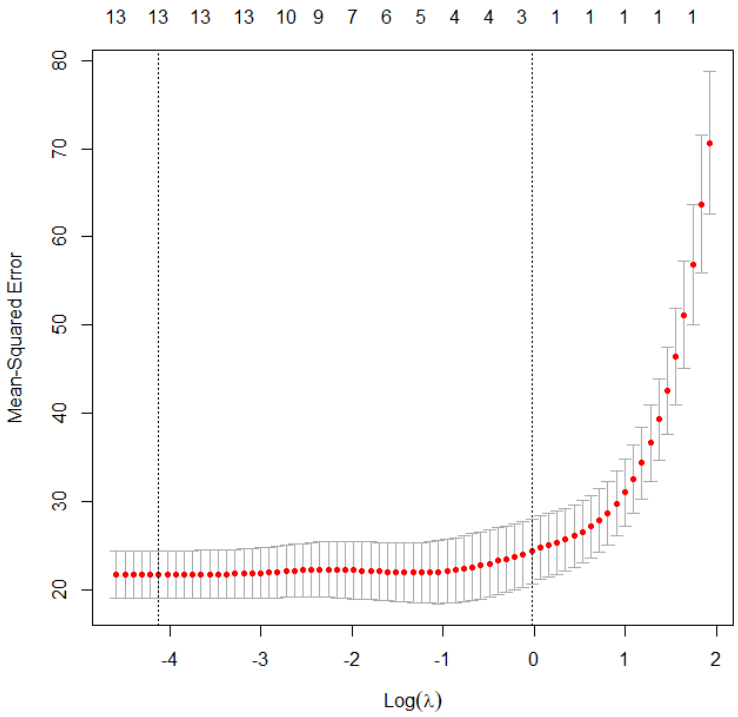

Fig.1 Relationship between MSE and Lambda

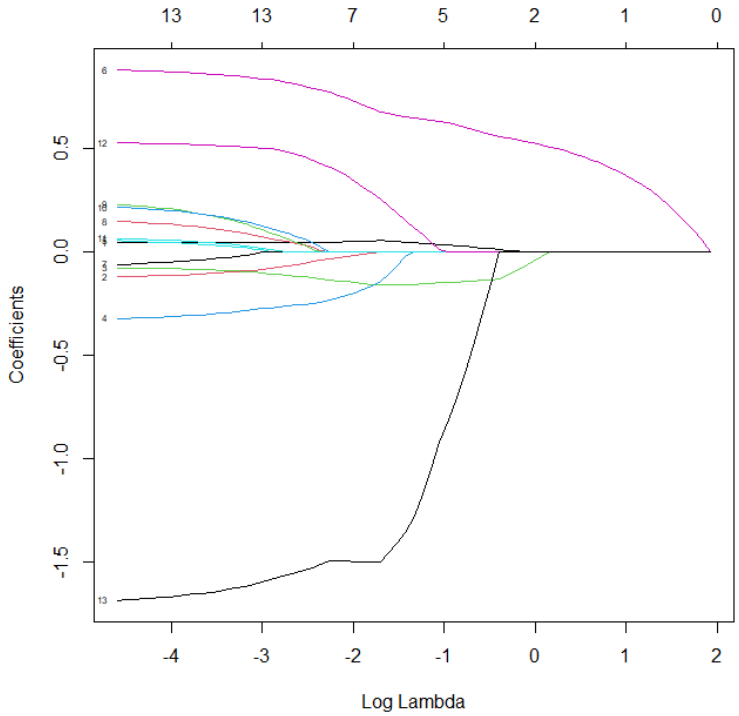

Fig. 2 Relationship between Coefficient and Lambda

Two lambdas were selected according to different criteria, and the two models were further constructed respectively. We used cross-validation to determine the optimal value of lambda. The two vertical lines were the for minimal MSE and 1SE rule. The MSE gave the min λ according to the figure. And the 1SE rule gave the most regularized model such that the error was within one standard error of the minimum. So we got two new models.

Table 2. Coefficients under two different Lambda selection criteria

	Lambda.min	*Lambda.1se*
(Intercept)	-32.643	-26.3
Age	0.047	-
Weight	-0.115	-
Height	-0.798	-0.047
Neck Circumference	-0.315	-
Chest Circumference	0.048	-
Abdomen Circumference	0.872	0.527
Hip Circumference	-0.052	-
Thigh Circumference	0.137	-
Knee Circumference	0.213	-
Ankle Circumference	0.204	-
Biceps Circumference	0.058	-
Forearm Circumference	0.523	-
Wrist Circumference	-1.67	-

3.1.3. Model Comparison

Table 3. R-square in different models

	MLR	Lasso.min	Lasso.1se
R-square	0.7479	0.7483	0.724
Adjusted R-square	0.7285	0.73	0.7162

We compared the R-square of these three models and got the best model. According to the table, it could be seen obviously that LASSO regression with MSE is the best model.

3.1.4. Remove Outliers

There were still some improvements that could be made to our model. The studentized residual was a tool to measure the outlier degree of our data. Usually, scatter points with studentized residuals over +2 or less than -2 could be defined as outliers. A horizontal axis over 0.2 or 0.3 had high leverage values. In addition, circle size was proportional to influence; points with large circles might have a strong influence with a disproportionate impact on the estimation of model parameters. These points have to be removed. Taking these factors into consideration, we finally removed points 25, 28, 61, and 162 and fitted a new model.

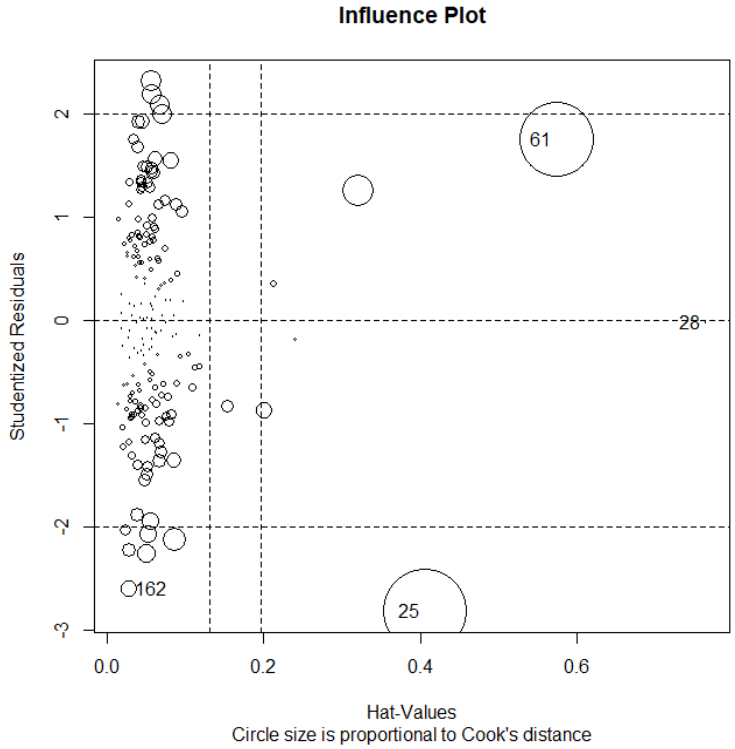

Fig. 3 Influence Plot of points

We did a summary of our new model, and it could be seen that the R-square was improved. Hence our final prediction model was as follows:

$$BodyFat = -3.60 + 0.05Age - 0.03Weight - 0.26Height + 0.78Abdomen - 0.01Hip + 0.17Thigh - 0.20Ankle + 0.05Biceps + 0.35Forearm - 1.95Wrist$$

Table 4. Statistical Analysis of Model 2

	Coefficient	Std.Error	t-value	P(>|t|)
Age	0.05	0.04	1.48	0.14
Weight	-0.03	0.06	-0.40	0.69
Height	-0.26	0.20	-1.34	0.18
Neck Circumference	-0.26	0.29	-0.91	0.36
Abdomen Circumference	0.78	0.10	7.84	<0.01
Hip Circumference	-0.01	0.17	-0.05	0.96
Thigh Circumference	0.17	0.16	1.08	0.28
Ankle Circumference	-0.20	0.42	-0.46	0.65
Biceps Circumference	0.05	0.19	0.27	0.78
Forearm Circumference	0.35	0.24	1.46	0.15
Wrist Circumference	-1.95	0.66	-2.97	<0.01
Inspection standards	Multiple R-squared		0.758	
	Adjusted R-squared		0.7418	
	p-value		<0.01	

3.2. Prediction of Obesity

Figure 4 showed a comparison result of prediction and observation of obesity. Through regression analysis of 13 independent variables, we finally screened out 11 independent variables to predict obesity, and we gave a specific multiple linear regression formula. Meanwhile, we found that abdomen circumference and wrist circumference have a significant impact on obesity, which is that for every 1 cm increase in the abdomen circumference, obesity will increase by 0.78, and for every 1

cm decrease in the wrist circumference, obesity will increase by 1.95, so people need to control the circumference of these two parts as much as possible to reduce the possibility of obesity. In addition, R-square was 0.742, which meant the model fitted well.

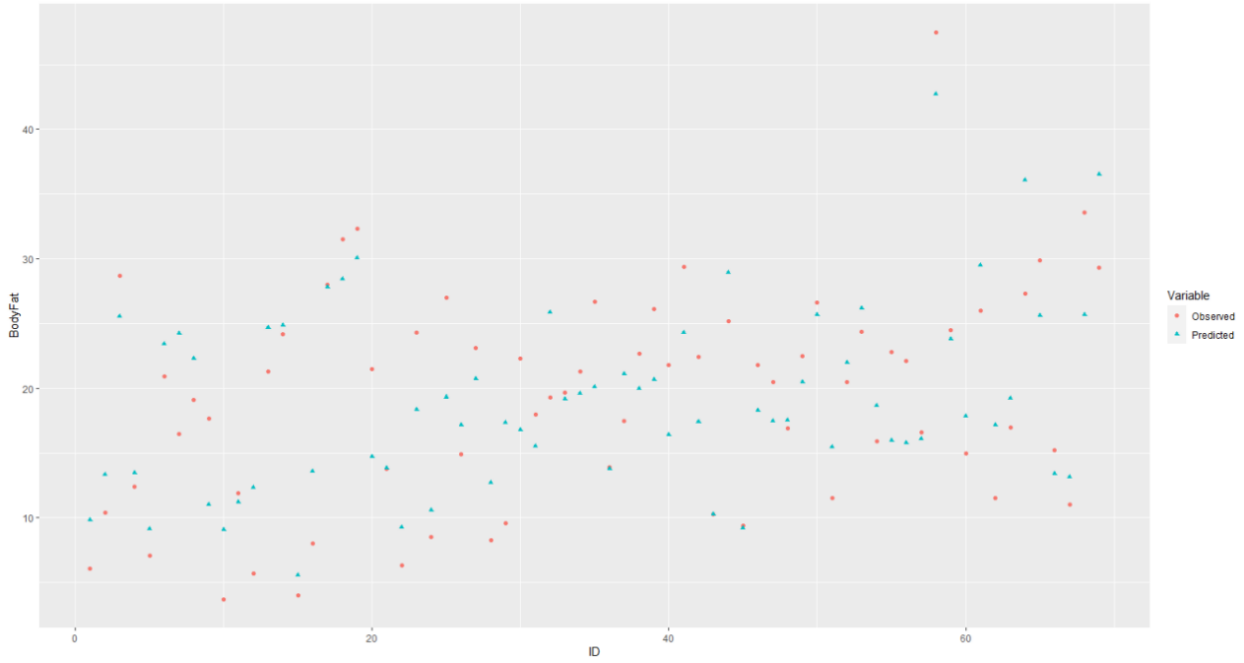

Fig. 4 Prediction and Observation of obesity

4. Discussion

In this paper, we showed that male obesity could be predicted by measuring the circumference of various parts of the body, such as abdominal girth and wrist girth. We determined to use a regression model when we designed this study to enable the final conclusion to be applied to daily life. However, it met some challenges, and the biggest one was a poor regression model arising from the low fitting degree, low prediction rate, and improper independent variable selection. To solve it, we made some changes to improve our models. Firstly, we built a multiple linear regression model to check each predictor's p-value and used LASSO to help us select predictors. Secondly, we checked the residual plot to see whether all the points are evenly distributed around the horizontal line: y equals 0, and whether we should change the form of the variable, such as quadratic, square root, and reciprocal. Thirdly, we used the *outlierTest()* function to check the p-value to ascertain whether there was an outlier and explore why this point was highly different from other points to decide whether to delete it. Eventually, we used R language to calculate R-square and mean square error of the model, concluding that our model was relatively accurate.

Although Siri's equation [11] only needs to measure body density to figure out obesity, the regression model we used is still more straightforward and more efficient in practice. Body density used in Siri's equation [11] is measured underwater since body volume is computed by measuring the weight in air and during water submersion [12]. Therefore, measuring body density requires complicated processes and has site-specific requirements, which means that it is impossible for people to monitor their obesity situation in their everyday life through this method. Nevertheless, the regression model we use only measures some simple data of one's body, such as the height, weight, and various girths, which are easily accessible to most people, removing several cumbersome steps. Therefore, the public can keep track of their obesity situation and improve a specific part of the body to reduce a series of risks arising from obesity. Simultaneously, our model does not neglect the accuracy of the predictions. When compared with the data source, we identify that most of the data are within acceptable margins of error. Additionally, compared with predicting obesity only by measuring neck circumference [13], waist circumference and BMI [14], or age and skin-fold

measurement [15], our regression model use multiple anthropometric data to avoid serious inaccurate prediction results arising from specific abnormal measurement data.

On the other hand, there are also some limitations to our study. First, the variable body density is difficult to properly handle when building the regression model. Due to Siri's equation [14], a strong inverse relationship between the dependent variable obesity and the predictor body density can be obtained, which means that if we do not count body density as an independent variable, the model's predictive rate will not be so accurate; however, if we count it, the other 13 independent variables will become meaningless or make a little contribution in the regression model. In addition, the lack of broad and representative data collection is another limitation. The data source only collects 252 male data and it does not mention the criteria for selection, which may be random or within a specific range, so there is no guarantee that the results we get can represent men.

Therefore, in the future study, we will consider studying the relationship between the variable body density and various body circumference measurements or illustrate the principle behind Siri's equation [11] to make the model more accurate. Simultaneously, we will optimize the method of selecting data collection objects and collect data from women, so as to further reveal the factor differences between male and female obesity.

5. Conclusion

This study has argued that it is necessary to find a reasonable and accurate model to explain the relationship between obesity and data on various body parts. We identified a multiple linear regression model to interpret the influence of independent variables except for body density in our data set on the dependent variable obesity. Predictions through our model approximately agree with the value calculated by Siri's equation [11]. Simultaneously, by comparing the coefficients of each independent variable, we discovered that abdominal girth and wrist girth had a significant impact on obesity. The ability to predict obesity accurately with only some readily available body data enables our regression model to be used in people's everyday lives rather than theoretically possible. Additionally, according to the different coefficients of the circumference of each body part in our model, individuals can also control the increase or decrease of these girths. In the future, the influence of the girth of specific body parts on body density can be further studied, and the regression model obtained in this paper can also provide a reference for other researchers concerned with predicting obesity. By implementing the multiple linear regression model, as well as measuring some simple and easily accessible body data, a relatively accurate obesity situation can be predicted, and then people can be alerted to their obesity status in an intuitive digital way, thereby reducing their risk of obesity-related diseases and improving the overall health of the population.

References

[1] WHO. (2021) WHO - Obesity and overweight. Fact sheet N°311. http://www.who.int/mediacentre/factsheets/fs311/en/.

[2] Mayo Clinic. (2021) Obesity - Symptoms and causes. https://www.mayoclinic.org/diseases-conditions/obesity/symptoms-causes/syc-20375742.

[3] Freedman, D.S., et al. (2007) Cardiovascular risk factors and excess adiposity among overweight children and adolescents. the Bogalusa Heart Study. Journal of Pediatrics 150(1): 12-17.e12.

[4] Speiser, P.W., et al. (2005) Childhood obesity. Journal of Clinical Endocrinology & Metabolism (3): 1871-1887.

[5] Daniels, S.R. (2009) Complications of obesity in children and adolescents. International Journal of Obesity 33(1): S60-S65.

[6] Freedman, D.S., et al. (2004) The relation of obesity throughout life to carotid intima-media thickness in adulthood: the Bogalusa Heart Study. International Journal of Obesity 28(1): 159-166.

[7] Muzumdar, H. and Rao, M. (2006) Pulmonary dysfunction and sleep apnea in morbid obesity. Pediatric Endocrinology Reviews Per 3 Suppl 4: 579-583.

[8] Finkelstein E.A., et al. (2009) Annual medical spending attributable to obesity: payer-and service-specific estimates. Health Affairs 28(5): 822-831.

[9] Cawley, J. and Meyerhoefer, C. (2012) The medical care costs of obesity: an instrumental variables approach. J Health Econ 31(1): 219–230.

[10] Frey, B. (2006) Statistics Hacks. Sebastopol. CA: O'Reilly Media Inc.

[11] Siri, W.E. (1956) The gross composition of the body. Adv Biol Med Phys 4(4): 239-280.

[12] Katch, F.I. and Mcardle, W.D. (1977) Nutrition, Weight Control, and Exercise. Houghton Mifflin Co., Boston.

[13] Sn, S., et al. (2019) Neck Circumference: A valid anthropometric tool to predict Obesity in Adults of Davanagere, South India Corresponding Author Citation Article Cycle. Indian Journal of Community Health 31(4): 457-463.

[14] Aparecida, S.E., et al. (2020) Accuracy of BMI and waist circumference cut-off points to predict obesity in older adults. Ciencia & saude coletiva 25(3): 1073-1082.

[15] Bailey. (1996) Smart exercise: Burning fat, getting fit, Smart exercise: burning fat, getting fit. Houghton Mifflin Co., Boston, pp. 179-187.