FNSPE CTU

ASM

DATASET N.6

Fitting Percentage of Body Fat to Simple Body Measurements

Author:
Vladislav BELOV

September 5, 2018

1 Outline

In this paper we will perform an analysis of the dataset which contains simple measurements of 252 men. Circumferences of body parts, age, weight and fat percentage are a part of the dataset. After providing descriptive statistics in section 2 and performing a more detailed examination of some selected variables in section 3, we will attempt to fit percentage of body fat to some of the body measurements in section 4. Body fat percentage can be measured using either Siri's equation or Brozek's equation:

$$\text{Siri: } \frac{457}{\text{Body Density}} - 414.2,$$

$$\text{Brozek: } \frac{495}{\text{Body Density}} - 450.$$

2 Descriptive Statistics

2.1 Numerical Analysis of Selected Variables

In this section a general numerical overview of the dataset population is be provided. Basic information about population's weight and height is available in Tab. 2.1. Age is also considered a continuous instance, nevertheless, for illustrative purposes we have categorized it, and results can be seen in Tab. 2.2.[1]

Variable	Min. Value	1st Quantile	Median	Mean Value	3rd Quantile	Max. Value
Total Weight, [kg]	53.75	72.12	80.06	81.16	89.36	164.72
Fat-Free Weight, [kg]	48.04	59.58	64.21	65.19	69.8	109.09
Height, [cm]	74.93	173.35	177.8	178.18	183.51	197.49

Table 2.1: Numerical descriptive statistics for weight and height of the population.

Age	Count	%
22-34	53	21.03
35-49	116	46.03
50-64	62	24.6
65-81	21	8.34

Table 2.2: Contingency table for age.

2.2 Graphical Analysis of Selected Variables

2.3 Fat Percentage Analysis

As body fat percentage is heavily dependent of body density, this section is opened by the histogram of it, see Fig. 2.1. Observing this diagram, we can speculate, that men with higher body weight have bodies with less density. Another noteworthy observation is that the histogram resembles normal distribution (Fig. 2.2). However, more detailed analysis of this fact is provided in section 3.

In Fig. 2.3 one can see the comparative diagram of both Siri's and Brozek's equations results. Similarities between them start to become noticeable. Looking at comparison of cumulative distribution functions (see Fig. 2.4), we can speculate, that they have similar distributions and, moreover, those distributions are normal, the fit is available in Fig. 2.5-2.6. Relevant tests are carried out in section 3.

[1] Inches were converted to centimeters, pounds to kilograms.

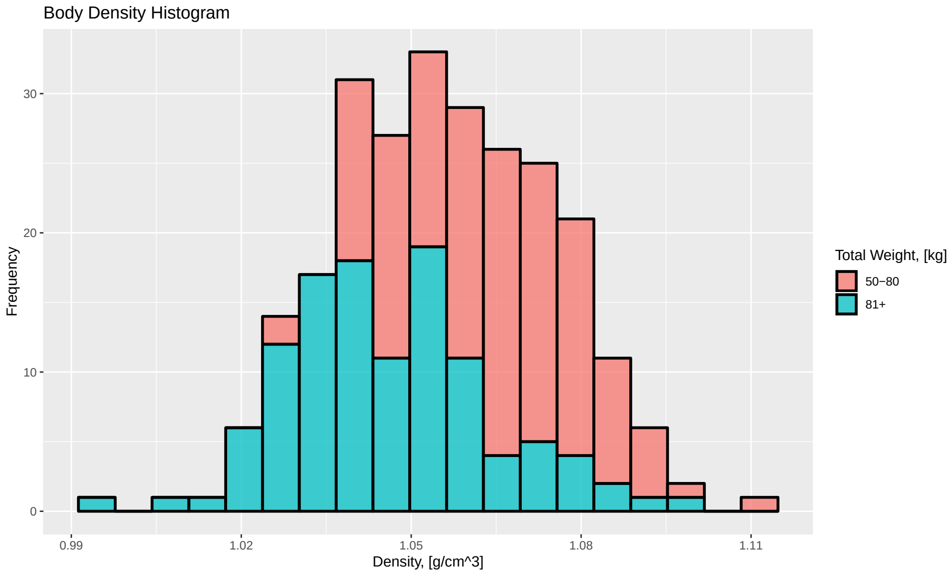


Figure 2.1: Body density histogram with categorized total weight for illustrative purposes.

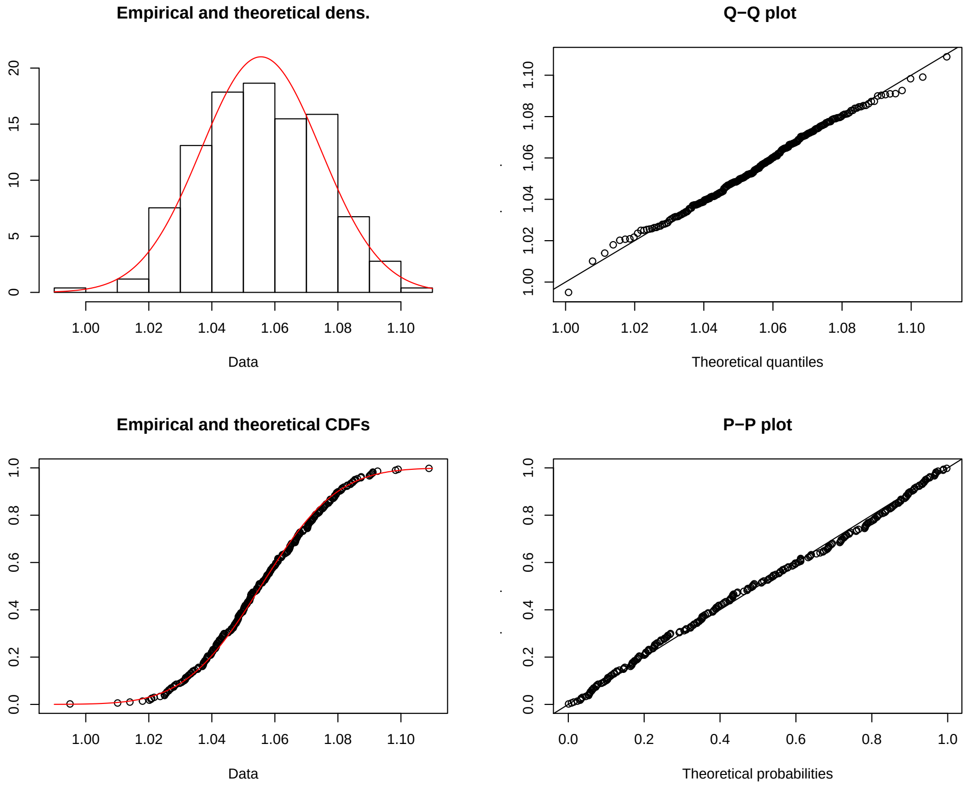


Figure 2.2: Normal distribution fit for body density.

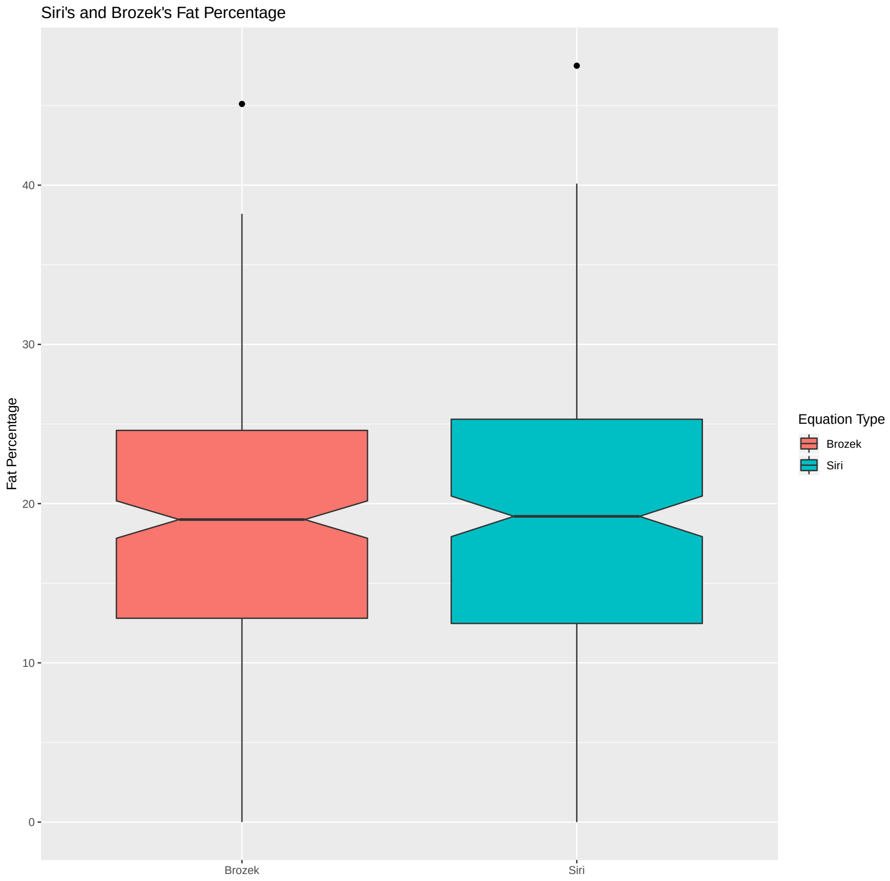


Figure 2.3: Comparative box plot of Siri’s and Brozek’s equations results.

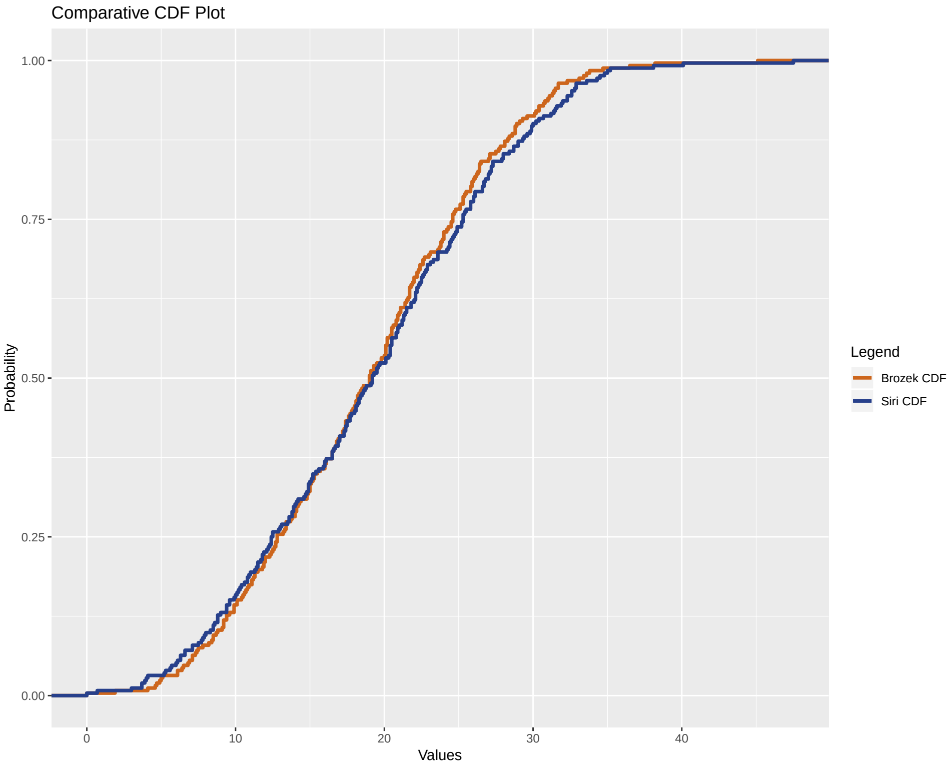


Figure 2.4: Comparative plot of cumulative distribution functions for Siri’s and Brozek’s equations results.

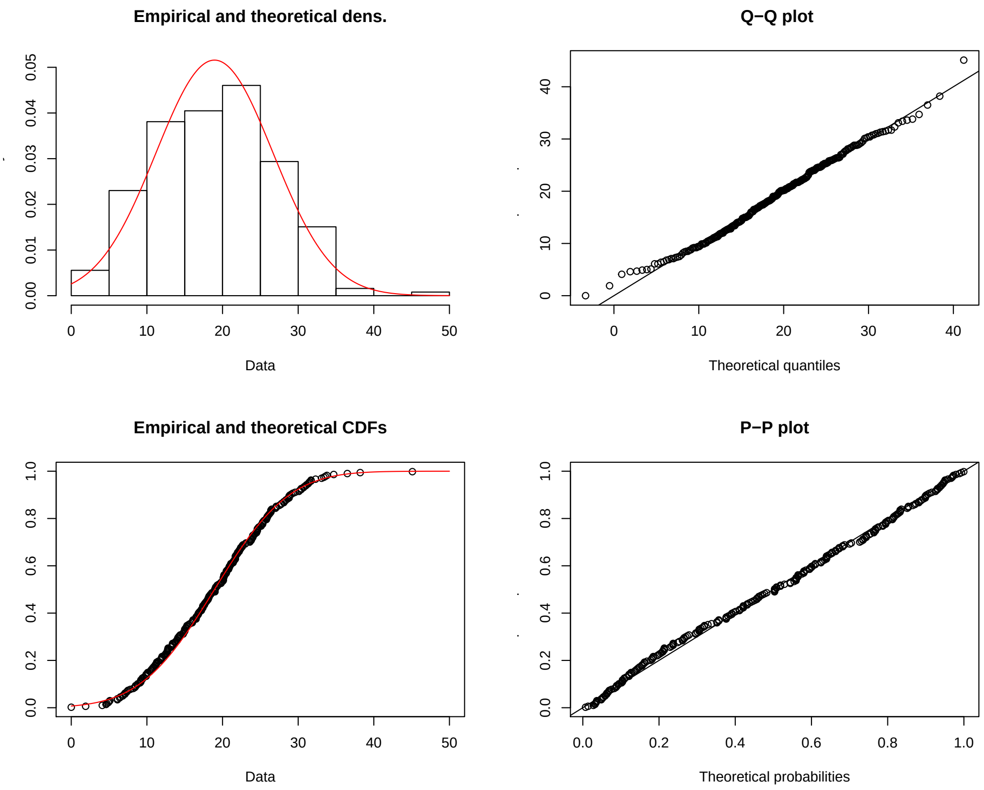


Figure 2.5: Normal distribution fit for Brozek's equation results.

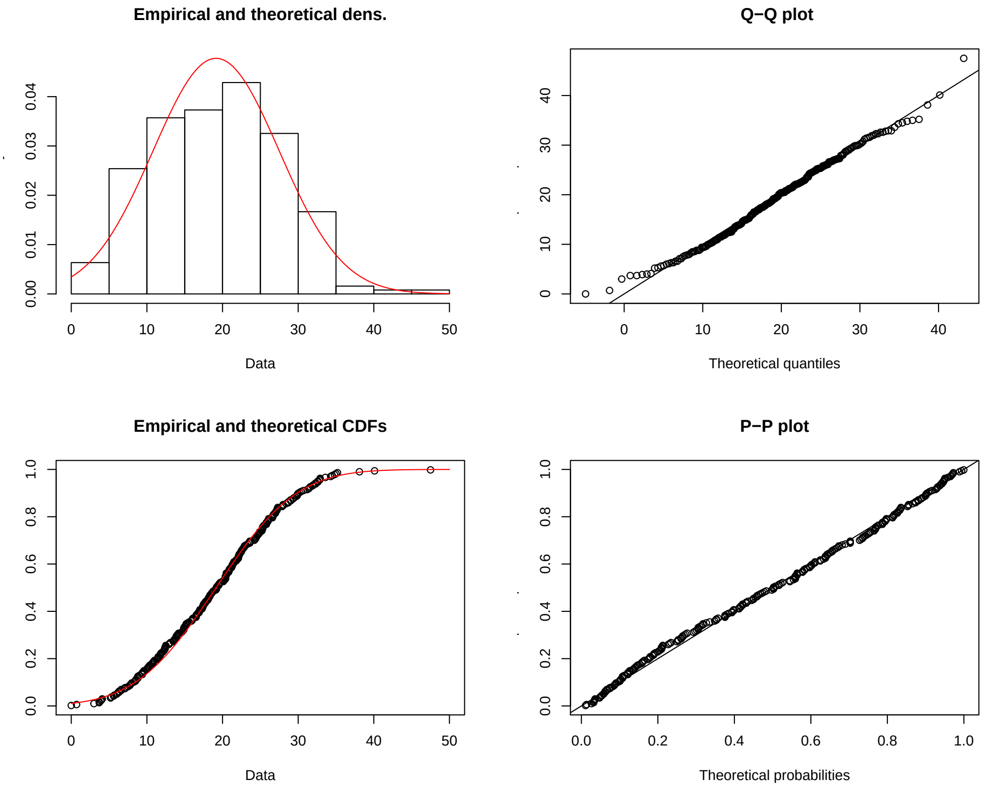


Figure 2.6: Normal distribution fit for Siri's equation results.

2.4 Total Body Weight Scatter Plots

In this section we will look at the behavior of total body weight in the scope of the provided dataset.

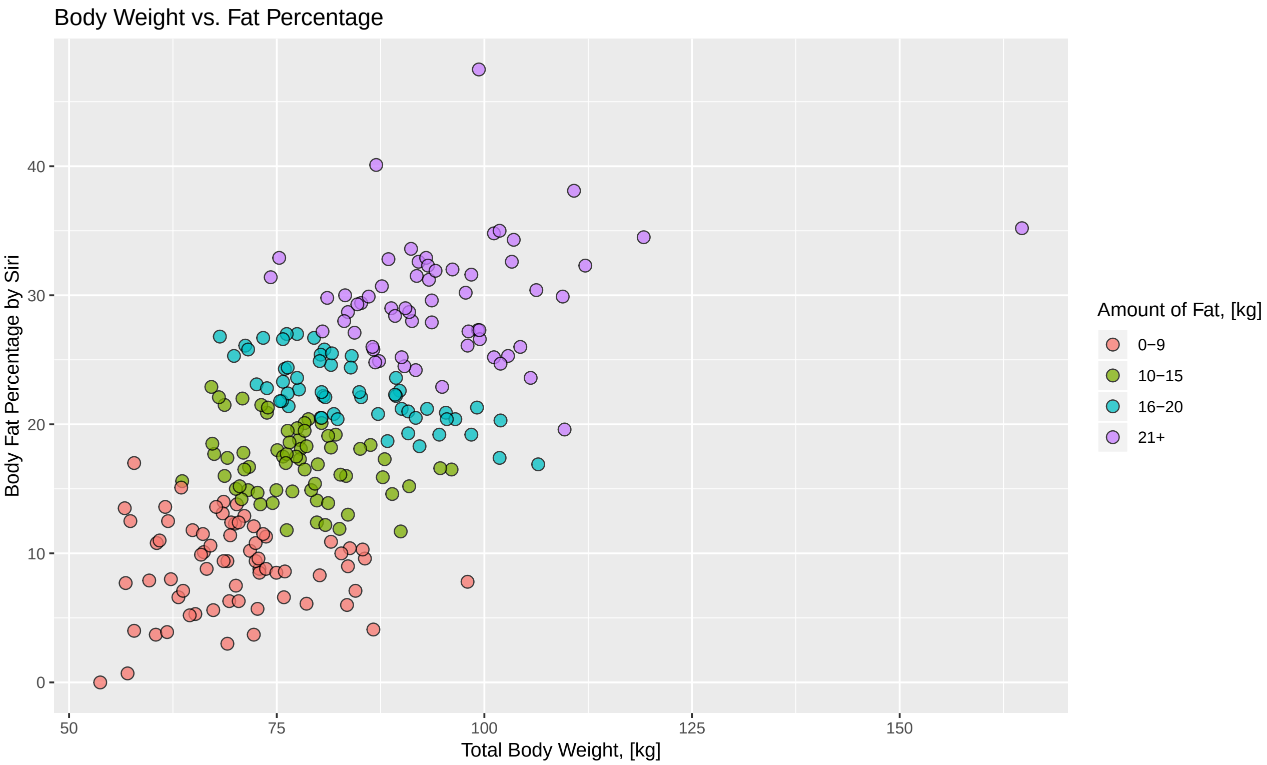


Figure 2.7: Influence of total body weight on the body fat percentage.

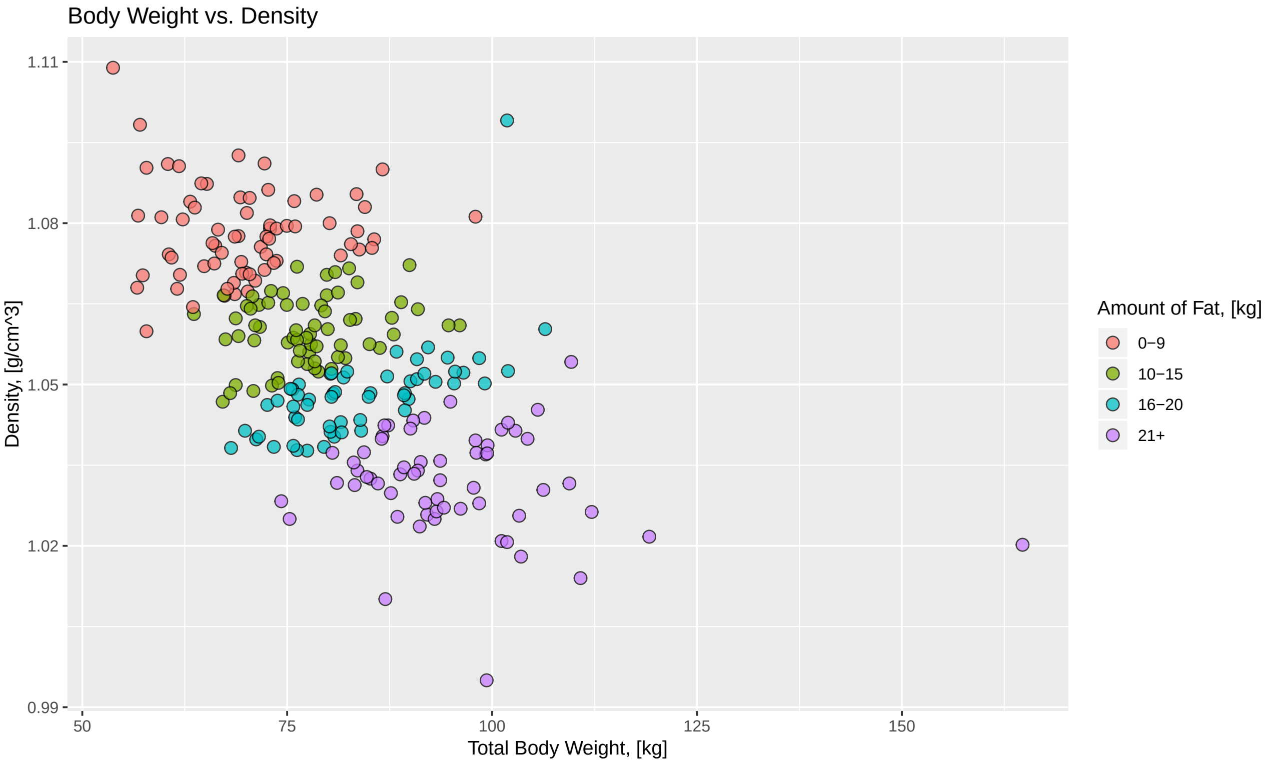


Figure 2.8: Influence of total body weight on body density.

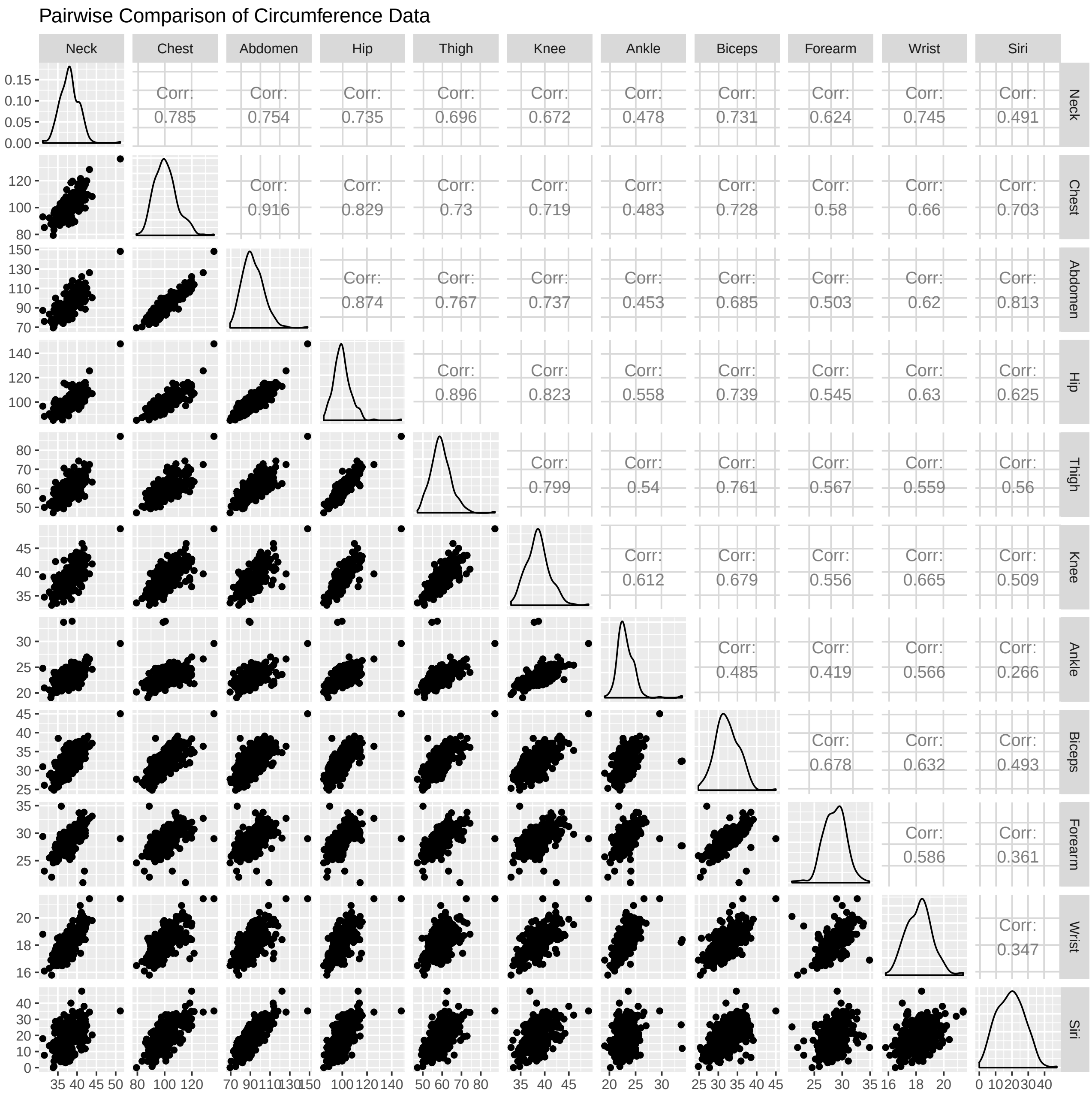
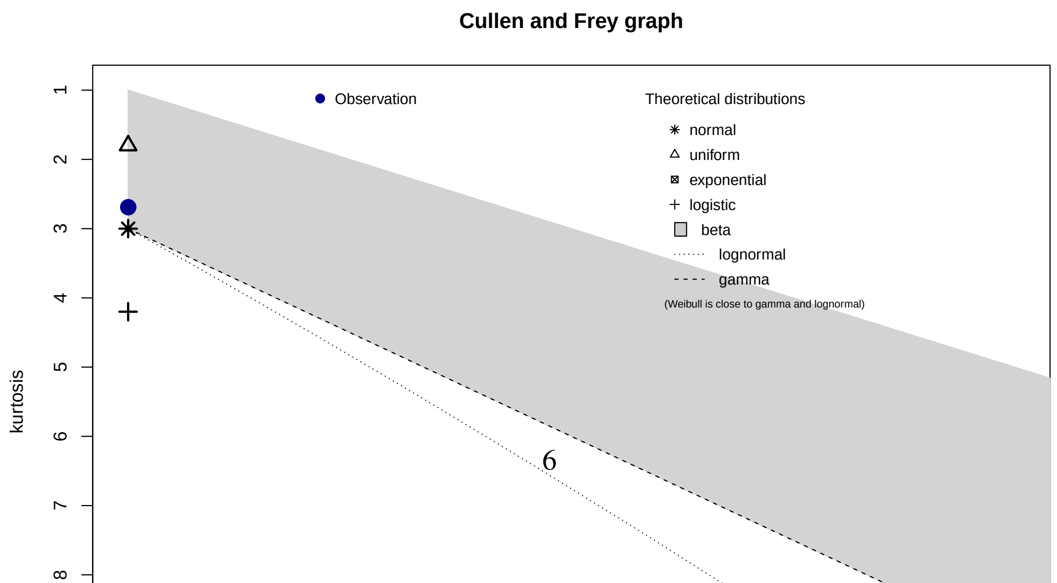


Figure 2.9

3 Analýza dat



4 Multivariate Linear Regression

In the scope of this section we will consider observation 39 with total weight equal to 164.7 kg as an outlier and remove it from the dataset.

4.1 2-variable Linear Regression

In this section we will assume, that the results of Siri's equation can be explained by two variables: abdomen circumference and total weight. Then the linear model takes the following form, $\forall i \in \{1, 2, \ldots, 251\}$:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + e_i \tag{1}$$

which, with dropped mathematical notation for the response and explanatory variables, results in

$$(\text{Body Fat Percentage by Siri}) = \beta_0 + \beta_1 \cdot (\text{Abdomen CC}) + \beta_2 \cdot (\text{Total Weight}).$$

Results of the estimation can be observed in the table below. The value of the R-squared statistic is equal to 0.72. This indicates, that 72% of the variability of the response data is explained around its mean which is quite adequate for our purposes.

Parameter	Estimated Value	Confidence Interval, 95%	
Intercept, β_0	-47.67	-52.86	-42.48
Abdomen CC, β_1	0.98	0.87	1.09
Total Weight, β_2	-0.29	-0.38	-0.2

Table 4.1: Estimated values for the 2-variable linear regression.

The fit function is displayed in Fig. 4.1. As can be seen in the Fig. 4.2, the residuals are distributed acceptably on the plane (see the first column of the figure), normal Q-Q plot also indicates normality of residuals. Assuredly, as normality of residuals is the absolute assumption to perform linear regression, a relevant test has to be performed to support this hypothesis.

$$\mathrm{H}_0 : res \in \{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\} \text{ vs. } \mathrm{H}_1 : res \notin \{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$$

According to the Lilliefors test with statistical significance set to 5% we cannot reject the null hypothesis, as the p-value is equal to 0.33.

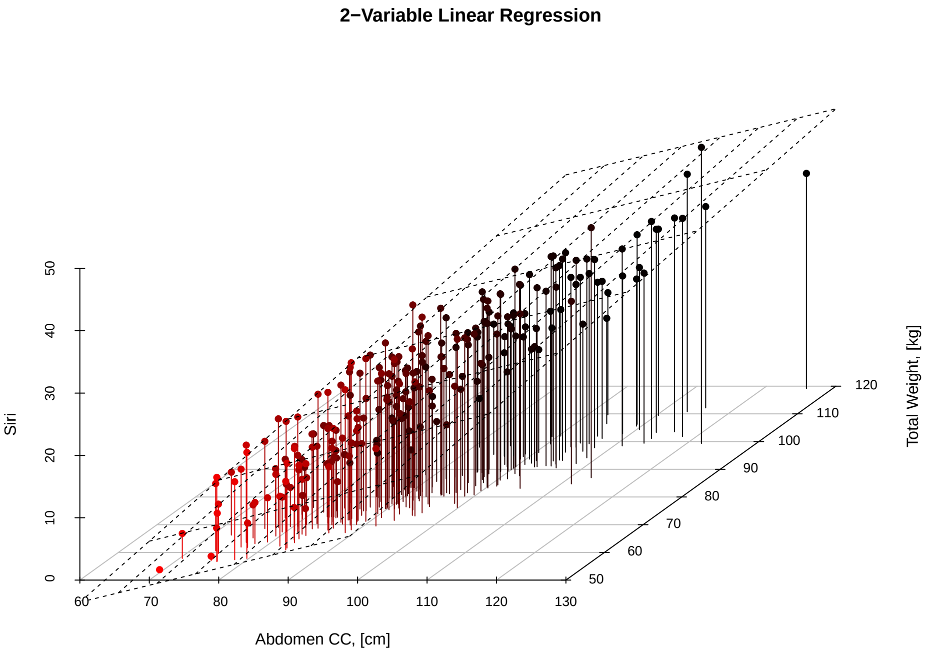


Figure 4.1: Fit of the linear model.

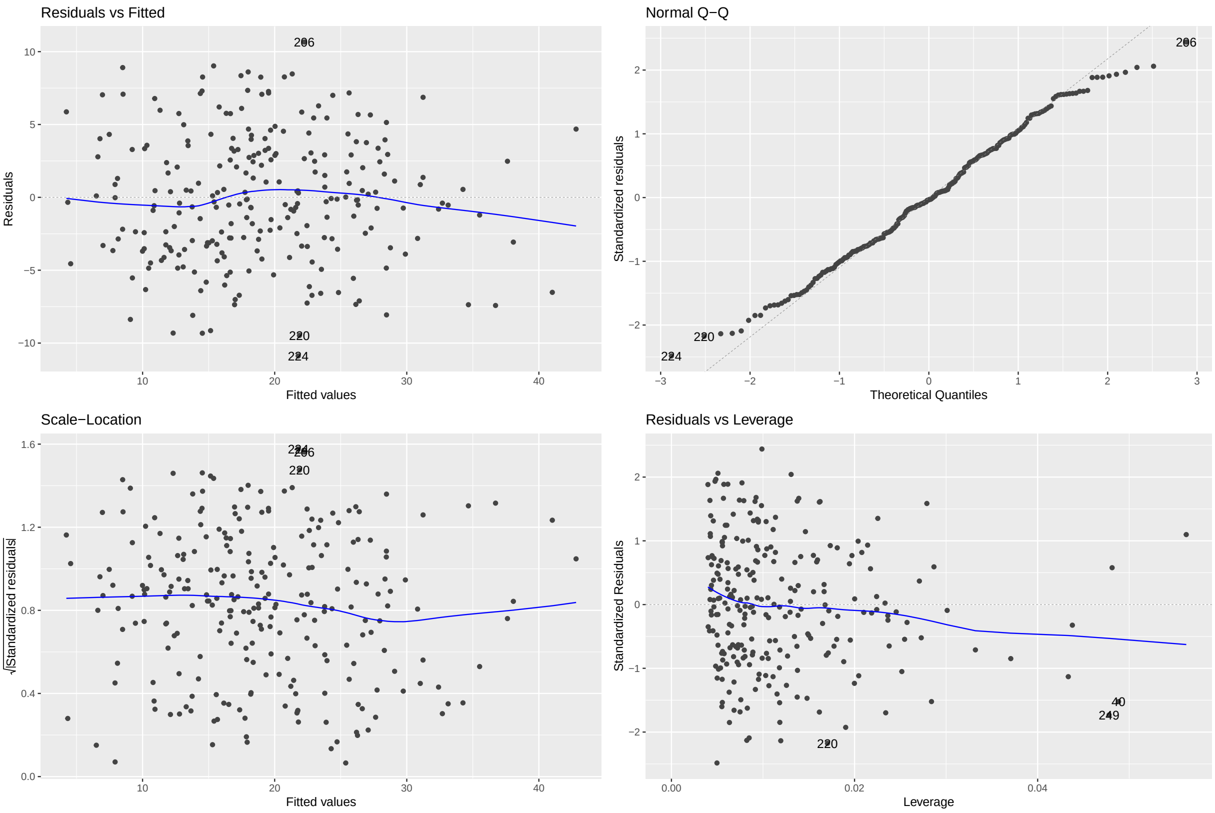


Figure 4.2: 2-variable linear regression results diagram.

4.2 Multiple Linear Regression

4.2.1 All Variables Included

Now we will consider the model with almost all available variables and once again attempt to explain the fat percentage given by Siri's equation. The model is as follows, $\forall i \in \{1,2,\dots,251\}$:[2]

$$Y_i = \sum_{j=1}^{15} \beta_j X_{i,j} + e_i \tag{2}$$

or $Y = X\beta + e$ in a matrix form. Estimated values are displayed in the Tab. 4.2. The R-squared statistic is quite high for the observed model and is equal to 0.995 which means, that almost all of the response data variability is explained. Nonetheless, a lot of confidence intervals for estimated parameters contain zero. Moreover, residuals are not normally distributed (see Fig. 4.3), as the null hypothesis in the Lilliefors test is rejected due to the p-value equal to $5.101 \cdot 10^{-7}$. Thus, linear regression cannot be used for this set of explanatory variables.

Parameter	Estimated Value	Confidence Interval, 95%	
Age, β_1	0.01	-0.01	0.04
Total Weight, β_2	0.89	0.84	0.95
Height, β_3	0.01	-0.01	0.04
Adiposity, β_4	-0.27	-0.47	-0.06
Fat-free weight, β_5	-1.23	-1.29	-1.18
Neck CC, β_6	0.05	-0.1	0.2
Chest CC, β_7	0.02	-0.04	0.09
Abdomen CC, β_8	0.1	0.03	0.18
Hip CC, β_9	-0.01	-0.09	0.07
Thigh CC, β_{10}	0.13	0.04	0.23
Knee CC, β_{11}	0.05	-0.1	0.21
Ankle CC, β_{12}	0.12	-0.02	0.26
Biceps CC, β_{13}	0.12	0.01	0.23
Forearm CC, β_{14}	0.08	-0.05	0.22
Wrist CC, β_{15}	-0.07	-0.42	0.28

Table 4.2: Estimated values for the multivariate linear regression with all parameters included.

[2] In contrast with the previous section, we also exclude the intercept from the model from by reason of common sense: with decreasing values of explanatory variables the fat percentage also decreases.

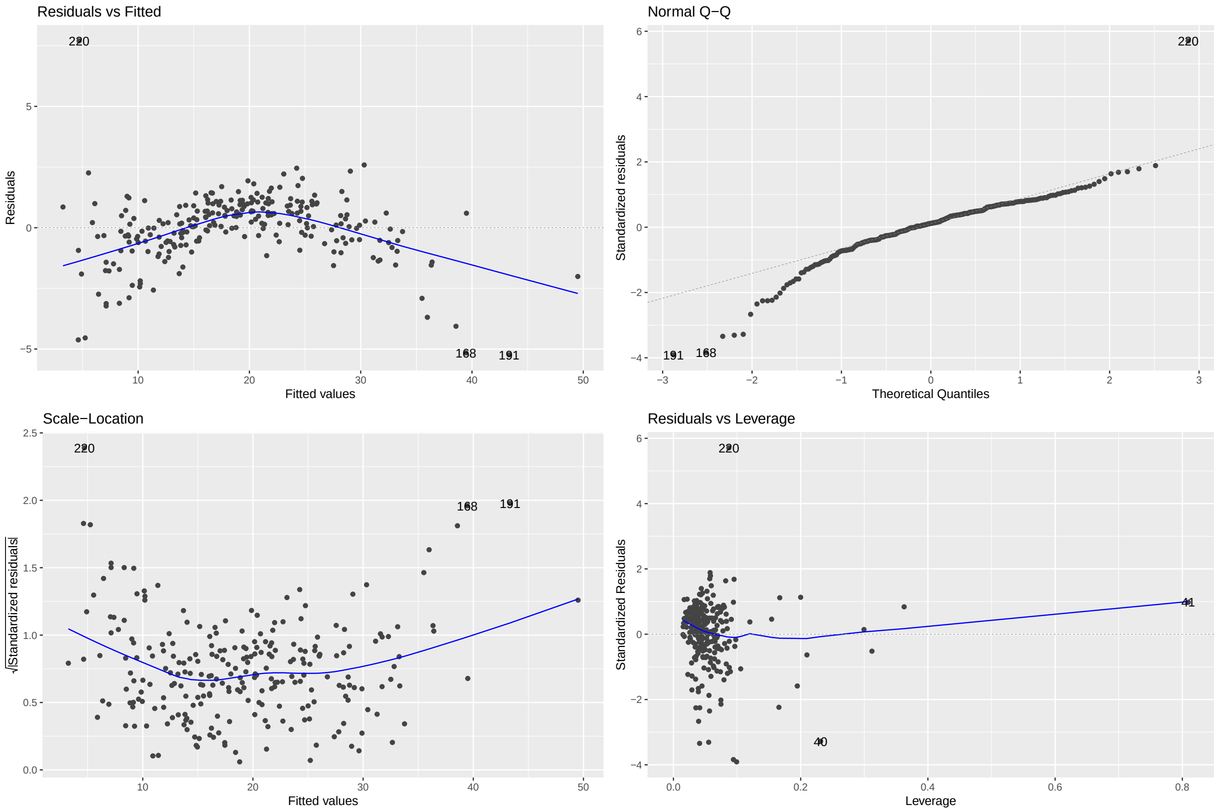


Figure 4.3: Multiple linear regression results diagram for the model with all variables included.

4.2.2 The Final Model

To make regression more sufficient, we exclude the total weight from the model due to its strong correlation (0.76) with the fat-free weight and run a stepwise algorithm to choose the the model by the Akaike Information Criterion. After excluding variables with low significance, we construct the final linear model using only 8 columns of the provided dataset: age, adiposity, fat-free weight, abdomen CC, hip CC, knee CC, biceps CC and wrist CC. Estimates of parameters are available in the Tab. 4.3. Fortunately, the R-squared statistic has remained high: $R^2 = 0.9766.$, thus, almost all of the variability of the response variable is accounted for by our model.

Parameter	Estimated Value	Confidence Interval, 95%	
Age, β_1	-0.05	-0.09	-0.01
Adiposity, β_2	0.52	0.19	0.85
Fat-free weight, β_3	-0.57	-0.65	-0.49
Abdomen CC, β_4	0.72	0.6	0.84
Hip CC, β_5	-0.19	-0.34	-0.03
Knee CC, β_6	0.34	0.02	0.66
Biceps CC, β_7	0.38	0.16	0.61
Wrist CC, β_8	-1.54	-2.13	-0.95

Table 4.3: Estimated values for the multivariate linear regression with all parameters included.

As can be seen, confidence intervals no longer contain zero values. As for residuals (see Fig. 4.4), their normality is evident. Moreover, the Lilliefors test with statistical significance set to 5% has resulted in our inability to reject the null hypothesis, as the p-value is equal to 0.4. Hence, residuals distribution may be considered normal and we are eligible to use this model to perform multivariate linear regression.

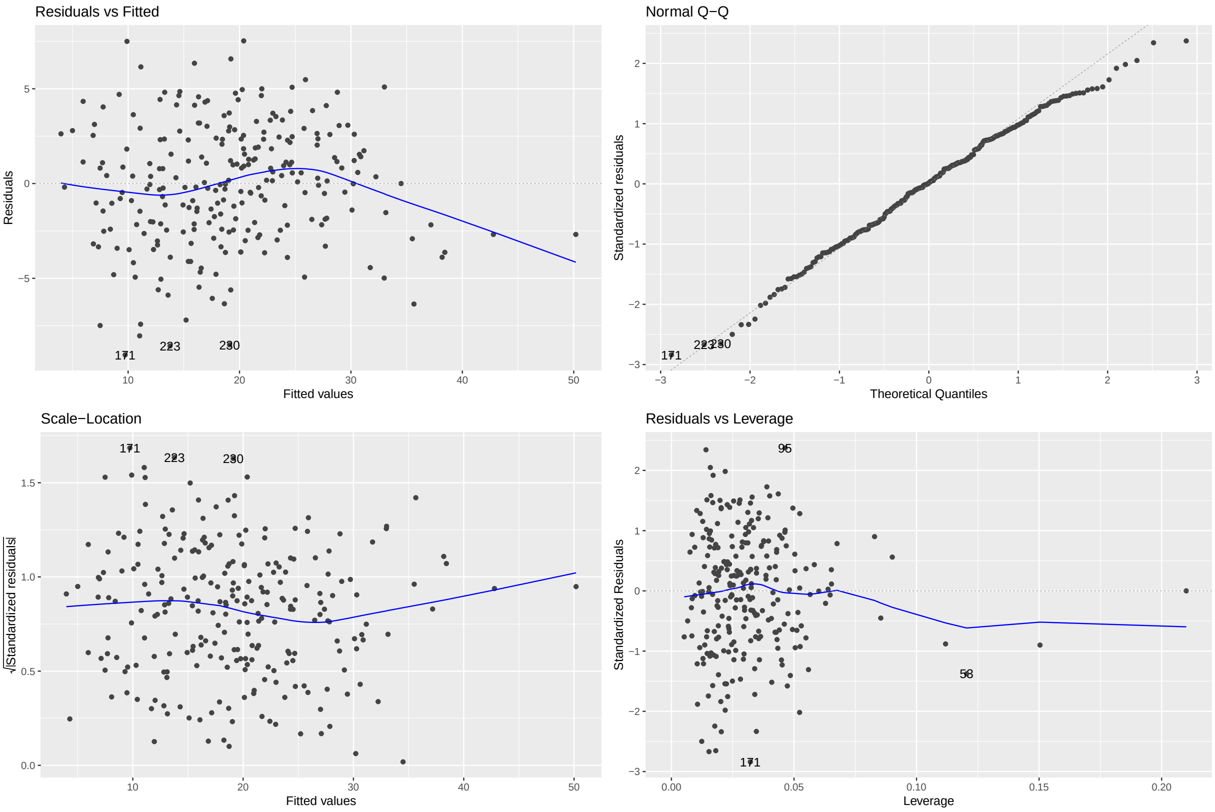


Figure 4.4: Multiple linear regression results diagram for the final model.