FNSPE CTU

ASM

DATASET N.6

Fitting Percentage of Body Fat to Simple Body Measurements

Author:
Vladislav BELOV

September 8, 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.[1] 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.[2] Age is also considered a continuous instance, nevertheless, for illustrative purposes we have categorized it, and results can be seen in Tab. 2.2.

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]	162.6	173.4	177.8	178.6	183.5	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 fat percentage given by both Siri's and Brozek's equations. Similarities between them start to become noticeable. Moreover, looking at comparison of empirical cumulative distribution functions (see Fig. 2.4), we can speculate, that they have similar distributions, besides, 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.

[2] Observation 42 had a suspicious value of height equal to 74.93 *cm*. The height was recalculated using the adiposity index and the total weight and was set to 176.349 *cm*.

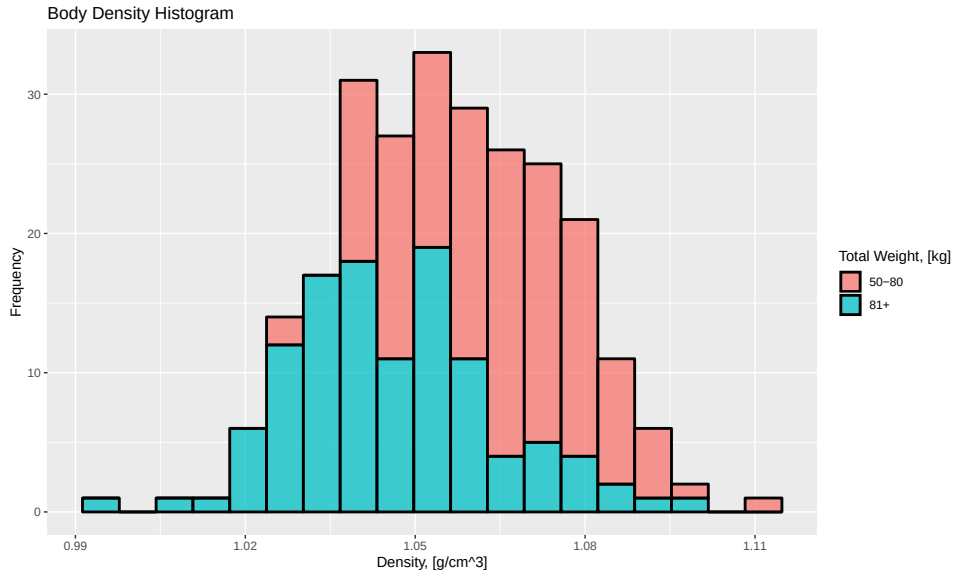

Figure 2.1: Body density histogram with categorized total weight for illustrative purposes.

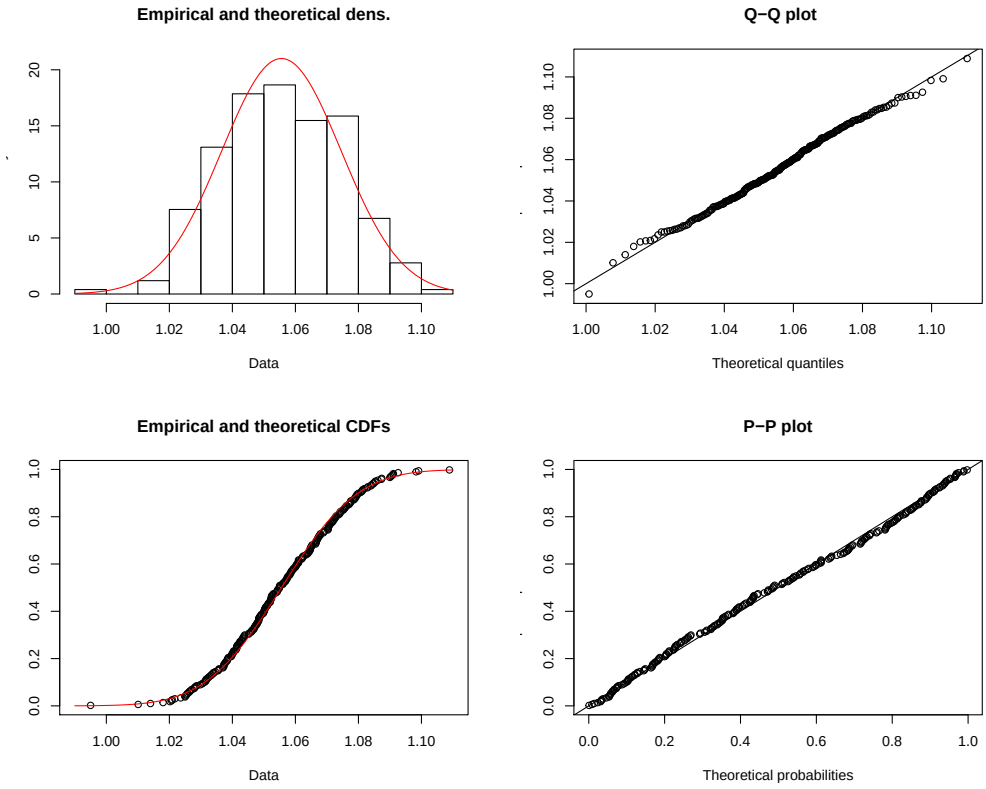

Figure 2.2: Normal distribution fit to the body density.

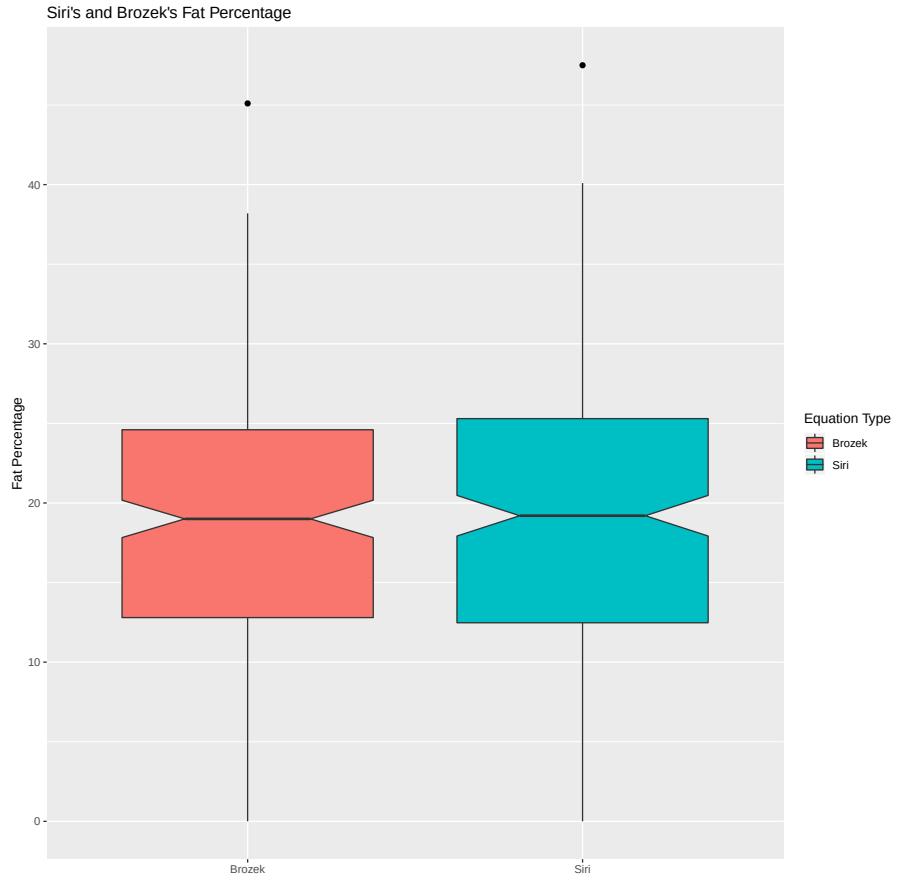

Figure 2.3: Comparative box plot of fat percentage given by Siri's and Brozek's equations.

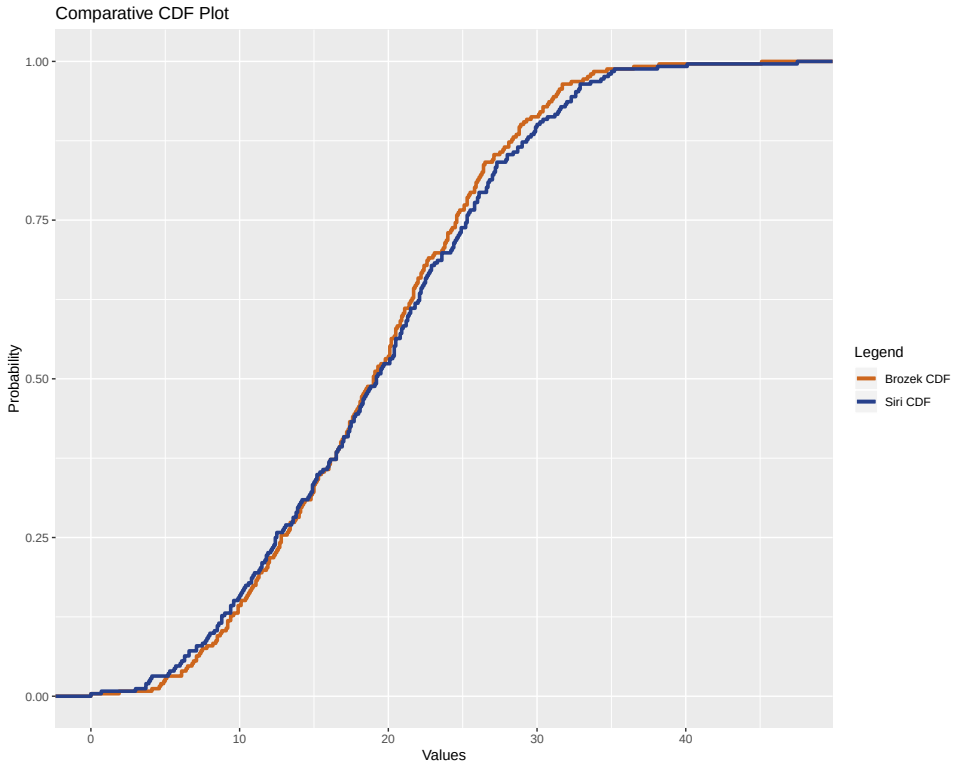

Figure 2.4: Comparative plot of ECDFs for fat percentage given by Siri's and Brozek's equations.

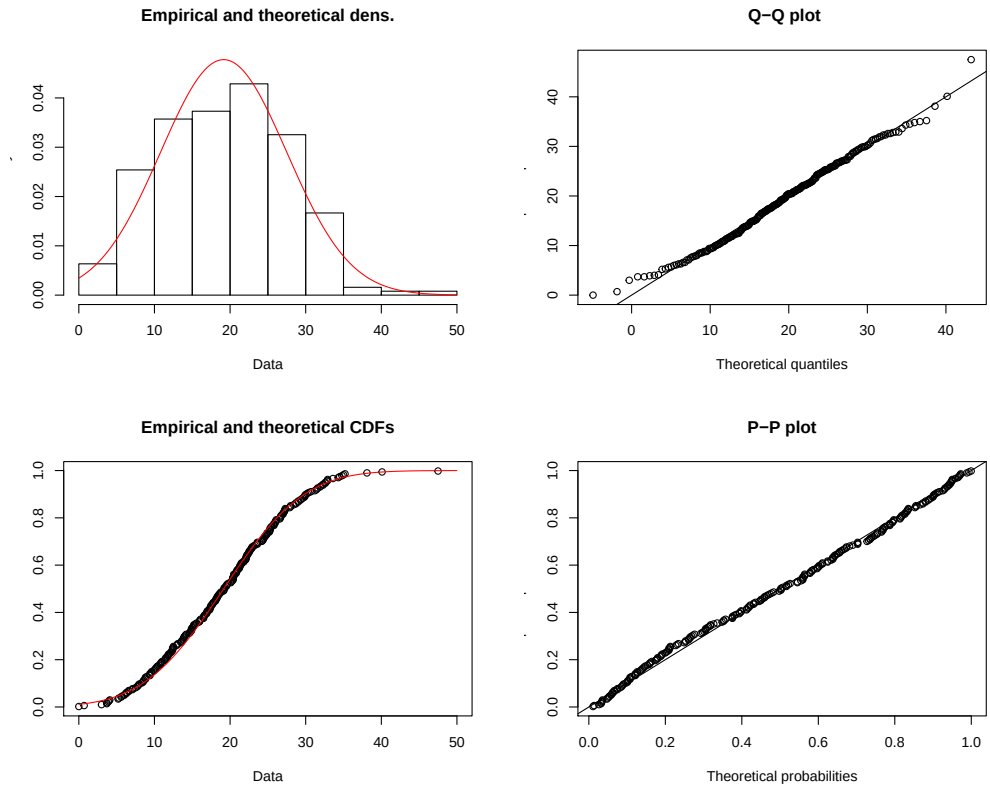

Figure 2.5: Normal distribution fit to fat percentage given by Siri's equation.

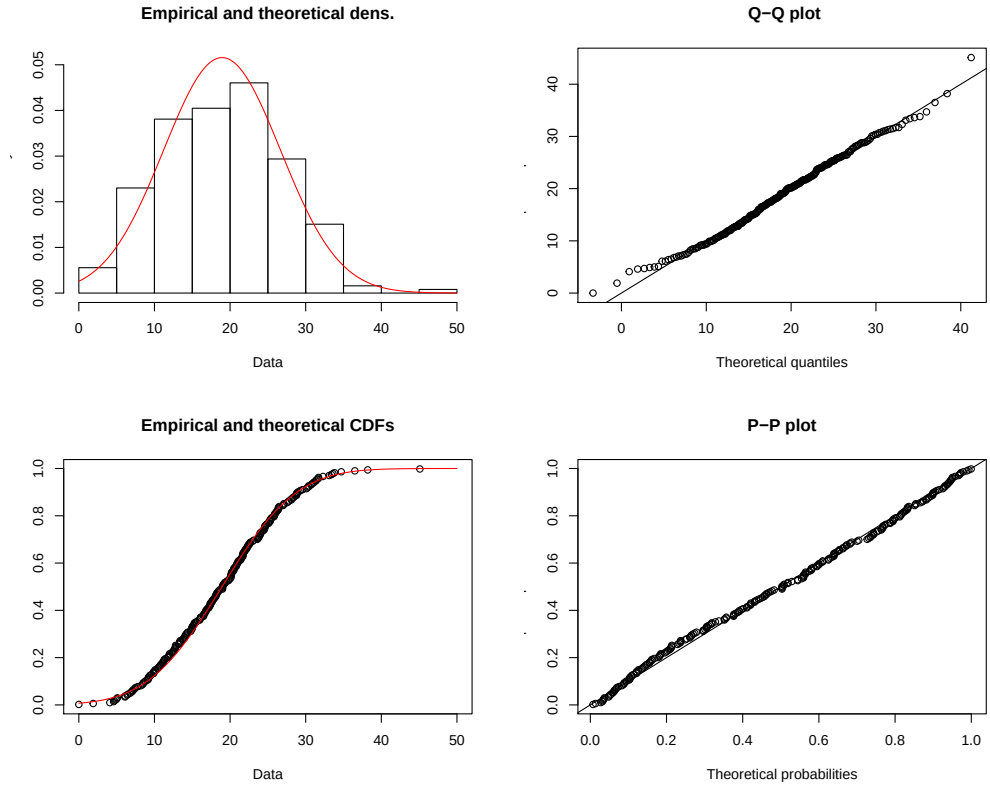

Figure 2.6: Normal distribution fit to fat percentage given by Brozek's equation.

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. In Fig. 2.7 a scatter plot against the body density can be seen. Once again our speculation about the fact, that the body density is decreasing with increasing body weight, is supported. Moreover, as we can see, the main reason for that is the increasing amount of fat.

Regarding the distribution of the total body weight, it is not a trivial task to perform a fit in this case. As one can see, both gamma (Fig. 2.8) and normal (Fig. 2.9) distributions seem to be able to describe the variable relatively well. To determine the true distribution a number of statistical tests, which are available in section 3, have to be carried out.

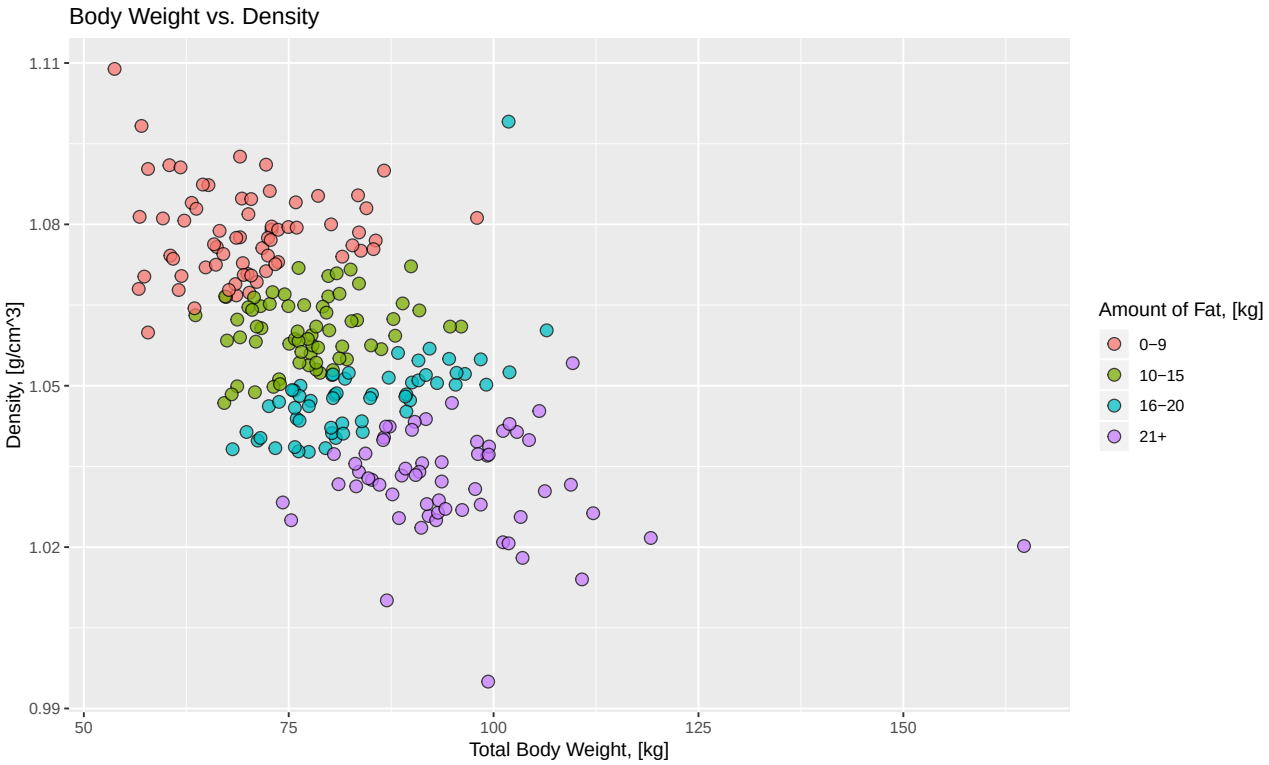

Figure 2.7: Influence of the total body weight on the body density.

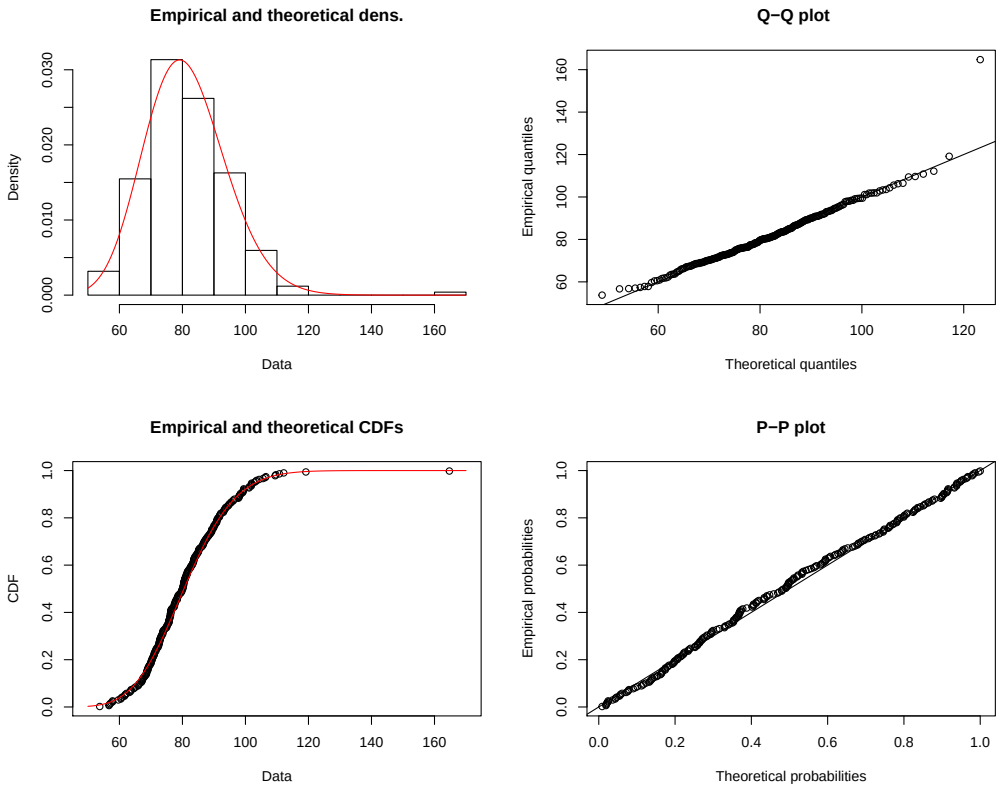

Figure 2.8: Gamma distribution fit to the total weight.

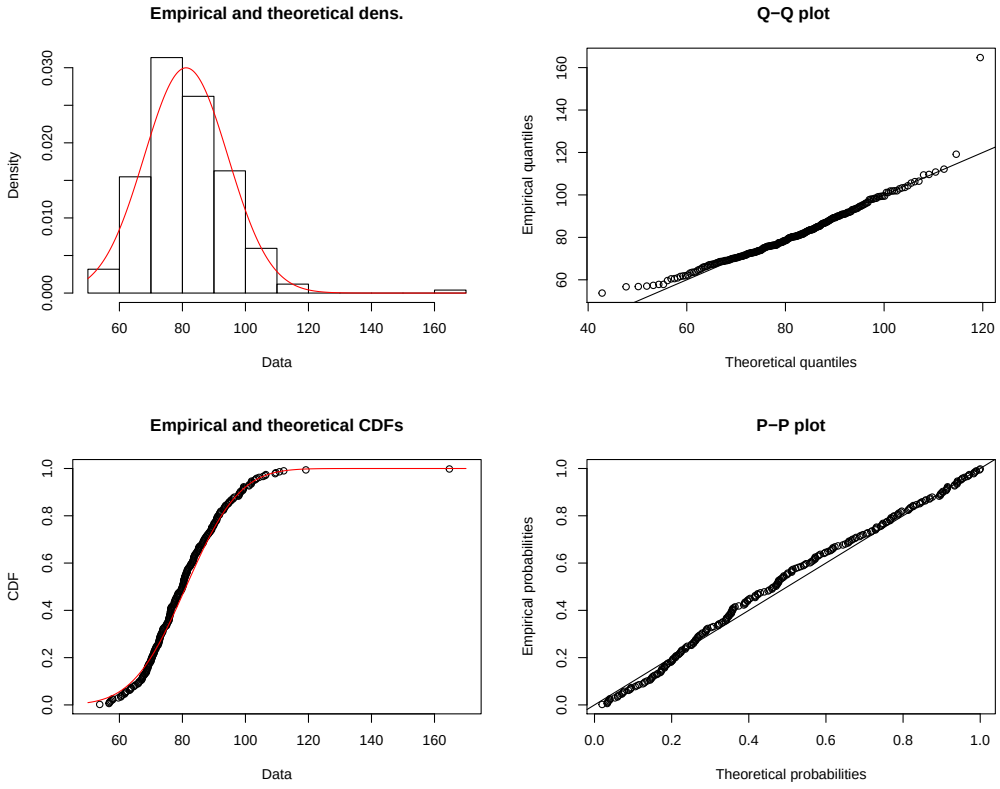

Figure 2.9: Normal distribution fit to the total weight.

3 Data Analysis

Note: all tests are performed with statistical significance $\alpha = 5\%$.

3.1 Body Density Distribution

In the previous section we have performed a fit of normal distribution to the body density, and, according to presented diagrams, it seemed to be a reasonable thing to do. In this section a number of normality tests will be performed to support our hypothesis. Lilliefors and Shapiro-Wilk normality tests will be of aid:

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

Results of tests are displayed in the table below:

Name of the Test	Value of the Statistic	p-valiue
Lilliefors	0.038407	0.4864
Shapiro-Wilk	0.9954	0.6571

According to both of the performed tests, no evidence against the null hypothesis is present. This entitles us to conclude, that the distribution of the body density is truly normal. Estimated parameter values and respective confidence intervals are available in the table below:

Parameter	Estimated Value	Confidence Interval, 95%	
Mean, μ_d	1.055574	1.053229	1.057919
Standard Deviation, σ_d	0.018994	0.017356	0.020631

Once again we cannot reject the null hypothesis, and the distribution of the body density is truly $\mathcal{N}(\mu_d, \sigma_d^2)$.

3.2 Analysis of Body Fat Percentage by Siri's and Brozek's Equations

3.2.1 Normality Tests

In this section we will perform normality tests for fat percentage given by Siri's and Brozek's equations in the same manner as in section 3.1. The results of Lilliefors and Shapiro-Wilk tests are as follows:

Variable	Name of the Test	Value of the Statistic	p-valiue
Siri	Lilliefors	0.044548	0.2584
	Shapiro-Wilk	0.99168	0.1649
Brozek	Lilliefors	0.039781	0.429
	Shapiro-Wilk	0.99292	0.2747

Tests entitle us to conclude, that the null hypothesis (normality of the distribution) cannot be rejected for both Siri's and Brozek's fat percentage - the distribution is truly normal. Its estimated parameters are displayed below:

Variable	Parameter	Estimated Value	Confidence Interval, 95%	
Siri	Mean, μ_s	19.150794	18.119590	20.182
	Standard Deviation, σ_s	8.352119	7.622948	9.08129
Brozek	Mean, μ_b	18.938492	17.983425	19.893560
	Standard Deviation, σ_b	7.735462	7.060127	8.410796

P-values allow us to conclude, that fat percentage given by both Siri's and Brozek's equations is normally distributed ($\mathcal{N}(\mu_s, \sigma_s^2)$ and $\mathcal{N}(\mu_b, \sigma_b^2)$, respectively).

3.2.2 Equality of Distributions

In this section we will perform the unpaired two-sample Kolmogorov-Smirnov test to determine, if distributions of fat percentage given by Siri's and Brozek's equations are equal. From the comparative diagram of cumulative distribution functions (Fig. 2.4) it is obvious, that no significant horizontal shifts are present. Thus, the KS-test will provide a sufficient and reliable result:

$$\mathrm{H}_0^{(\mathrm{KS})} : F = G \text{ vs. } \mathrm{H}_1^{(\mathrm{KS})} : F \neq G$$

where F and G are distribution functions of fat percentage calculated using Siri's and Brozek's equations, respectively.[3] Results of the test are positive, as the p-value is equal to 0.9375 which means, that no strong evidence against the null hypothesis $\mathrm{H}_0^{(\mathrm{KS})}$ is existent - distributions are similar.[4] Now that we know, that two normal distributions are equal, the t-test will be performed to confirm, that their mean values are equal:

$$\mathrm{H}_0^{(\mathrm{t})} : \mu_s = \mu_b \text{ vs. } \mathrm{H}_1^{(\mathrm{t})} : \mu_s \neq \mu_b.$$

The p-value of the t-test is equal to 0.7678, hence, the null hypothesis $\mathrm{H}_0^{(\mathrm{t})}$ cannot be rejected. Distributions are truly equal.

3.3 Total Weight Distribution

In the scope of this section the distribution of the total weight of the population will be determined. As seen in section 2.4 we are unable to determine which distribution fits to the total weight better: gamma or normal. Taking a closer look at the Q-Q plot (Fig. 3.1) gives us a clue, that fitting with gamma distribution appears to be slightly more competent.[5]

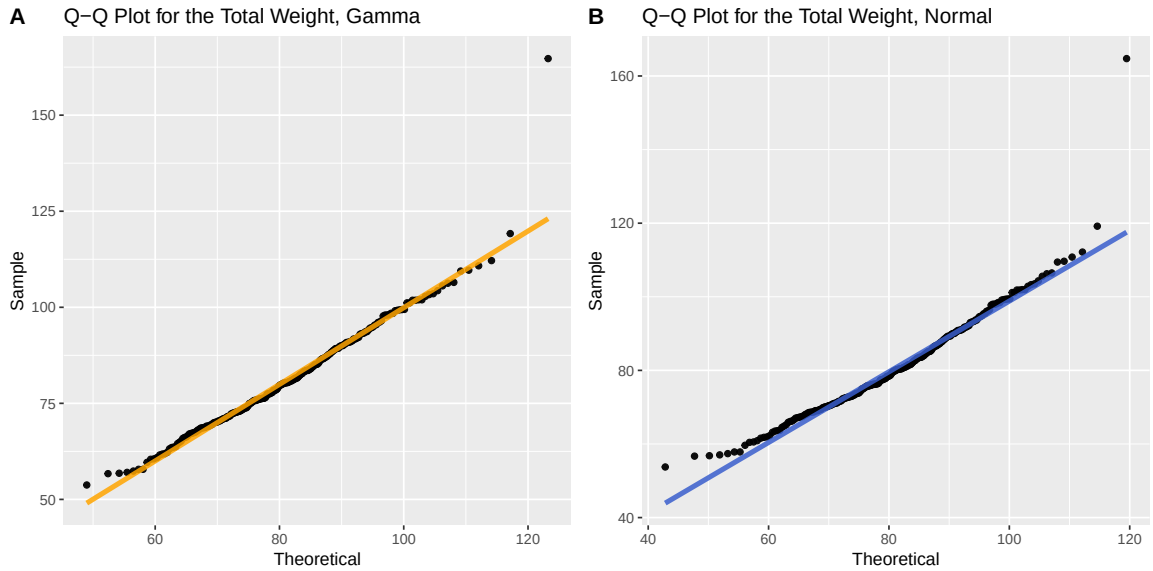

Figure 3.1: Q-Q plot for estimated Gamma and Normal distributions for the total weight.

However, we will test both distributions using the chi-squared goodness of fit test:

$$\mathrm{H}_0^{(G)} : \mathit{weight} = \mathit{Gamma}(\alpha_w, \beta_w) \text{ vs. } \mathrm{H}_1^{(G)} : \mathit{weight} \neq \mathit{Gamma}(\alpha_w, \beta_w),$$
$$\mathrm{H}_0^{(N)} : \mathit{weight} = \mathcal{N}(\mu_w, \sigma_w^2) \text{ vs. } \mathrm{H}_1^{(N)} : \mathit{weight} \neq \mathcal{N}(\mu_w, \sigma_w^2).$$

[3] We already know, that those distributions are normal, and their parameters have already been estimated. However, the unpaired two-sample KS-test is nonparametric.

[4] The unpaired two-sample Wilcoxon test works well to detect horizontal shifts in compared CDFs. We did not use this test to reach the conclusion, however, its results are also positive: p-value = 0.7789.

[5] The point at the top-right part of both diagrams is an obvious outlier. It will be removed from the dataset in the scope of section 4.

Tested Distribution	Value of the Statistic	p-value
Gamma	2.9615	0.9975
Normal	7.3248	0.8881

P-values of tests do not suggest to reject either $H_0^{(G)}$ or $H_0^{(N)}$. On the other hand, they favor gamma distribution (p-value$^{(G)}$ is greater than p-value$^{(N)}$ by 0.1094). That is the main reason behind us choosing gamma distribution as the best fit to the total weight of the population. In the table below estimated values of distribution parameters are displayed:

Parameter	Estimated Value	Confidence Interval, 95%	
Shape, α_w	39.731517	32.824526	46.638507
Scale, β_w	0.489557	0.403914	0.575201

To sum up, we have statistically proved, that the distribution of the total weight within the population is $Gamma(\alpha_w, \beta_w)$.

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 Two-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%		p-value
Intercept, β_0	-47.67	-52.86	-42.48	$3.16 \cdot 10^{-47}$
Abdomen CC, β_1	0.98	0.87	1.09	$3.63 \cdot 10^{-45}$
Total Weight, β_2	-0.29	-0.38	-0.2	$1.5 \cdot 10^{-9}$

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

The fit 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.

$$H_0 : res \in \{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\} \text{ vs. } H_1 : res \notin \{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$$

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 - usage of the linear model is permitted.

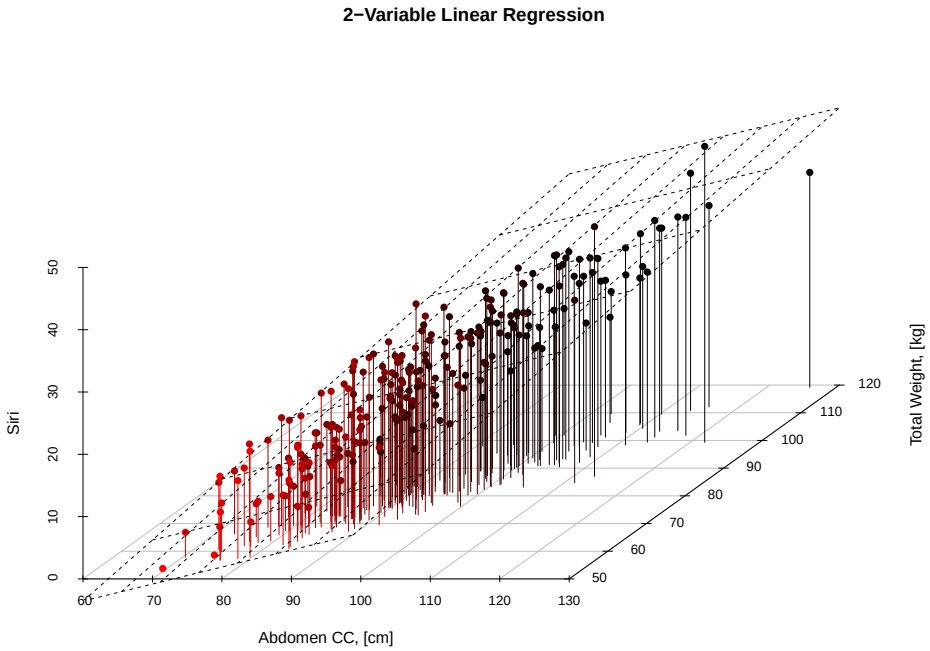

Figure 4.1: Fit of the linear model with two explanatory variables.

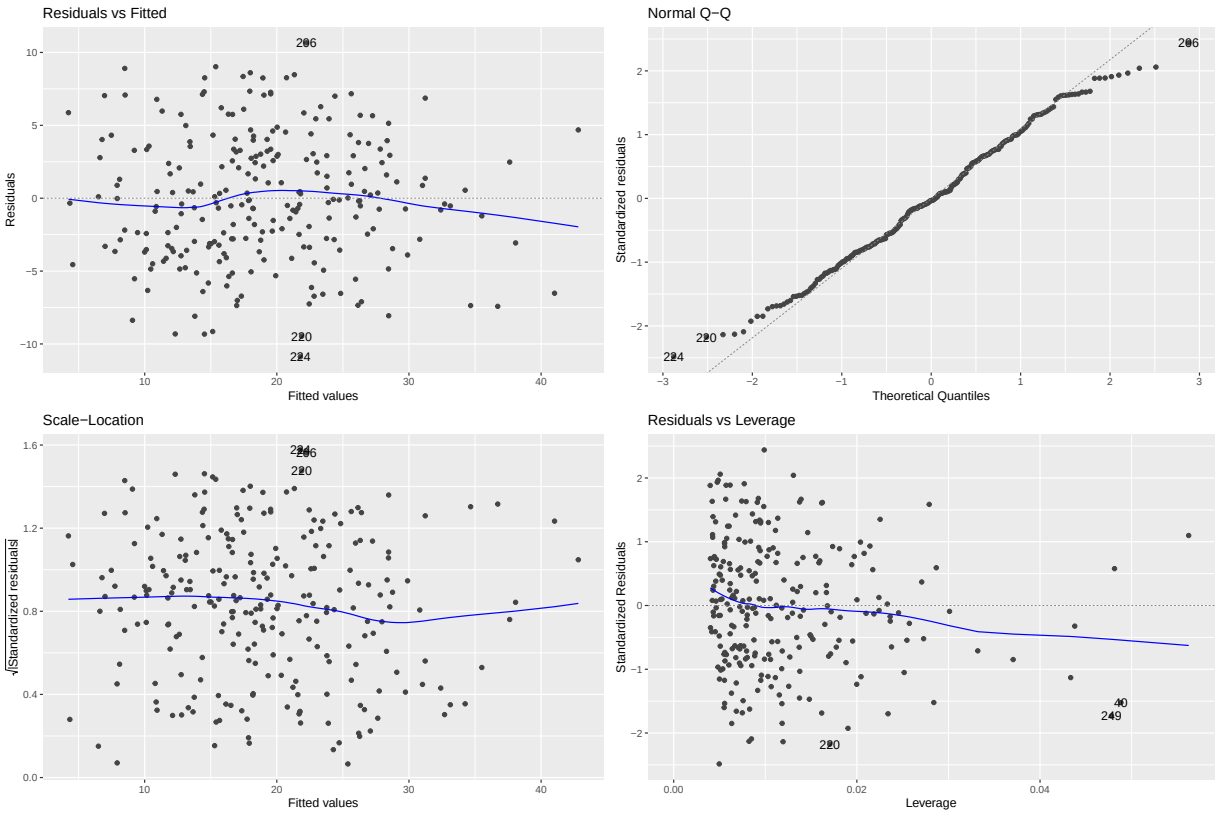

Figure 4.2: Two-variable linear regression results diagram.

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, $\forall i \in \{1, 2, \ldots, 251\}$:[6]

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

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 0. 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 $1.461 \cdot 10^{-6}$. Thus, linear regression cannot be used for this set of explanatory variables.

Parameter	Estimated Value	Confidence Interval, 95%		p-value
Age, β_1	0.015	-0.006	0.036	0.16
Total Weight, β_2	0.893	0.838	0.949	$6.46 \cdot 10^{-87}$
Height, β_3	0.038	-0.016	0.092	0.17
Adiposity, β_4	-0.159	-0.457	0.139	0.3
Fat-free weight, β_5	-1.233	-1.288	-1.177	$3.34 \cdot 10^{-115}$
Neck CC, β_6	0.035	-0.117	0.188	0.65
Chest CC, β_7	0.011	-0.06	0.083	0.75
Abdomen CC, β_8	0.104	0.033	0.174	0.004
Hip CC, β_9	-0.039	-0.134	0.056	0.42
Thigh CC, β_{10}	0.13	0.036	0.225	0.007
Knee CC, β_{11}	0.031	-0.129	0.191	0.7
Ankle CC, β_{12}	0.102	-0.043	0.247	0.17
Biceps CC, β_{13}	0.114	0.003	0.225	0.04
Forearm CC, β_{14}	0.072	-0.064	0.208	0.3
Wrist CC, β_{15}	-0.087	-0.44	0.267	0.63

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

4.2.2 The Final Model

To make regression sufficient and authorized to be used, 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 best model by the Akaike Information Criterion. After excluding variables with low significance, we construct the final linear model using only 7 columns of the provided dataset: age, height, fat-free weight, abdomen CC, Knee CC, biceps CC and wrist CC. Estimates of parameters are available in Tab. 4.3. Fortunately, the R-squared statistic has remained high: $R^2 = 0.9762$, thus, almost all of the variability of the response variable is accounted for by our model.

As can be seen, confidence intervals no longer contain zeros. 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 0.3947. Hence, residuals distribution may be considered normal and we are eligible to use this model to perform multivariate linear regression.

[6] 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.

Parameter	Estimated Value	Confidence Interval, 95%		p-value
Age, β_1	-0.051	-0.091	-0.012	0.012
Height, β_2	-0.117	-0.174	-0.06	$6.86 \cdot 10^{-5}$
Fat-free weight, β_3	-0.558	-0.638	-0.478	$3.48 \cdot 10^{-32}$
Abdomen CC, β_4	0.747	0.681	0.814	$2.5 \cdot 10^{-60}$
Knee CC, β_6	0.397	0.083	0.71	0.013
Biceps CC, β_7	0.442	0.228	0.656	$6.5 \cdot 10^{-5}$
Wrist CC, β_8	-1.09	-1.749	-0.434	0.001

Table 4.3: Estimated values for the multivariate linear regression in the final model.

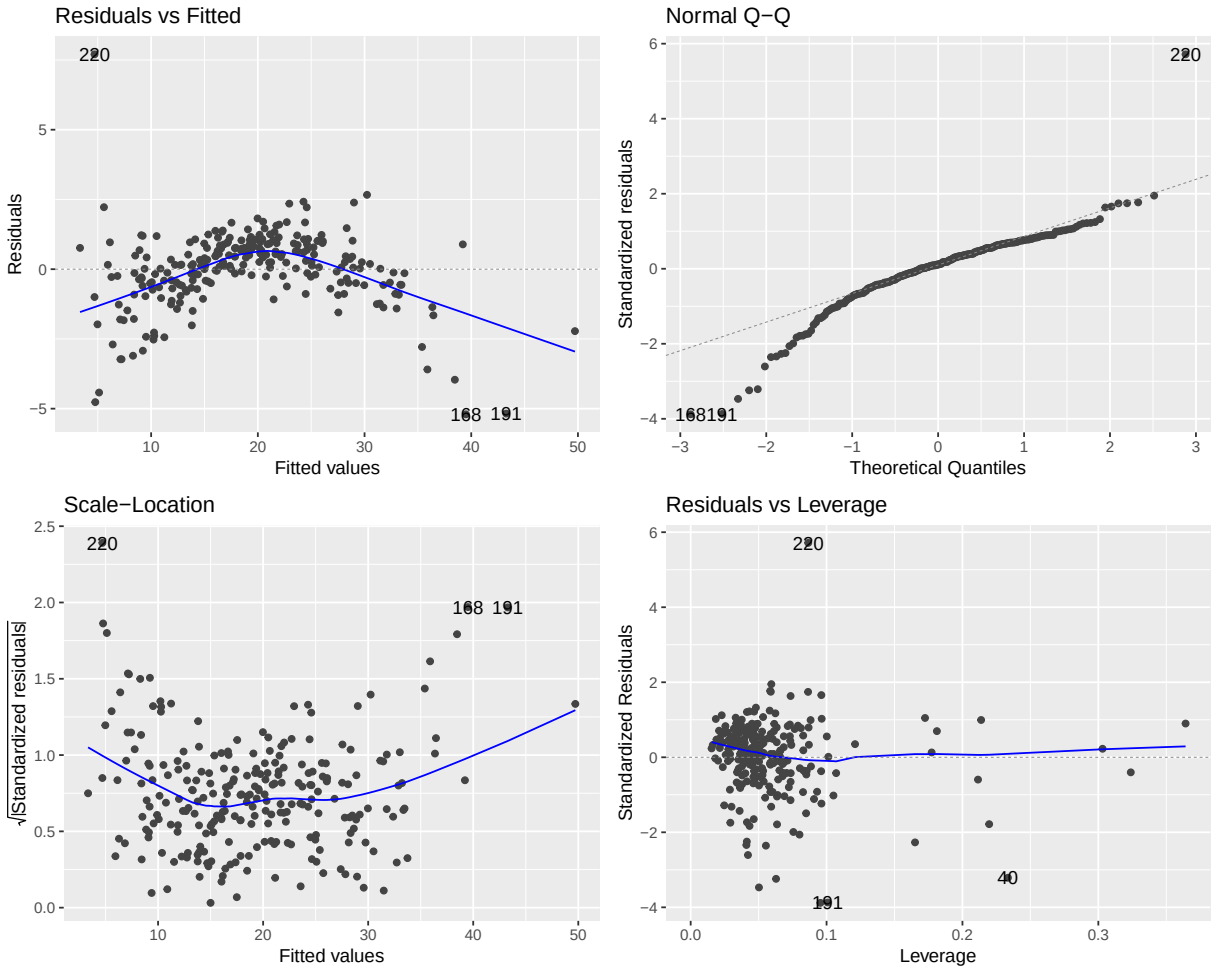

Figure 4.3: Multivariate linear regression results diagram for the model with all variables included.

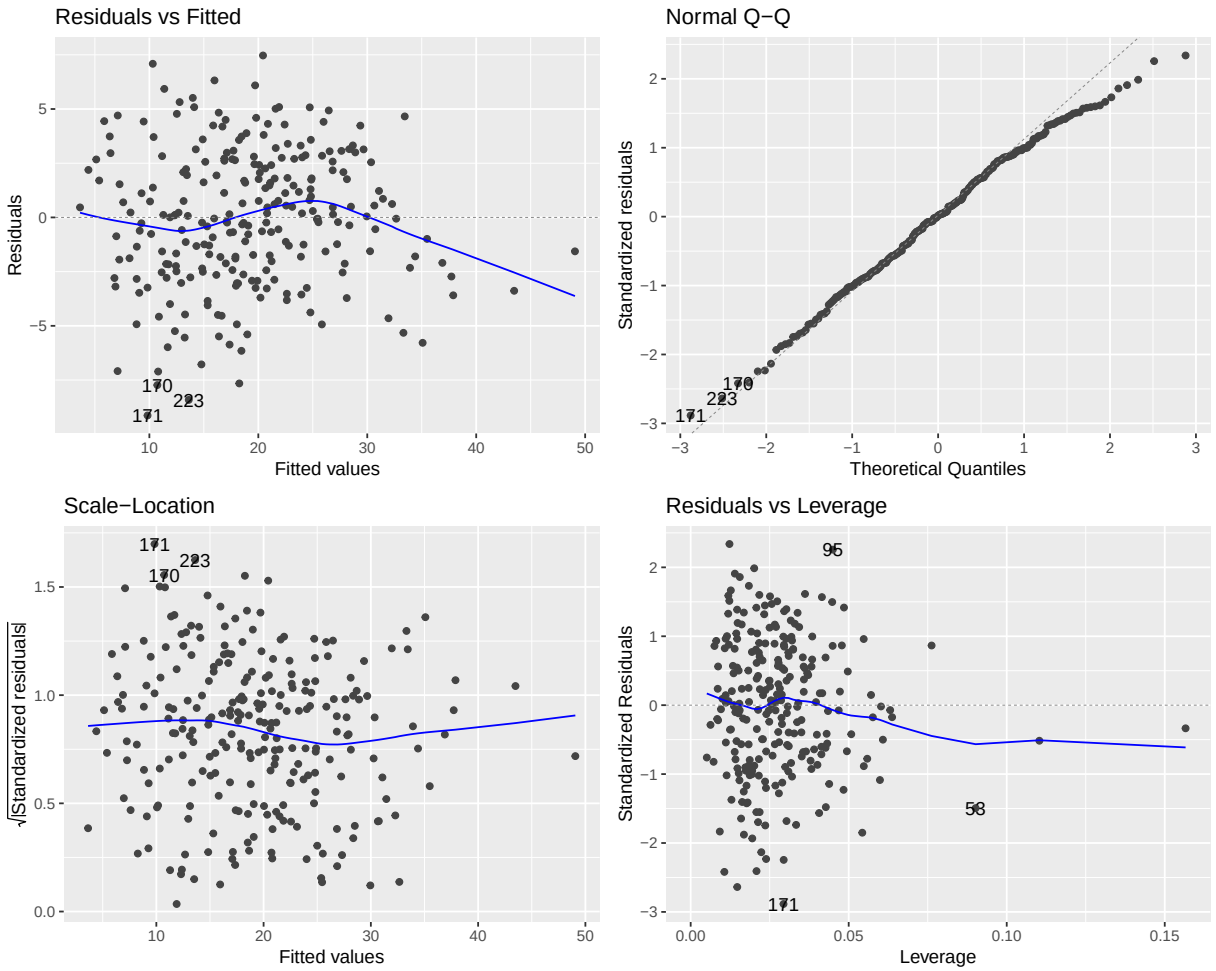

Figure 4.4: Multivariate linear regression results diagram for the final model.

5 Conclusion

In this paper an analysis of the population of 252 men was performed based on their body parts circumference, weight, age and other parameters.

- We have provided descriptive statistics and detailed analysis of selected variables. Results are the following:
  - Distribution of the body density is $\mathcal{N}(\mu_d, \sigma_d^2)$;
  - Distribution of the total weight is $Gamma(\alpha_w, \beta_w)$;
  - Distributions of body fat percentage given by Siri's and Brozek's equations are equal to each other $\mathcal{N}(\mu_s, \sigma_s^2)$ and $\mathcal{N}(\mu_b, \sigma_b^2)$, respectively.
- Fits to body fat percentage using multivariate linear regression were carried out:
  In the final model estimated parameters for some of explanatory variables are negative (namely, age, height, fat-free weight, wrist CC). Increasing age, height and muscle mass (part of the fat-free weight) obviously have a negative effect on the fat level in the body. Fortunately a logical explanation for the wrist circumference may also be existent, as wrists of working out men (consequently, men who have relatively small fat percentage) are usually grown in size due to weight lifting.