Gender Differences in Body Weight and Weightlifting Performance: A Comparative Analysis

Aditya Sagave

2023-06-03

Abstract

This statistical report examines the relationship between gender, age, body weight, and weight lifted among weightlifters. The research questions address whether there is a difference in the average age between female and male weightlifters and the relation between body weight and weight lifted, considering gender as a factor. The analysis involves hypothesis testing, linear regression, and diagnostic assessments. The results provide insights into the average age differences and the relationship between body weight and weight lifted among weightlifters.

Introduction

Weightlifting is a popular sport that requires strength, power, and precision. Understanding the factors that contribute to weightlifting performance is of great interest to athletes, trainers, and researchers. In this statistical report, we aim to investigate two research questions related to weightlifters' characteristics and their performance:

1. Is there any difference in the average age of female and male weightlifters?
2. What is the relation between the body weight of weightlifters and the weight lifted?

To address the first research question, we will examine if there is a significant difference in the average age between female and male weightlifters. This analysis will help us understand if age plays a role in weightlifting performance and if there are any gender-specific differences in age distribution within the weightlifting population.

For the second research question, we will explore the relationship between the body weight of weightlifters and the weight lifted. We will consider the role of gender and investigate whether there is a consistent relationship between body weight and weight lifted for both females and males. This analysis will provide insights into the impact of body weight on weightlifting performance and whether there are gender-specific variations in this relationship.

The dataset used in this report consists of information on weightlifters' gender, body weight, age, and the maximum weight lifted. We will analyze this dataset using statistical methods learned in the unit to draw meaningful conclusions and provide valuable insights into weightlifting performance.

In the following sections, we will describe the methods used for data analysis, present the results of our analyses, and discuss the implications of our findings.

Methods

The research questions in this report were addressed using the following methods:

Data Collection and Description

The dataset used in this analysis consists of information on weightlifters, including their gender, body weight, age, and the maximum weight lifted. The data were collected from a random sample of 201 weightlifters. Please refer to the "weightlifters.csv" file provided with this report for the complete dataset.

To gain an understanding of the dataset, we performed initial data exploration. This involved examining the structure of the dataset, checking for missing values or outliers, and calculating summary statistics. The distribution of variables was visualized using histograms, bar plots, and boxplots to identify any patterns or anomalies.

Research Question 1: Average Age Difference between Female and Male Weightlifters

To investigate the difference in the average age of female and male weightlifters, we performed a two-sample t-test. This test compares the means of the age variable for females and males, assuming that the samples are independent and follow approximately normal distributions with equal variances.

To check the assumptions of the t-test, we verified that the sample sizes were large enough for the Central Limit Theorem to apply, ensuring that the sampling distribution of the sample mean would be approximately normal. We also assessed the assumption of equal variances between the two groups.

Research Question 2: Relationship between Body Weight and Weight Lifted

To study the relationship between body weight and weight lifted, we considered the role of gender. We examined if the relationship differed between females and males and whether separate analyses were necessary.

For both females and males, we performed simple linear regression analyses to model the relationship between body weight (independent variable) and weight lifted (dependent variable). We assessed the linearity assumption by inspecting scatter plots and checked other assumptions such as independence of errors, normality of residuals, and homoscedasticity.

Additionally, we explored potential interactions between body weight and gender by including interaction terms in the regression models. This allowed us to assess whether the relationship between body weight and weight lifted differed significantly between females and males.

The statistical analyses were conducted using the R programming language. The built-in plotting functions in R, such as hist(), barplot(), and boxplot(), were used for data visualization. The lm() function in R was employed for regression analysis.

Please proceed to the next section for the presentation of the results obtained from the data analysis.

Results 1 — Preliminary Data Exploration

To provide a comprehensive understanding of the dataset and its variables, we conducted preliminary data exploration. This involved generating graphs to visualize the frequency distributions of the variables, as well as comparative and bivariate relationships. In addition, we included brief numerical summaries in a table, along with accompanying comments to summarize the main features of the graph and provide insights into the data.

First let us import the dataset in RStudio

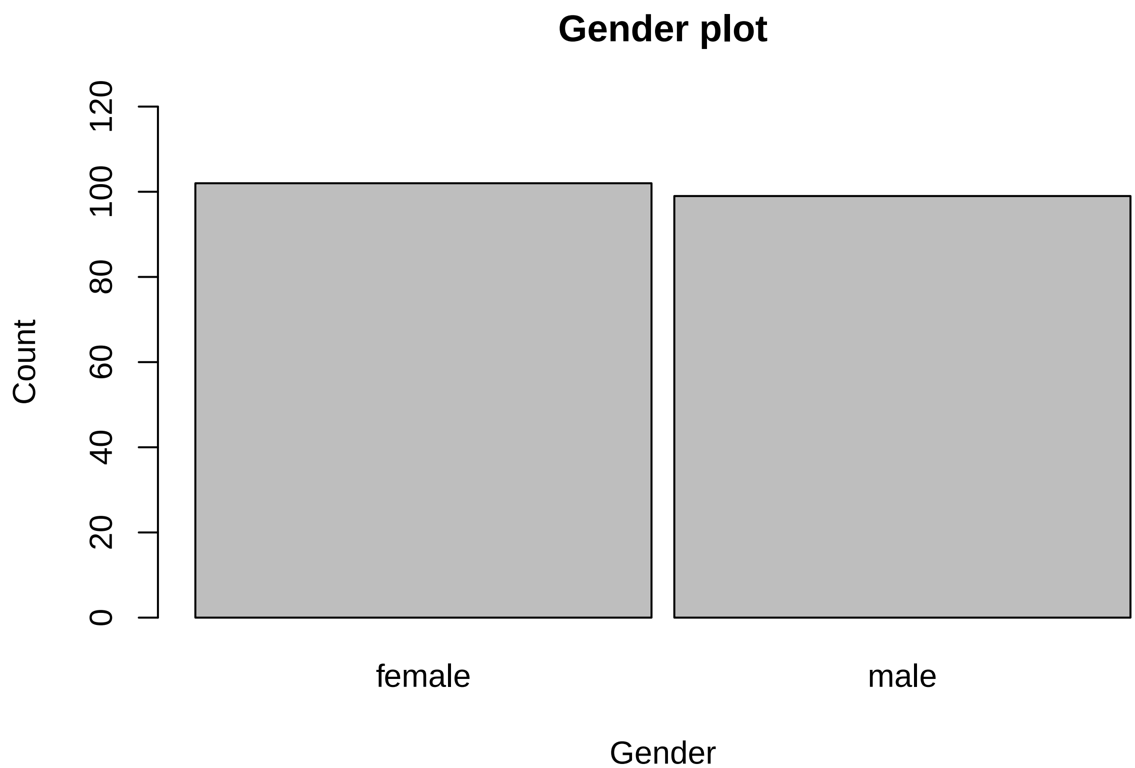
```
dat <- read.csv("./data/dataset.csv", header = TRUE)
```

Now let us have a look at first few rows of the dataset

```
##      ID gender bodyweight   age weightlifted
## 1 subj1   male      113.8 31.45        177.0
## 2 subj2   male      110.9 24.87        175.1
## 3 subj3 female       66.9 23.89        106.9
## 4 subj4   male      115.4 23.66        179.7
## 5 subj5 female       84.5 32.62        113.2
## 6 subj6   male      108.3 31.32        178.6
```

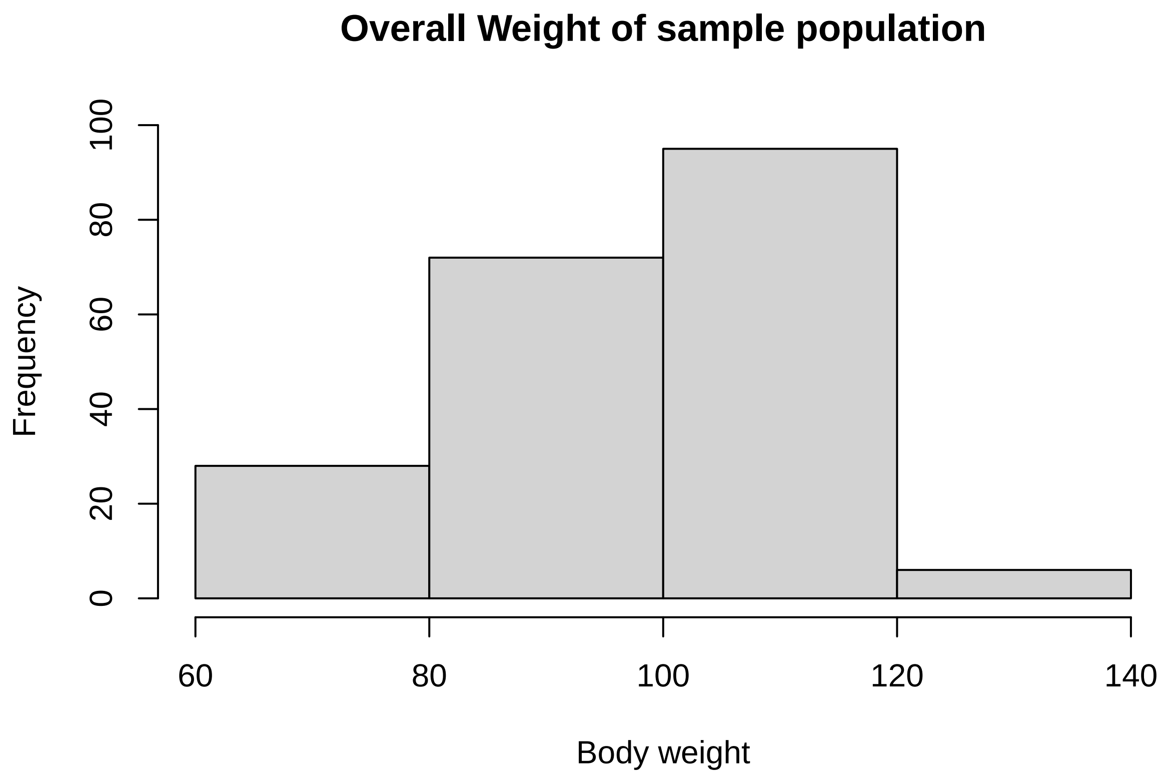
Frequency Distributions:

1. Gender:
   - Bar Chat:



Gender	Total Number
Female	102
Male	99

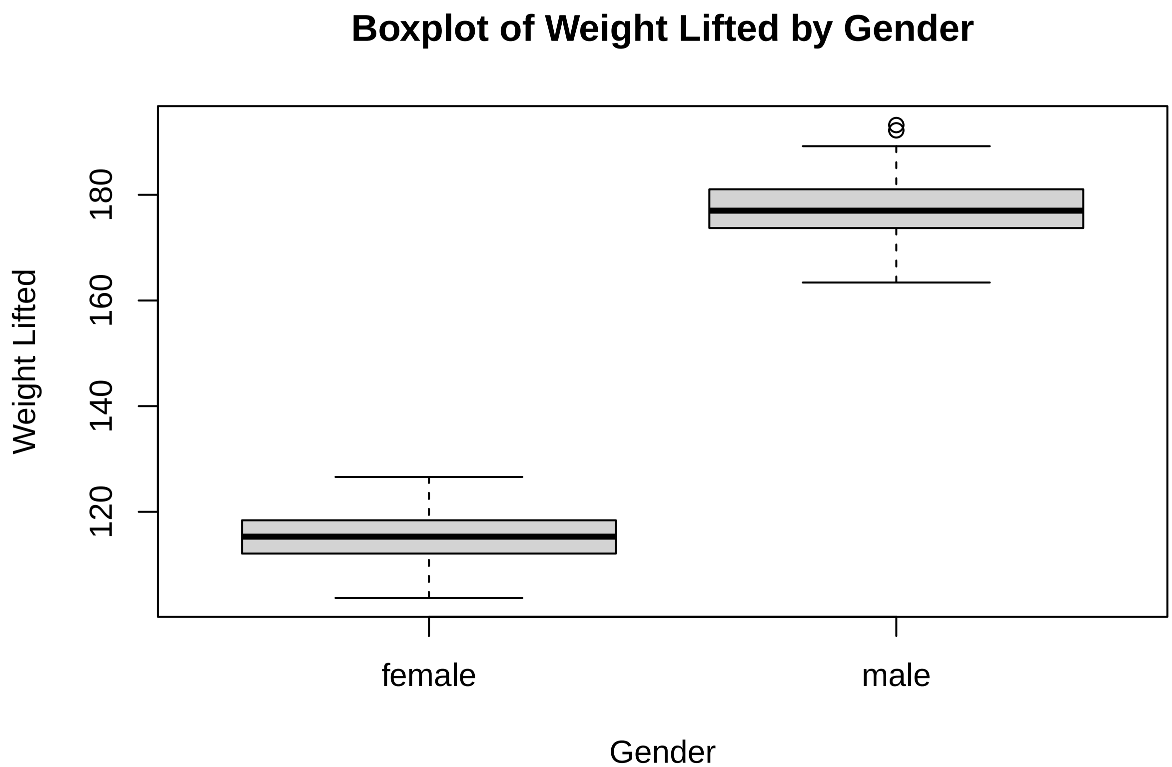
2. Body Weight:
   - Histogram:



Comment:

The histogram illustrates the distribution of body weights among weightlifters. The data is divided into four intervals: 60-80, 80-100, 100-120, and 120-140 units. The histogram shows that approximately 30 weightlifters have body weights between 60 and 80 units, around 70 weightlifters fall within the range of 80 to 100 units, approximately 90 weightlifters have body weights between 100 and 120 units, and only around 10 weightlifters have body weights between 120 and 140 units. The distribution appears to be slightly right-skewed, with the majority of weightlifters concentrated in the 80-100 and 100-120 unit intervals.

3. Gender <> Weight Lifted
    - Boxplot:



Comment:

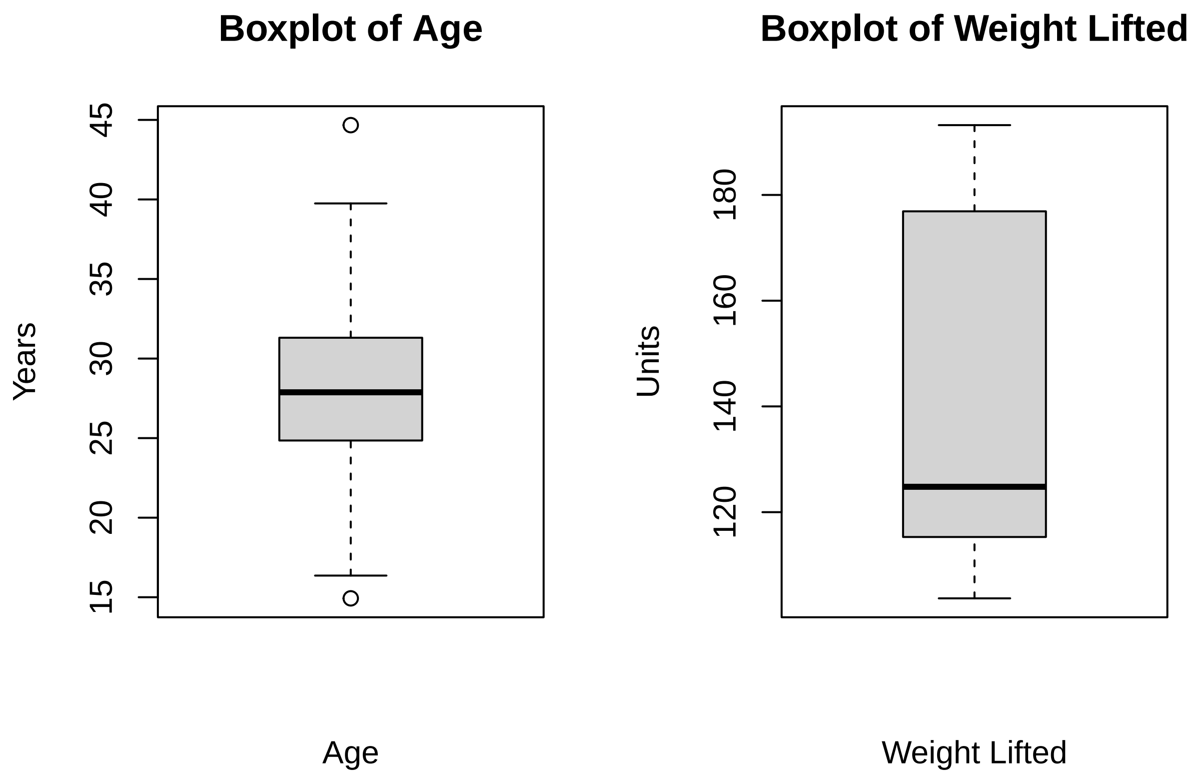
The boxplot of weight lifted by gender reveals distinct differences between male and female weightlifters. The median weight lifted for female weightlifters is around 120, while for male weightlifters, it is approximately 180. This substantial difference suggests that, on average, male weightlifters lift significantly heavier weights compared to their female counterparts.

Regarding the dispersion, both male and female weightlifters show relatively low variability as indicated by the boxplot's narrow width. It suggests that the majority of weightlifters, regardless of gender, have weights lifted that are relatively close to the median.

In the case of female weightlifters, the boxplot indicates an equal distribution of data points above and below the median, demonstrating a symmetrical distribution within the interquartile range. On the other hand, the boxplot for male weightlifters appears to be slightly right-skewed, indicating that some weightlifters lifted exceptionally heavy weights, possibly due to measurement errors or outliers.

These observations highlight the significant differences in weightlifting performance between male and female weightlifters, with males generally lifting higher weights. Additionally, the boxplot provides insights into the dispersion and distribution characteristics of weight lifted for each gender.

4. Age <> Weight Lifted
    - Boxplot:



Comment:

The boxplot of age reveals that the median age of weightlifters is between 25 and 30 years. The minimum age falls between 15-20 years, while the maximum age reaches around 40 years. Additionally, there are two outliers—one at 15 years and the other at 45 years—indicating weightlifters who are significantly younger or older compared to the majority of the sample. The spread of data points is symmetrically distributed above and below the median, suggesting no apparent skew in the age distribution.

In contrast, the boxplot of weight lifted shows a right-skewed distribution. The median weight lifted is approximately between 120-130 units. The absence of outliers in the boxplot indicates that there are no extreme values for weight lifted. This suggests that weightlifters generally fall within a certain range of weight lifted, with no exceptionally high or low values.

These observations highlight the distribution characteristics of age and weight lifted among weightlifters. The age distribution shows no skewness and exhibits a symmetrical spread of data, while the weight lifted distribution is right-skewed, indicating a concentration of weightlifted values towards the lower end with no outliers.

Results 2 — Analyses

In this section, we present the outcomes of the analyses conducted to address the research questions. We incorporate statistical models and test statistics, along with their corresponding p-values, to provide relevant insights. We also evaluate the extent to which the assumptions of the models or tests have been satisfied and discuss any discrepancies in the data, such as outliers, and how they were handled.

Research Question 1:

Is there any difference in the average age of female and male weightlifters?

The null hypothesis for Research Question 1 would be: "There is no difference in the average age of female and male weightlifters."

Symbolically, this can be represented as: $H_0 : \mu_{female} = \mu_{male}$

The alternative hypothesis would be that there is a difference in the age of female weightlifters and male weightlifters.

Symbolically, this can be represented as: $H_1 : \mu_{female} \neq \mu_{male}$

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.04   25.46   27.66   28.37   31.10   44.67
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.93   23.85   28.27   27.58   31.32   39.75
```

Summary of the age of female and male weightlifters

Gender	Min.	1st Qu.	Median	Mean μ	3rd Qu.	Max.
Female	18.04	25.46	27.66	28.37	31.10	44.67
Male	14.93	23.85	28.27	27.58	31.32	39.75

Let's have a look at total number (n), mean ($\bar{y}$) and standard deviation (s)

Gender	total number (n)	mean ($\bar{y}$)	standard deviation (s)
Female	102	28.37	4.95
Male	99	27.58	5.17

Considering the sample size of our data, which is greater than 25, we can invoke the Central Limit Theorem (CLT). The CLT states that when the sample size is large enough, the sampling distribution of the sample mean, denoted as $\bar{y}$, will be approximately normally distributed, regardless of the shape of the underlying population distribution. In our case, as the sample size is greater than 25, we can assume that the distribution of the sample mean of age will be approximately normal.

Since the population standard deviation, σ, is unknown, we utilize the t-distribution for the test statistic. The t-distribution is a suitable choice when dealing with small sample sizes or when the population standard deviation is unknown. By utilizing the t-distribution, we account for the uncertainty associated with estimating the population standard deviation based on the sample.

By acknowledging the CLT and utilizing the t-distribution, we adhere to the appropriate statistical principles and ensure the validity of our analysis in comparing the average age between female and male weightlifters.

Let's do a t-test on age and gender by Female and Male. Using the t.test() method.

Code:

```
# Perform the t-test
t_testResult <- t.test(age ~ gender, data = dat, var.equal = TRUE)
```

Results:

```
##
##  Two Sample t-test
##
## data:  age by gender
## t = 1.1035, df = 199, p-value = 0.2711
## alternative hypothesis: true difference in means between group female and group male is not equal to
## 95 percent confidence interval:
##  -0.6199726  2.1956351
## sample estimates:
## mean in group female   mean in group male
##             28.37157             27.58374
```

	t	df	p-value
Value	1.1035	199	0.2711

Based on the t-test results, the analysis comparing the average age between female and male weightlifters yielded the following findings:

The t-test statistic was calculated to be 1.1035, with degrees of freedom (df) equal to 199. The corresponding p-value obtained from the test was 0.2711. This p-value represents the probability of observing a difference in average age as extreme as the one observed, assuming the null hypothesis is true.

Since the p-value (0.2711) is greater than the chosen significance level (e.g., 0.05), we do not have sufficient evidence to reject the null hypothesis. Consequently, we cannot conclude that there is a statistically significant difference in the average age between female and male weightlifters.

Cross checking the results from t-test manually:

Consider:

$\bar{y_1} = \bar{y}_{female} = 28.37$

$\bar{y_2} = \bar{y}_{male} = 27.58$

$n_1 = n_{female}$

$n_2 = n_{male}$

$s_1 = s_{female}$

$s_2 = s_{male}$

Calculations:

$s_p - >$ Standard error

$s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$ $s_p = \sqrt{\frac{(102-1)4.95^2+(99-1)5.17^2}{102+99-2}}$ $s_p = \sqrt{25.72824}$ $s_p = 5.07230$

$t = \frac{\bar{y_1}-\bar{y_2}}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ $t = \frac{28.37-27.58}{5.07230\sqrt{\frac{1}{102}+\frac{1}{99}}}$ $t = \frac{28.37-27.58}{5.07230*0.14108}$ $t = \frac{0.78783}{0.715624}$ $t = 1.100899$

$p = 0.2723$

Based on the t-test results you obtained for Research Question 1, where the t-value is 1.100899 and the p-value is 0.2723, with a significance level of 0.05, we can interpret the results as follows:

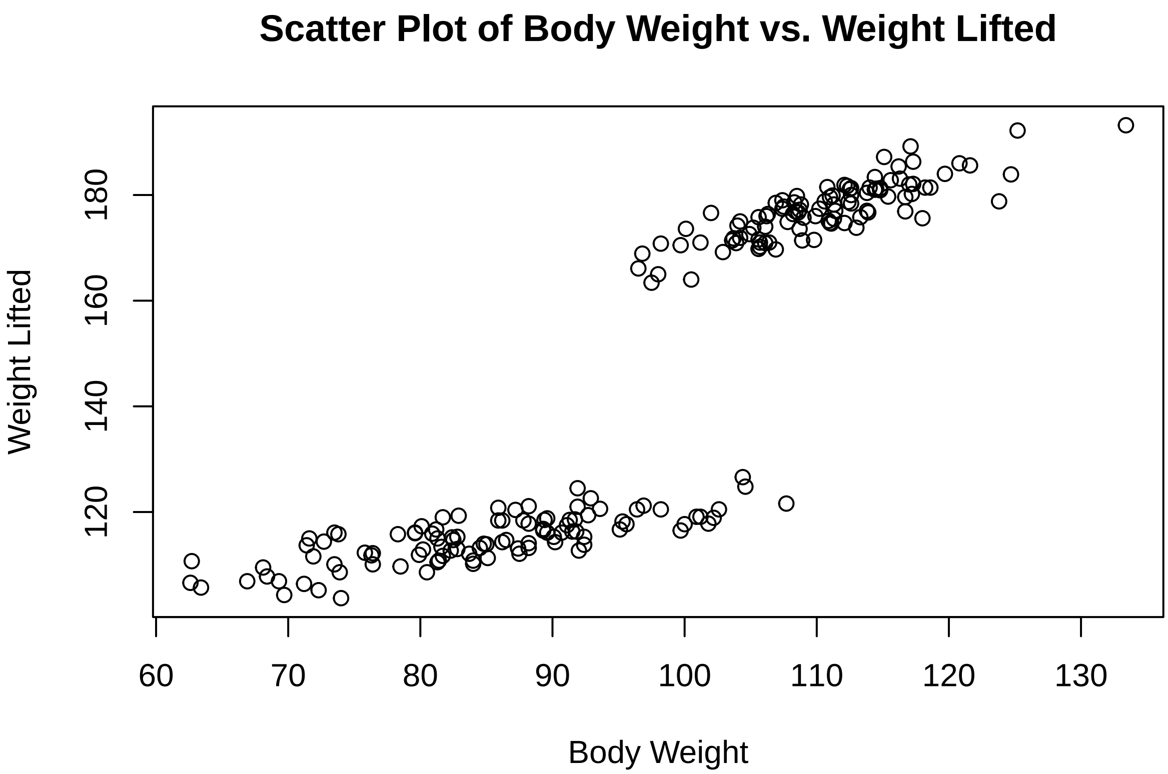
Since the p-value (0.2723) is greater than the chosen significance level of 0.05, we do not have sufficient evidence to reject the null hypothesis. Therefore, we cannot conclude that there is a statistically significant difference in the average age between female and male weightlifters at the 0.05 significance level.

The t-value of 1.100899 indicates the magnitude and direction of the difference between the average age of female and male weightlifters. Since the t-value is positive, it suggests that the average age of female weightlifters tends to be slightly higher than that of male weightlifters, although this difference is not statistically significant.

Research Question 2:

What is the relation between the body weight of weightlifters and the weight lifted?

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   103.7   115.3   124.8   145.6   176.9   193.2
```



The scatter plot reveals the presence of two distinct groups. The first group, represented by a range of body weights between 60 and 110 on the x-axis, displays a linear relationship with weight lifted, which increases from approximately 100 to 130 on the y-axis. This group demonstrates a positive correlation between body weight and weight lifted."

"The second group, characterized by body weights ranging from 95 to 135 on the x-axis, exhibits a strong linear relationship with weight lifted. Within this group, weight lifted demonstrates a significant increase from around 160 to 190 on the y-axis. The positive correlation between body weight and weight lifted in this group is more pronounced.

Since the scatter plot show to distinctive groups(Female, Male) which follow different linear direction. It is best to split the whole dataset into two i.e Female and Male.

Code:

```r
# Linear regression for females
female_model <- lm(weightlifted ~ bodyweight, data = female_data)

# Linear regression for males
male_model <- lm(weightlifted ~ bodyweight, data = male_data)
```

Results:

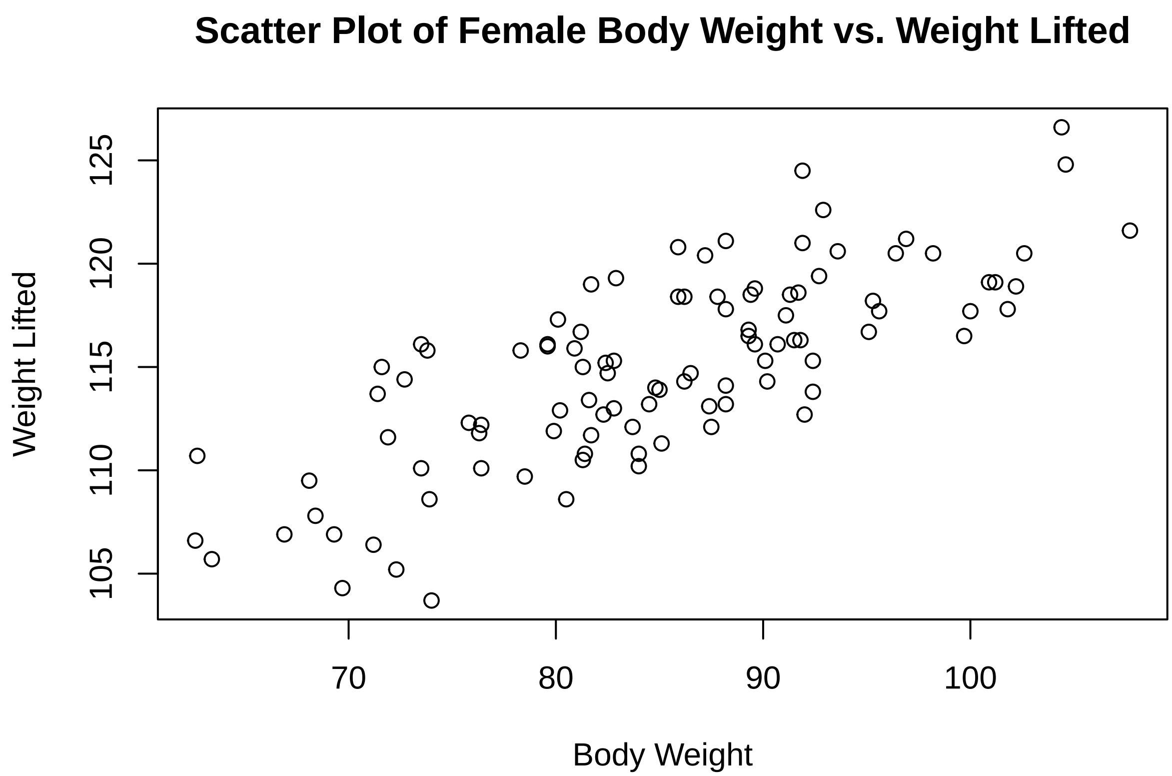
1. **Female subset:**

```
##
## Call:
## lm(formula = weightlifted ~ bodyweight, data = female_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3762 -2.0944 -0.4913  2.3196  7.0544
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 84.74469    2.54142   33.34   <2e-16 ***
## bodyweight   0.35583    0.02973   11.97   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.973 on 100 degrees of freedom
## Multiple R-squared:  0.589,  Adjusted R-squared:  0.5849
## F-statistic: 143.3 on 1 and 100 DF,  p-value: < 2.2e-16
```

2. **Male subset:**

```
##
## Call:
## lm(formula = weightlifted ~ bodyweight, data = male_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.9752 -2.2733  0.4114  2.1027  7.2565
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 97.49694    5.09263   19.14   <2e-16 ***
## bodyweight   0.72115    0.04603   15.66   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.003 on 97 degrees of freedom
## Multiple R-squared:  0.7167, Adjusted R-squared:  0.7138
## F-statistic: 245.4 on 1 and 97 DF,  p-value: < 2.2e-16
```

Female subset:



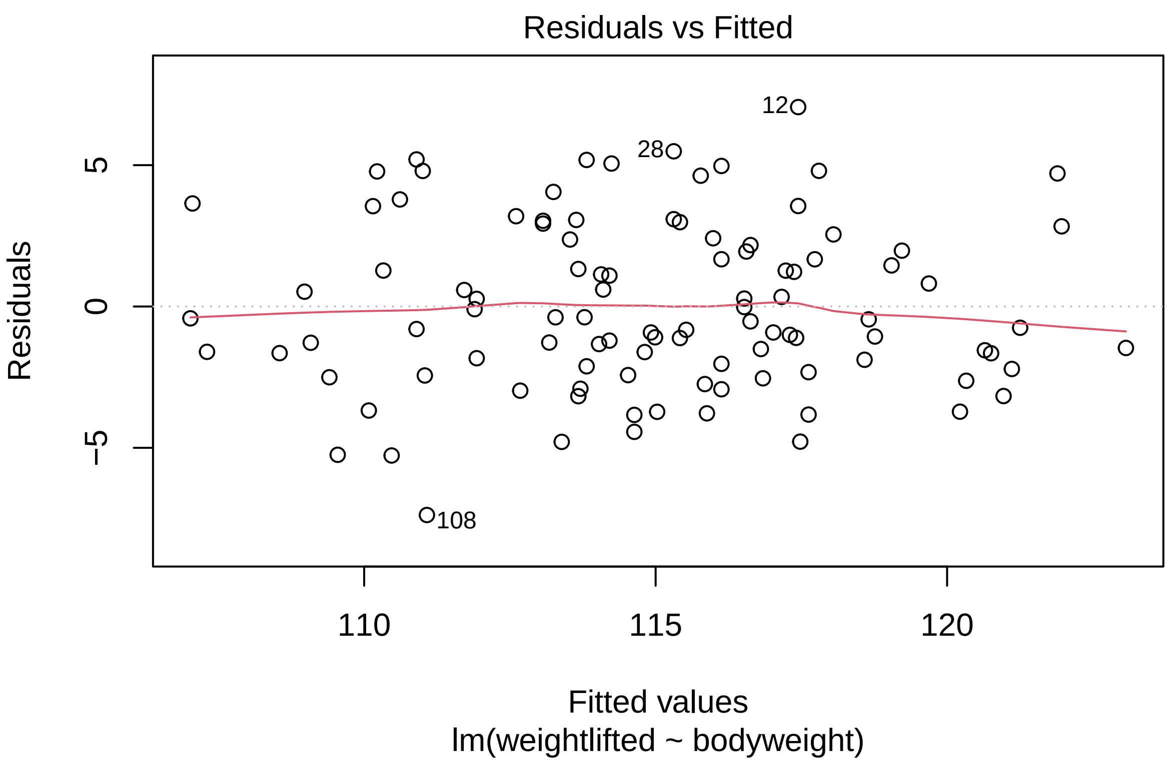
Summary Output:

Regression Statistics	
Multiple R	0.77
R Square	0.59
Adjusted R Square	0.58
Standard Error	2.94
Observations	102

ANOVA					
	df	SS	MS	F	Significance F
Regression (bodyweight)	1	1266.22	1266.22	143.29	$5.0897237 \times 10^{-21}$
Residual	100	883.68	8.84		
Total	101	2149.9			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	84.74	2.54	33.35	0	79.76	89.72
BodyWeight	0.36	0.03	11.97	0	0.3	0.42

Residual vs Fitted values:



Equation for the Least Squares Regression line for Female weight lifters:

$$\text{weight}\hat{-}\text{liffted} = 84.74 + (0.36 * BodyWeight)$$

Outline of the HATPDC framework for the regression analysis of the female subset, incorporating the information provided by Summary Output:

H - Hypothesis tests:

- Null hypothesis (H0): There is no relationship between body weight and weight lifted for female weightlifters.
- Alternative hypothesis (HA): There is a relationship between body weight and weight lifted for female weightlifters.

A - Assumptions:

- Linearity: There is a linear relationship between body weight and weight lifted for female weightlifters.
- Independence of errors: The observations of weightlifters within the female subset are independent of each other.

- Normality of residuals: The residuals of the regression model for the female subset follow a normal distribution.
- Homoscedasticity: The variance of the residuals is constant across different levels of body weight for the female subset.

T - t-value and df:

- The t-value for the body weight coefficient is 11.97.
- The degrees of freedom (df) for the t-test is 100.

P - p-value:

- The p-value for the body weight coefficient is $< 2e\text{-}16$.

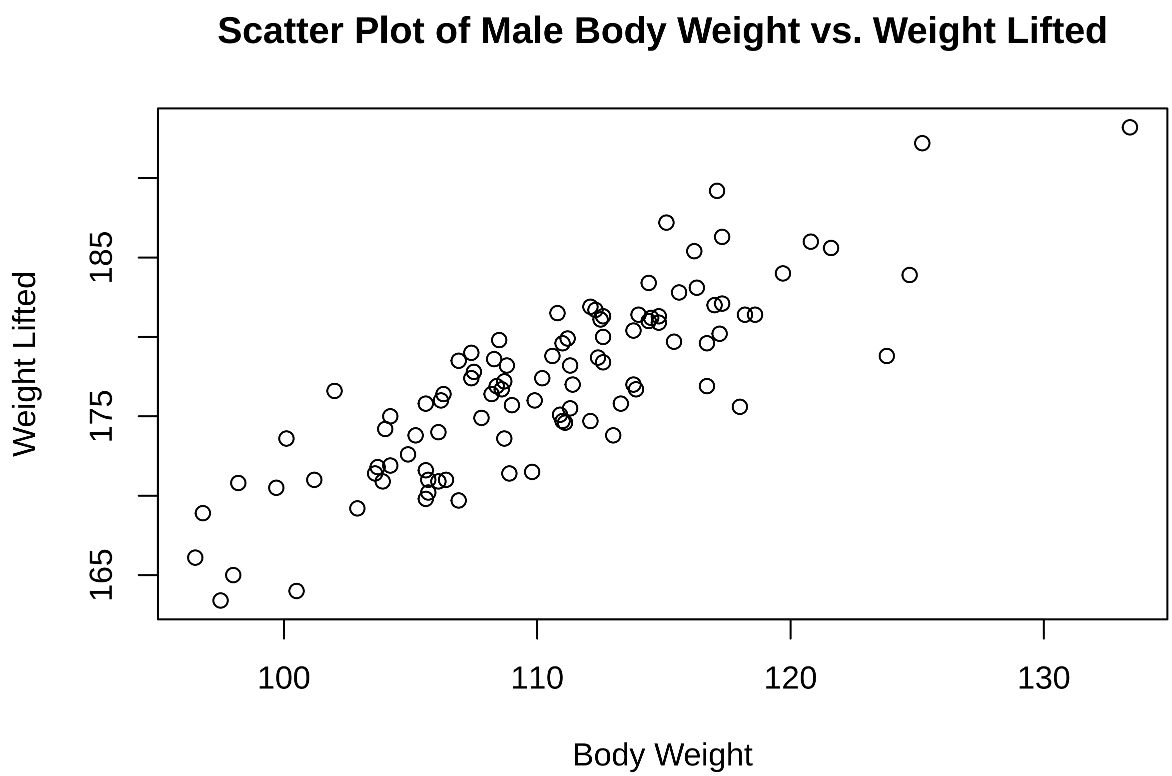
D - Decision on hypothesis:

- The p-value ($< 2e\text{-}16$) is less than the chosen significance level (e.g., 0.05). Therefore, we reject the null hypothesis.

C - Conclusion:

- The regression analysis of the female subset indicates a statistically significant relationship between body weight and weight lifted ($p < 0.05$). As body weight increases, weight lifted tends to increase for female weightlifters. The coefficient estimate of 0.35583 suggests that, on average, for every one-unit increase in body weight, weight lifted increases by approximately 0.356 units. The regression model explains about 58.9% of the variance in weight lifted among female weightlifters.

Male subset:



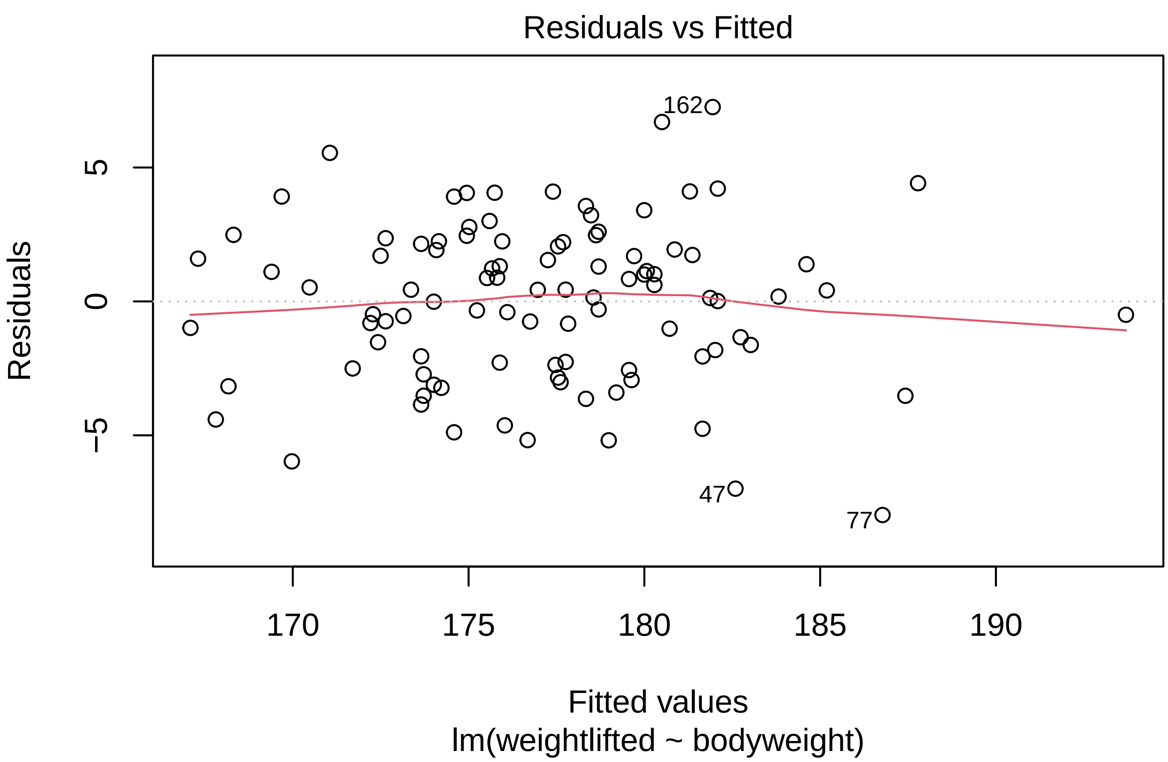
Summary Output:

Regression Statistics	
Multiple R	0.85
R Square	0.72
Adjusted R Square	0.71
Standard Error	2.97
Observations	99

ANOVA					
	df	SS	MS	F	Significance F
Regression (bodyweight)	1	2213.09	2213.09	245.4	$2.5792609 \times 10^{-28}$
Residual	97	874.77	9.02		
Total	98	3087.86			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	97.5	5.09	19.14	0	79.76	89.72
BodyWeight	0.72	0.05	15.67	0	0.3	0.42

Residual vs Fitted values:



Equation for the Least Squares Regression line for Male weight lifters:

$$\widehat{\text{weight-liffted}} = 97.5 + (0.72 * BodyWeight)$$

Outline of the HATPDC framework for the regression analysis of the male subset, incorporating the information provided by Summary Output:

H - Hypothesis tests:

- Null hypothesis (H0): There is no relationship between body weight and weight lifted for male weightlifters.
- Alternative hypothesis (HA): There is a relationship between body weight and weight lifted for male weightlifters.

A - Assumptions:

- Linearity: There is a linear relationship between body weight and weight lifted for male weightlifters.
- Independence of errors: The observations of weightlifters within the male subset are independent of each other.

- Normality of residuals: The residuals of the regression model for the male subset follow a normal distribution.
- Homoscedasticity: The variance of the residuals is constant across different levels of body weight for the male subset.

T - t-value and df:

- The t-value for the body weight coefficient is 15.66.
- The degrees of freedom (df) for the t-test is 97.

P - p-value:

- The p-value for the body weight coefficient is $< 2e\text{-}16$.

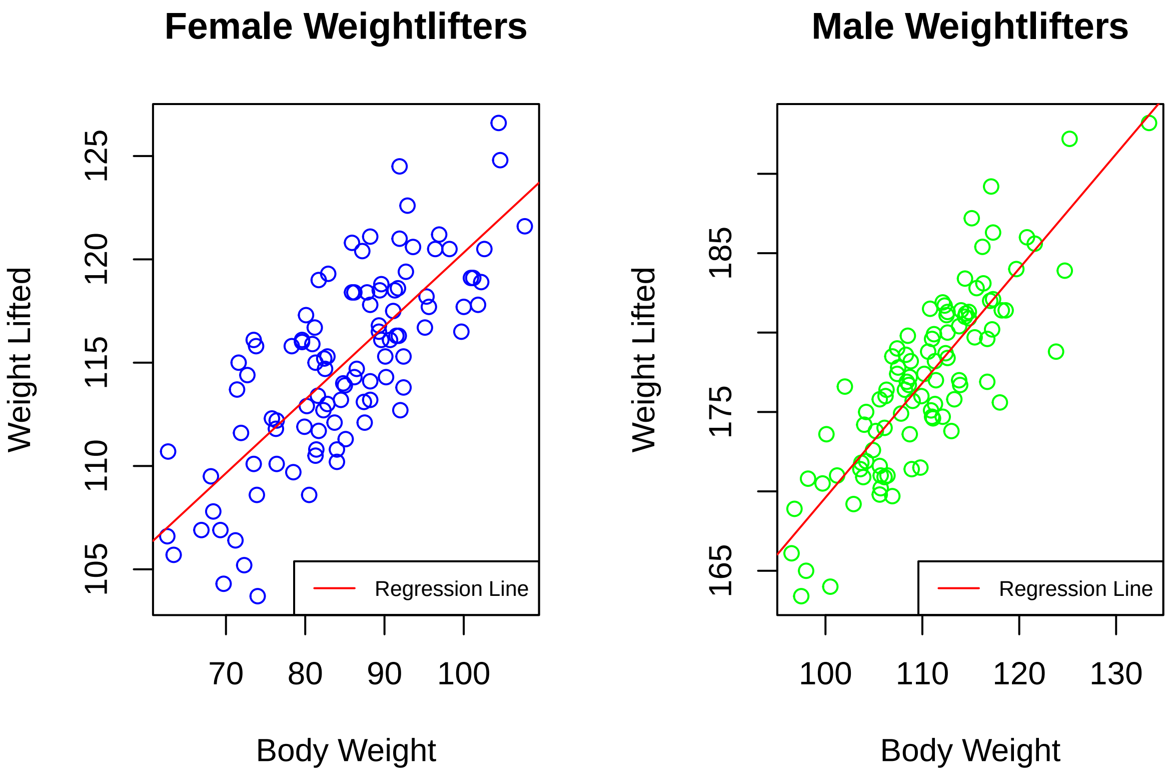
D - Decision on hypothesis:

- The p-value ($< 2e\text{-}16$) is less than the chosen significance level (e.g., 0.05). Therefore, we reject the null hypothesis.

C - Conclusion:

- The regression analysis of the male subset indicates a statistically significant relationship between body weight and weight lifted ($p < 0.05$). As body weight increases, weight lifted tends to increase for male weightlifters. The coefficient estimate of 0.72115 suggests that, on average, for every one-unit increase in body weight, weight lifted increases by approximately 0.721 units. The regression model explains about 71.7% of the variance in weight lifted among male weightlifters.

Comparing both genders:



Appendix

Appendix A: Descriptive Statistics

Table A.1: Total number of females and males

Gender	Total Number
Female	102
Male	99

Table A.2: Summary of the age of female and male weightlifters

Gender	Min.	1st Qu.	Median	Mean μ	3rd Qu.	Max.
Female	18.04	25.46	27.66	28.37	31.10	44.67
Male	14.93	23.85	28.27	27.58	31.32	39.75

Table A.3: Total number (n), mean ($\bar{y}$) and standard deviation (s)

Gender	total number (n)	mean ($\bar{y}$)	standard deviation (s)
Female	102	28.37	4.95
Male	99	27.58	5.17

Table A.4: t-test of Age and Gender

	t	df	p-value
Value	1.1035	199	0.2711

Appendix B: Linear Regression Results for Female weight lifters

Table B.1: Regression Statistics of female weight lifters

Regression Statistics	
Multiple R	0.77
R Square	0.59
Adjusted R Square	0.58
Standard Error	2.94
Observations	102

Table B.2: ANOVA table female weight lifters

ANOVA					
	df	SS	MS	F	Significance F
Regression (bodyweight)	1	1266.22	1266.22	143.29	$5.0897237 \times 10^{-21}$
Residual	100	883.68	8.84		
Total	101	2149.9			

Table B.3: Coefficient Table of female weight lifters

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	84.74	2.54	33.35	0	79.76	89.72
BodyWeight	0.36	0.03	11.97	0	0.3	0.42

Appendix C: Linear Regression Results for Male weight lifters

Table C.1: Regression Statistics of male weight lifters

Regression Statistics	
Multiple R	0.85
R Square	0.72
Adjusted R Square	0.71
Standard Error	2.97
Observations	99

Table C.2: ANOVA table male weight lifters

ANOVA					
	df	SS	MS	F	Significance F
Regression (bodyweight)	1	2213.09	2213.09	245.4	$2.5792609 \times 10^{-28}$
Residual	97	874.77	9.02		
Total	98	3087.86			

Table C.3: Coefficient Table of male weight lifters

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	97.5	5.09	19.14	0	79.76	89.72
BodyWeight	0.72	0.05	15.67	0	0.3	0.42

Appendix D: Diagnostic Plots

Figure D.1: Bar plot female vs. male

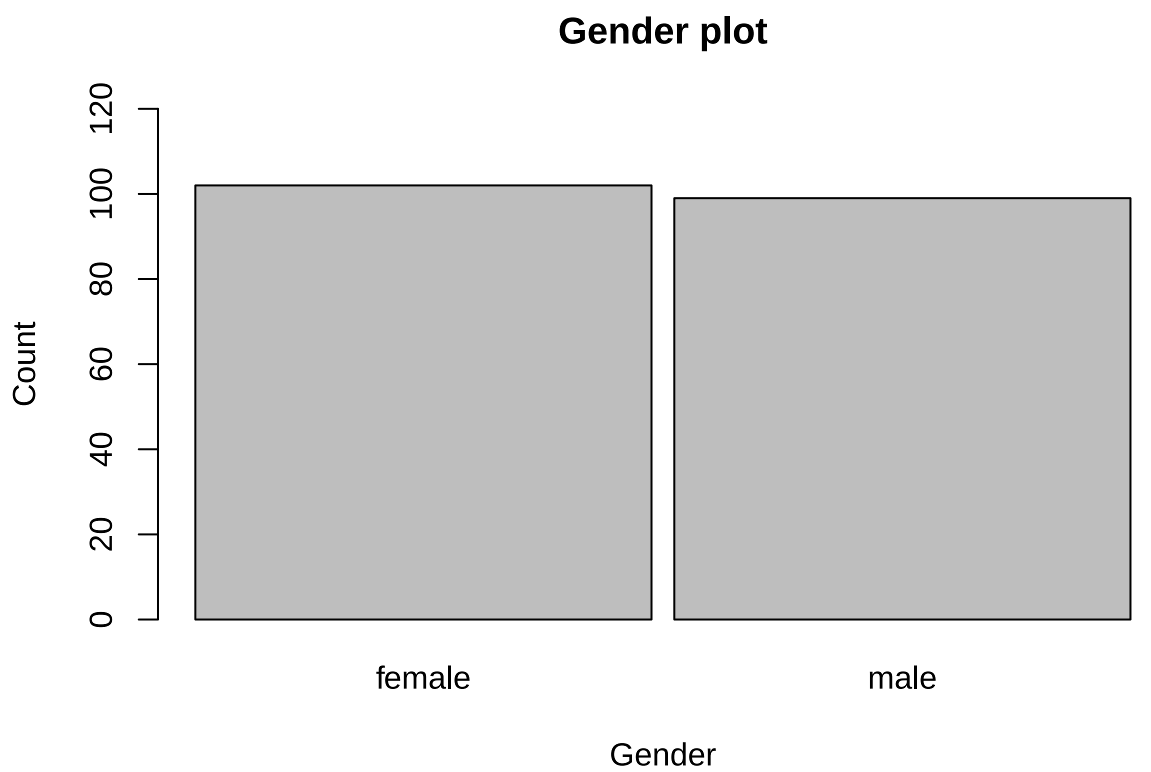


Figure D.2: Overall Histogram of Weight of population

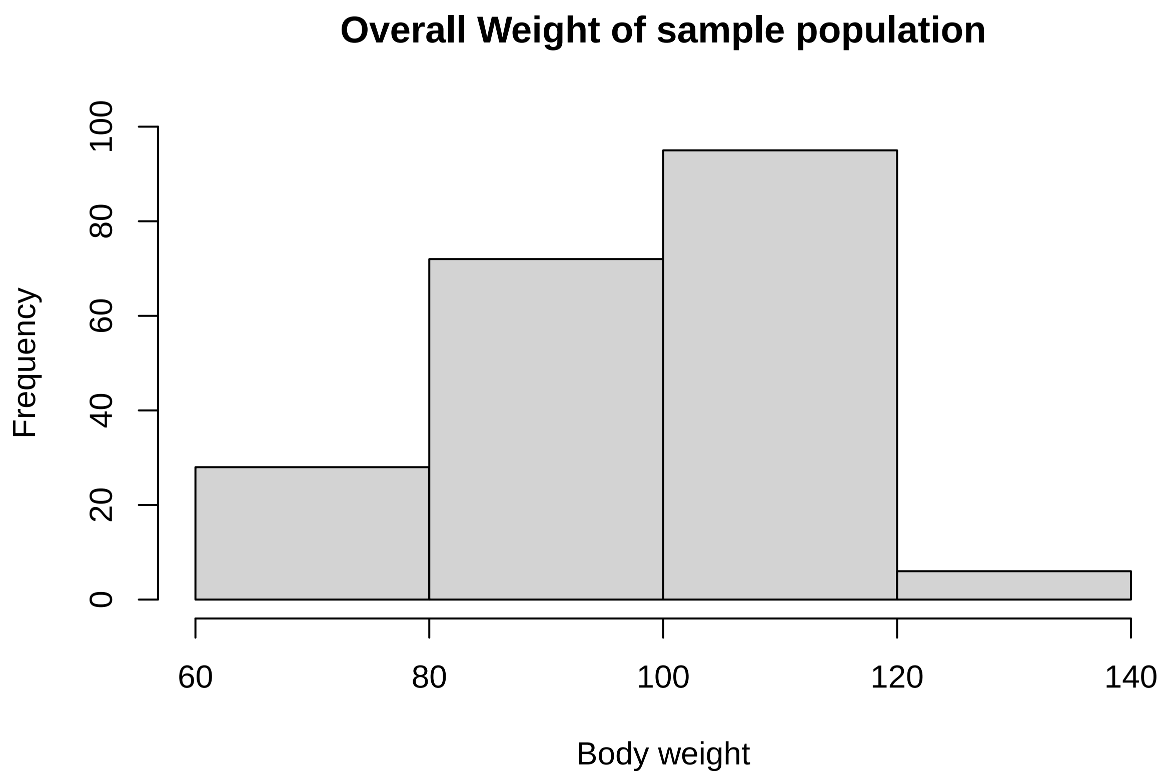


Figure D.3: Boxplot of Weight Lifted by Gender

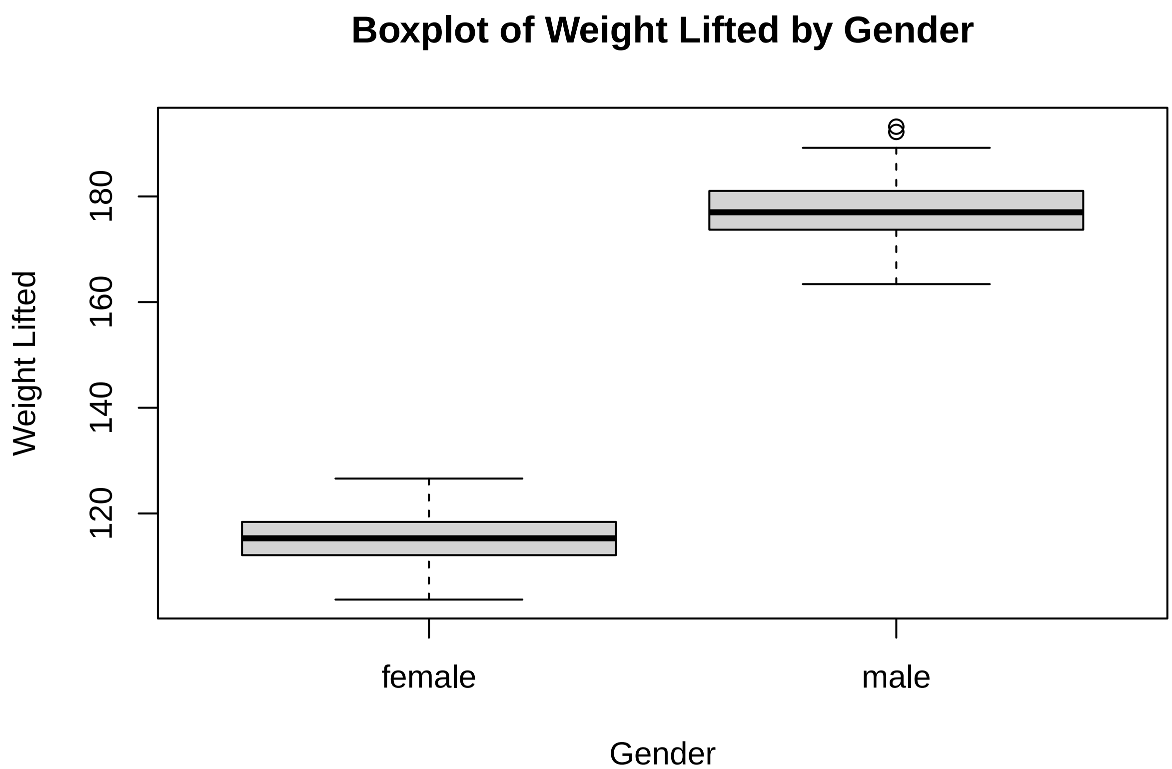


Figure D.4: Boxplot Age vs. Weight Lifted

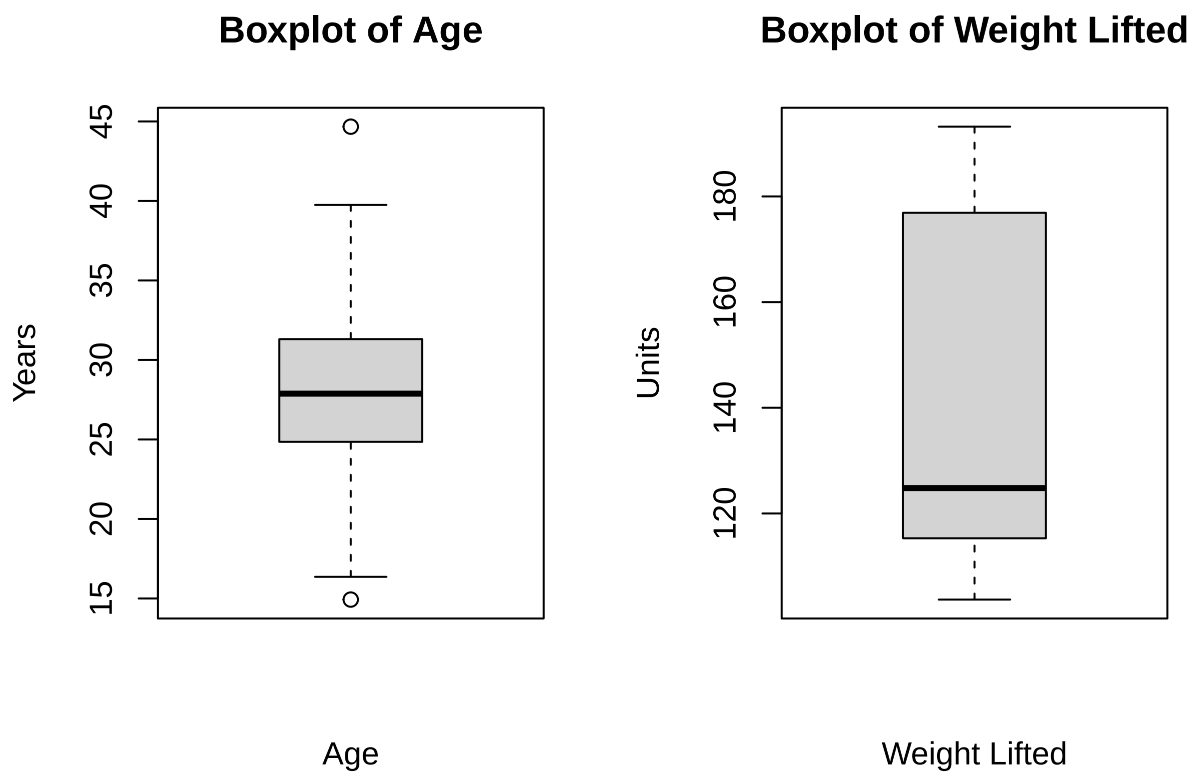


Figure D.5: Scatter Plot of Body Weight vs. Weight Lifted

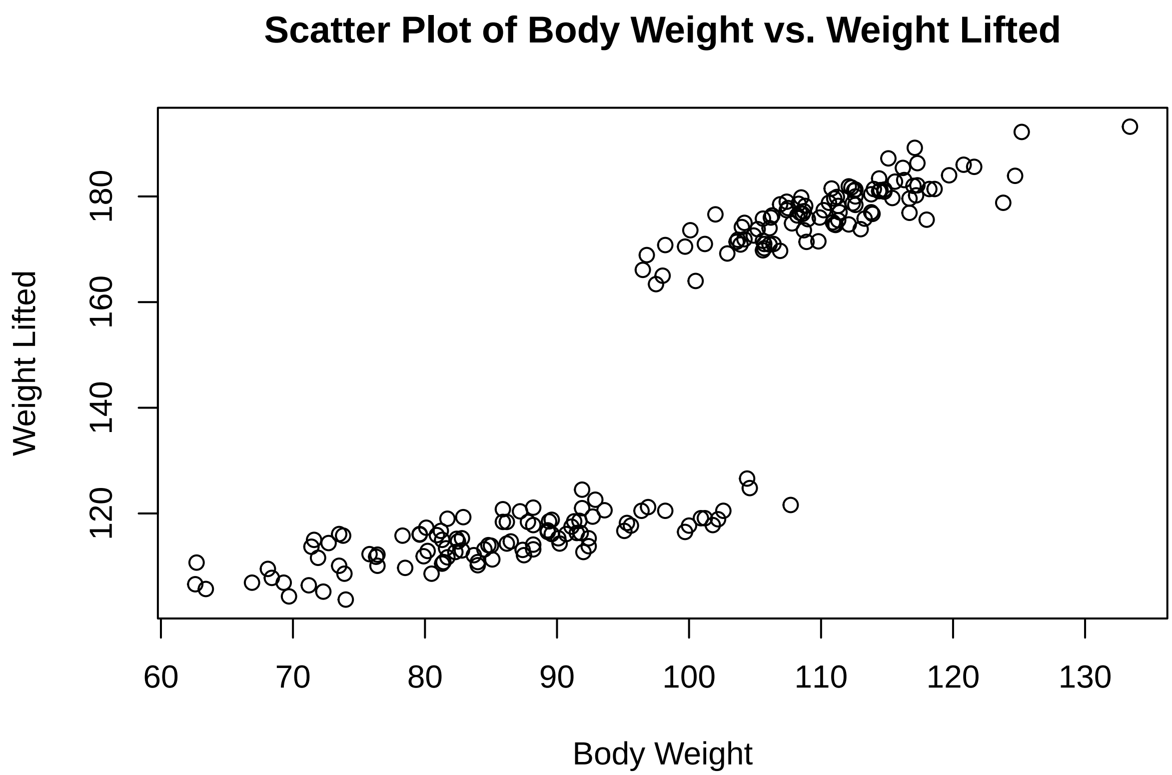


Figure D.6: Scatter Plot of Female Body Weight vs. Weight Lifted

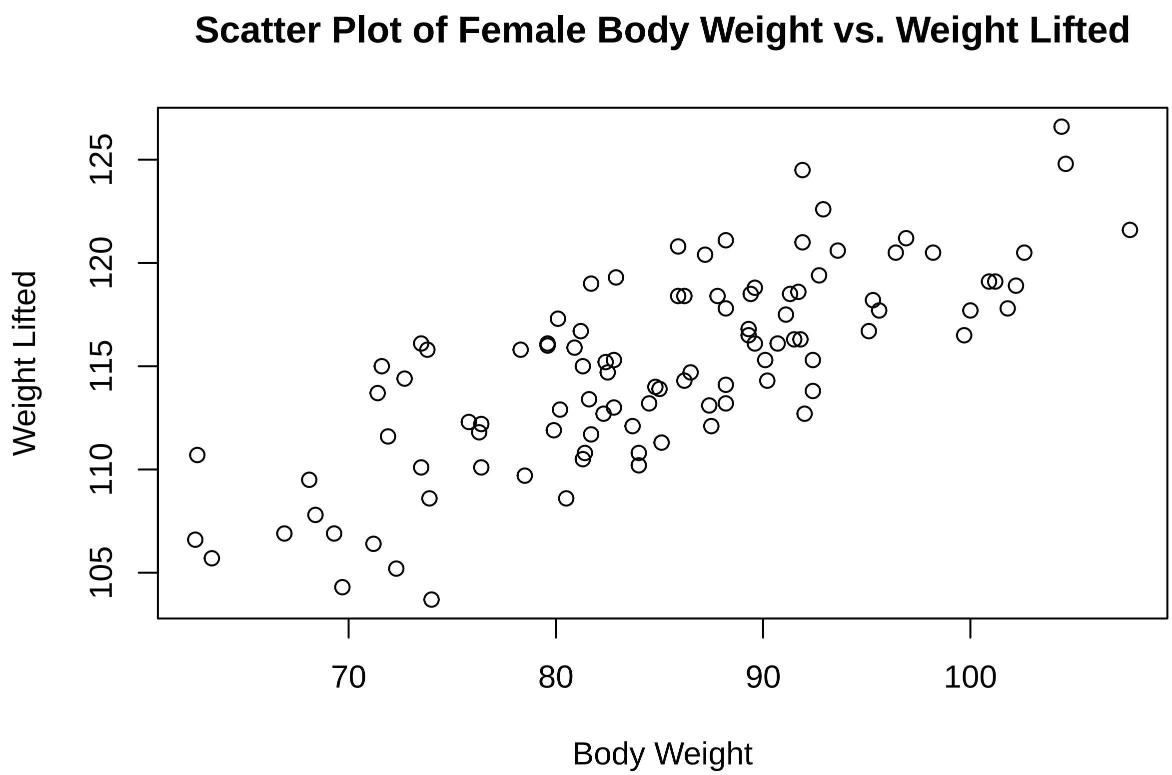


Figure D.7: Scatter plot of Residuals vs. Fitted Values (Female Subset)

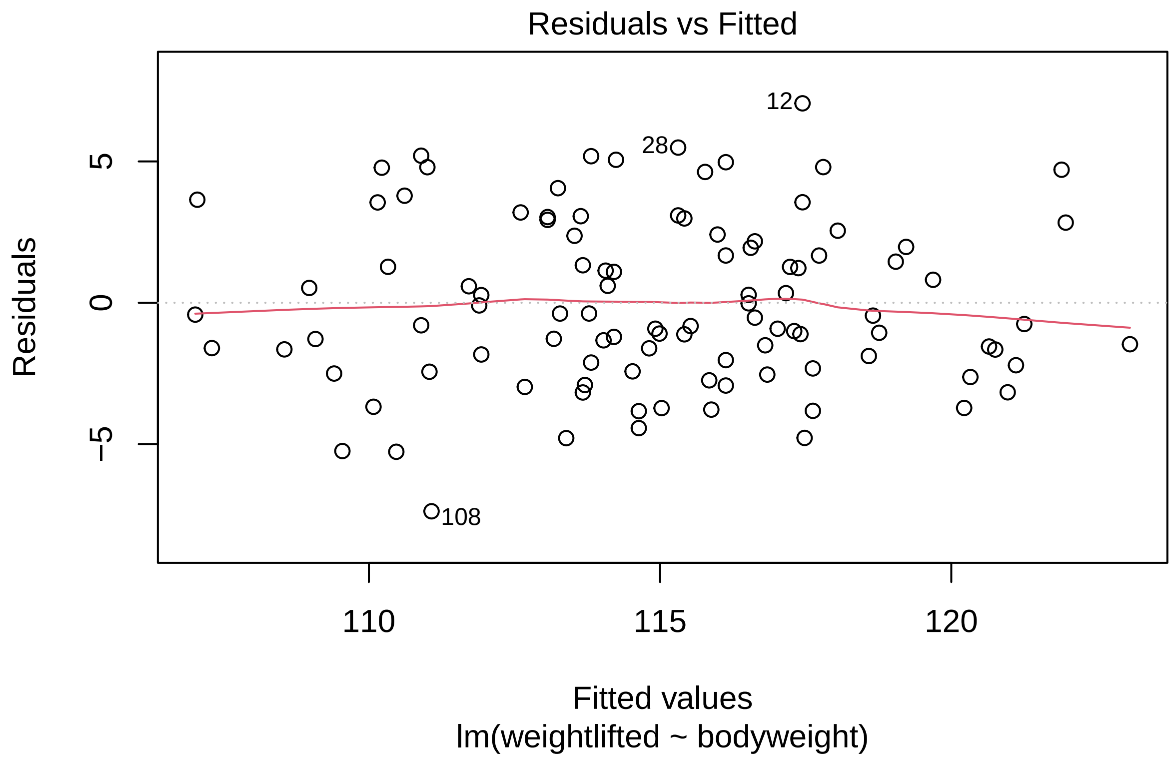


Figure D.8: Scatter Plot of Male Body Weight vs. Weight Lifted

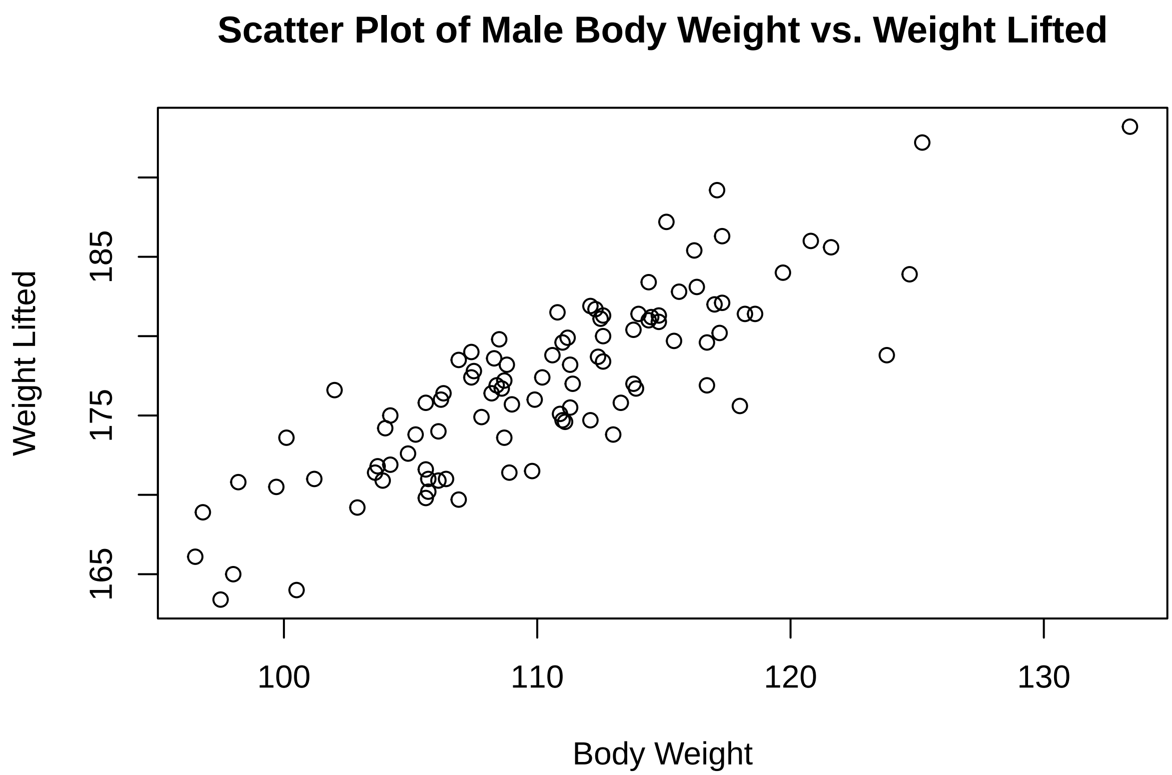


Figure D.9: Scatter plot of Residuals vs. Fitted Values (Male Subset)

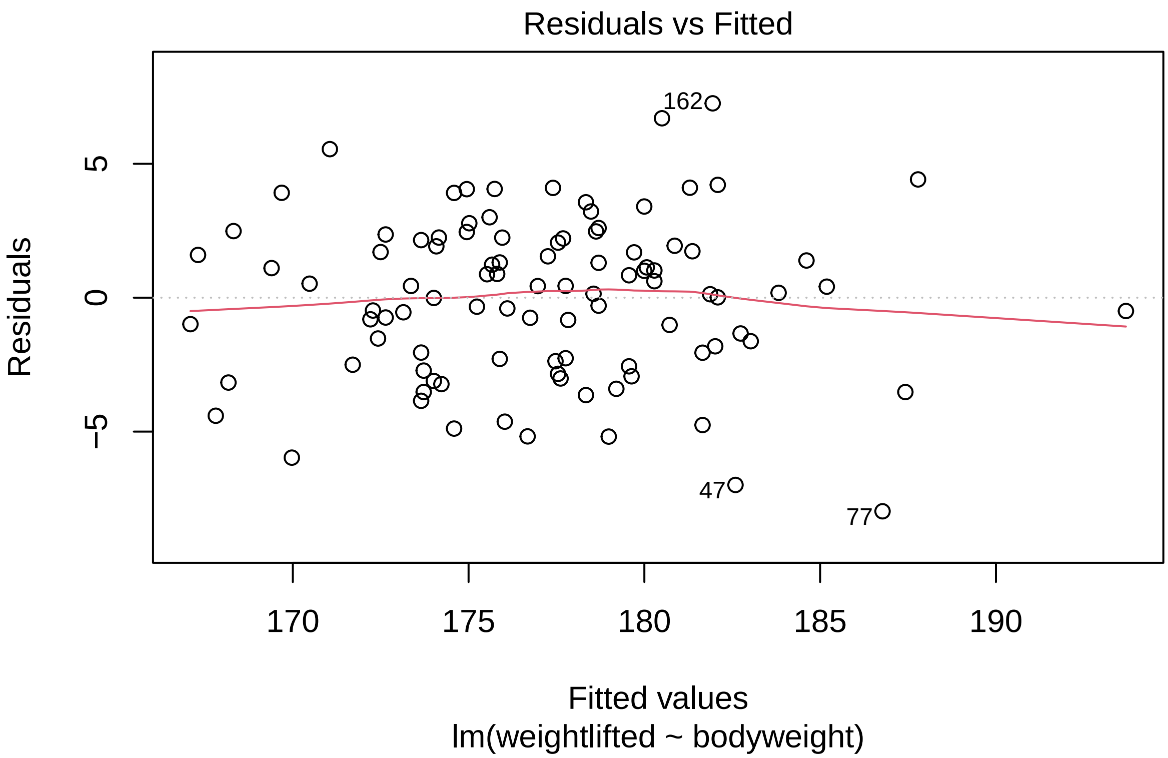
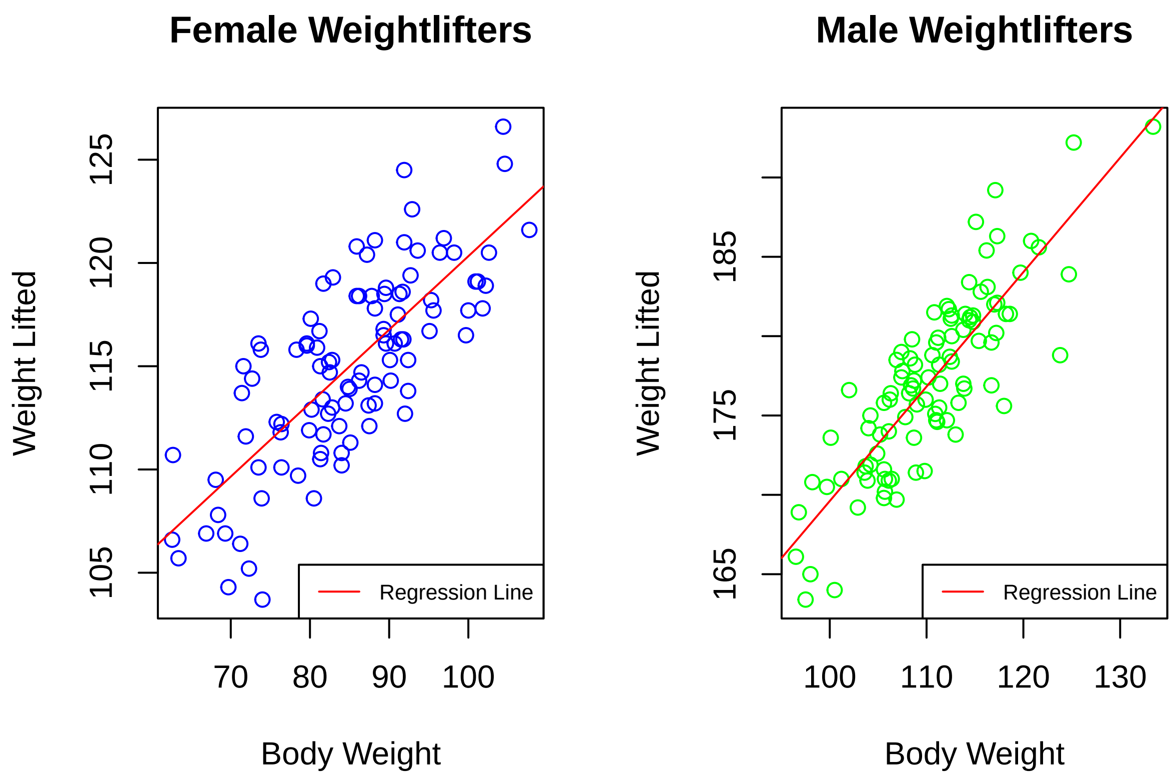


Figure D.10: Scatter plot fir regression line of Female vs. Male Weightlifters



Appendix E: R Code Chunkss

Code Chunk D.1: Data Import and Subset

```
dat <- read.csv("./data/dataset.csv", header = TRUE)

# Subset the dataset for females
female_data <- subset(dat, gender == "female")

# Subset the dataset for males
male_data <- subset(dat, gender == "male")
```

Code Chunk D.2: Linear Regression for Females

```
# Linear regression for females
female_model <- lm(weightlifted ~ bodyweight, data = female_data)
summary(female_model)
```

Code Chunk D.3: Linear Regression for Males

```
# Linear regression for males
male_model <- lm(weightlifted ~ bodyweight, data = male_data)
summary(male_model)
```

Code Chunk D.4: Diagnostic Plots for Females

```
# Residuals vs. Fitted Values
plot(female_model, which = 1)

# Normal Q-Q Plot
plot(female_model, which = 2)

# Scale-Location Plot
plot(female_model, which = 3)

# Residuals vs. Leverage
plot(female_model, which = 5)
```

Code Chunk D.5: Diagnostic Plots for Males

```
# Residuals vs. Fitted Values
plot(male_model, which = 1)

# Normal Q-Q Plot
plot(male_model, which = 2)

# Scale-Location Plot
plot(male_model, which = 3)

# Residuals vs. Leverage
plot(male_model, which = 5)
```