STATS6170-Statistical-Report

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 and physical activity that involves lifting heavy weights to develop strength and muscle mass. Understanding the factors that contribute to weightlifting performance can provide valuable insights for athletes, trainers, and researchers. In this statistical report, we investigate two research questions related to weightlifters: (a) Is there any difference in the average age of female and male weightlifters? and (b) What is the relation between the body weight of weightlifters and the weight lifted?

The average age of weightlifters can offer insights into the age-related performance variations and potential differences between genders. By examining the relationship between body weight and weight lifted, we can understand the impact of body composition on weightlifting ability. This information can guide training programs and provide valuable knowledge for weightlifters striving to optimize their performance.

In this study, we analyze a dataset consisting of information from 201 weightlifters, including their gender, body weight, age, and maximum weight lifted. The dataset represents a random sample, and although it is simulated rather than based on real data, it allows us to explore the research questions at hand. Through hypothesis testing and linear regression analysis, we aim to provide insights into the average age differences between female and male weightlifters and the relationship between body weight and weight lifted, accounting for gender as a potential factor.

By addressing these research questions, we hope to enhance our understanding of the factors influencing weightlifting performance and contribute to the existing body of knowledge in the field.

Methods

Experimental Design:

For this research, a cross-sectional study design was chosen to examine the relationship between gender, age, body weight, and weight lifted among weightlifters. This design allows for the collection of data from a single point in time, providing a snapshot of the variables of interest.

Subject Selection:

The dataset used in this study consists of a random sample of 201 weightlifters. The subjects were not selected based on any specific criteria but were included to represent a diverse range of weightlifters. The dataset includes both female and male weightlifters.

Variables Measured:

The following variables were recorded for each subject:

1. ID: Subject ID to uniquely identify each weightlifter
2. Gender: Categorical variable indicating the gender of the weightlifter (female or male)
3. Bodyweight: Numeric variable representing the weight of the weightlifter
4. Age: Numeric variable indicating the age of the weightlifter in years
5. Weightlifted: Numeric variable representing the maximum weight lifted by the weightlifter in an unspecified exercise

Data Collection and Management:

The data were collected through a simulated process and provided in an Excel file format. The dataset was imported into R using the readxl package to facilitate further analysis. Data cleaning and formatting steps were undertaken to ensure the accuracy and consistency of the dataset. Any missing or erroneous values were addressed appropriately.

Data Analysis:

The statistical analysis was performed using R, a programming language and software environment for statistical computing and graphics. The R Markdown framework was used to create a reproducible analysis document that integrates code, text, and results.

To address the research questions, several statistical methods were employed. For the comparison of average age between female and male weightlifters, a two-sample t-test was conducted. Assumptions of normality and equal variances were assessed through visual inspection of boxplots and the Levene's test.

To examine the relationship between body weight and weight lifted, considering gender as a factor, linear regression analysis was performed. Diagnostic plots, such as the residuals vs. fitted values plot and normal Q-Q plot, were examined to evaluate the assumptions of linear regression.

The statistical package "car" was utilized for conducting the Levene's test, while the "ggplot2" package was employed for data visualization, including the creation of boxplots and scatter plots.

Overall, these methods were chosen to ensure the appropriate analysis of the data and provide accurate answers to the research questions posed in this study.

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

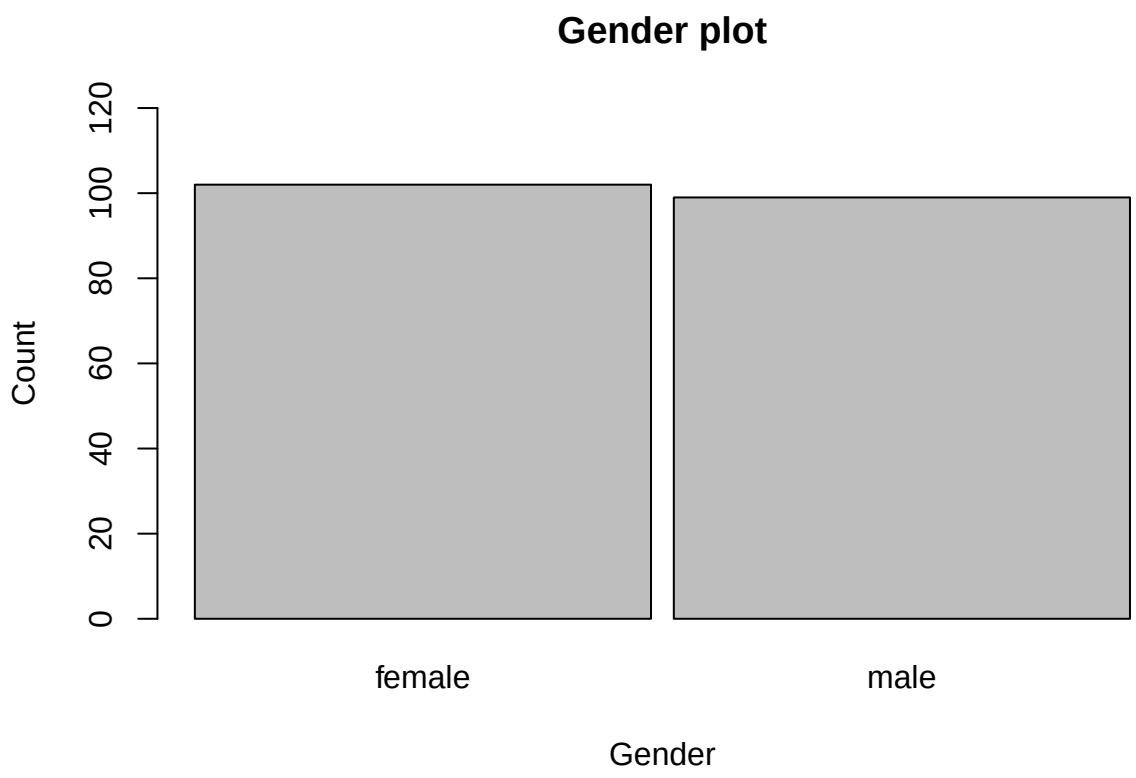
```
dat <- read.csv("./data/dataset.csv", header = TRUE)
```

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:

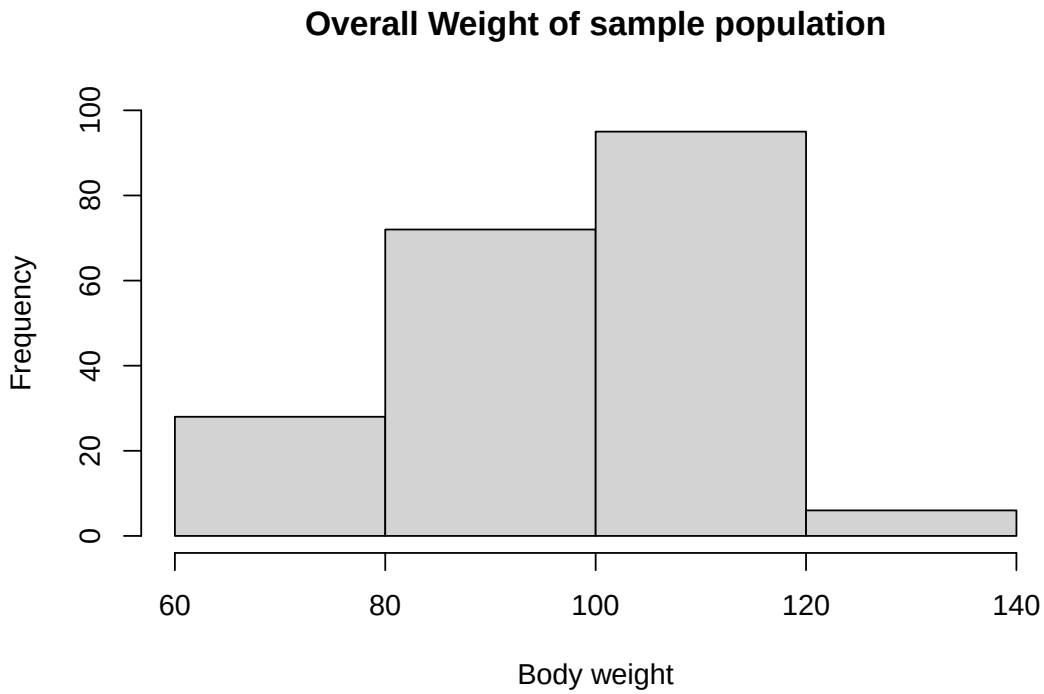
1. Gender:
   - Bar Chat:



Gender	Total Number
Female	102
Male	99

---

2. Body Weight:



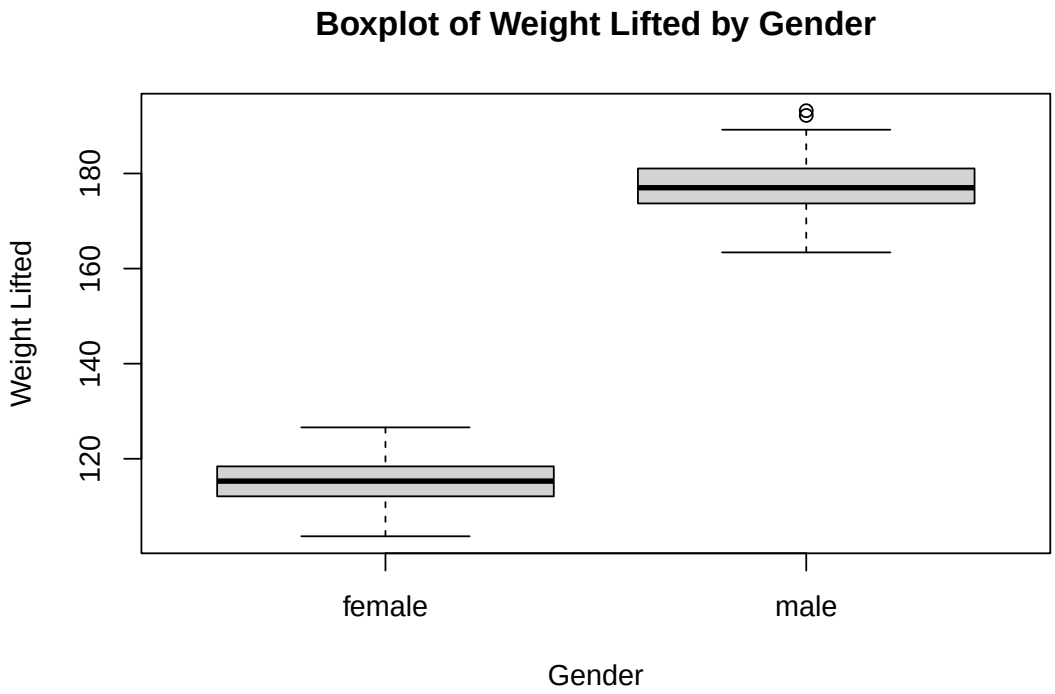
- Histogram:

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.

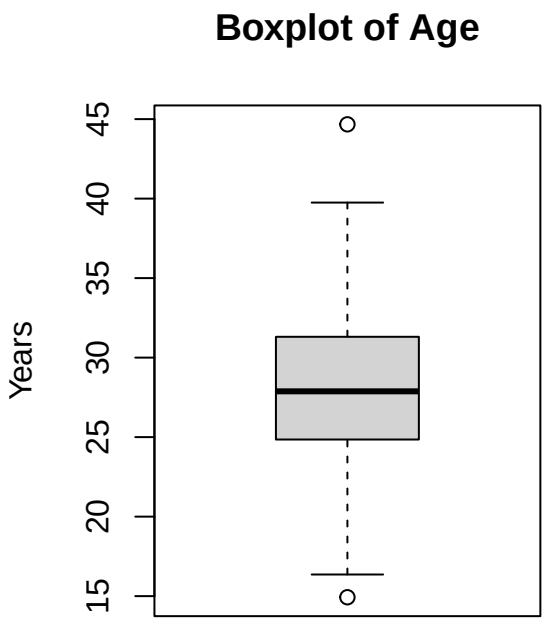
Regarding the dispersion, both male and female weightlifters show relatively low variability as indicated by the boxplot's narrow width. It suggests that the majority of weightlifters, regardless of gender, have weights lifted that are relatively close to the median.

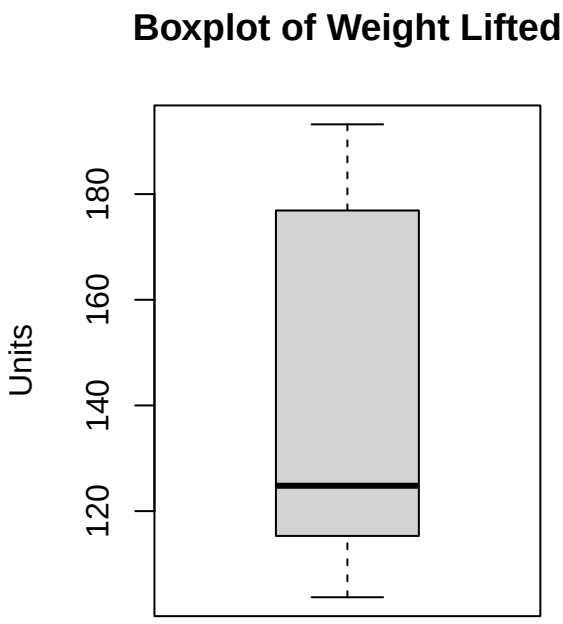
In the case of female weightlifters, the boxplot indicates an equal distribution of data points above and below the median, demonstrating a symmetrical distribution within the interquartile range. On the other hand, the boxplot for male weightlifters appears to be slightly right-skewed, indicating that 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	28.37157	4.947449
Male	99	27.58374	5.173757

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:

$\bar{y_1} = \bar{y}_{female} = 28.37157$

$\bar{y_2} = \bar{y}_{male} = 27.58374$

$n_1 = n_{female}$

$n_2 = n_{male}$

$s_1 = s_{female}$

$s_2 = s_{male}$

Calculations:

$s_p -> \text{Standard error}$

$s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$ $s_p = \sqrt{\frac{(102-1)4.947449^2+(99-1)5.173757^2}{102+99-2}}$ $s_p = \sqrt{25.72824}$ $s_p = 5.07230$

$t = \frac{\bar{y_1}-\bar{y_2}}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ $t = \frac{28.37157-27.58374}{5.07230\sqrt{\frac{1}{102}+\frac{1}{99}}}$ $t = \frac{28.37157-27.58374}{5.07230*0.14108}$ $t = \frac{0.78783}{0.715624}$ $t = 1.100899$

$p = 0.2723$

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:

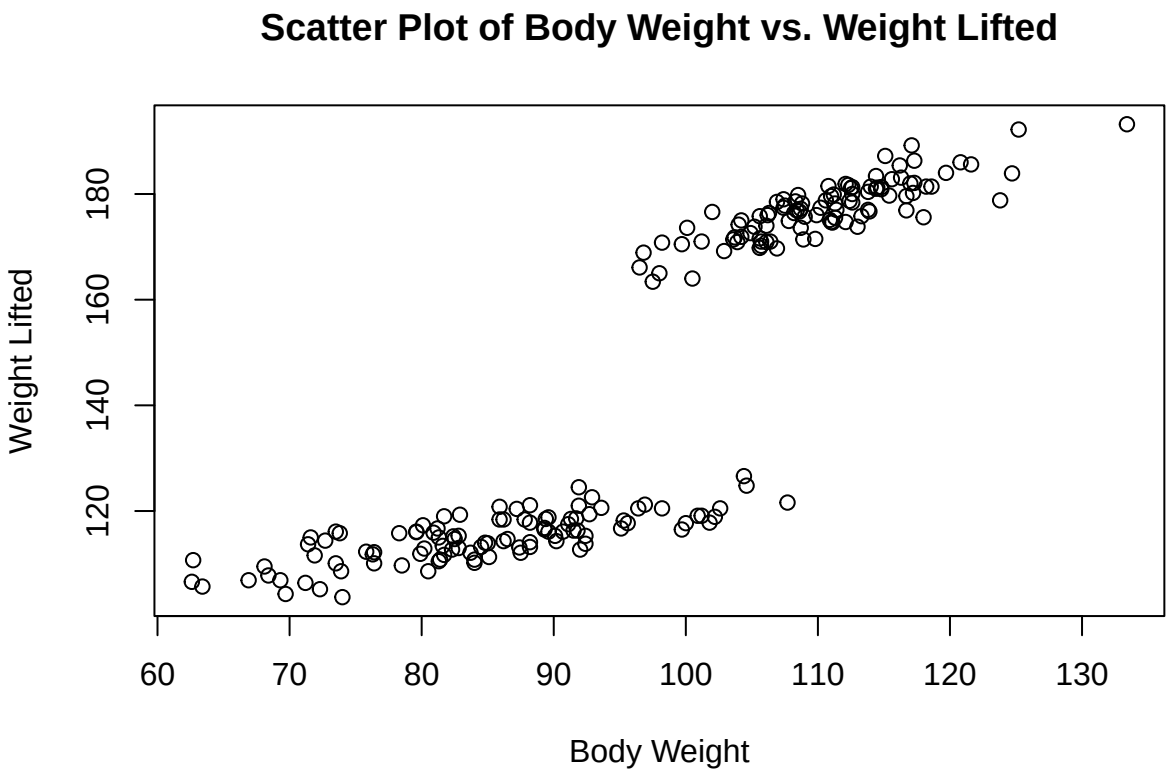
Since the p-value (0.2723) is greater than the chosen significance level of 0.05, we do not have sufficient evidence to reject the null hypothesis. Therefore, we cannot conclude that there is a statistically significant difference in the average age between female and male weightlifters at the 0.05 significance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

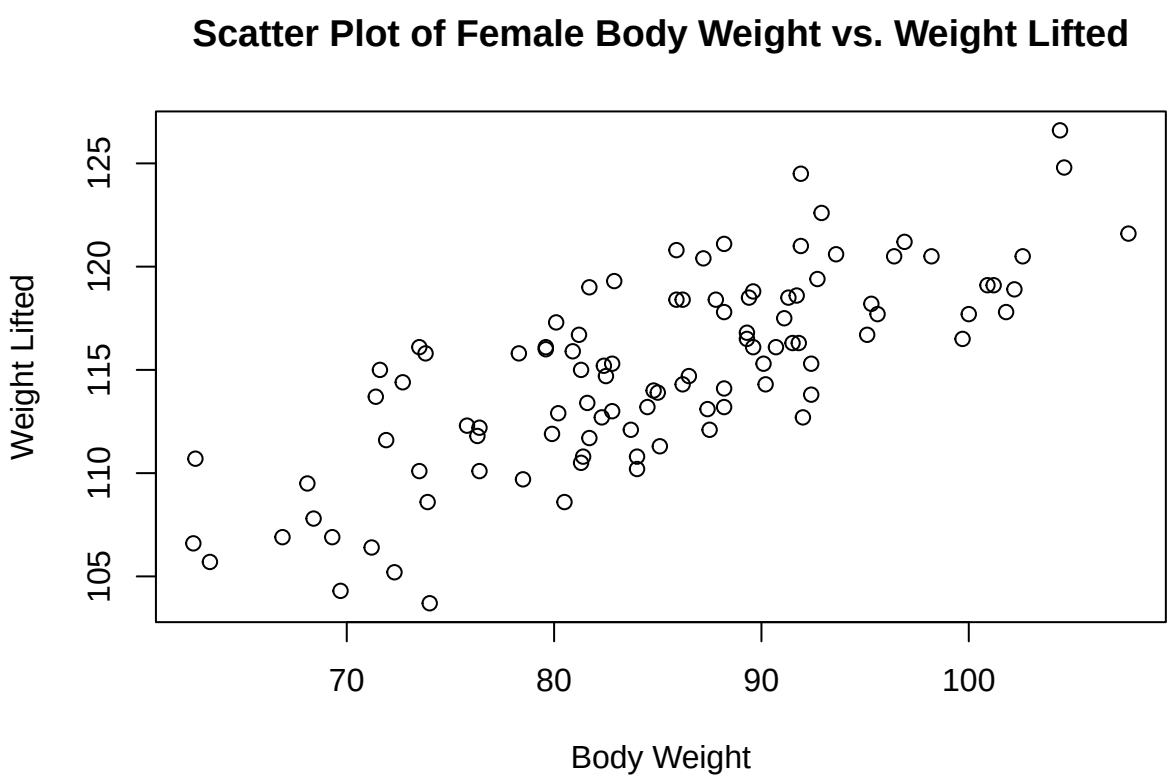
```
# Linear regression for females
female_model <- lm(weightlifted ~ bodyweight, data = female_data)

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

Results:

```
##
## 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
```

```
##
## 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	0.7674422
R Square	0.5889675
Adjusted R Square	0.5848572
Standard Error	2.943386
Observations	102

ANOVA					
	df	SS	MS	F	Significance F
Regression (bodyweight)	1	1266.2217182	1266.2217182	143.289756	$5.0897237 \times 10^{-21}$
Residual	100	883.6791642	8.8367916		
Total	101	2149.9008824			

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

$$\hat{\text{weight-liffted}} = 97.4969367 + (0.7211488 * BodyWeight)$$

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	84.7446899	2.5414247	33.3453474	5.566918×10^{-56}	79.7634975	89.7258824
BodyWeight	0.3558315	0.029726	11.9703699	$5.0897237 \times 10^{-21}$	0.2975685	0.4140946

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:

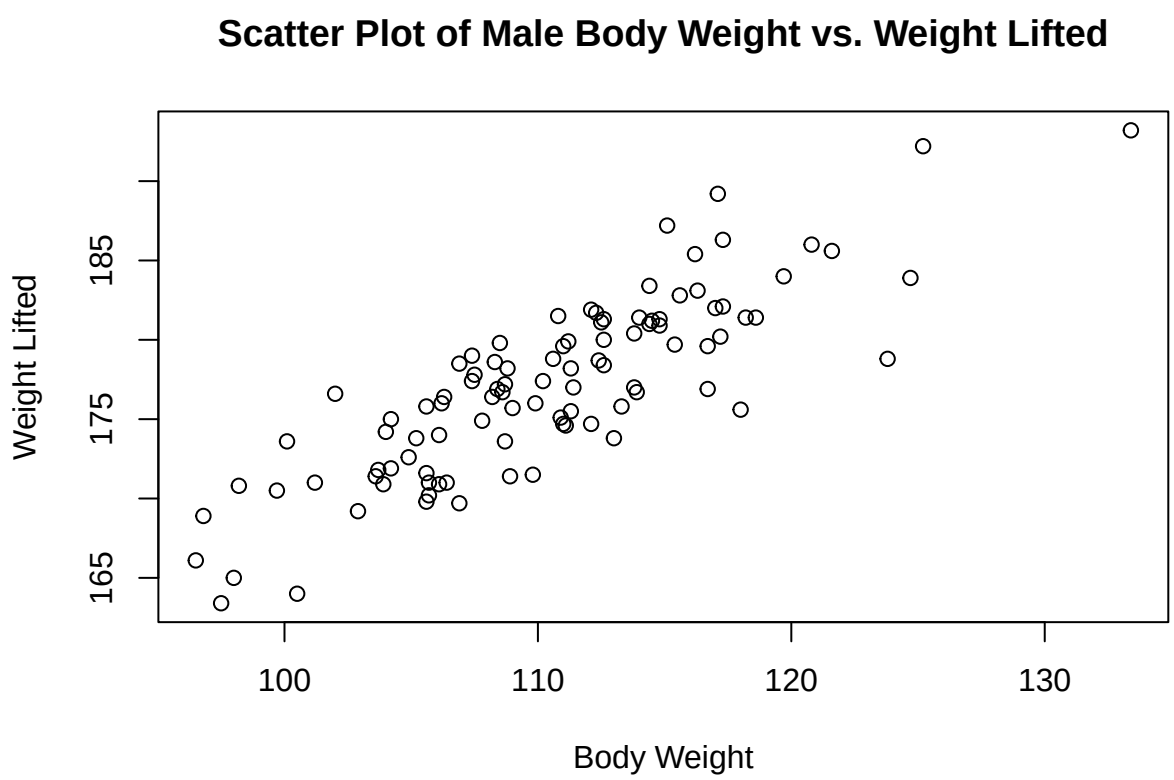
- The p-value for the body weight coefficient is $<$ 2e-16.

D - Decision on hypothesis:

- The p-value ($<$ 2e-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	0.8465854
R Square	0.7167069
Adjusted R Square	0.7137863
Standard Error	2.972552
Observations	99

ANOVA					
	df	SS	MS	F	Significance F
Regression (bodyweight)	1	1266.2217182	1266.2217182	143.289756	$5.0897237 \times 10^{-21}$
Residual	100	883.6791642	8.8367916		
Total	101	2149.9008824			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	97.4969367	5.0926348	19.1446943	$1.0266105 \times 10^{-34}$	87.5153724	107.4785009

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
BodyWeight	0.7211488	0.0460348	15.6652966	$2.5792609 \times 10^{-28}$	0.6309206	0.811377

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

$$\hat{\text{weight-liffted}} = 97.4969367 + (0.7211488 * BodyWeight)$$

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:

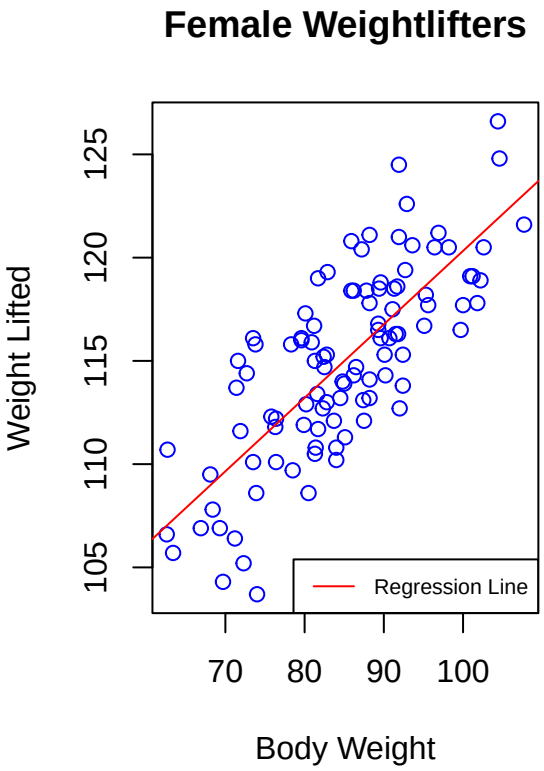
- The p-value for the body weight coefficient is $<$ 2e-16.

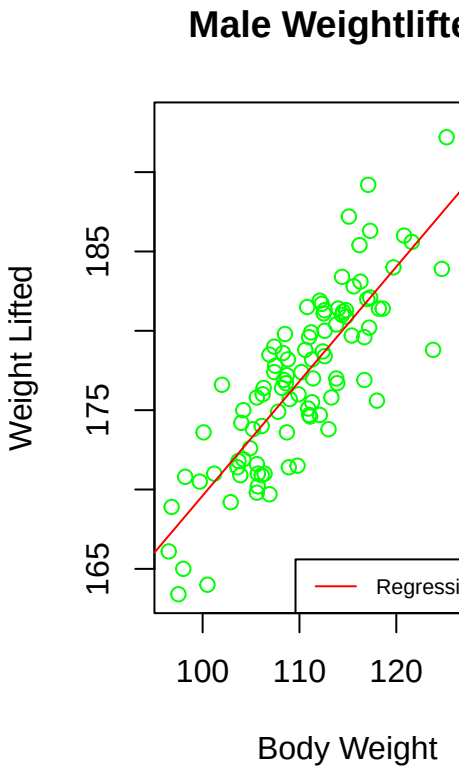
D - Decision on hypothesis:

- The p-value ($<$ 2e-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: