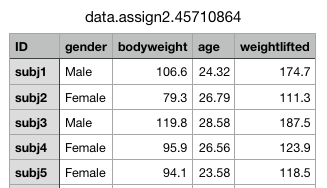
Statistical Project Report

Introduction

The dataset of this assignment has been collected from some physical and performance data that contains random sample of 289 weightlifters.



These are the first 5 records of our dataset.

The variables recorded for each subject are listed below:

**Variable Description**

Id Subject ID   
gender Either "Male" or "Female"

Bodyweight the weight of the subject

age age of subject in years

weightlifted The maximum weight lifted by the subject in a specific exercise [not specified]

From this data, I got information about the physical performance of 289 weightlifters. The data illustrates the Id, body weight, age, and weight lifted by either a female or male.

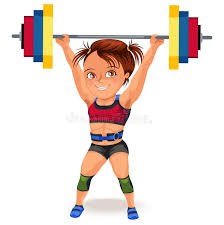
(a)

Research Question: Is there any difference in the average ages of female and male weightlifters?

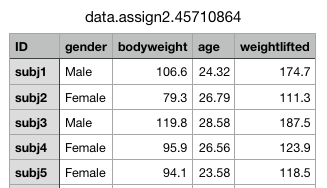
Here, I considered that the given research question is concerned about the difference between the average ages for two independent populations i.e. female and male.

We could not use paired 𝑡-tests to address this question as there is no meaningful way to match up subjects in the this sample. Given Research Question compares the average age of female and male weightlifters. This question could be answered using two-sample 𝑡-tests (assuming any test assumptions are satisfied).

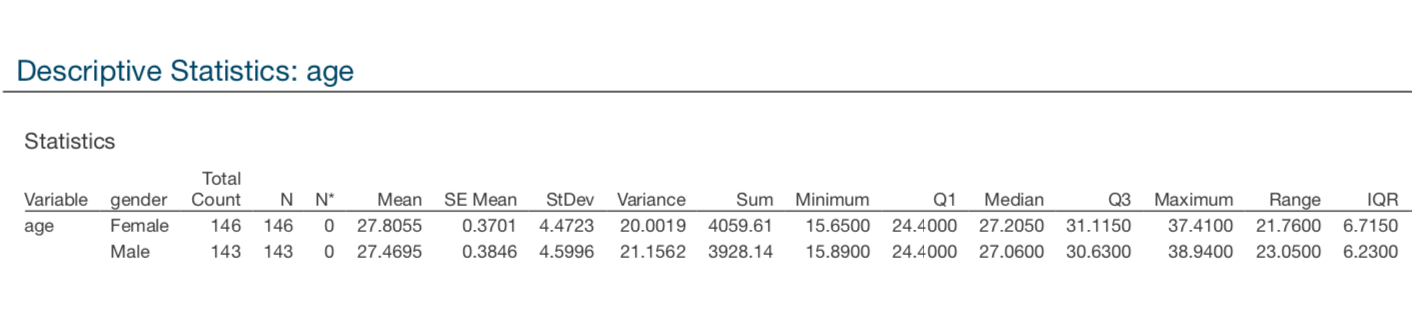
Male weightlifters Vs Female weightlifters



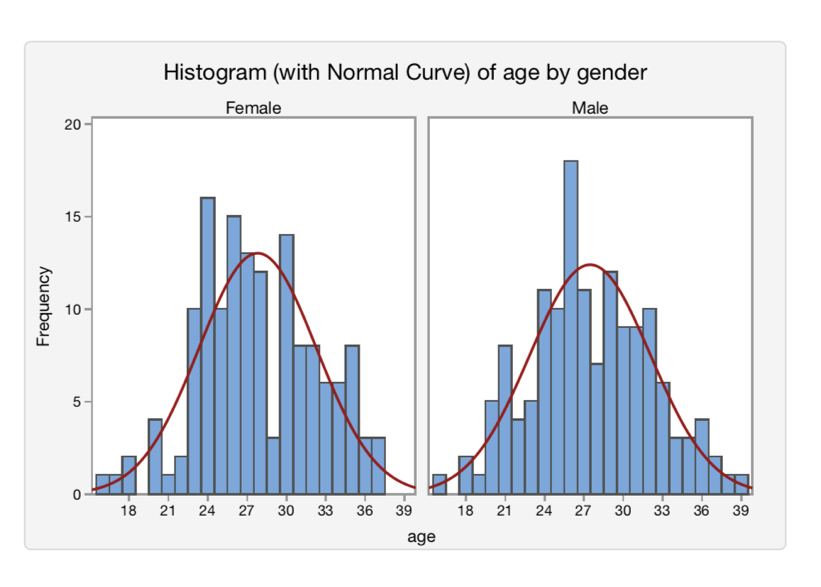
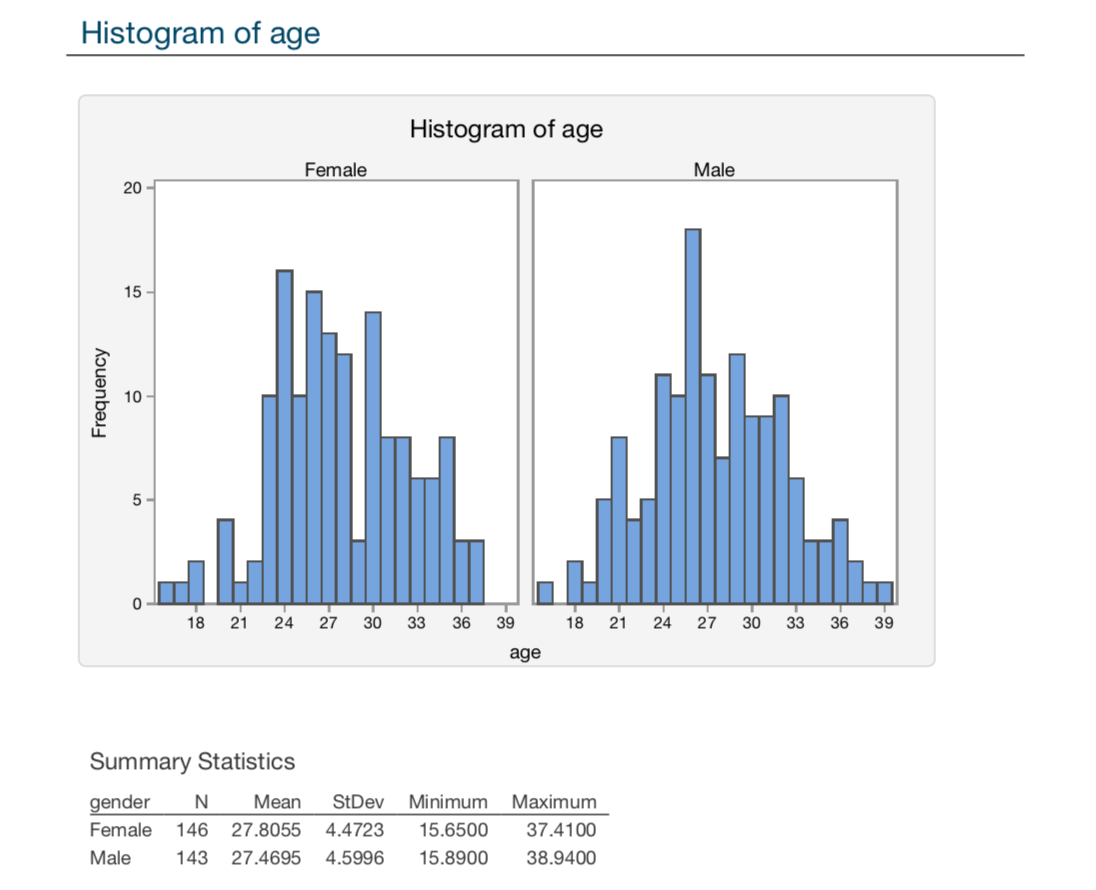
Some Information about the Sample



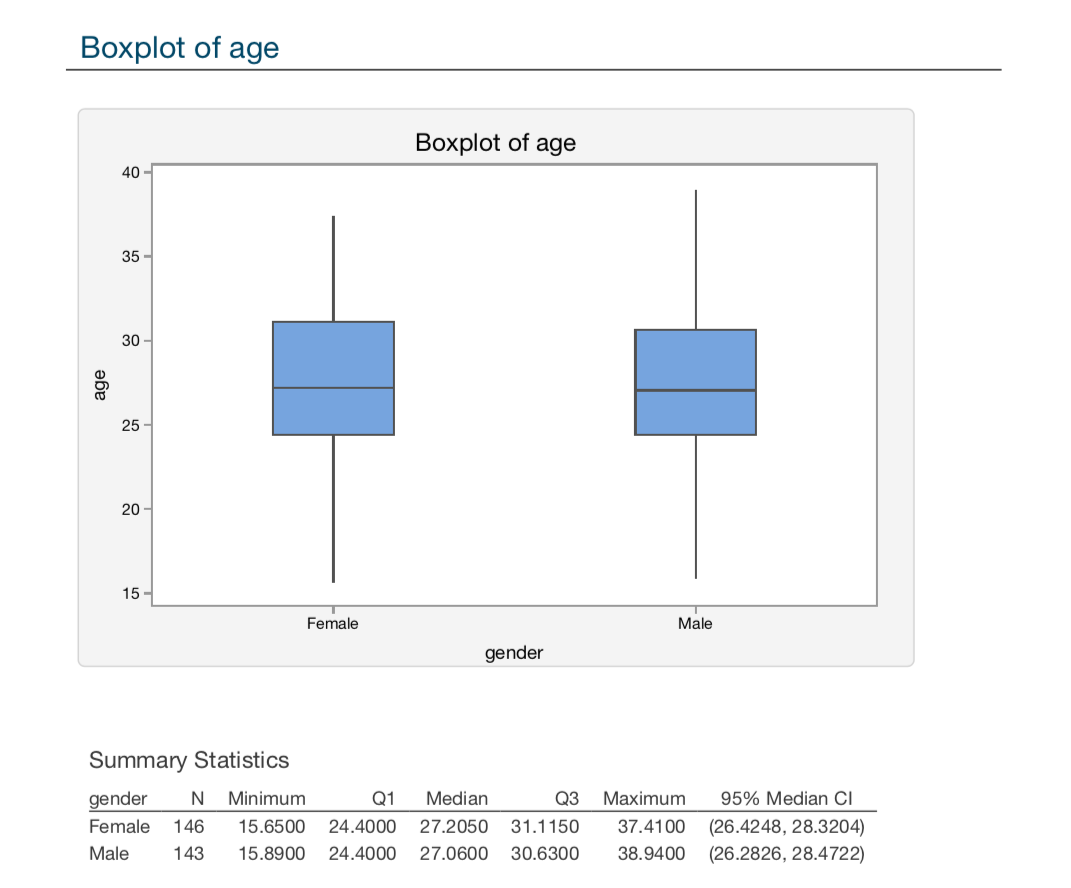
These are the first 5 records of our dataset.

This data could either be provided as two samples, describing the variables.

𝑌1 = ages of female weightlifters  
𝑌2 = ages of male weightlifters

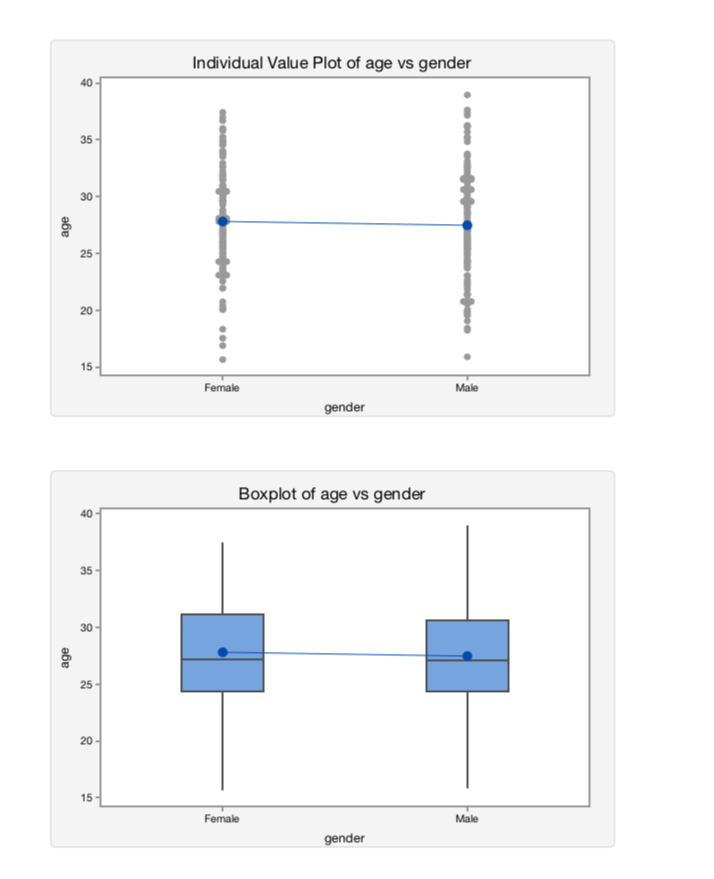
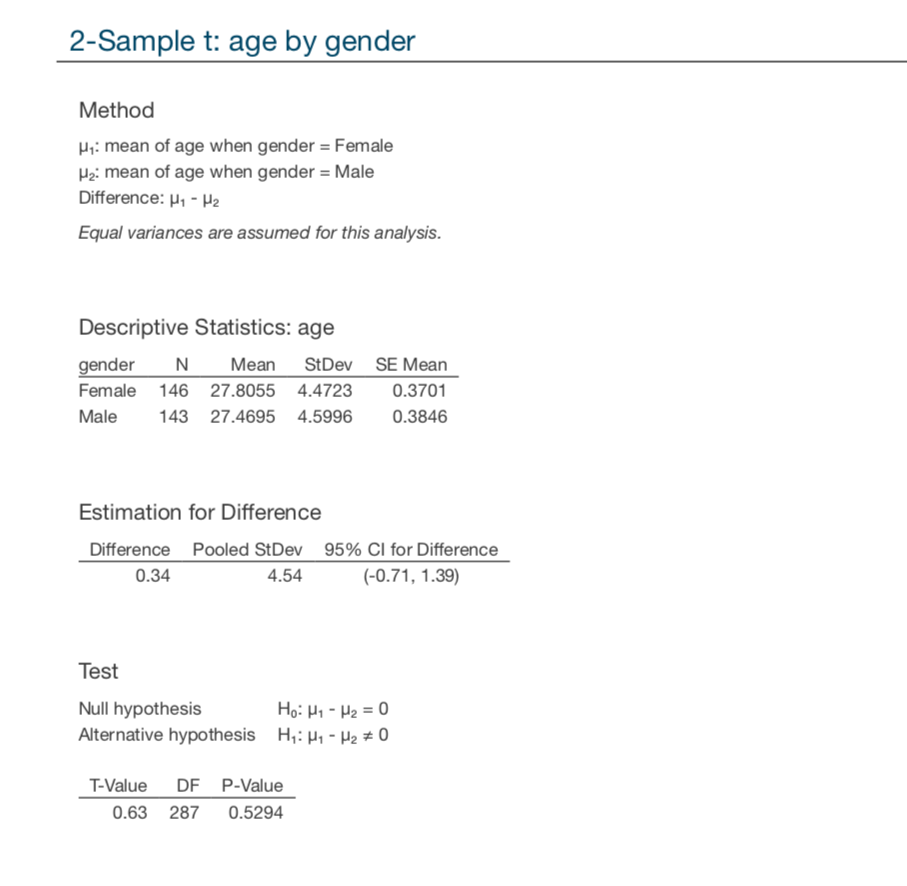


Interpretation: Above Histograms illustrate that ages of both male and female bodybuilders show a normal distribution and seem almost equal.



Interpretation: Above Boxplots illustrate that the variation in age is fairly similar for both female and male. In addition, there are no outliers, i.e., people with unusually high or less ages.

Two-sample 𝒕-test- female ages Vs male ages



Interpretation: Above Boxplots illustrate that the spread/ variation is almost equal of the ages of both male and female bodybuilders.

Hypothesis Test:

Testing for a two-sample 𝑡-test

H: Hypothesis (stating the null & the alternative hypothesis)

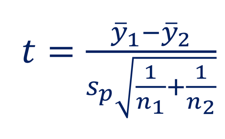
𝐻0:μ1 =μ2 The null hypothesis will be that there is no difference between the ages of female and male.

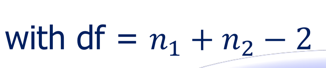
𝐻1:μ1 ≠μ2 The alternative hypothesis is that there is a difference between the ages of female and male.

A: Assumptions (checking the assumptions of the test)

The above mentioned Histograms suggest that ages of both female and male follows normal distributions **and** boxplots indicate that the spread in each ages could be equal (i.e., we assume σ1=σ2) , as required.

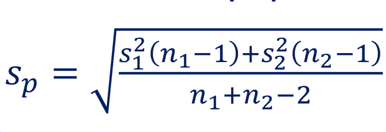
T: Test Statistic (calculating the test statistic)





Here, we have the mean (27.8055 & 27.4695) and sample size (146 & 143) for each location. We need to calculate 𝑠𝑝 (the pooled standard deviation). df is the degree of freedom.

* The pooled variance is a weighted average of the two sample variances.
* We use a pooled standard deviation because we are assuming the two samples are from populations with the same standard deviation, i.e., we assume σ1 = σ2.
* So we calculate 𝑠𝑝 to estimate σ𝑃 where σ𝑃 is the common population standard deviation.

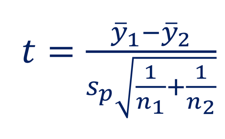


𝑠𝑝= [((4.4723)2(146-1)+(4.5996)2(143-1))/(146+143-2)]0.5

𝑠𝑝= ((2900.212757+3004.197463)/287)0.5

𝑠𝑝= 4.54

Hence, 𝑠1< 𝑠𝑝< 𝑠2



t = ( 27.8055 - 27.4695 ) / (4.54( 0.006849315 + 0.006993006 )0.5 )

t = 0.63 with df = 287

P: 𝒑-value (Obtaining the 𝑝-value for the test from the distribution of the test statistic)

p = 0.5295

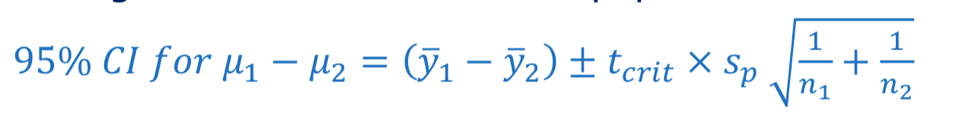
D: Decision (If the 𝑝-value is less than 0.05 (the significance level), reject the null hypothesis. If the 𝑝-value is not less than 0.05, do not reject the null hypothesis)

Since 𝑝-value ≥ 0.05, do not reject 𝐻0.

C: Conclusion (Write a conclusion to the original research question in terms of the target population)

There is no evidence to suggest that there is any difference in the average age of female and male weightlifters. The mean ages could be the same.

95% confidence interval for 𝜇1−𝜇2



Here, 𝑡𝑐𝑟𝑖𝑡 is the critical value which cuts off 5% in the two tails of the 𝑡-distribution with 𝑛1 + 𝑛2 − 2 (287) 𝑑𝑓. We cannot find the value of tcrit by the t- table because it does not contain the value for 287 𝑑𝑓.

To estimate the difference in the ages of female and male weightlifters:

 (-0.71 , 1.39)

We can be 95% confident that the mean age of female weightlifters is between -0.71 less and 1.39 more than the mean age of male weightlifters.

Research Answer: Hence, there is no difference in the average ages of male and female weightlifters, they could be the same.

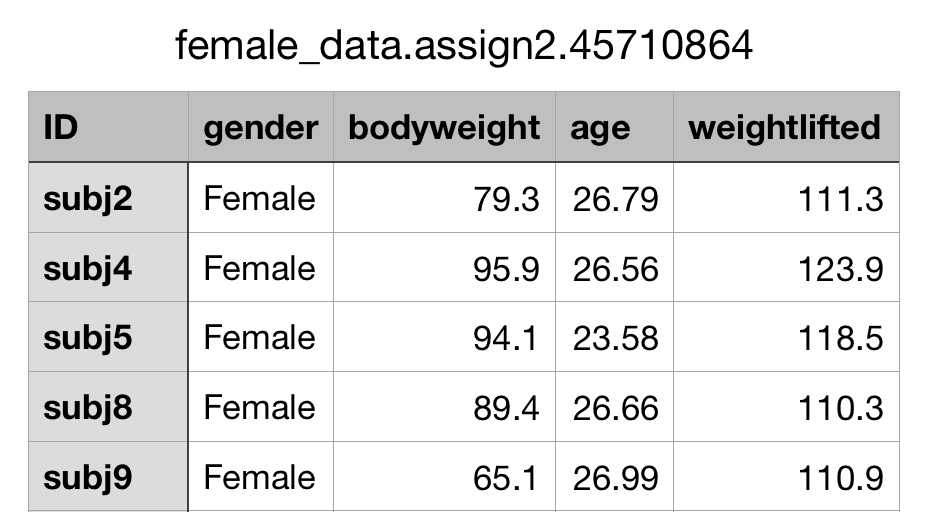
(b)

Research Question: Is there any relation between the body weight of weightlifters and the maximum weight they can lift (by taking care of gender i.e. male & female)?.

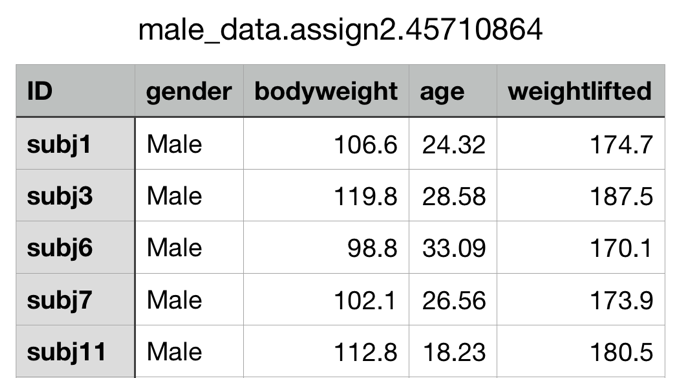
Here, we can study the data of males and females together, but in this approach we would not be able to differentiate between the two different genders (male and female) perfectly while checking the scatter plots in Minitab. The plots will form two different clusters for male and females which is not showing the specific information between bodyweight and weightlifted of male and female bodybuilders.

In the first question, I have applied two sample t-test that means there are two independent sample populations. So, both of them cannot be represented clearly in a scatter plot of Multiple regression. Hence, I used two simple linear regressions for 2 different samples of population according to the gender the weightlifters i.e. male and female.

So, I splitted the given dataset into two datasets (for female and male weightlifters) and made two simple linear regressions instead of a single multiple regression.



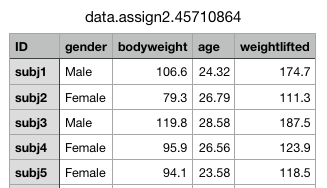
These are the first 5 records of our splitted dataset for female weightlifters.



These are the first 5 records of our splitted dataset for male weightlifters.

So, we could not use Multiple Regression to address this question as there is no meaningful way to differentiate between the gender of weightlifters in this sample.

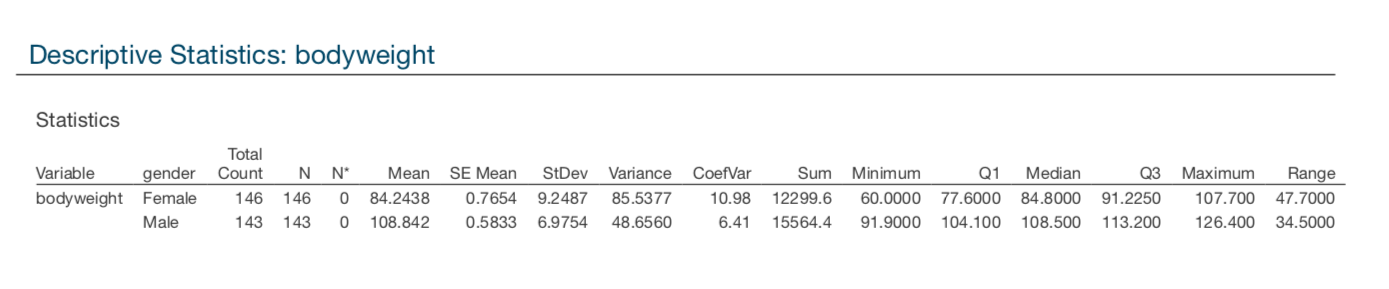
Some Information about the initially given sample of weightlifters

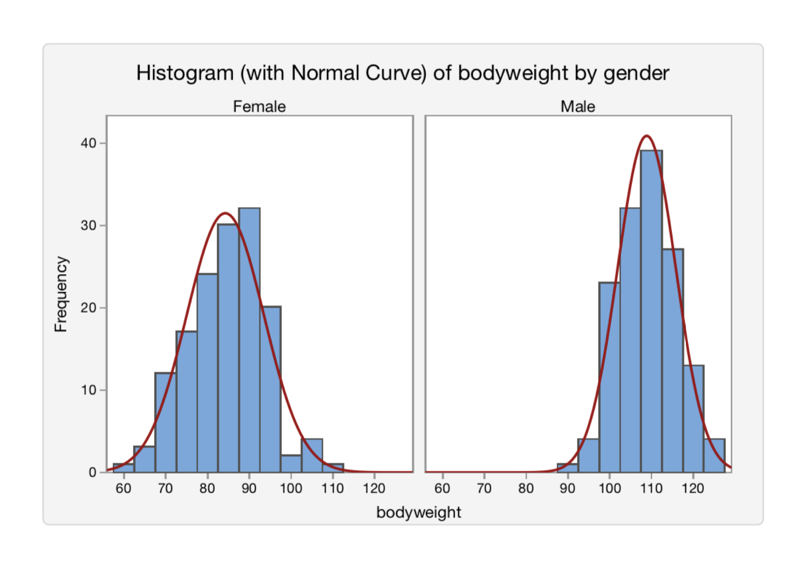


These are the first 5 records of our initial dataset.

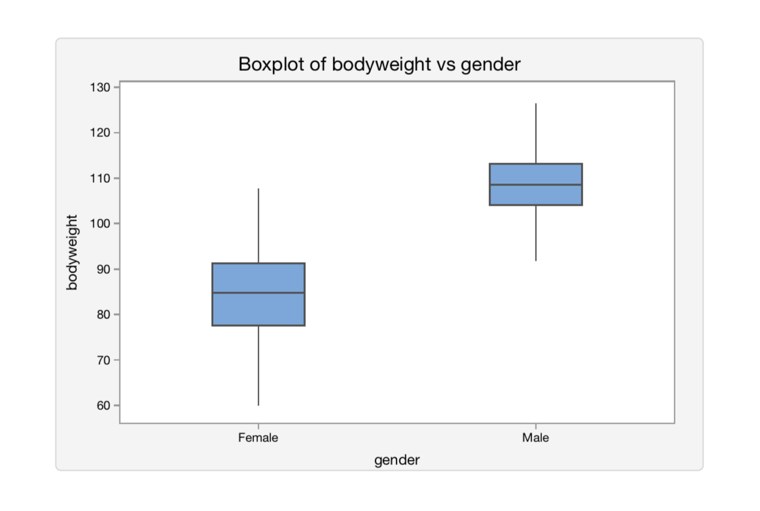
X= bodyweight

Y= weightlifted

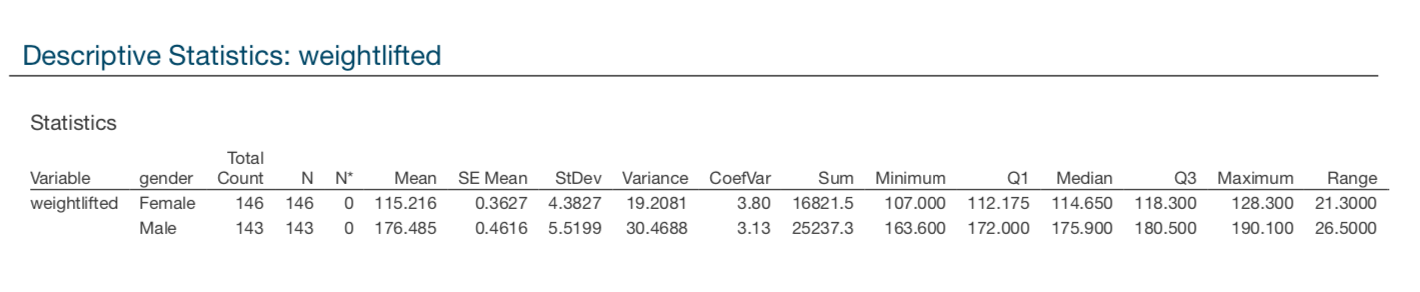


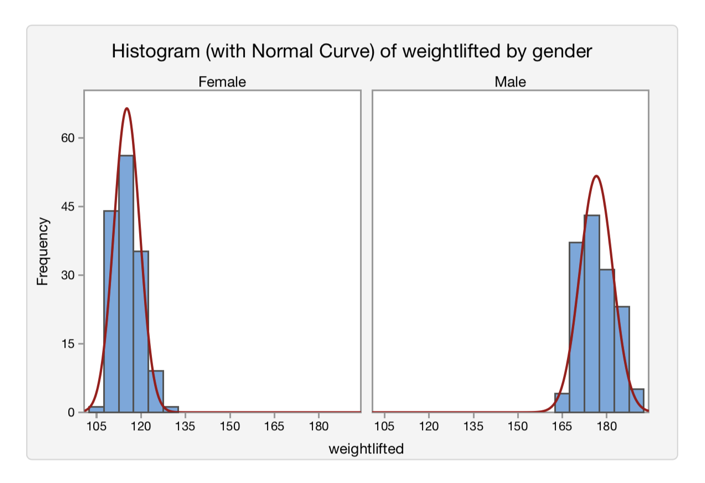


Interpretation: Above Histograms illustrate that the bodyweight of both male and female bodybuilders are approximately normally distributed. However, bodyweight of males are relatively higher than females.

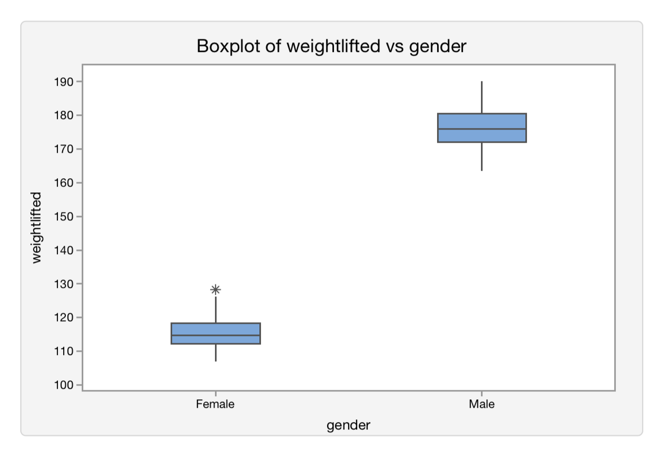


Interpretation: Above Boxplots illustrate that the variation in female bodyweight is higher than male. In addition, there are no outliers, i.e., people with unusually high or less bodyweights.



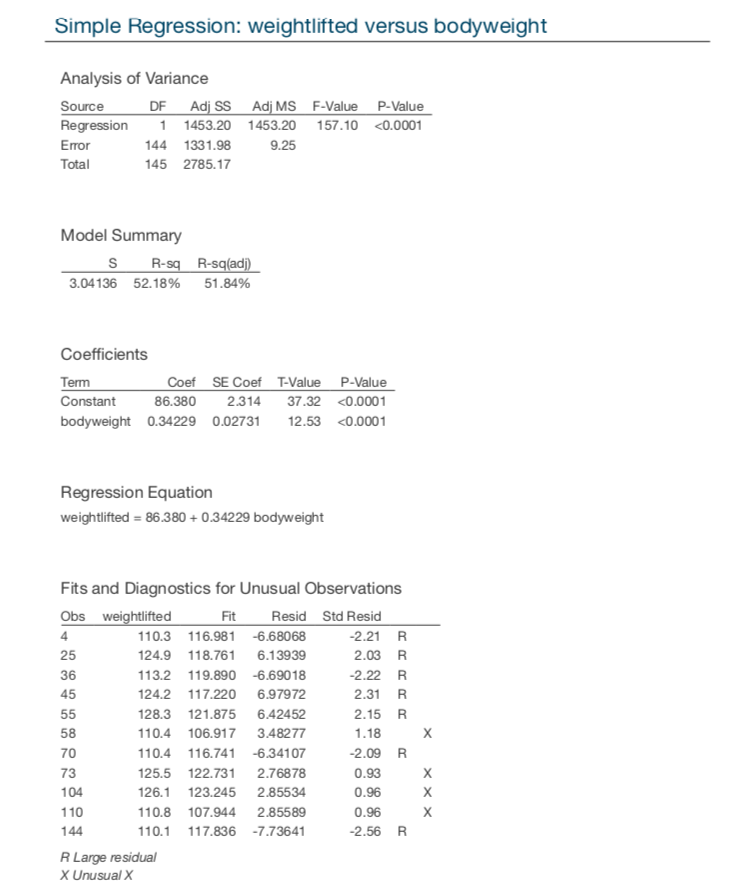


Interpretation: Above Histograms illustrate that the weightlifted by both male and female bodybuilders are approximately normally distributed. However, weightlifted by males are relatively higher than females.



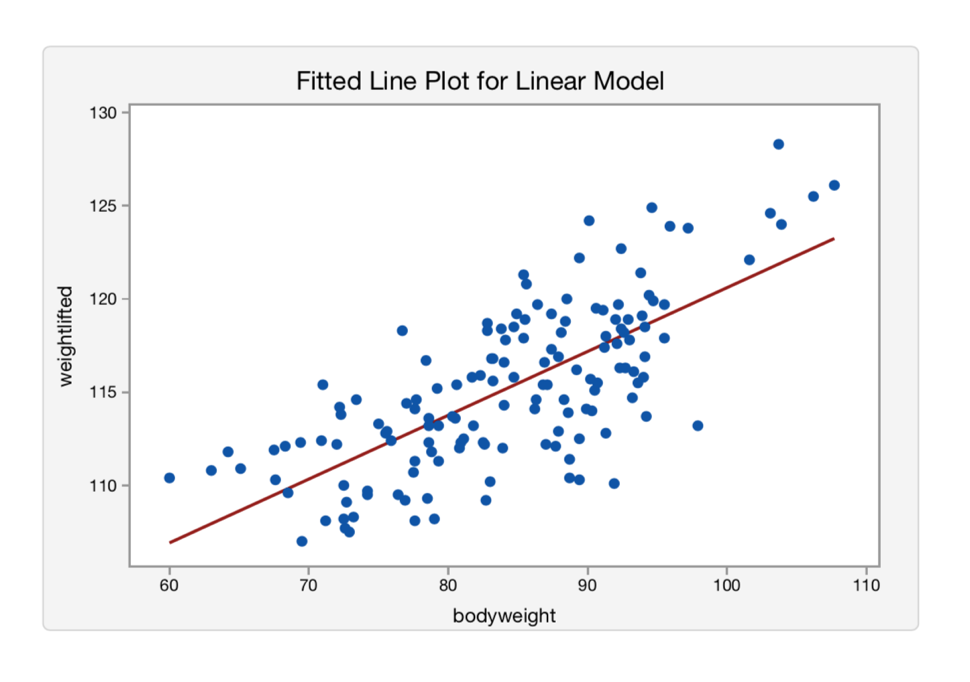
Interpretation: Above Boxplots illustrate that the variation in male weightlifted is slightly higher than female. In addition, there is an outlier in the female weightlifted, i.e., female with unusually high weightlifting capacity.

Simple Linear Regression for female weightlifters



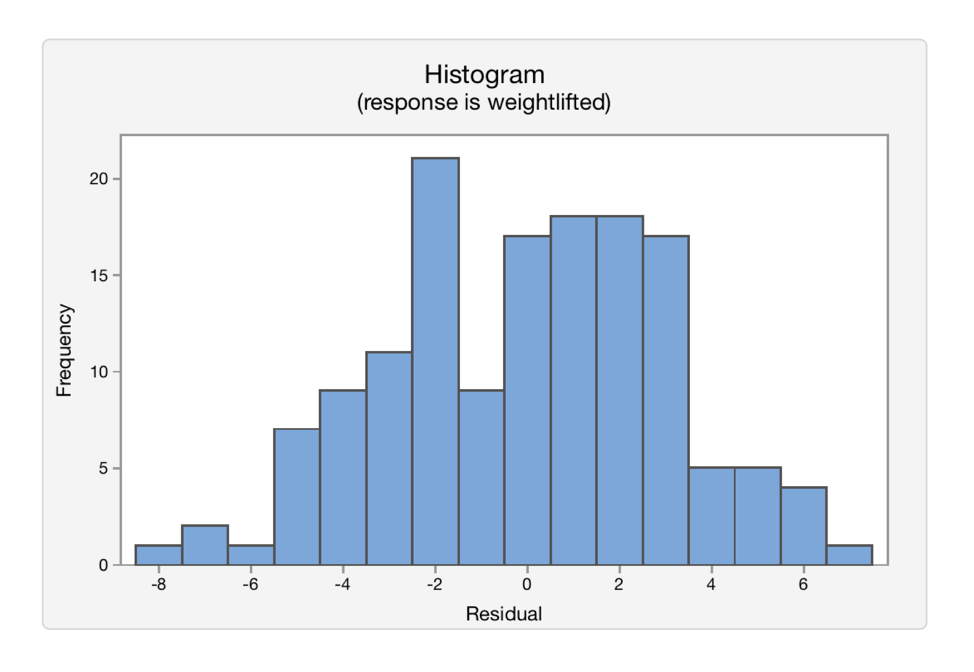
Interpretation: weightlifted = 86.38 + 0.34229 bodyweight

This is a ‘BEST’ line for female weightlifters and Linear Regression is the process of fitting the best straight line to summarise the relation between two numerical variables. The ‘least-squares regression line’ is the line that makes the total squared residual distances as small as possible.

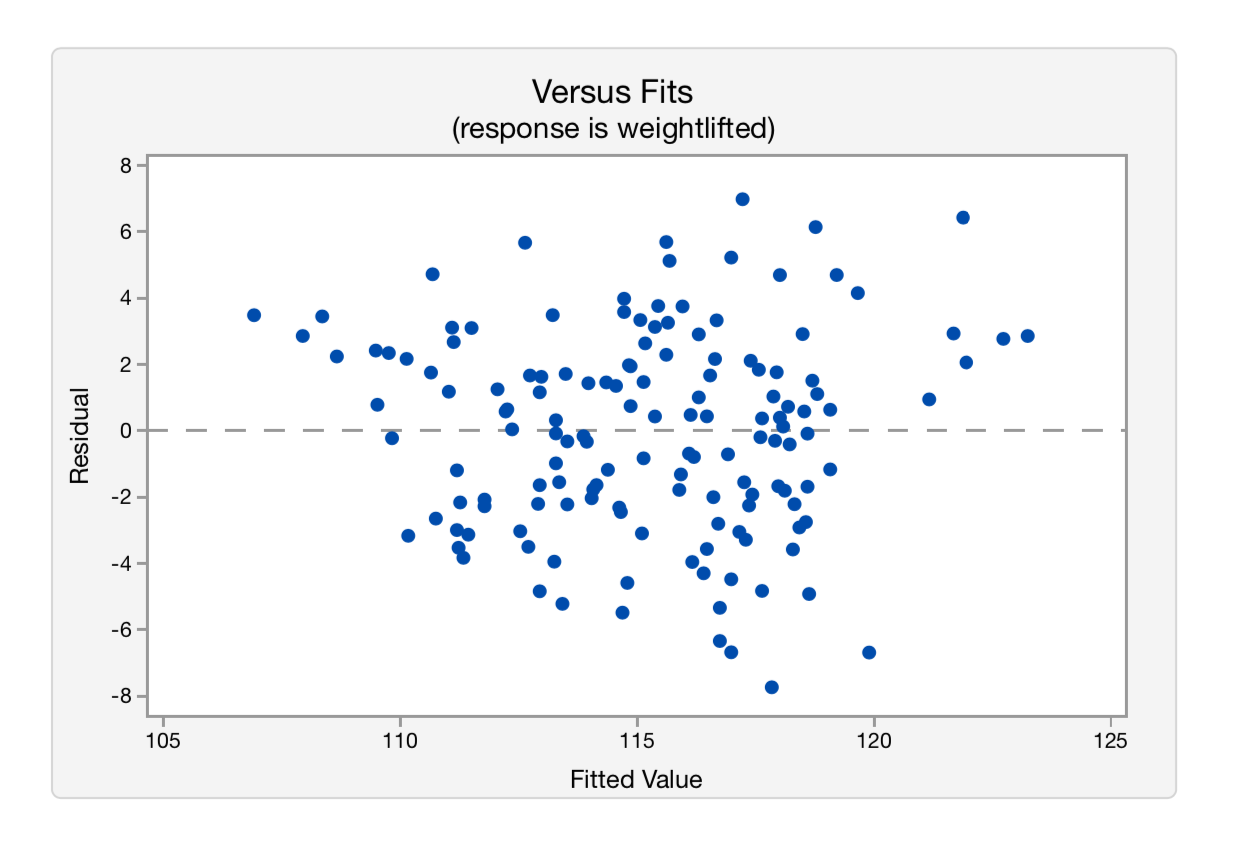


weightlifted = 86.38 + 0.34229 bodyweight

Interpretation: The general trend appears to be positive. As the bodyweight increases, the weightlifted by the female also rises. So, the relation appears to be positive linear.

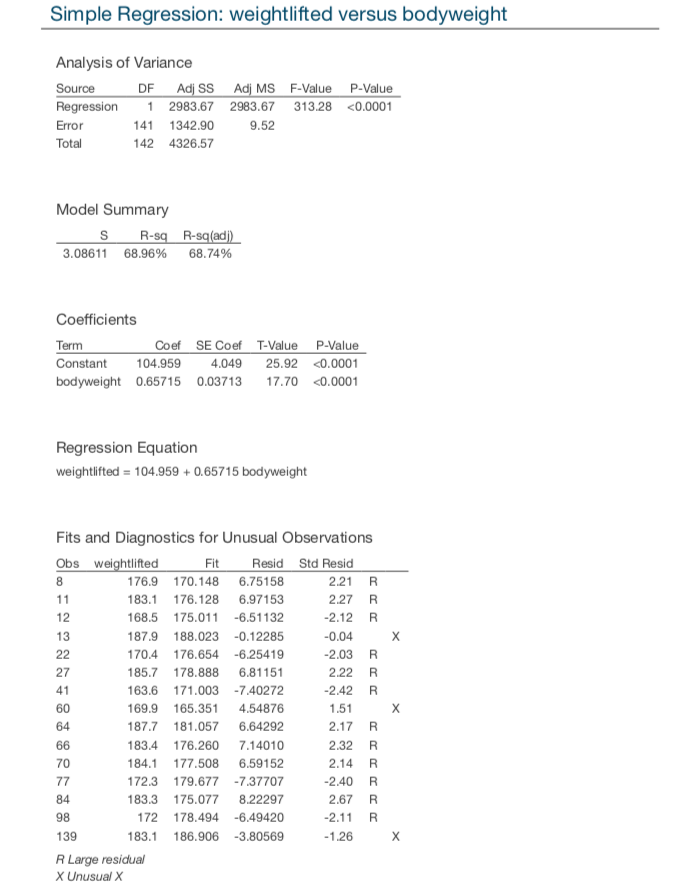


Interpretation: The above histogram of the residuals indicates that for the female weightlifters data, it is reasonable to assume that the residuals are drawn from a normal distribution.



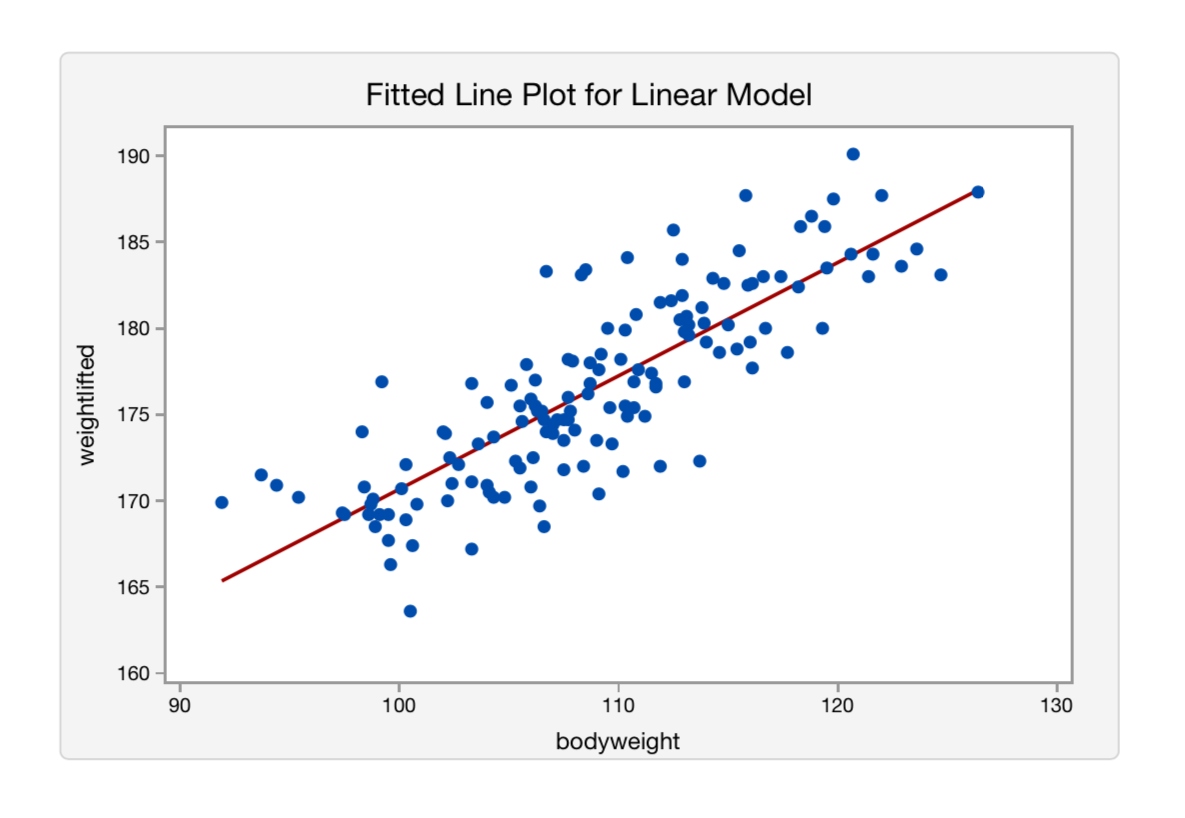
Interpretation: The above plot shows the residuals vs the fitted values for the female weightlifters data. The residuals are fairly evenly scattered around zero across the range of the fitted values. There may be a little more scatter at the high end than the low end but it is probably still reasonable to assume that the residuals will be fairly evenly spread across the range of the female weightlifters data, indicating that the residuals seem to have a constant standard deviation.

Simple Linear Regression for male weightlifters

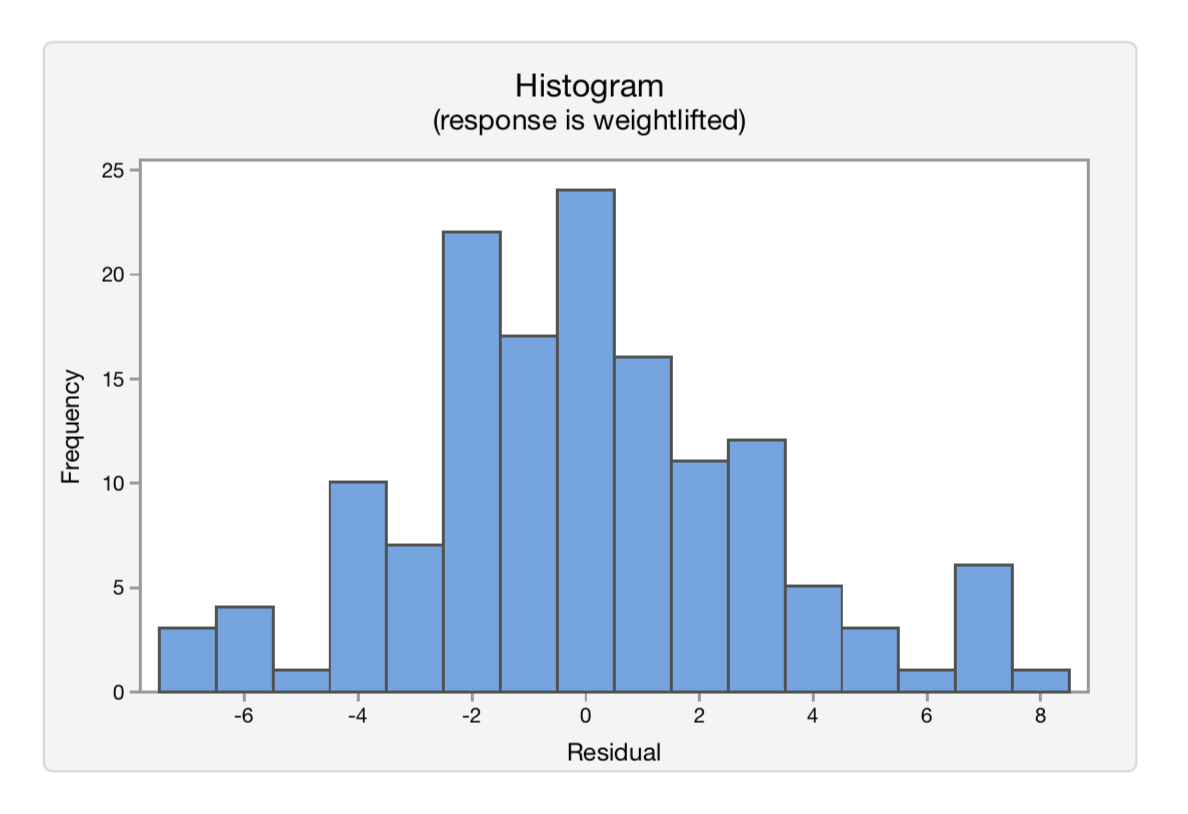


Interpretation: weightlifted = 104.959 + 0.65715 bodyweight

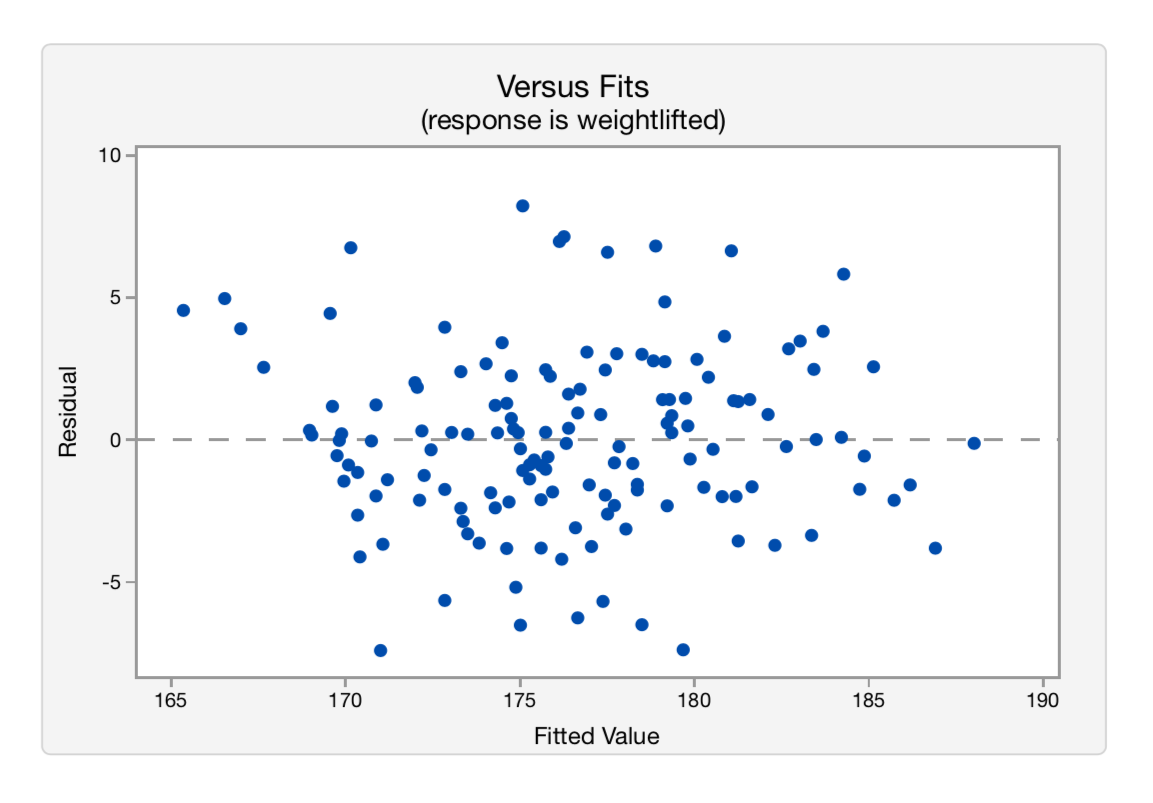
This is a ‘BEST’ line for female weightlifters and Linear Regression is the process of fitting the best straight line to summarise the relation between two numerical variables. The ‘least-squares regression line’ is the line that makes the total squared residual distances as small as possible.

 weightlifted = 104.959 + 0.65715 bodyweight

Interpretation: The general trend appears to be positive. As the bodyweight increases, the weightlifted by the male also rises. So, the relation appears to be positive linear.



Interpretation: The above histogram of the residuals indicates that for the male weightlifters data, it is reasonable to assume that the residuals are drawn from a normal distribution.



Interpretation: The above plot shows the residuals vs the fitted values for the male weightlifters data. The residuals are fairly evenly scattered around zero across the range of the fitted values. There may be a little more scatter at the high end than the low end but it is probably still reasonable to assume that the residuals will be fairly evenly spread across the range of the male weightlifters data, indicating that the residuals seem to have a constant standard deviation.

Hypothesis Test:

Testing for Linear Regression

H: Hypothesis (stating the null & the alternative hypothesis)

𝐻0: β = 𝟎 (i.e., population slope = 0) The null hypothesis will be that there is no significant linear relation between 𝑋 and 𝑌, i.e., 𝑋 is not a useful predictor of 𝑌.

𝐻1: β ≠ 𝟎 (i.e., population slope ≠ 0) The alternative hypothesis is that there is a significant linear relation between 𝑋 and 𝑌, i.e., 𝑋 is a useful predictor of 𝑌.

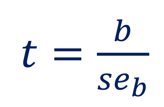
A: Assumptions (checking the assumptions of the test)

The scatter plots of both female and male illustrate that the relation between bodyweight and weightlifted may be linear. The histograms indicate that the residuals are from a normal distribution. The ‘Residuals vs. fits’ plots which plots the residuals against the fitted values (i.e., the predicted values) could have a constant standard deviation.

These 3 assumptions are for both the linear regressions of male and female weightlifters.

T: Test Statistic (calculating the test statistic)





with df= n-2

Here, we will calculate two t-values for both female and male. t1 will be the t-value of female weightlifters and t2 will be the t-value of male weightlifters.

t1 = b1 / (seb1) = 0.34229 / (0.02731) = 12.5335 with df1= n1 – 2 = 146-2= 144

t2 = b2 / (seb2) = 0.64715 / (0.03713) = 17.6986 with df2= n2 – 2 = 143-2= 141

P: 𝒑-value (Obtaining the 𝑝-value for the test from the distribution of the test statistic)

As we have two t- values, so we will get two p- values i.e. p1 and p2

p1 ≈0

p2 ≈0

D: Decision (If the 𝑝-value is less than 0.05 (the significance level), reject the null hypothesis. If the 𝑝-value is not less than 0.05, do not reject the null hypothesis)

Since, 𝑝-value ≥ 0.05, do not reject 𝐻0.

C: Conclusion (Write a conclusion to the original research question in terms of the target population)

There is evidence of a positive linear relation between the bodyweights and weightlifted by female and male weightlifters.

95% confidence interval for β

Minitab does not give this confidence interval with the regression output, but it is useful because it estimates the average change in the 𝑌 for each one unit change in 𝑋.

A 95% CI for 𝛽 is 𝑏 ± 𝑡𝑐𝑟𝑖𝑡 × 𝑠𝑒𝑏, where 𝑡𝑐𝑟𝑖𝑡 is the critical value which cuts off 5% in the two tails of the 𝑡-distribution with 𝑛 − 2 df, and 𝑠𝑒𝑏 is the estimated standard error for the slope, which we will get from Minitab.

To estimate the slope of the regression line in the target population of female, 95% CI for 𝛽 is 0.34229 ± 1.97658 × 0.02731 = (0.29, 0.4) .

To estimate the slope of the regression line in the target population of male, 95% CI for 𝛽 is 0.64715 ± 1.97693 × 0.03713 = (0.57, 0.72) .

We can be 95% confident that, for female weightlifters, the value of weightlifted increases between 0.29 and 0.4 .

We can be 95% confident that, for male weightlifters, the value of weightlifted increases between 0.57 and 0.72 .

(Note that the CI does not contain 0, confirming the decision to reject the null hypothesis that the population slope was 0.)

Research Answer: Hence, there is a positive linear relation between the body weight of weightlifters and the maximum weight they can lift (by taking care of gender i.e. male & female).