1. In the dataset linked below, some physical and performance data from a random sample of 222 weightlifters have been collected. 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]

The data set (whose first 5 records are shown below) can be found on iLearn under Resources in the file:
data.assign2.30.44468423.222.csv

ID	gender	bodyweight	age	weightlifted
subj1	Female	86.9	24.3	110.7
subj2	Male	113.0	32.2	180.3
subj3	Female	91.6	33.9	113.4
subj4	Male	104.9	29.5	169.9
subj5	Female	84.1	29.3	116.3

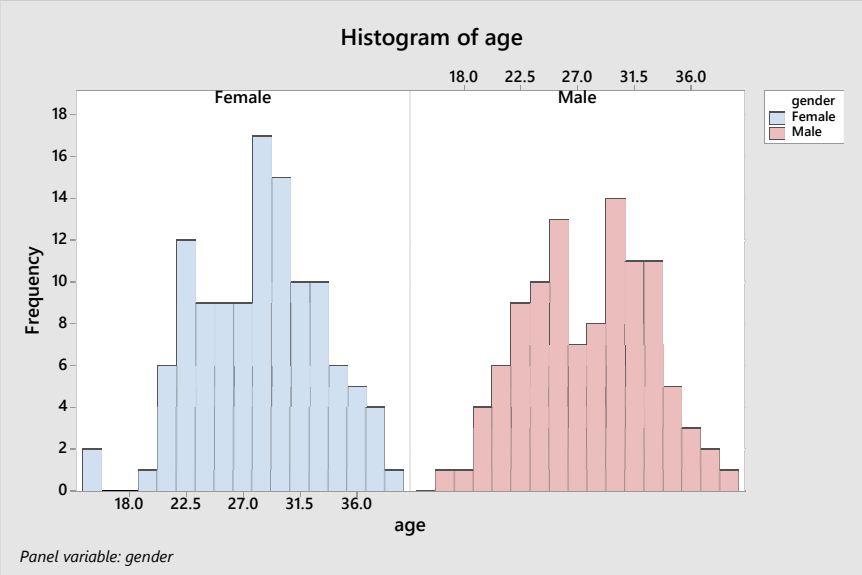
(a) Is there any difference in the average age of male and female weightlifters?

Introduction:

To determine whether there is any difference in the average age of male and female weightlifters from the physical and performance data collected from a random sample of 222 weightlifters.

Methods:

A histogram of age is first drawn by gender separately for female and male. Both the histograms indicate that the normality assumption seems to have been met as the sample shows a normal distribution.



A boxplot of age is then drawn by gender to check the variability and it shows equal spread for both female and male.

Since we have two independent samples with females and males, a Two-Sample T-Test is performed to verify whether the distribution of age has the same mean for females and males assuming that the samples show normal distribution with equal variances.

The null hypothesis would be that there is no difference between the means of the two samples whereas the alternative hypothesis would be the opposite that there is a difference between the two samples.

Result:

Two-Sample T-Test and CI: age, gender

Method

μ_1: mean of age when gender = Female

μ_2: mean of age when gender = Male

Difference: $\mu_1 - \mu_2$

Equal variances are assumed for this analysis.

Descriptive Statistics: age

Gender	N	Mean	StDev	SE Mean
Female	116	28.34	4.89	0.45
Male	106	27.88	4.87	0.47

Estimation for Difference

Difference	Pooled StDev	95% CI for Difference
0.462	4.882	(-0.831, 1.755)

Test

Null hypothesis	$H_0: \mu_1 - \mu_2 = 0$
Alternative hypothesis	$H_1: \mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
0.70	220	0.482

The test statistic is z = 0.70 and the P-Value is 0.482. Since P-Value is greater than 0.05, the null hypothesis holds true may not be rejected.

Conclusion:

There is insufficient evidence to suggest that the means for females and males are not same. With 95% confidence, we can state that the distribution of age has the same means for females and males.

b) What is the relation between the body weight of weightlifters and the maximum weight they can lift?

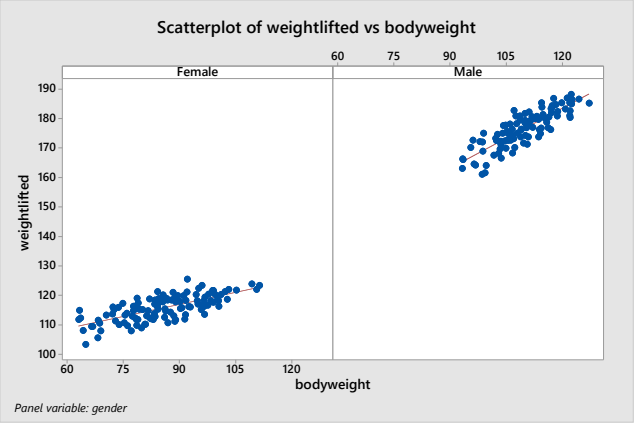
Introduction:

To determine whether there is any relationship between the body weight of weightlifters and the maximum weight they can lift.

Methods:

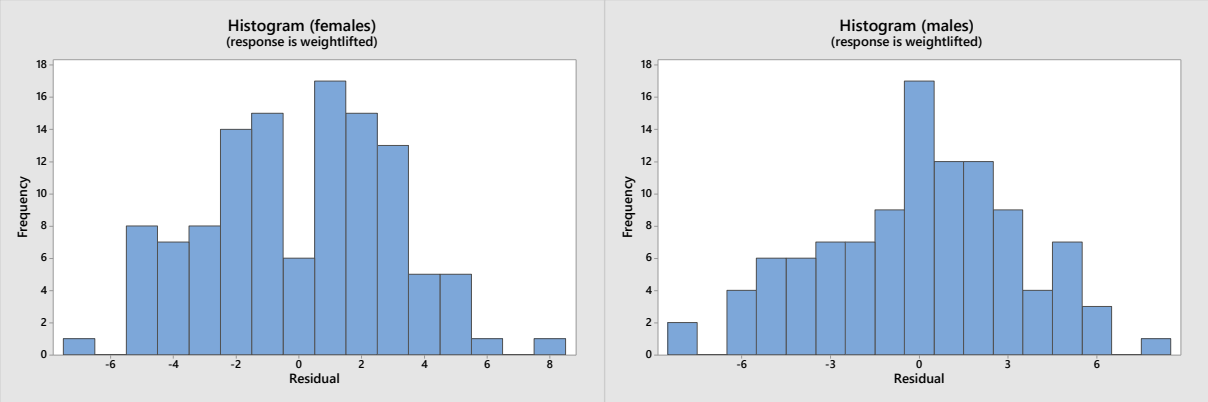
In order to study the relationship between the body weight of weightlifters and the maximum weight they can lift, the data set is split by gender into female and male as they are two independent samples.

When considering the relation between two numerical variables such as weightlifted and bodyweight, we should always look at a scatter plot before any further analysis.

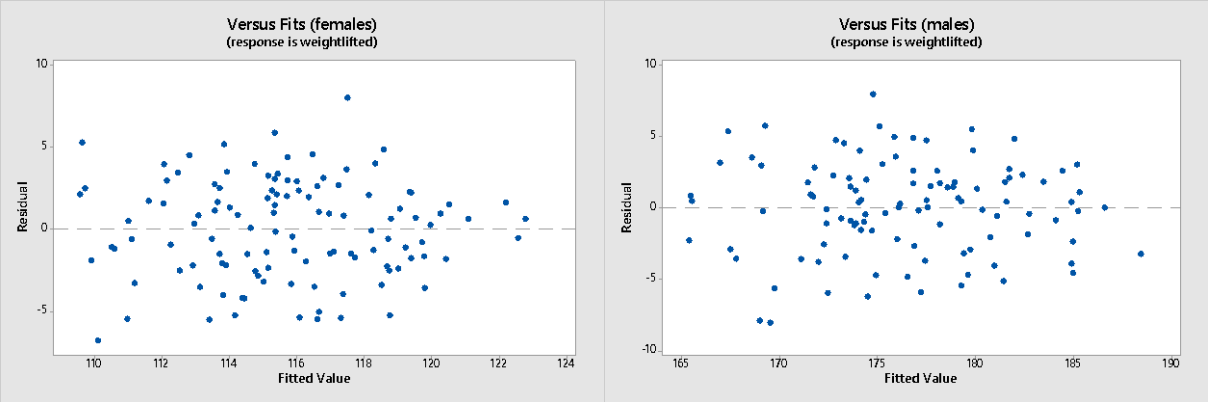


The scatterplots for both females and males shows the residuals grouped closely around the best fitted line suggesting a linear relationship between weightlifted and bodyweight.

This linear relationship satisfies the first condition for linear regression but to perform the regression analysis, the residuals should show a normal distribution and have a constant standard deviation for which Histogram of Residuals and Residuals versus Fits are plotted separately for females and males.



The histogram of the residuals for both females and males show a normal distribution satisfying the second condition for linear regression.



The plot of the residuals versus the fitted values for both females and males show an even scatter of points around zero across the range of the fitted values, indicating that the residuals seem to have a constant standard deviation which satisfies the third and final condition for linear regression.

Result:

Upon satisfying the above three conditions, regression analysis is performed separately for both females and males with the variable weightlifted as Responses and bodyweight as Continuous Predictors.

(Continued in next page)

Regression Analysis: weightlifted versus bodyweight (females)

Analysis of Variance (females)

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	1022.56	1022.56	114.32	0.000
bodyweight	1	1022.56	1022.56	114.32	0.000
Error	114	1019.73	8.95		
Lack-of-Fit	105	924.46	8.80	0.83	0.699
Pure Error	9	95.27	10.59		
Total	115	2042.29			

Model Summary (females)

S	R-sq	R-sq(adj)	R-sq(pred)
2.99082	50.07%	49.63%	48.42%

Coefficients (females)

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	92.34	2.20	41.92	0.000	
Bodyweight	0.2732	0.0255	10.69	0.000	1.00

Regression Equation (females)

Weightlifted = 92.34 + 0.2732 bodyweight

Fits and Diagnostics for Unusual Observations (females)

Obs	weightlifted	Fit	Resid	Std Resid		
50	125.500	117.497	8.003	2.69	R	
74	122.000	122.551	-0.551	-0.19		X
110	103.300	110.094	-6.794	-2.32	R	
111	123.400	122.769	0.631	0.22		X

R Large residual
X Unusual X

(Continued in next page)

Regression Analysis: weightlifted versus bodyweight (males)

Analysis of Variance (males)

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	2830.88	2830.88	265.03	0.000
bodyweight	1	2830.88	2830.88	265.03	0.000
Error	104	1110.85	10.68		
Lack-of-Fit	92	1054.90	11.47	2.46	0.042
Pure Error	12	55.95	4.66		
Total	105	3941.73			

Model Summary (males)

S	R-sq	R-sq(adj)	R-sq(pred)
3.26822	71.82%	71.55%	70.72%

Coefficients (males)

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	102.05	4.58	22.29	0.000	
Bodyweight	0.6796	0.0417	16.28	0.000	1.00

Regression Equation (males)

Weightlifted = 102.05 + 0.6796 bodyweight

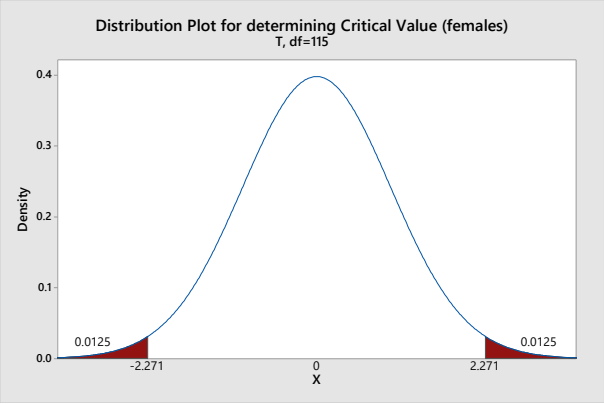
Fits and Diagnostics for Unusual Observations (males)

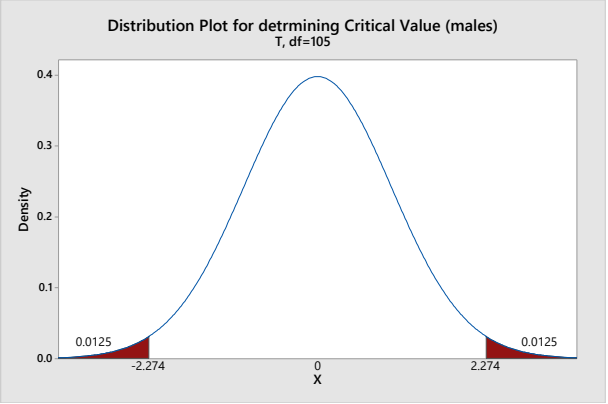
Obs	weightlifted	Fit	Resid	Std Resid	
11	185.200	188.427	-3.227	-1.02	X
24	182.700	174.767	7.933	2.44	R
32	161.500	169.533	-8.033	-2.49	R
69	161.100	168.990	-7.890	-2.45	R

R Large residual
X Unusual X

The P-Value obtained for both females and males is 0. As the P-Value is less than 0.05, the null hypothesis may be rejected.

(Continued in next page)





95% Confidence Interval (Females) = Coef of Bodyweight ± Critical Value × SE Coef of Bodyweight

= 0.2732 ± 2.271 × 0.0255

= (0.22, 0.33)

95% Confidence Interval (Males) = Coef of Bodyweight ± Critical Value × SE Coef of Bodyweight

= 0.6796 ± 2.274 × 0.0417

= (0.58, 0.77)

The confidence intervals for both females and males does not contain 0, confirming the decision to reject the null hypothesis that the slope was 0.

Conclusion:

There is a negative linear reaction between the body weight of weightlifters and the maximum weight they can lift for both females and males.