Statistical Report

Sean Spence

26/10/2021

Abstract

The project consisted of three research questions that had to be answered using three different statistical test. The research questions and answers are,

- Is there a linear relationship between height and weight? This test was conducted using a linear model test and tested to see if the slope was significant. It was concluded that there was a positive relationship between height and weight.
- Is the mean height of male and female the same? This test was conducted using a two sample t test on the heights of males and females. It was concluded that the mean height for male and female are different.
- Is there any association between gender and the amount of physical activity? This test was conducted using the chi square test of independence to test if there is any association between gender and physical activity. It was concluded that there were no association between gender and the level of physical activity.

Introduction

Three research questions have been given and the purpose of this report is to do statistical testing is done is to answer those questions. The research questions that are given are:

1. Is there a linear relationship between height and weight?
2. Is the mean height of male and female the same?
3. Is there any association between gender and the amount of physical activity?

The data given contains 1000 observations with male and female aged 26-45 with information on the following variables:

- ID
- Gender
- Height (in cm)
- Weight (in kg)
- Physical Activity (None, Moderate, Intense)

Data

Here is a small preview of data that we are working with,

Table 1: First six rows of data

ID	gender	height	weight	phys
ID1	Male	183.59	84.01	None
ID2	Male	177.78	61.26	Moderate
ID3	Female	167.25	68.03	None
ID4	Male	173.96	67.93	Moderate
ID5	Male	174.99	63.65	Moderate
ID6	Female	167.38	68.59	Moderate

Methods

Is there a linear relationship between height and weight?

This test requires the data to be fit into a linear model to observe whether there is a relationship between height and weight. The function that is used in R Studio is the $lm()$ function which would give the values for β_0 and β_1. By using $summary(lm())$, other important values can be analysed such as if the coefficients are significant and how good the fit is to the data through the R^2 coefficient. Hence the equation to the is fit would be Wikipedia (2021) :

$$Height = \beta_0 + \beta_1 * weight + \epsilon$$

A sample of this data used to conduct this test can be seen in the data below,

Table 2: Data for height and weight used for linear test

height	weight
183.59	84.01
177.78	61.26
167.25	68.03
173.96	67.93
174.99	63.65
167.38	68.59

Is the mean height of male and female the same?

This question is asking to compare whether the average height male and female are the same. The most appropriate test to answer this question is to conduct a two sample t-test for equal means. It is able to see if there is any significant difference between the height of both genders by using this test Handbook (2021) .

Table 3: Data on females and their heights for t-test

gender	height
Female	167.25
Female	167.38
Female	164.17
Female	164.83
Female	175.11
Female	172.28

Table 4: Data on males and their heights for t-test

gender	height
Male	183.59
Male	177.78
Male	173.96
Male	174.99
Male	182.80
Male	178.46

Is there any association between gender and the amount of physical activity?

This question will be conducted by a Chi-square test for independence. This test will be able to see if there is association on the gender the level of physical activity. The data will count how many male are doing no, moderate and intensive physical activity and will put the count in that specific cell. The same process will be done for females. Again, a sample data will be shown to see what we are dealing with thebmj (2021).

Table 5: Data on gender and physical level for chi-square test

gender	height	weight	phys
Male	183.59	84.01	None
Male	177.78	61.26	Moderate
Female	167.25	68.03	None
Male	173.96	67.93	Moderate
Male	174.99	63.65	Moderate
Female	167.38	68.59	Moderate

Results

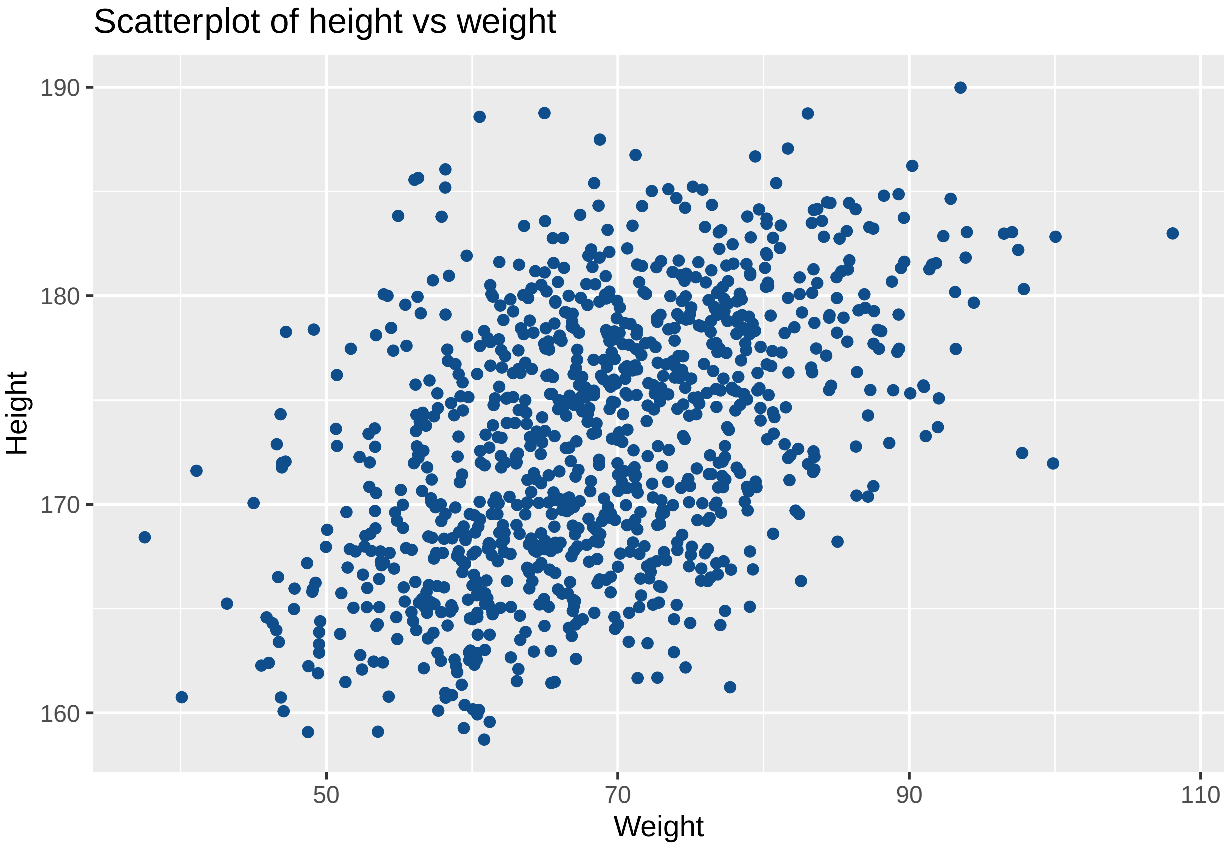
Is there a linear relationship between height and weight?

Hypothesis - Lineaer

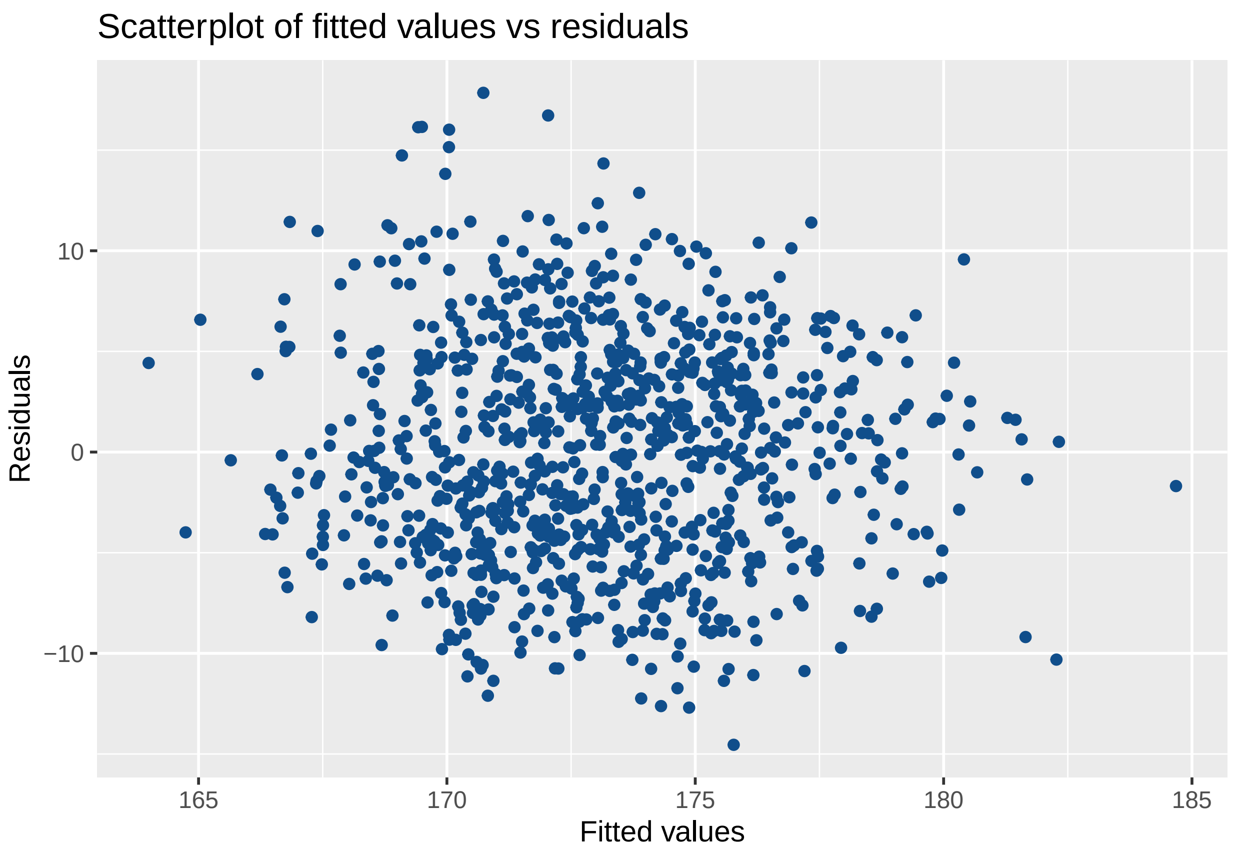
```
## HYPOTHESIS
## The null hypothesis H0 is beta = 0, and the alternative hypothesis
##        H1 is beta does not equal to 0 at a significance level of 5%. Beta is the true slope paramete
##        model height = alpha + beta weight + epsilon.
```

Assumptions - Linear

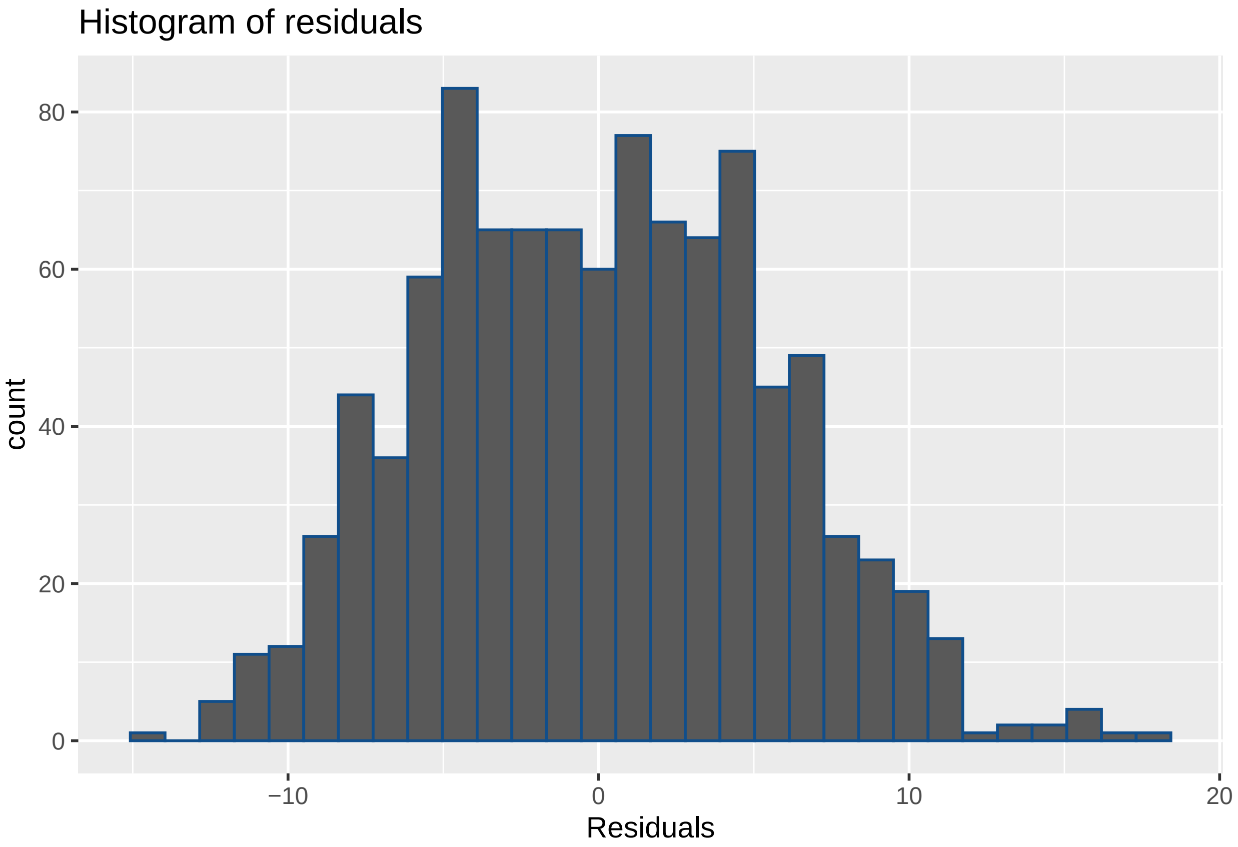
The linear model assumptions should be acknowledged before proceeding to the test values to see that if the linear model is an appropriate fit to the data. There are three assumptions that can be look at graphically and one that can be explained JMP (2021).



1. This scatter plot shows the height vs. weight and it can be observed that there is a linear trend from this plot.



2 .The next assumption to check whether the residuals are around zero. In the case, it looks like it is centered around zero. Also what it aims to look like is like a gunshot spray, meaning no pattern and looks random. In this case that is what it looks like.



3. The next assumption is that we have to check whether the residuals are normally distributed. In this case, there seems to a pattern in the distribution of residuals and it looks like it is normally distributed.
4. This assumption cannot be graphically show but we have to assume that each observation is independent from one another.

Test Results - Linear

```
##
## Call:
## lm(formula = height ~ weight, data = project)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.5439  -4.2003  -0.0516   4.0516  17.8465
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 152.98837    1.15625  132.31   <2e-16 ***
## weight        0.29321    0.01667   17.59   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.519 on 998 degrees of freedom
## Multiple R-squared:  0.2366, Adjusted R-squared:  0.2358
## F-statistic: 309.3 on 1 and 998 DF,  p-value: < 2.2e-16
```

```
##
## FIT - Linear Regression
```

```
## beta_hat =  0.2932118
```

```
## p_value =  1.627962e-60
```

From the summary statistic it shows there are three three stars next to the weight coefficient (which is the β_1 in our case) which shows that it a significance variable in explaining height. This is furthered confirmed by the very small p-value which is less than 0.05 hence making β_1 significant. The β_1 is shown in the output above.

Decision - Linear

```
## DECISION
## Reject NULL hypothesis
```

Conclusion - Linear

```
## CONCLUSION
## There is evidence that the slope (beta) is different
## than 0. There is a significant linear relationship
## between height and weight. For each unit-increse
## in weight, height increases by
## 0.2932 at a 5% significance level.
```

Is the mean height of male and female the same?

Hypothesis - t Test

```
## HYPOTHESIS
## The Null hypothesis H0 is meu1=meu2 where the means in the
##       height of male and female are the same and the alternate hypothesis H1 is
##       meu1 does not equal meu2 where the height of male does not equal to height
##       of female.
```

Assumptions - t Test

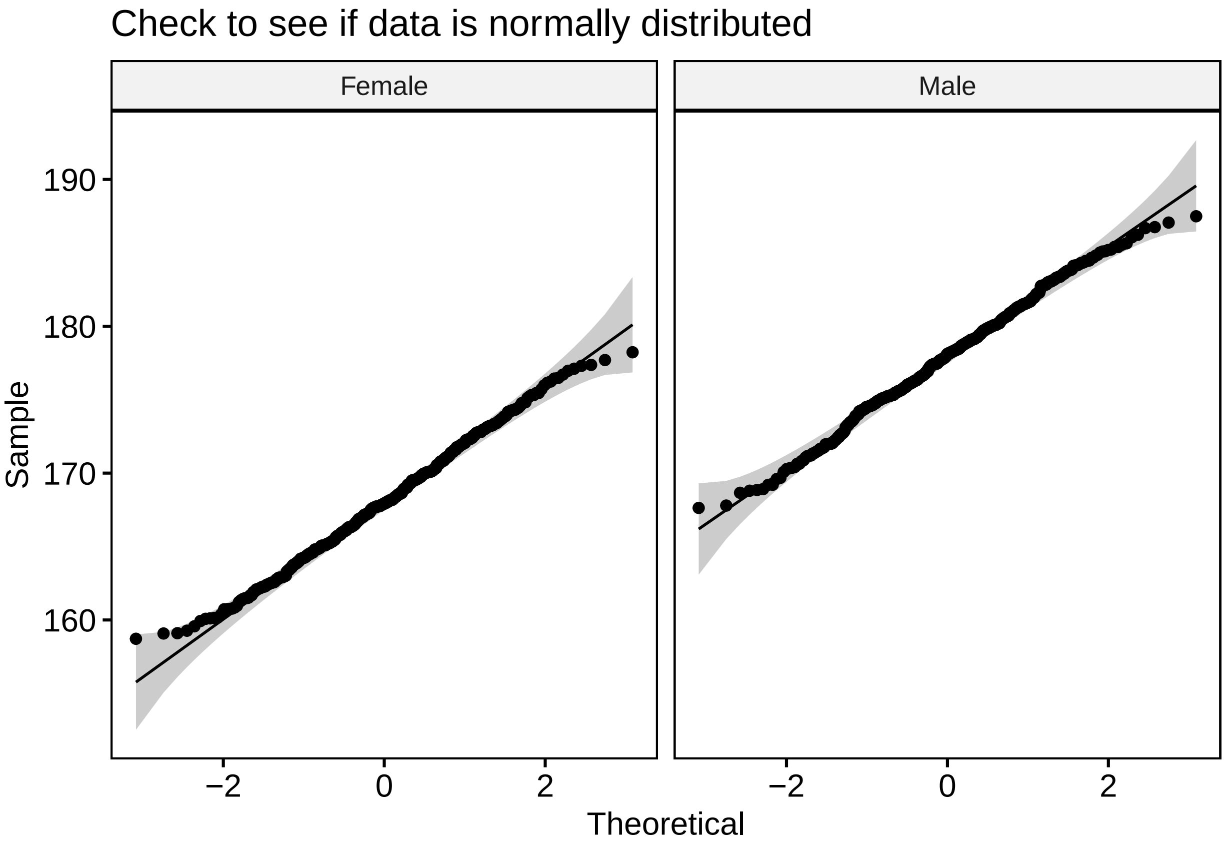
The two sample t test has four assumptions that needs to be address before proceeding DataNova (2021).

```
## [1] 1000    7
```

```
## [1] 993   7
```

1. The first assumption is that there are no outliers in the data. By sorting the data out between male and female height, there were 7 outliers that were removed in order to do the two sample t test. As you can see in the dimension in the R output, the original data had 1000 observations but reduced to 993 observations.

```
## # A tibble: 2 x 4
##   gender variable statistic     p
##   <chr>  <chr>        <dbl> <dbl>
## 1 Female height       0.996 0.250
## 2 Male   height       0.996 0.201
```



2. The next assumption is to check whether the data for male and female heights are normally distributed. In the graph, we can see that most of the data are along the solid line which gives an indication that the data is normally distributed
3. We are assuming that there are equal variance between make and female height
4. We are also assuming that male and female observations are independent from each other

Test results - t Test

The two sample t test statistic is calculated from the formula below.

$$t = (\overline{x}_1 - \overline{x}_2)/\sqrt(s_p^2(1/n_1 + 1/n_2))$$

The s_p^2 is the pooled sample variance and is calculated from the formula below,

$$s_p^2 = ((n_1 - 1)s_1^2 + (n_2 - 1)s_2^2)/(n_1 + n_2 - 2)$$

where

- $\overline{x}_1$ and $\overline{x}_2$ are mean of male and female,
- n_1 and n_2 are the total number of males and females

```
##
## T-Test
```

```
## p_value = 1.58e-215
```

```
## t_value = 40.97931
```

```
## t_critical_value = 1.962361
```

Decision - t Test

```
## DECISION
## Reject NULL hypothesis
```

Conclusion - t Test

```
## CONCLUSION
## There is evidence that the mean for male and female
## height are different at a 5% signifance level.
```

Is there any association between gender and the amount of physical activity?

Hypothesis - Chi square

```
## HYPOTHESIS
## The Null Hypothesis H0 is where there is no relationship betweeen
##       gender and the level of physical activity. The alternate hypothesis H1 is
##       there is a relationship between gender and the amount of physical activity.
```

Assumptions - Chi Square

The Chi square test has some assumptions that need to be address. There are six assumptions that the chi square test of independence has to make sure the test is appropriate to use McHugh (2013) .

1. The first assumption is that each cell expected value of each cell should be greater than five and no cell expected value should be less than three.
2. There are two variables that are categorical. It can be ordinal that has been grouped into ordinal categories.
3. The test group must be independent.
4. The data in the cells should be frequencies our count.
5. The categories are mutually exclusive
6. Each subject may contribute data to one and only one cell in the chi square.

Test results - Chi square

Table 6: Oberseved Values

	None	Moderate	Intese	Total Row
Male	127	255	125	507
Female	116	242	135	493
Total Column	243	497	260	1000

The Chi-square statistic can be calculated using approximation formula below,

$$\chi^2_{(r-1)(c-1)} = \sum_{i=1}^{r}\sum_{j=1}^{c}(O_{i,j} - E_{i,j})^2/E_{i,j}$$

where

- the degrees of freedom is $v = (r-1)(c-1)$
- O_{ij} is the observed value on the ith row and jth column.
- E_{ij} is the expected value on the ith row and jth column. To calucate E_{ij}, it is the sum of the ith row and sum of the jth row and dividing by the total number of observations, $E_{ij} = (\sum_{i=1}^{r} O_{i,j} * \sum_{j=1}^{c} O_{i,j})/N$ where
  - $\sum_{i=1}^{r} O_{i,j}$ is the sum of the ith row
  - $\sum_{j=1}^{c} O_{i,j}$ is the sum of the jth column

The expected values can then be calculated and is given in the table below,

Table 7: Expected Values

	None	Moderate	Intense
Male	123.201	251.979	131.82
Female	119.799	245.021	128.18

That being said, it would be more quicker and accurate to let R compute the p-value, chi statistic and the critical value.

```
##
##  Pearson's Chi-squared test
##
## data:  df2
## X-squared = 1.0268, df = 6, p-value = 0.9846

##
## CHI SQUARE VALUES

## p_value =  0.985

## chi_sq_statistic = 1.026799

## critical_value = 12.59159
```

Decision Chi square

```
## Do not reject NULL hypothesis
```

Conclusion Chi square

```
## CONCLUSION
```

```
## There is no evidence that there is a relationship
## between gender and the level of physical activity
## at a 5% signficance
## level.
```

Conclusions

From our data set, we were able to conduct three different statistical test to answer three research questions. The statistical tests has shown that,

1. There is a positive linear relationship between height and weight.
2. The mean of male and female heights are not the same.
3. There are no association between the genders and physical activities.

References

Wikipedia (2021) Handbook (2021) thebmj (2021) JMP (2021) DataNova (2021) McHugh (2013)

DataNova. 2021. "Independent t-Test Assumptions." URL:https://www.datanovia.com/en/lessons/t-test-assumptions/independent-t-test-assumptions/.

Handbook, Engineering Statistic. 2021. "Two-Sample t-Test for Equal Means." URL:https://www.itl.nist.gov/div898/handbook/eda/section3/eda353.htm.

JMP. 2021. "Regression Model Assumptions." URL:https://www.jmp.com/en_au/statistics-knowledge-portal/what-is-regression/simple-linear-regression-assumptions.html.

McHugh, Mary L. 2013. "The Chi-Square Test of Independence." URL:https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900058/.

thebmj. 2021. "The Chi Squared Tests." URL:https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/8-chi-squared-tests.

Wikipedia. 2021. "Simple Linear Regression." URL:https://en.wikipedia.org/wiki/Simple_linear_regression.