STAT1378: Project

Manjiri Ramesh Pendharkar

2022-11-04

1 Abstract

The main aim of this report is to conduct an exploratory and statistical tests analysis to see if there is any evidence of a relationship between the variables height, weight, gender and levels of physical activity using project2022.csv dataset containing observations of men and women aged 26-45. Firstly, a linear regression model was used in order to investigate whether a linear relationship exists between height and weight. Secondly, a t-test was utilised in order to do compare mean height of males and females,i.e., to investigate whether gender has an affect on the height. Lastly, a chi-squared test is utilised to delve into whether there is any association between gender and the amount of physical activity.

The report addresses and summarises the results of the stated research questions by using appropriate statistical tests.

2 Introduction

In this statistical report, we provide answers to the following research questions:

1. **Is there a linear relationship between height and weight?**
   Linear regression attempts to model the relationship between X = height and Y = weight, by fitting a linear equation to see whether we can predict weight of a person by their height.
2. **Is the mean height of male and female the same?**
   In this scenario, we assume equal variances between male and female heights.
   Since we assume equal variances, we have conducted two sample t-test to investigate whether gender has an effect on average heights of males and females.
3. **Is there any association between gender and the amount of physical activity?**
   The Chi-square test of independence test is used to determine whether gender and the amount of physical activity are likely to be related or not.

3 Data

In this report, **project2022.csv** dataset is used to answer research questions defined above.

The datset contains observations of men and women aged 26-45 with information on the following variables:

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

4 Methodology

4.1 Linear regresion test to investigate the relationship between height and weight

According to Boslaugh (2012), the analysis of relationship between independent and dependent variable is investigated with the aid of linear regression. We set height as the independent variable (x-axis) and weight as the dependent variable (y-axis).

Hypothesis

- Null hypothesis: There is no significant linear relation between height and weight (slope is equal to zero).
  $H_0 : \beta = 0$
- Alternate hypothesis: There is a significant linear relation between height and weight (slope is not equal to zero).
  $H_a : \beta \neq 0$

Assumptions

Ross and Willson (2017) states that for linear regression, the following assumptions should be met:

- The relation in the population is linear.
  Scatter plot is used to see whether the relation between the height and weight is linear or not.
- The residuals (errors) are drawn from a normal distribution.
  Histogram is used to check if the residuals follow a bell-shaped distribution.
- The residuals have a constant standard deviation.
  We plot the residuals against the fitted values (i.e., the predicted values). If the residuals have a constant standard deviation, the spread of the residuals around zero will be constant across the range of the predicted values.
- The two samples are independent and the observations in each sample are independent of each other

'Least-squares regression line

$$\hat{weight} = a + b * height$$

where a = y-intercept, b = slope of the line

Test statistic

$$t = \frac{b}{se_b}$$

where b=slope and se= standard error of the slope

Degrees of freedom

$$df = n - 2$$

Conducting linear regression test in R

We can conduct linear regression in R using lm function. We pass the weight as dependent and height as independent variable in the lm function.

Decision of test

We observe the p value of the lm test to decide the outcome of the test.

- We reject the null hypothesis if the p value is $<= 0.05$.
- We do not reject the null hypothesis if the p value is > 0.05.

4.2 Two sample t-test to explore differences in height between males and females

According to Ross and Willson (2017), the Two sample t-test was used to determine if the means of two sets of data are significantly different from each other or not. The Two sample t-test was conducted to analyze whether the mean height for males and females are same or not.

Hypothesis

- Null hypothesis: There is no difference between the average height of males and females.
  $H_0 : \mu_{mh} = \mu_{fh}$
- Alternate hypothesis: There is a difference between the average height in males and females
  $H_a : \mu_{mh} \neq \mu_{fh}$

Assumptions

Ross and Willson (2017) states that for two-sample t-test, the following assumptions should be met:

- The sample means being compared for two populations follow normal distribution.
  Normal Quantile-Quantile plot are used to give us a subjective measure of how closely our data (female heights and male heights) match a Normal distribution. The further the plot is from a straight line, the less confident we are about if our data are Normally distributed.
- The two samples are from populations with equal/same variances.
  Boxplots are used to see if two samples have equal spread or not.
- The two samples are independent and the observations in each sample are independent of each other

Pooled Variance

According to Xuemao and Zoe (2019), in two sample t-test, we use pooled variance because we assume that the two samples are from populations with equal variances.

$$s_p = \sqrt{\frac{(n_{mh}-1)s_{mh}^2 + (n_{fh}-1)s_{fh}^2}{n_{mh}+n_{fh}-2}}.$$

where n_{mh}, n_{fh} are the sample size for the two populations and s_{mh}^2 & s_{fh}^2 are the two sample variances.

Degrees of freedom

$$df = n_{mh} + n_{fh}$$

Test statistics

$$t_{obs} = \frac{\bar{x}_{mh} - \bar{x}_{fh}}{s_p\sqrt{1/n_{mh} + 1/n_{fh}}}$$

where $\overline{x}_{mh}$, $\overline{x}_{fh}$ are the sample mean for the two populations and s_p is the pooled variance.

Conducting two sample t-test in R

We can conduct two sample t-test in R using t.test() function. We pass the gender and height to the t.test() function and specify var.equal=TRUE as we assume equal variances between our two populations.

Decision of test

We observe the p value of the t-test to decide the outcome of the test.

- We reject the null hypothesis if the p value is <= 0.05.
- We do not reject the null hypothesis if the p value is > 0.05.

4.3 Chi-squared test to investigate the association between the amount of physical activity and gender

According to Xuemao and Zoe (2019), the chi-square test of independence evaluates whether there is an association between the categories of the two variables. The chi-squared test is used to determine whether there is any association between gender and the amount of physical activity.

Hypothesis

- Null hypothesis: The gender and level of physical activity are independent of each other
- Alternate hypothesis: The levels of physical activity are dependent on gender

Assumptions

The chi-squared test is only valid if all expected counts are >= 5.

Expected values

$$E = \frac{row.sum * column.sum}{grand.total}$$

Degrees of freedom

$$df = (rows - 1) * (columns - 1)$$

Chi-squared statistics

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

where O=observed value and E=expected value

Conducting chi-squared test in R

We can conduct chi-squared test in R using chisq.test function and pass the below data about gender and amount of phyical activity.

Decision of test

We observe the p value of the chi-squared to decide the outcome of the test.

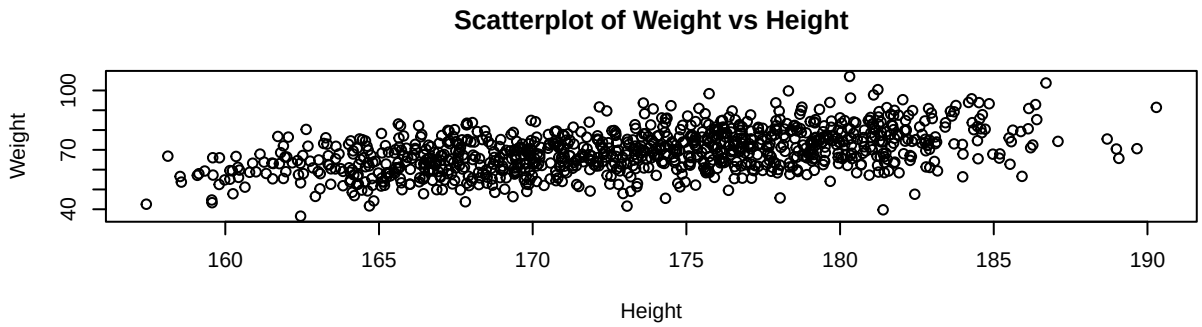
- We reject the null hypothesis if the p value is <= 0.05.
- We do not reject the null hypothesis if the p value is > 0.05.

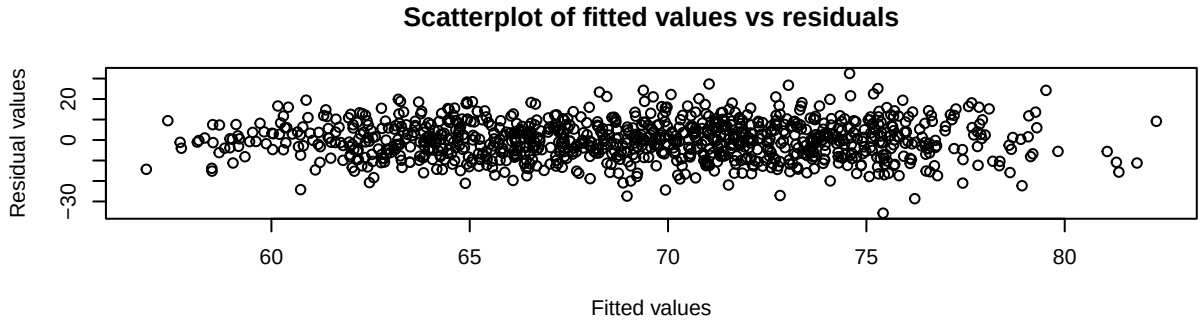
The dplyr and broom package of Robinson, Hayes, and Couch (2022) and Wickham et al. (2022) is used in the following section to display the test results.

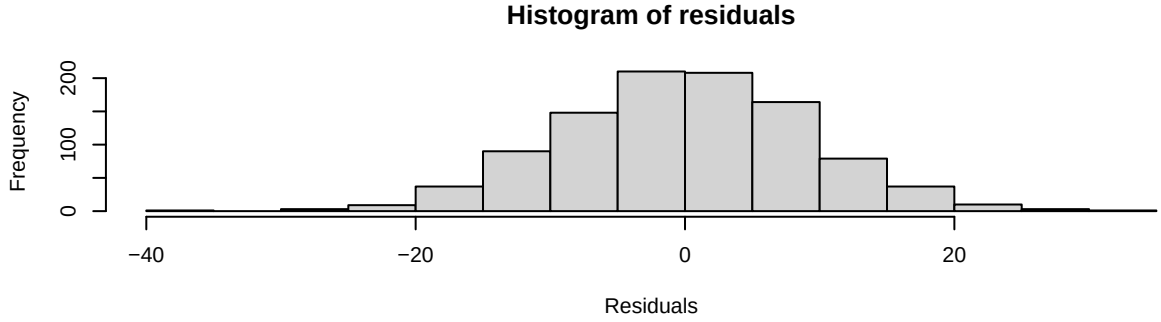
5 Results

5.1 Results of Linear regresion test to investigate the relationship between height and weight

```
## The script is going to read the file project2022.csv and perform a linear regression to see
##   if there is a linear relationship between height and weight.
##
##
##   HYPOTHESIS:
##  Null hypothesis is H0 : slope = 0
##  Alternative hypothesis is H1 : slope != 0
##
## ASSUMPTIONS:
##  Please check the plots.
```



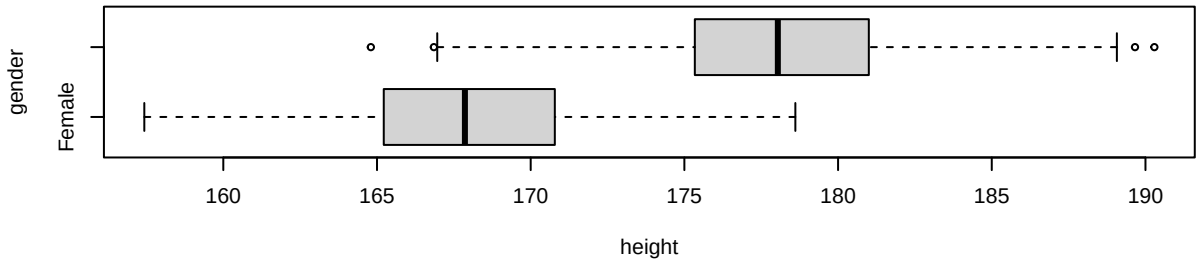


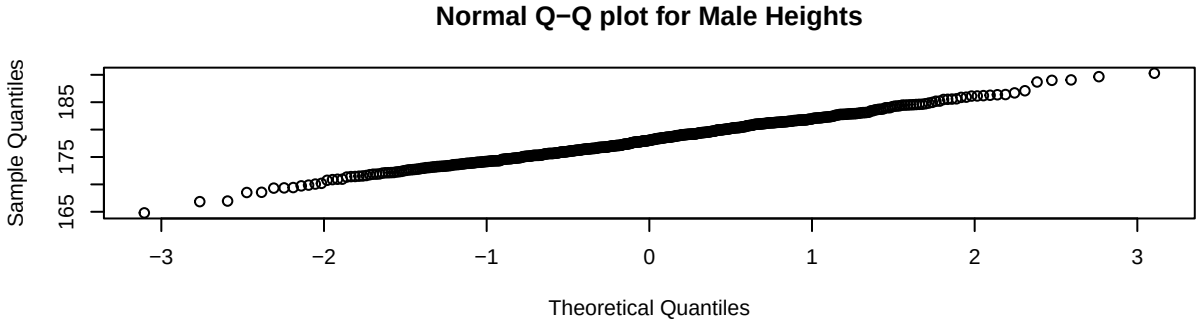


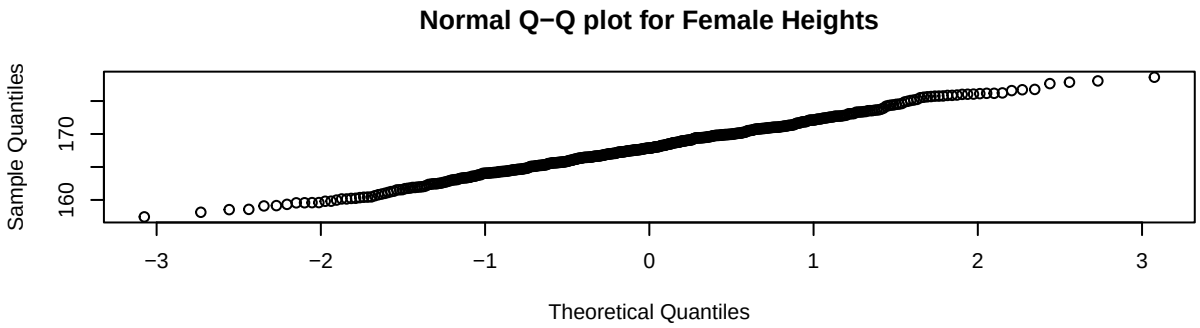

```
## LINEAR MODEL SUMMARY REPORT:
##  Estimated Regression line : Weight =  -65.15096  + ( 0.7749303  * Height)
##  Intercept =  -65.15096
##  Estimated slope =  0.7749303
##  t value =  17.33413
##  degree of freedom =  998
##  95% CIs  =  0.6872029  ,  0.8626578
##  p-value =  4.807171e-59
##
##  DECISION:
##  Since the p-value is  4.807171e-59  which is less than 0.05, we reject the null hypothesis.
##
##  CONCLUSION:
##  As the p-value is  4.807171e-59  and estimated slope is  0.7749303
##  there is a statistically      significant evidence of a positive linear relationship between
##     the predictor variable Height and the response variable Weight. For each unit-increase in
##     Height, Weight increases by 0.7749303
```

5.2 Results of Two sample t-test to explore differences in height between males and females

```
##  The script is going to read the file project2022.csv and perform two sample t test to see
##   if the mean height of male and female are same.
##
##
##   HYPOTHESIS:
##  Null hypothesis: There is no difference between the average height in males and females.
##  Alternate hypothesis: There is a difference between the average height in males and females
##
## ASSUMPTIONS:
##  Please check the plots.
```








```
## TWO SAMPLE T-TEST SUMMARY REPORT:
##  t statistics =  -39.84779
```

```
##  degree of freedom =  998
##  95% CI=  -10.64981  ,  -9.650119
##  p-value =  1.531954e-208
##  Estimated female height =  167.98
##  Estimated male height =  178.1299
##
##  DECISION:
##  Since the p-value is  1.531954e-208  which is less than 0.05, we reject the null hypothesis.
##
##  CONCLUSION:
##  As the p-value is  1.531954e-208
##  there is statistically significant evidence that the        mean height of females
##     is less than the mean height of males.
##  Estimated female height =  167.98
##  Estimated male height     =  178.1299
```

5.3 Results of Chi-squared test to investigate the association between the amount of physical activity and gender

```
## The script is going to read the file project2022.csv and perform chi squared test to see
##   if there is any association between gender and the amount of physical activity.
##
## HYPOTHESIS:
##  Null hypothesis: The gender and level of physical activity are independent of each other.
##  Alternate hypothesis: The levels of physical activity are dependent on gender
##
## ASSUMPTIONS: Checking if the expected values are >= 5
## Expected number of males and females participating in each level of physical activity
##            Female   Male
##   Intense  118.75 131.25
##   Moderate 233.70 258.30
##   None     122.55 135.45
##
##  CHI SQUARED TEST SUMMARY REPORT:
##  statistics =  3.226111
##  degree of freedom =  2
##  p-value =  0.1992778
##
##  DECISION:
##  Since the p-value is  0.1992778  which is greater than 0.05, we do not reject the null hypothesis.
##
##  CONCLUSION:
##  As the p-value is  0.1992778
##  there is statistically significant evidence that gender and levels of physical activity are
## independent of each other. Meaning that there exists no association between the gender
## and level of physical activity.
```

6 Conclusion

6.1 Conclusion of Linear regresion test to investigate the relationship between height and weight

```
##
##
##  CONCLUSION:
##  As the p-value is  4.807171e-59  and estimated slope is  0.7749303
##  there is a statistically      significant evidence of a positive linear relationship between
##     the predictor variable Height and the response variable Weight. For each unit-increase in
##     Height, Weight increases by 0.7749303
```

6.2 Conclusion of Two sample t-test to explore differences in height between males and females

```
##
##
##  CONCLUSION:
##  As the p-value is  1.531954e-208
##  there is statistically significant evidence that the        mean height of females
##     is less than the mean height of males.
##  Estimated female height =  167.98
##  Estimated male height     =  178.1299
```

6.3 Conclusion of Chi-squared test to investigate the association between the amount of physical activity and gender

```
##
##
##  CONCLUSION:
##  As the p-value is  0.1992778
##  there is statistically significant evidence that gender and levels of physical activity are
## independent of each other. Meaning that there exists no association between the gender
## and level of physical activity.
```

References

Boslaugh, S. 2012. *Statistics in a Nutshell*. 2nd ed. O'Reilly Media.

Robinson, D., A. Hayes, and S. Couch. 2022. *Broom: Convert Statistical Objects into Tidy Tibbles*. URL: https://CRAN.R-project.org/package=broom.

Ross, A., and V. Willson. 2017. *Basic and Advanced Statistical Tests*. 1st ed. Sense Publishers.

Wickham, H., R. Francois, L. Henry, and K. Muller. 2022. *Dplyr: A Grammar of Data Manipulation*. URL: https://CRAN.R-project.org/package=dplyr.

Xuemao, Z., and M. Zoe. 2019. "Using r in Teaching Introductory Statistics." *International Electronic Journal of Mathematics Education* 14 (3): 59–61.