Report

Nicholas Vella

2022-11-03

Abstract

This report investigates trends between physical attributes of people. The research compares data between height and weight, gender and height, and gender and physical activity using statistical tests. The results find a linear relationship between height and weight, a difference in height between genders, and a lack of association between gender and physical activity.

Introduction

The project aims to answer the following research questions:

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

These research questions are important as relationships in physical attributes have many use cases. Trends in physical attributes between genders can be used to determine average and normal ranges in medical, or help determine the size of clothing for different genders. Trends between height and weight are useful as they are used to diagnose obesity with the BMI. The data used in this report were sampled randomly.

Methods

The experiment used hypothesis testing with the appropriate statistical analysis to address each research question. This design was chosen as the research questions were different, requiring different statistical analysis top assess the probability of the hypotheses. The subjects were 1000 randomly sampled people. The data was provided in the "project2022" data-set and did not give light to any methodology to the selection process of the subjects.The data was given in a tidy format that did required very little wrangling. The data was managed by creating a package to store the data set, and it was used as given to input into the functions required to analysed the data.

Five variables were measured:

- ID of the subject
- Gender of the subject
- Height of the subject
- Weight of the subject
- Intensity of physical exercise completed by the subject ranked as:
  - Intense
  - Moderate
  - None

The first research question was addressed using a linear regression statistical test, as this is the appropriate test to find linear relationships between variables. The second research question was addressed by the use of the Student's T-test, as this is appropriate to test the means of two samples. This was used instead of the Welch T-test, as the variances of the means were assumed to be equal. The third research question was addressed with a chi-squared test of independence, as this is the appropriate statistical test to check for associations between variables in categorical data.

Each test for the three research questions following the following format, and used R to complete all the calculations:

- Stating null and alternate hypotheses
- Checking the assumptions
- Providing descriptive statistics
- Providing a "decision" of the test
- Providing a "conclusion" of the test

The following packages were used in R to provide this output in the making of the "project" package:

- Stats package to run the statistical tests (R Core Team 2022)
- ggplot2 package to provide graphical assumptions (Wickham 2016)
- dplyr package to filter data to input into test/graphs and provide tabular output (Wickham et al. 2022)
- glue package to interpret conclusions and decisions from the input data-set (Hester and Bryan 2022)
- magrittr package for the pipe operator (Bache and Wickham 2022)
- broom package to make statistical output tidy (Robinson, Hayes, and Couch 2022)

Results

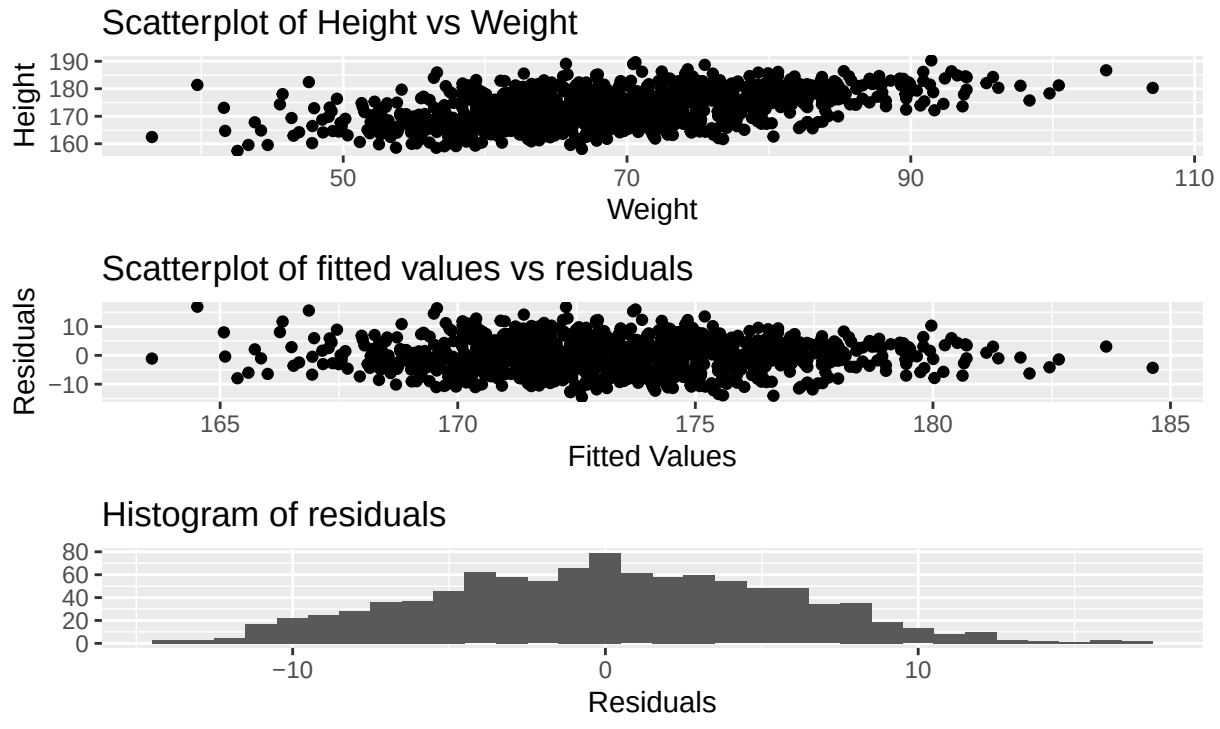
Exploratory Data Analysis

Hypotheses for research question 1,2, and 3 respectively (generated in R, using the package "project"):

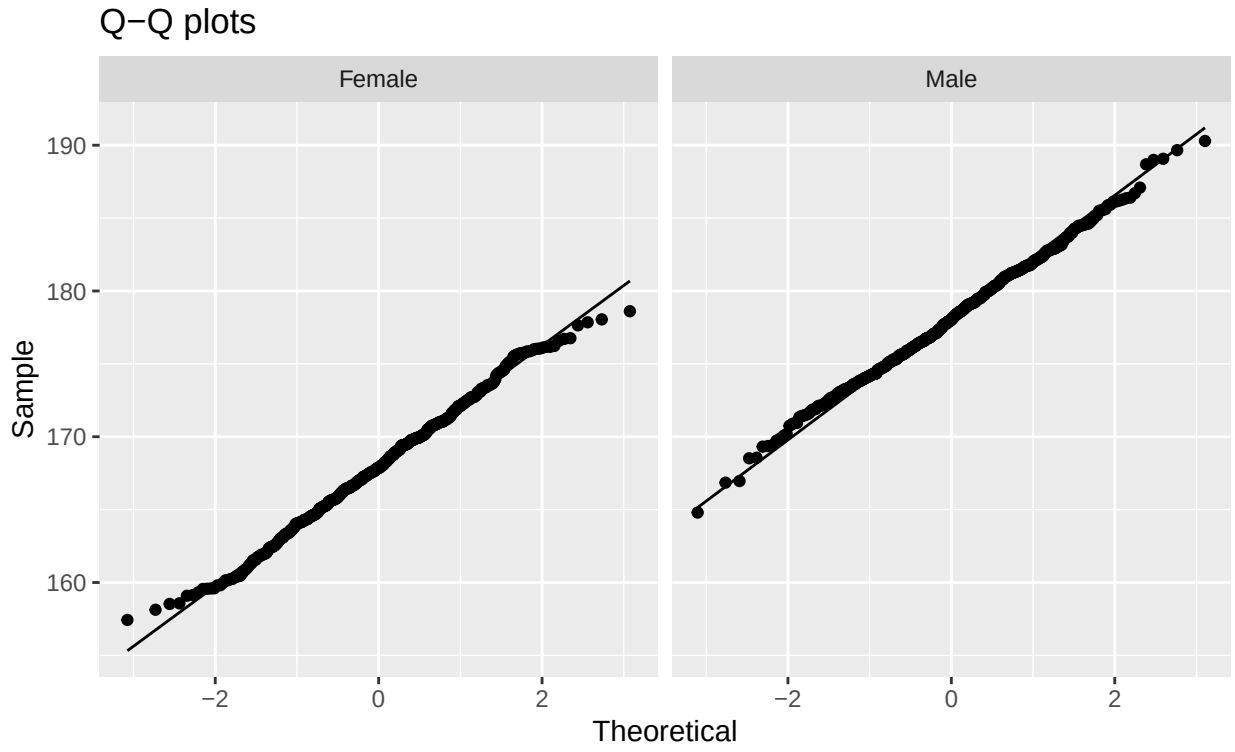
```
##
## HYPOTHESIS
## Null Hypothesis: H0 is beta is equal to zero.
## Alternate Hypothesis: H1 is beta is not equal to zero.
## (Where beta is the slope parameter in the model)

##
## HYPOTHESIS
## Null Hypothesis: H0 is beta is the sample means are equal.
## Alternate Hypothesis: H1 is beta is the sample means are not equal

##
## HYPOTHESIS
## Null Hypothesis: H0 is the two variables are independent.
## Alternate Hypothesis: H1 is there is a relationship between the variables.
```

Assumptions for research questions 1,2, and 3 respectively (generated by R). The ggplot2 package (Wickham 2016) was used to generate graphs. The dplyr (Wickham et al. 2022) and magrittr packages (Bache and Wickham 2022) were used to manipulate the data:

```
##
## ASSUMPTIONS
```




```
##
## ASSUMPTIONS
```




```
##   Male Variance Female Variance
##        15.56823          16.85590

##
## ASSUMPTIONS
## Expected Values:
##
##           Intense Moderate   None
##   Female   118.75    233.7 122.55
##   Male     131.25    258.3 135.45
```

Checking the assumptions for the linear model (research question 1):

- The observations are independent and identically distributed
- Normality of residuals. This can be seen from the distribution of the histogram of residuals from the R output.
- Uncorrelated residuals and fitted values. This can be seen from "fitted vs residuals" scatterplot.

Checking the assumptions for the t-test (research question 2):

- Equal variances: the variances of the samples clearly satisfy the rule of thumb
- Normality of observations is shown by the points following the line in the q-q plots.
- There are 1000 observations and therefore enough to satisfy the Central Limit Theorem. However, a power calcualtions has not been done.

Checking the assumptions for the chi-squared test (research question 3):

- From the expected values output, each expected value is >5 in all cells.
- Both variables are categorical (gender, physical activity).

Outcomes

Results of the statistical tests (calculated by R). The stats package (R Core Team 2022) was used to run the statistical tests:

```
##
## MODEL - Linear Regression
## Beta Hat: 0.2986132
## df: 998
## P-value: 4.807171e-59
## 95% CI: (0.264808122696613,0.332418278591899)

##
## DECISION
## Reject NULL hypothesis: YES

##
## MODEL
## Method: Two Sample t-test
## Female Estimate: 167.98
## Male Estimate: 178.1299
## Test Statistic: -39.84779
## df: 998
## 95% CI: (-10.6498088725359,-9.65011894701302)
## P-value: 1.531954e-208

##
## DECISION
## Reject NULL hypothesis: YES

##
## MODEL
## Method: Pearson's Chi-squared test
## Test Statistic: 3.226111
## df: 2
## P-value: 0.1992778

##
## DECISION
## Reject NULL Hypothesis: NO
```

Conclusion

Conclusions to research questions 1,2, and 3 respectively (as calculated in R). The (Hester and Bryan 2022) package was used in the functions to interpret the conclusion from the data-set:

```
##
## CONCLUSION
## In Summary: There is evidence that the slope (beta) is different than 0.
    There is a significant linear relationship between height and weight. For
    each unit-increase in weight, height increases by 0.2986.
```

```
##
## CONCLUSION
## In Summary: There is evidence that the sample means are different. There is
     a difference in height between genders.
```

```
##
## CONCLUSION
## In Summary: There is no evidence that the variables are dependent.There is
    no association between gender and phys.
```

Limitations

One limitation of the study was the sample size. There was no power study conducted before sampling the subjects to determine the required sample size for the tests. The external validity of the results may be limited as the results may not be representative of the general population. The categorising of physical activity to “intense”, “moderate”, “none” provides very little information, as many other studies would take other measurements, such as “duration” for example. It was not also not clear how the physical activity was recorded. Therefore the ability to reproduce the results of this chi-squared, using the same methodology would be very limited.

References

Bache, Stefan Milton, and Hadley Wickham. 2022. *Magrittr: A Forward-Pipe Operator for r*. https://CRAN.R-project.org/package=magrittr.

Hester, Jim, and Jennifer Bryan. 2022. *Glue: Interpreted String Literals*. https://CRAN.R-project.org/package=glue.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, David, Alex Hayes, and Simon Couch. 2022. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2022. *Dplyr: A Grammar of Data Manipulation*.