STAT1378 Part I Report

ADAB GILL

Abstract

This report examines the relationships among height, weight, gender, and physical activity levels within a sample dataset of men and women aged 26–45. The analysis utilizes linear regression, t-tests, and chi-square tests to answer questions about linear relationships, mean differences, and associations (Xie, Allaire, and Grolemund 2015).

Introduction

The study investigates three main questions: 1. Is there a linear relationship between height and weight? 2. Does mean height differ between males and females? 3. Is there an association between gender and physical activity level?

Statistical methods like linear regression, t-tests, and chi-squared tests are employed, aiming to answer these questions rigorously.

Methods

Data Description

The dataset includes 1000 observations, each with the following variables: - **ID**: Unique identifier for each participant. - **Gender**: Male or Female. - **Height**: Participant height in centimeters. - **Weight**: Participant weight in kilograms. - **Physical Activity**: Level of physical activity (None, Moderate, Intense).

Statistical Tests

- **Linear Regression**: To investigate the relationship between height and weight.
- **t-Test**: To compare mean heights of males and females.
- **Chi-squared Test**: To examine the association between gender and physical activity.

Analysis and R Code

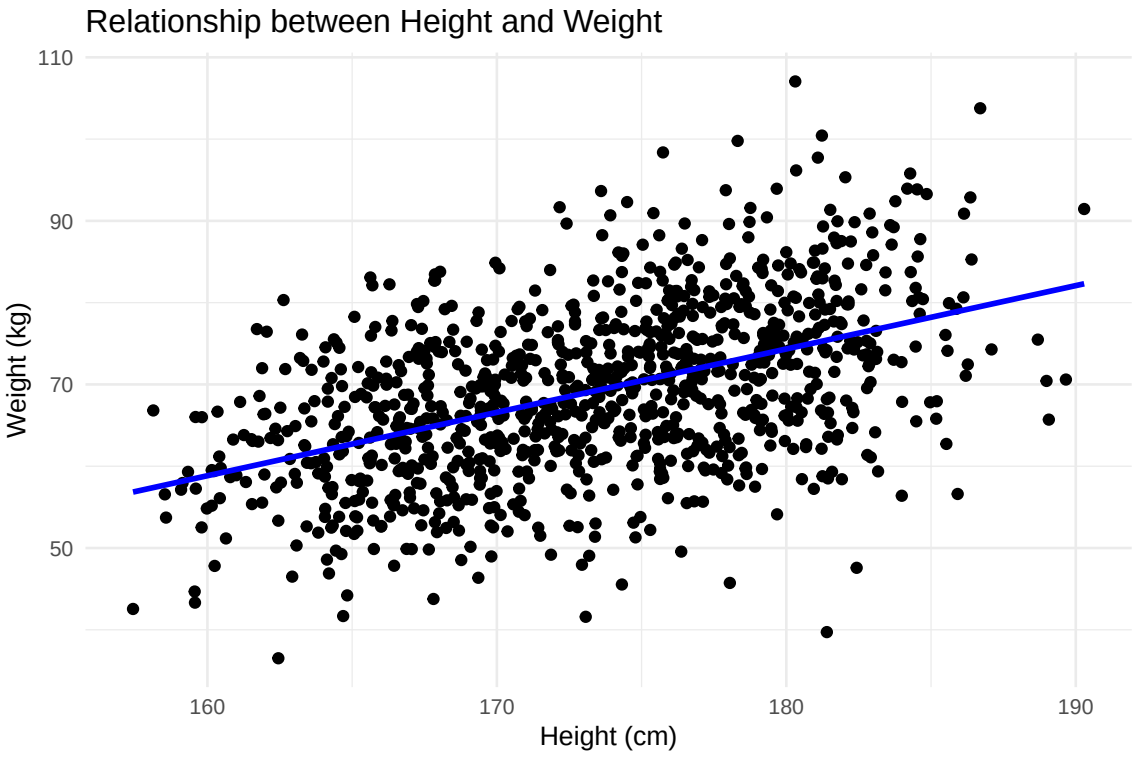
Question 1: Linear Relationship between Height and Weight

```
##
## Call:
## lm(formula = weight ~ height, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.701  -6.331   0.033   6.320  32.473
##
## Coefficients:
```

Table 1: Linear Regression Summary for Height and Weight

	Estimate	Std_Error	t_value	p_value
(Intercept)	-65.1509590	7.7532387	-8.403064	0
height	0.7749303	0.0447055	17.334135	0

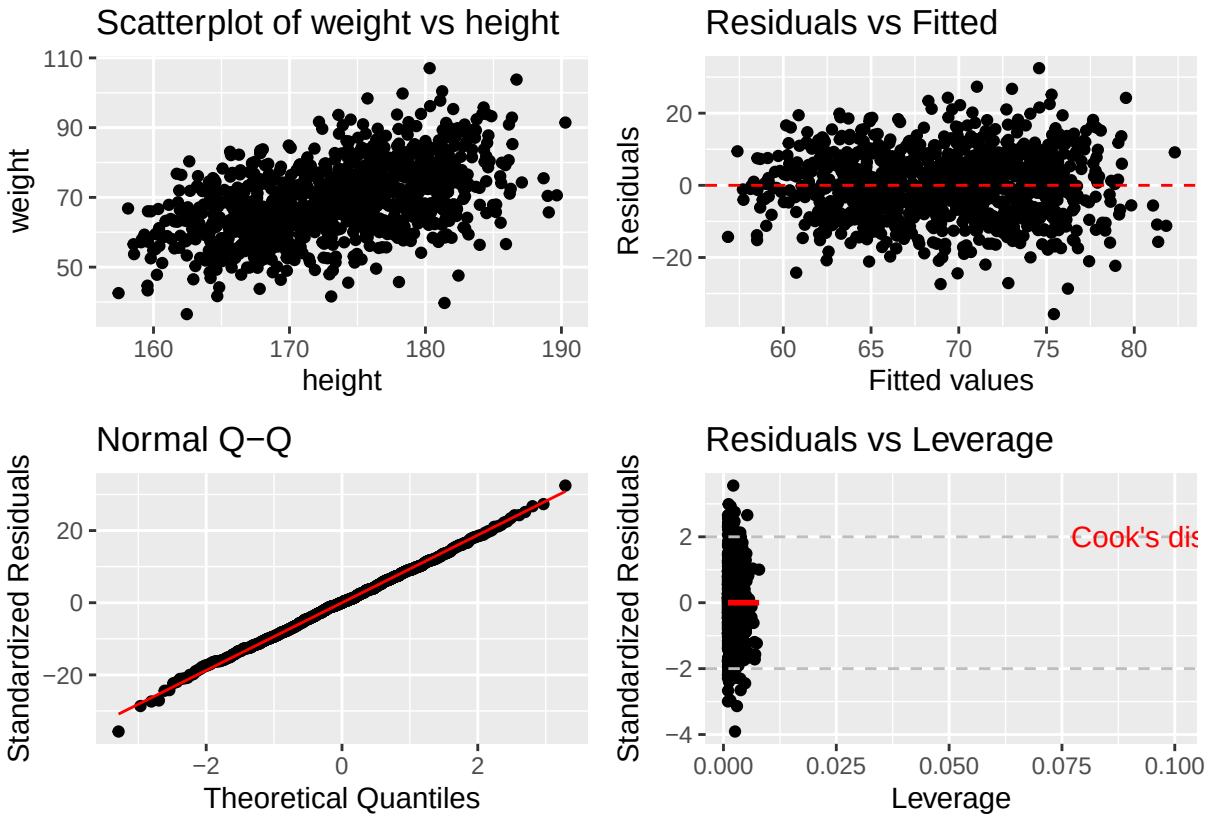
```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -65.15096    7.75324  -8.403   <2e-16 ***
## height        0.77493    0.04471  17.334   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.144 on 998 degrees of freedom
## Multiple R-squared:  0.2314, Adjusted R-squared:  0.2306
## F-statistic: 300.5 on 1 and 998 DF,  p-value: < 2.2e-16
```



Results for Linear Regression

Based on the linear regression analysis, a significant relationship was found between height and weight, with a p-value of $4.8071712 \times 10^{-59}$. The slope of the line suggests that with each centimeter increase in height, weight changes by approximately 0.77 kg (Moore, McCabe, and Craig 2016).

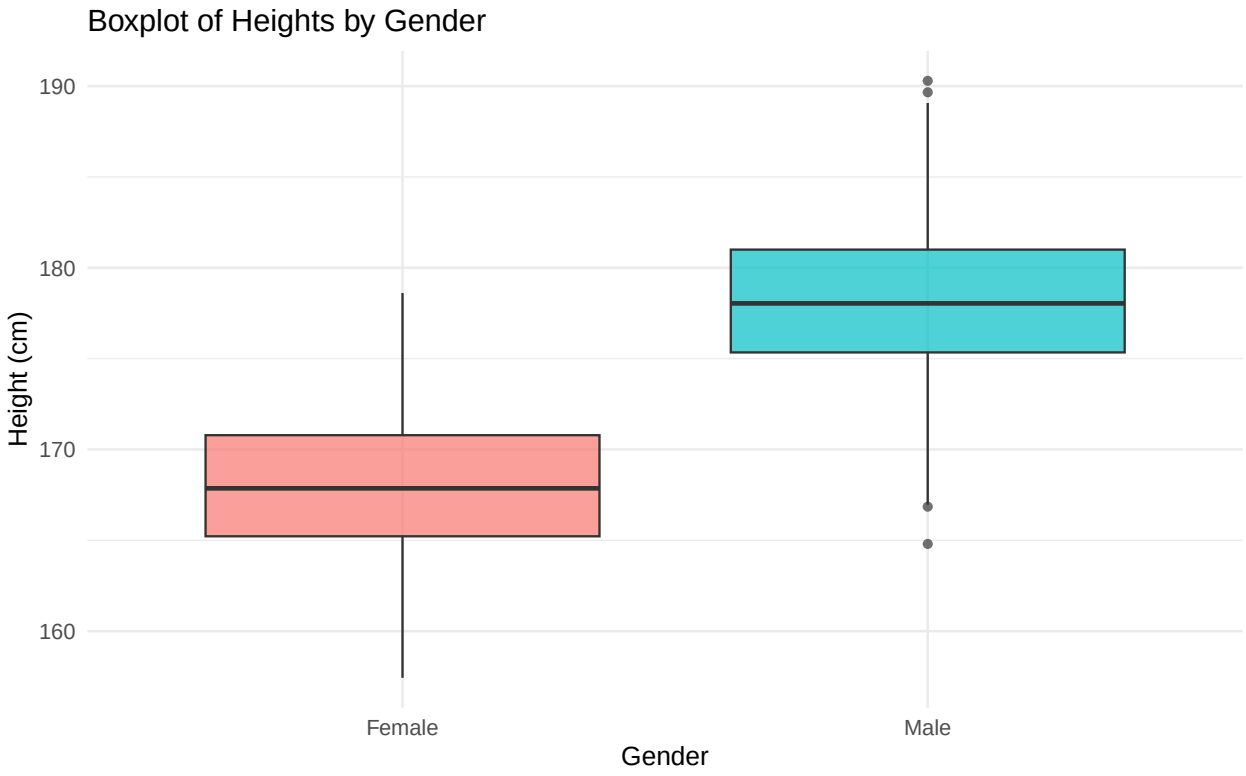
Assumptions



Question 2: Difference in Mean Height between Males and Females

Table 2: t-Test Summary for Mean Height Comparison

	Statistic	Parameter	p_value	Confidence_Interval	Mean_Female	Mean_Male
t	-39.76867	978.8659	0	-10.65 - -9.65	167.98	178.1299



Assumptions

```
# Your function call here
assumptions_ttest <- function(file_name, group_var, outcome) {
  data <- read.csv(file_name)

# Normality check using Q-Q plots for each group
plot_qq <- ggplot(data, aes(sample = get(outcome))) +
  stat_qq() +
  stat_qq_line(color = "red") +
  facet_wrap(as.formula(glue("~ {group_var}"))) +
  labs(title = glue("Q-Q Plot for {outcome} by {group_var} Levels"), x = "Theoretical Quantiles", y = "S

# Variance check using boxplot for each group
plot_box <- ggplot(data, aes_string(x = group_var, y = outcome, fill = group_var)) +
  geom_boxplot(alpha = 0.7) +
  labs(title = glue("Boxplot of {outcome} by {group_var} Levels"), x = group_var, y = outcome) +
  theme_minimal()

file_name <- "project.csv"  # Replace with the path to your file
predictor <- "height"       # Column name for the predictor variable in your data
outcome <- "weight"         # Column name for the outcome variable in your data

# Display plots in a 2x1 layout
print(plot_qq)
print(plot_box)
}
```

Results for t-Test

The Welch Two Sample t-test shows a statistically significant difference in mean height between males and females, with a t-value of -39.77 and a p-value less than 2.2e-16. This result suggests that the mean height of males (178.13 cm) is significantly greater than the mean height of females (167.98 cm). The 95% confidence interval for the difference in means is between -10.65 and -9.65 cm, indicating that males are, on average, taller than females in this sample. Since the p-value is much smaller than 0.05, we reject the null hypothesis and conclude that there is a statistically significant difference in height by gender. The negative t-value and confidence interval (both in the negative range) confirm that males are, on average, taller than females. The interval provides an estimated range for the difference in means, reinforcing that males are taller on average (Agresti 2018).

The hypotheses for the chi-squared test are:

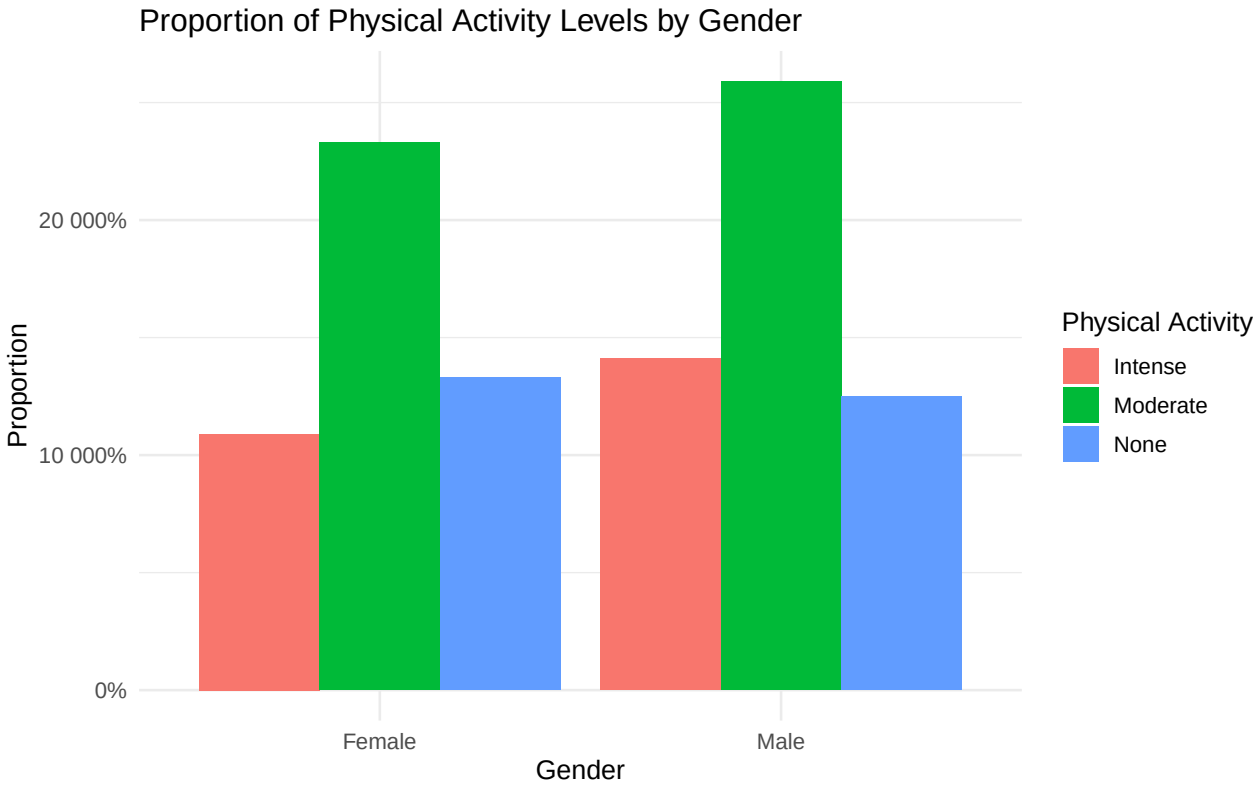
$$H_0 : \text{Gender and Physical Activity Level are independent}$$

$$H_1 : \text{Gender and Physical Activity Level are not independent}$$

Question 3: Association between Gender and Physical Activity Level

Table 3: Chi-Squared Test Summary for Gender and Physical Activity

	Statistic	Degrees_of_Freedom	p_value
X-squared	3.226111	2	0.1992778



Assumptions

Results for Chi-squared Test

The chi-squared test for association between gender and physical activity level shows a chi-squared value of 3.23 with 2 degrees of freedom and a p-value of 0.1993. Since the p-value is greater than 0.05, we fail to reject the null hypothesis. This suggests that there is no statistically significant association between gender and physical activity level in this sample.

The p-value of 0.1993 is above the typical significance level of 0.05, indicating that any observed differences in physical activity level by gender are likely due to random chance. Failing to reject the null hypothesis implies that gender does not have a significant association with physical activity level in this data set.

Conclusion

1. **Height and Weight Relationship**: The linear regression analysis indicates a statistically significant relationship between height and weight. This suggests that taller individuals in the sample tend to have higher weights, supporting a positive association between these two variables.
2. **Gender Differences in Height**: The t-test results show a statistically significant difference in mean height between males and females, with males being taller on average. This finding is consistent with general observations of height differences by gender.
3. **Gender and Physical Activity Level**: The chi-squared test for association between gender and physical activity level does not indicate a statistically significant relationship. This implies that, within this sample, physical activity levels do not differ significantly between males and females.

References

Agresti, Alan. 2018. *Statistical Methods for the Social Sciences*. 5th ed. Pearson.

Moore, David S., George P. McCabe, and Bruce A. Craig. 2016. *Introduction to the Practice of Statistics*. 9th ed. W.H. Freeman. https://www.macmillanlearning.com/college/us/product/Introduction-to-the-Practice-of-Statistics/p/1319013384.

Xie, Yihui, J. J. Allaire, and Garrett Grolemund. 2015. *R Markdown: The Definitive Guide*. Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown/.

Agresti, Alan. 2018. *Statistical Methods for the Social Sciences*. 5th ed. Pearson.

Moore, David S., George P. McCabe, and Bruce A. Craig. 2016. *Introduction to the Practice of Statistics*. 9th ed. W.H. Freeman. https://www.macmillanlearning.com/college/us/product/Introduction-to-the-Practice-of-Statistics/p/1319013384.

Xie, Yihui, J. J. Allaire, and Garrett Grolemund. 2015. *R Markdown: The Definitive Guide*. Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown/.