Problem Set 2

Matt Krasnow

Due Friday, September 20, 2024 at 11:59pm

Problem set policies. *Please provide concise, clear answers for each question while making sure to fully explain your reasoning. For problems asking for calculations, show your work in addition to clearly indicating the final answer. For problems involving* `R`*, be sure to include the code and output in your solution.*

Please submit the PDF of your knit solutions to Gradescope and be sure to assign which pages of your solution correspond to each problem. Make sure that the PDF is fully readable to the graders; e.g., make sure that lines don't run off the page margin.

We encourage you to discuss problems with other students (and, of course, with the teaching team), but you must write your final answer in your own words. Solutions prepared "in committee" are not acceptable. If you do collaborate with classmates on a problem, please list your collaborators on your solution. Be aware that simply copying answers found online, whether human-generated or machine-generated, is a violation of the Honor Code.

Question 1: Creating our Dataset

We will create our own dataset to try to replicate the analyses we did in class with the human height data (the application in which the method of regression was developed). Start by filling out the questionnaire here:

https://forms.gle/VQJJsNz4Vjs9RFAk6

Question 2: Some Algebra

Let $x_1, x_2, \ldots, x_n$ be any numbers and define $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$. Prove the following:

(a) $\text{argmin}_a \sum_{i=1}^n (x_i - a)^2 = \bar{x}$

(b) $(n-1)s^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$

(c) $\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}$

2. Prove the following:

(a)

$$\arg\min_a \sum_{i=1}^n (x_i - a)^2 = \bar{x}$$

Proof:

To find the value of a that minimizes the sum of squared differences, we differentiate with respect to a and set the derivative to zero:

$$\frac{d}{da}\sum_{i=1}^n (x_i - a)^2 = 0$$

Applying the chain rule:

$$\sum_{i=1}^n \frac{d}{da}(x_i - a)^2 = \sum_{i=1}^n -2(x_i - a) = 0$$

Simplifying:

$$-2\sum_{i=1}^n (x_i - a) = 0$$

$$-2\left(\sum_{i=1}^n x_i - na\right) = 0$$

Dividing both sides by -2:

$$\sum_{i=1}^{n} x_i - na = 0$$

$$n\bar{x} - na = 0$$

Dividing both sides by n:

$$\bar{x} - a = 0$$

$$\implies a = \bar{x}$$

To confirm this critical point is a minimum, we compute the second derivative:

$$\frac{d^2}{da^2} \sum_{i=1}^{n} (x_i - a)^2 = \sum_{i=1}^{n} \frac{d}{da} \left[-2(x_i - a)\right] = \sum_{i=1}^{n} 2 = 2n > 0$$

Since the second derivative is positive, $a = \bar{x}$ indeed minimizes the sum.

Therefore, we have proven that:

$$\arg\min_a \sum_{i=1}^{n} (x_i - a)^2 = \bar{x}$$

(b) Prove that:

$$(n-1)s^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

Proof:

We begin with the definition of the sample variance s^2:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Multiplying both sides by $(n-1)$:

$$(n-1)s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Expanding the right-hand side:

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i^2 - 2x_i\bar{x} + \bar{x}^2)$$

$$= \sum_{i=1}^{n} x_i^2 - 2\bar{x} \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \bar{x}^2$$

Using the fact that $\sum_{i=1}^{n} x_i = n\bar{x}$ and $\sum_{i=1}^{n} \bar{x}^2 = n\bar{x}^2$:

$$= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2$$

$$= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

Therefore, we have proven that:

$$(n-1)s^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

(c) Prove that:

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

Proof:

We start by expanding the left-hand side of the equation:

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} (x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x}\bar{y})$$

Distributing the summation:

$$= \sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i - \bar{x} \sum_{i=1}^{n} y_i + \sum_{i=1}^{n} \bar{x}\bar{y}$$

Using the properties of the arithmetic mean, we know that $\sum_{i=1}^{n} x_i = n\bar{x}$ and $\sum_{i=1}^{n} y_i = n\bar{y}$. Also, $\sum_{i=1}^{n} \bar{x}\bar{y} = n\bar{x}\bar{y}$. Substituting these:

$$= \sum_{i=1}^{n} x_i y_i - n\bar{y}\bar{x} - n\bar{x}\bar{y} + n\bar{x}\bar{y}$$

Simplifying:

$$= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

Thus, we have proven that:

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

Question 3: Deriving the OLS Estimates

Assume the following population regression model:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i = 1, \ldots n$$

where the ELIH assumptions we learned in class hold. Derive the Ordinary Least Squares (OLS) estimates of β_0 and β_1 using calculus, as described in Lecture 3. That is, show:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

The facts you proved in Question 2 will help you.

$$

Solution

Step 1: Set up the Objective Function

We aim to minimize the sum of squared residuals (SSR):

$$\text{Minimize } SSR = \sum_{i=1}^{n}[Y_i - (\beta_0 + \beta_1 x_i)]^2$$

Step 2: Take Derivatives with Respect to β_0 and β_1

Compute the partial derivatives:

1. With respect to β_0:

$$\frac{\partial SSR}{\partial \beta_0} = -2\sum_{i=1}^{n}[Y_i - (\beta_0 + \beta_1 x_i)] = 0$$

2. With respect to β_1:

$$\frac{\partial SSR}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i[Y_i - (\beta_0 + \beta_1 x_i)] = 0$$

Step 3: Simplify the Equations

1. From the derivative with respect to β_0:

$$-2\sum_{i=1}^{n}[Y_i - \beta_0 - \beta_1 x_i] = 0$$

Simplify:

$$\sum_{i=1}^{n} Y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i = 0$$

Divide both sides by n:

$$\bar{Y} - \beta_0 - \beta_1 \bar{x} = 0$$

Solve for β_0:

$$\beta_0 = \bar{Y} - \beta_1 \bar{x}$$

2. From the derivative with respect to β_1:

$$-2\sum_{i=1}^{n} x_i[Y_i - \beta_0 - \beta_1 x_i] = 0$$

Simplify:

$$\sum_{i=1}^{n} x_i Y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0$$

Substitute $\beta_0 = \bar{Y} - \beta_1 \bar{x}$:

$$\sum_{i=1}^{n} x_i Y_i - (\bar{Y} - \beta_1 \bar{x}) \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0$$

Simplify:

$$\sum_{i=1}^{n} x_i Y_i - \bar{Y} \sum_{i=1}^{n} x_i + \beta_1 \bar{x} \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\sum_{i=1}^{n} x_i Y_i - n\bar{Y}\bar{x} + \beta_1 n\bar{x}^2 - \beta_1 \sum_{i=1}^{n} x_i^2 = 0$$

Step 4: Solve for β_1

Rewriting the equation:

$$\sum_{i=1}^{n} x_i Y_i - n\bar{Y}\bar{x} + \beta_1(n\bar{x}^2 - \sum_{i=1}^{n} x_i^2) = 0$$

Recognize that:

$$\sum_{i=1}^{n} x_i Y_i = n\overline{xY}, \quad \sum_{i=1}^{n} x_i^2 = n\overline{x^2}$$

Substitute:

$$n\overline{xY} - n\bar{Y}\bar{x} + \beta_1(n\bar{x}^2 - n\overline{x^2}) = 0$$

Divide both sides by n:

$$\overline{xY} - \bar{Y}\bar{x} + \beta_1(\bar{x}^2 - \overline{x^2}) = 0$$

Note that $\bar{x}^2 - \overline{x^2} = -\operatorname{Var}(x)$ (since $\operatorname{Var}(x) = \overline{x^2} - \bar{x}^2$). Therefore:

$$\overline{xY} - \bar{Y}\bar{x} - \beta_1 \operatorname{Var}(x) = 0$$

Recognize that $\operatorname{Cov}(x, Y) = \overline{xY} - \bar{x}\bar{Y}$. So:

$$\operatorname{Cov}(x, Y) - \beta_1 \operatorname{Var}(x) = 0$$

Solve for β_1:

$$\beta_1 = \frac{\operatorname{Cov}(x, Y)}{\operatorname{Var}(x)} = \frac{\overline{xY} - \bar{x}\bar{Y}}{\overline{x^2} - \bar{x}^2}$$

Expressed in summation form:

$$\beta_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i Y_i - \bar{x}\bar{Y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Step 5: Obtain β_0

Using $\beta_0 = \bar{Y} - \beta_1\bar{x}$:

$$\beta_0 = \bar{Y} - \left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)\bar{x}$$

Conclusion

We have derived the OLS estimates:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

and

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

$$

Problem 4. Correlation of OLS Estimates

Assume the model in Problem 3 holds, but add the assumption that $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2) \quad i = 1, 2, ..., n$:

(a) Show that $\bar{Y}$ and $\hat{\beta}_1$ are independent. This will help you in part (b).

We can show that $\bar{Y}$ and $\hat{\beta}_1$ are normally distributed. Thus, if we show that they are also uncorrelated, by properties of the normal distribution, they must be independent.

We know that $X_1, ..., X_n$ are numbers (scalars), our model is:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Since Y_i is a linear combination of normals, it must be normal. In this model, β_0 and β_1 are also constants (perhaps unknown, but still constant). However, $\hat{\beta}_0$ and $\hat{\beta}_1$ are distributed normally because $\hat{\beta}_1$ is a linear function of $\bar{Y}$ and $\bar{X}$ (which is a linear combination of Y_i).

Since the Y_i are normally distributed, a linear function of normals, $\hat{\beta}_0$ and $\hat{\beta}_1$ must be normally distributed (multivariate normal rules).

Now, we will attempt to show that they are uncorrelated.

We must show:

$$\text{Cov}(\bar{Y}, \hat{\beta}_1) = 0$$

$$\text{Cov}(\bar{Y}, \hat{\beta}_1) = E[\bar{Y}\hat{\beta}_1] - E[\bar{Y}]E[\hat{\beta}_1]$$

Because $\hat{\beta}_1$ is normally distributed about β_1, we have:

$$E[\hat{\beta}_1] = \beta_1$$

Now, $E[\bar{Y}\hat{\beta}_1] = E[\beta_0 + \beta_1 \bar{X}] = 0$

$$E[(\beta_0 + \beta_1 \bar{X})\hat{\beta}_1] - [\beta_0 + \beta_1 \bar{X}]\beta_1 = 0$$

$$0 = 0$$

Since they are uncorrelated, $\bar{Y}$ and $\hat{\beta}_1$ are independent.

(b) Derive the covariance and correlation between $\hat{\beta}_1$ and $\hat{\beta}_0$:

$$\text{Cov}[\hat{\beta}_1, \hat{\beta}_0] = \mathbb{E}[\hat{\beta}_1\hat{\beta}_0] - \mathbb{E}[\hat{\beta}_1]\mathbb{E}[\hat{\beta}_0]$$

$$= \mathbb{E}[\hat{\beta}_1\hat{\beta}_0] - \beta_1\beta_0$$

Substitute:

$$y = \beta_0 + \beta_1 X + \epsilon$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{X} + \epsilon$$

$$\mathbb{E}[\hat{\beta}_1\hat{\beta}_0] = \mathbb{E}[\hat{\beta}_1(\bar{y} - \hat{\beta}_1\bar{X} + \epsilon)]$$

$$= \mathbb{E}[\hat{\beta}_1\bar{y}] - \mathbb{E}[\hat{\beta}_1{}^2\bar{X}] + 0 - \beta_1\beta_0$$

Since $\bar{y}$ and $\hat{\beta}_1$ are independent, we know:

$$\mathbb{E}[\bar{y}\hat{\beta}_1] = \mathbb{E}[\bar{y}]\mathbb{E}[\hat{\beta}_1]$$

$$= [\mathbb{E}[\bar{y}] + \mathbb{E}[\hat{\beta}_1\bar{X}]] - \mathbb{E}[\hat{\beta}_1{}^2\bar{X}] - \beta_1\beta_0$$

$$\text{Cov}[\hat{\beta}_1, \hat{\beta}_0] = \beta_0\mathbb{E}[\hat{\beta}_1] + \bar{X}[\mathbb{E}[\hat{\beta}_1] - \mathbb{E}[\hat{\beta}_1{}^2\bar{X}]] - \beta_1\beta_0$$

$$\text{Cov}[\hat{\beta}_1, \hat{\beta}_0] = \mathbb{E}[\hat{\beta}_1\bar{X}] - \mathbb{E}[\hat{\beta}_1]\mathbb{E}[\bar{X}] - \mathbb{E}[\hat{\beta}_1{}^2\bar{X}]$$

By the definition of correlation:

$$\text{Corr}[x, y] = \frac{\text{Cov}[x, y]}{\sqrt{\text{Var}[x]\text{Var}[y]}}$$

Now, we need to find $\text{Var}[\hat{\beta}_1]$ and $\text{Var}[\hat{\beta}_0]$. We know:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{X})^2}$$

Substitute $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{X})[(\beta_0 + \beta_1 x_i + \epsilon_i) - \bar{y}]}{\sum_{i=1}^n (x_i - \bar{X})^2}$$

This simplifies to:

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{X})\epsilon_i}{S_{xx}}$$

$$\text{Var}[\hat{\beta}_1] = \sigma^2 + \text{Var}\left[\frac{\sum_{i=1}^n (x_i - \bar{X})\epsilon_i}{S_{xx}}\right] = \frac{\sigma^2 \sum_{i=1}^n (x_i - \bar{X})^2}{S_{xx}^2} = \frac{\sigma^2 S_{xx}}{S_{xx}^2}$$

$$\text{Var}[\hat{\beta}_1] = \frac{\sigma^2}{S_{xx}}$$

Now, find the variance of $\hat{\beta}_0$:

$$\text{Var}[\hat{\beta}_0] = \text{Var}\left[\bar{y} - \hat{\beta}_1 \bar{X}\right]$$

Since $\bar{y}$ and $\hat{\beta}_1$ are uncorrelated (part a),

$$\text{Var}[\hat{\beta}_0] = \text{Var}[\bar{y}] + \text{Var}[\hat{\beta}_1 \bar{X}] = \frac{\sigma^2}{n} + \bar{X}^2 \frac{\sigma^2}{S_{xx}}$$

$$\text{Var}[\hat{\beta}_0] = \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right]$$

Thus, the covariance between $\hat{\beta}_1$ and $\hat{\beta}_0$ is:

$$\text{Cov}[\hat{\beta}_1, \hat{\beta}_0] = \bar{X}\mathbb{E}[\hat{\beta}_1{}^2] - \mathbb{E}[\hat{\beta}_1]\mathbb{E}[\hat{\beta}_0]$$

$$\text{Cov}[\hat{\beta}_1, \hat{\beta}_0] = -\bar{X}\text{Var}[\hat{\beta}_1]$$

$$\text{Cov}[\hat{\beta}_1, \hat{\beta}_0] = -\frac{\bar{X}\sigma^2}{S_{xx}}$$

$$\text{Corr}[\hat{\beta}_1, \hat{\beta}_0] = \frac{-\frac{\bar{X}\sigma^2}{S_{xx}}}{\sqrt{\frac{\sigma^2}{S_{xx}}\left[\frac{\sigma^2}{n} + \frac{\bar{X}^2\sigma^2}{S_{xx}}\right]}} = \frac{-\bar{X}}{\sqrt{S_{xx}\left[\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right]}}$$

$$\text{Corr}[\hat{\beta}_1, \hat{\beta}_0] = \frac{-\bar{X}}{\sqrt{\frac{S_{xx}}{n} + \bar{X}^2}}$$

(c) In terms of $\bar{x}$, when will this correlation be positive? When will it be negative? In 1-2 sentences, justify why this makes sense if $\bar{x} > 0$ (think where the scatterplot and regression line lies on the coordinate system).

When $\overline{X} > 0$, correlation is negative, $\overline{X} < 0$, correlation is positive, $\overline{X} = 0$, correlation is 0.
Due to $-\overline{X}$ in the numerator. When $\overline{X} > 0$, that means that most of the points are to the right of the origin on
That means that the slope increases—To account for this, $\hat{\beta}_0$, the intercept, must reduce to still fit the data.
This indicates an inverse relationship between $\hat{\beta}_1$ and $\hat{\beta}_0$, which is represented by the equation.

Question 5: Fitting the Model

All empirical work begins with some data "cleaning", including ensuring the data are in the right form. James will do this for you this time. After he cleans the data you provided, he will provide you with a dataset with these three variables:

- `studentheight`: your heights (in inches, to the nearest half inch)
- `maternalheight`: the height of your mothers (in inches, to the nearest half inch)
- `paternalheight`: the height of your fathers (in inches, to the nearest half inch)

(a) Next, it's important to conduct exploratory data analysis (EDA) to ensure the integrity of your data. This includes summarizing your variables, including the extent of missingness, checking for outliers and inconsistencies, and potentially addressing any data entry errors. Provide an appropriate EDA (e.g., appropriate figures and/or a table), and provide commentary. Also, justify any decisions you make about how you choose to handle any suspect data in your analysis.

```r
# List of packages to install and load
packages <- c("skimr", "ggplot2", "gridExtra")

# Function to install and load packages
install_and_load <- function(packages) {
  new_packages <- packages[!(packages %in% installed.packages()[,"Package"])]
  if(length(new_packages)) install.packages(new_packages)
  invisible(lapply(packages, library, character.only = TRUE))
}

# Install and load the packages
install_and_load(packages)
library(tidyr)
```

```r
heights_data <- read.csv("heights_139.csv")
head(heights_data)
```

```
##   studentheight maternalheight paternalheight
## 1            61             63             66
## 2          67.5             62             70
## 3          70.5             61             72
## 4          66.5             69           69.5
## 5            65             67
## 6            61             60             64
```

```r
str(heights_data)
```

```
## 'data.frame':    70 obs. of  3 variables:
##  $ studentheight : chr  "61" "67.5" "70.5" "66.5" ...
##  $ maternalheight: chr  "63" "62" "61" "69" ...
##  $ paternalheight: chr  "66" "70" "72" "69.5" ...
```

```r
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:gridExtra':
##
##     combine
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(stringr)

# Import the data
heights_data <- read_csv("heights_139.csv", col_types = cols(.default = "c"))

# Function to convert height to inches
convert_to_inches <- function(height) {
  if (is.na(height)) return(NA)

  # Convert '5'10' format to inches
  if (str_detect(height, "^\\d+'\\d+(\\.\\d+)?$")) {
    parts <- str_split(height, "'", simplify = TRUE)
    return(as.numeric(parts[1]) * 12 + as.numeric(parts[2]))
  }

  # Remove any non-numeric characters
  height <- str_replace_all(height, "[^0-9.]", "")

  # Convert to numeric
  height <- as.numeric(height)

  # Convert cm to inches if greater than 100
  if (!is.na(height) && height > 100) {
    height <- height / 2.54
  }

  return(height)
}

# Apply cleaning to each column
heights_data <- heights_data %>%
  mutate(across(everything(), ~sapply(., convert_to_inches)))

# Drop rows with NA values
heights_data <- heights_data %>% drop_na()

# Display the first few rows and structure of the cleaned data
print(head(heights_data))
```

```
## # A tibble: 6 x 3
##   studentheight maternalheight paternalheight
##           <dbl>          <dbl>          <dbl>
## 1          61              63           66
## 2          67.5            62           70
## 3          70.5            61           72
## 4          66.5            69           69.5
## 5          61              60           64
```

```
## 6          70                64           68
```

```
print(str(heights_data))
```

```
## tibble [61 x 3] (S3: tbl_df/tbl/data.frame)
##  $ studentheight : Named num [1:61] 61 67.5 70.5 66.5 61 70 68.5 69.5 71.5 71 ...
##   ..- attr(*, "names")= chr [1:61] "61" "67.5" "70.5" "66.5" ...
##  $ maternalheight: Named num [1:61] 63 62 61 69 60 64 64 62 63 60 ...
##   ..- attr(*, "names")= chr [1:61] "63" "62" "61" "69" ...
##  $ paternalheight: Named num [1:61] 66 70 72 69.5 64 68 73 72 69 67 ...
##   ..- attr(*, "names")= chr [1:61] "66" "70" "72" "69.5" ...
## NULL
```

```
# Summary statistics of cleaned data
summary(heights_data)
```

```
##  studentheight   maternalheight  paternalheight
##  Min.   :59.00   Min.   :57.00   Min.   :59.00
##  1st Qu.:65.50   1st Qu.:62.00   1st Qu.:67.50
##  Median :68.50   Median :63.00   Median :69.00
##  Mean   :67.96   Mean   :63.61   Mean   :69.26
##  3rd Qu.:71.00   3rd Qu.:65.00   3rd Qu.:71.00
##  Max.   :76.00   Max.   :69.00   Max.   :75.50
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v purrr     1.0.2
## v lubridate 1.9.3     v tibble    3.2.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::combine() masks gridExtra::combine()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to bec
```

```
library(skimr)
library(ggplot2)
library(gridExtra)

summary_stats <- summary(heights_data)
print(summary_stats)
```

```
##  studentheight   maternalheight  paternalheight
##  Min.   :59.00   Min.   :57.00   Min.   :59.00
##  1st Qu.:65.50   1st Qu.:62.00   1st Qu.:67.50
##  Median :68.50   Median :63.00   Median :69.00
##  Mean   :67.96   Mean   :63.61   Mean   :69.26
##  3rd Qu.:71.00   3rd Qu.:65.00   3rd Qu.:71.00
##  Max.   :76.00   Max.   :69.00   Max.   :75.50
```

```r
# Check for missing values
missing_values <- colSums(is.na(heights_data))
print(missing_values) # none!!
```

```
##  studentheight maternalheight paternalheight
##              0              0              0
```

```r
skim(heights_data)
```

Table 1: Data summary

Name	heights_data
Number of rows	61
Number of columns	3
_____	
Column type frequency:	
numeric	3
_____	
Group variables	None

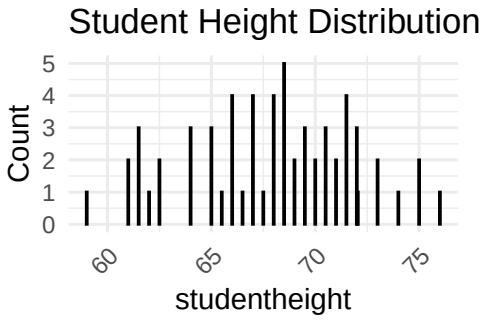
Variable type: numeric

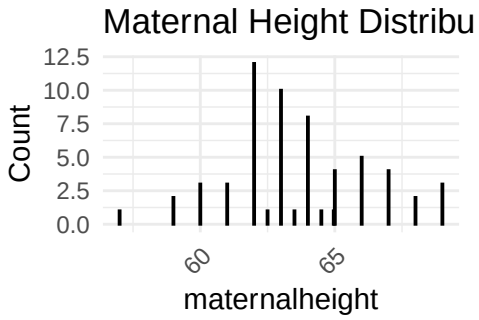
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
studentheight	0	1	67.96	3.92	59	65.5	68.5	71	76.0	
maternalheight	0	1	63.61	2.56	57	62.0	63.0	65	69.0	
paternalheight	0	1	69.26	2.86	59	67.5	69.0	71	75.5	

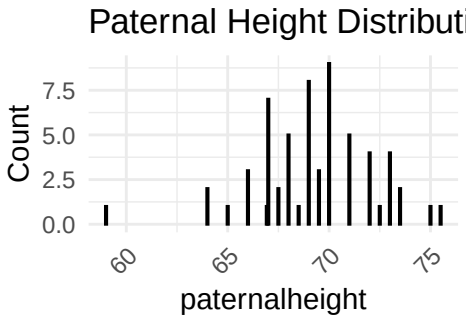
```r
# Function to create a bar chart
create_bar_chart <- function(data, column, title) {
  ggplot(data, aes(x = !!sym(column))) +
    geom_bar(fill = "skyblue", color = "black") +
    labs(title = title, x = column, y = "Count") +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
}

# Create bar charts
p1 <- create_bar_chart(heights_data, "studentheight", "Student Height Distribution")
p2 <- create_bar_chart(heights_data, "maternalheight", "Maternal Height Distribution")
p3 <- create_bar_chart(heights_data, "paternalheight", "Paternal Height Distribution")

# Arrange the plots in a grid
grid.arrange(p1, p2, p3, ncol = 2)
```

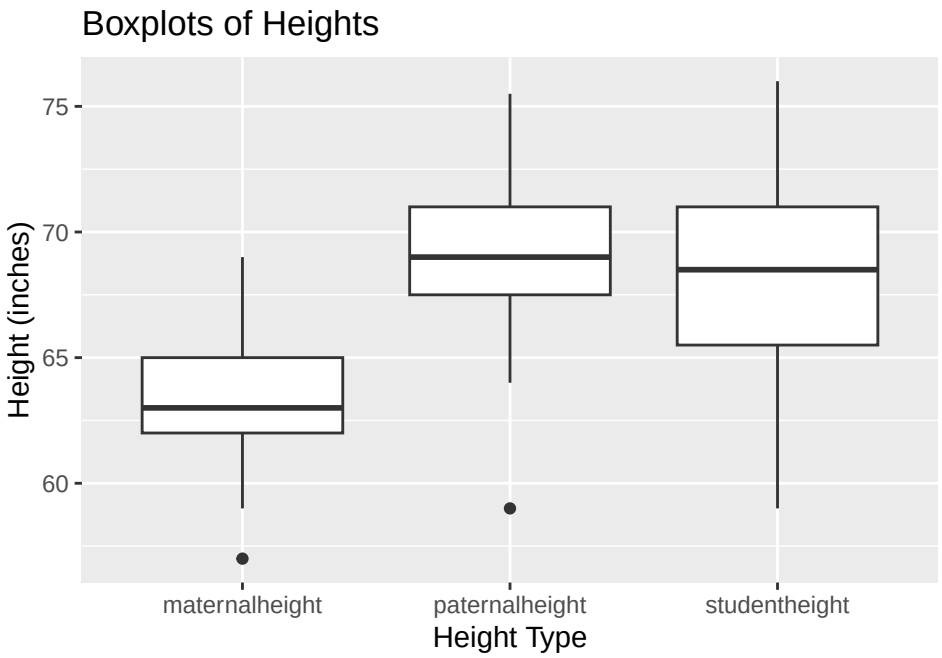







```
heights_long <- pivot_longer(heights_data,
                             cols = c(studentheight, maternalheight, paternalheight),
                             names_to = "height_type",
                             values_to = "height")

ggplot(heights_long, aes(x = height_type, y = height)) +
  geom_boxplot() +
  labs(title = "Boxplots of Heights", x = "Height Type", y = "Height (inches)")
```

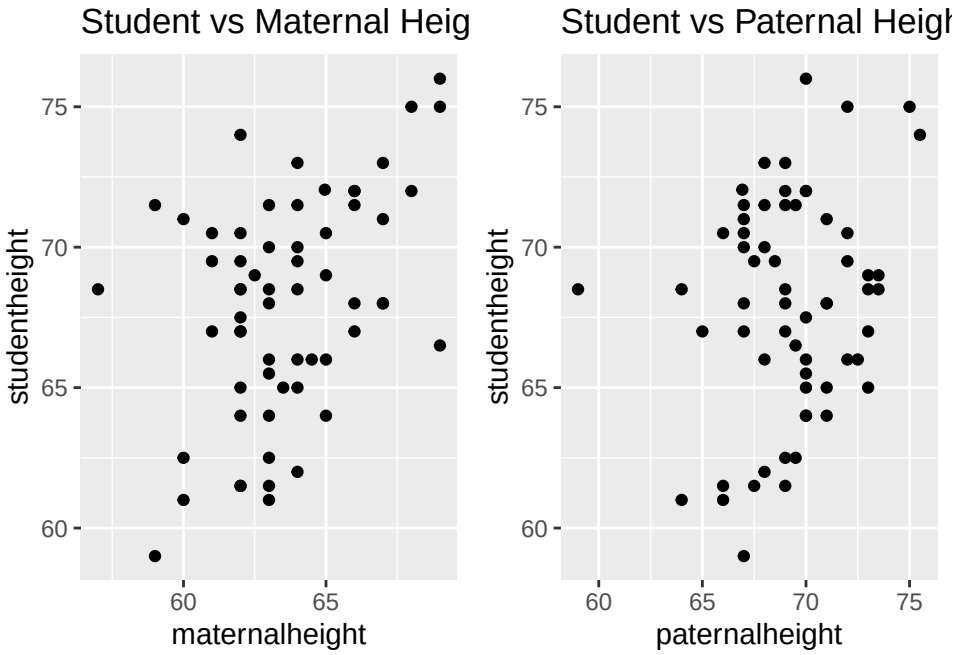



```
p1 <- ggplot(heights_data, aes(x = maternalheight, y = studentheight)) +
  geom_point() +
```

```
  labs(title = "Student vs Maternal Height")

p2 <- ggplot(heights_data, aes(x = paternalheight, y = studentheight)) +
  geom_point() +
  labs(title = "Student vs Paternal Height")

grid.arrange(p1, p2, ncol = 2)
```



Commentary: Based on the EDA, we can roughly say that: - the data for all columns is roughly symmetrical - there appears to be a weak positive linear correlation between child's height and parent's height (stratified on gender) - There appears to be two outliers in the data, both lower than most of the data - the data is clean – ther are no null values and the data is standardized.

Next steps: - We should further analyze the relationship between the parents' heights and the child's height - We should investigate how much of the relationship in the data can be explained linearly by the parents' heights.

(b) Create the following new variables in your dataset:

- `midparentheight` = `paternalheight` + 1.08×`maternalheight`
- `tallparents` = 1 if `midparentheight` is greater than or equal to the median of `midparentheight` and 0 if not. That is, create an *indicator variable* of whether `midparentheight` is greater than or equal to the median.

```
str(heights_data)
```

```
## tibble [61 x 3] (S3: tbl_df/tbl/data.frame)
##  $ studentheight : Named num [1:61] 61 67.5 70.5 66.5 61 70 68.5 69.5 71.5 71 ...
##   ..- attr(*, "names")= chr [1:61] "61" "67.5" "70.5" "66.5" ...
##  $ maternalheight: Named num [1:61] 63 62 61 69 60 64 64 62 63 60 ...
##   ..- attr(*, "names")= chr [1:61] "63" "62" "61" "69" ...
```

```
##  $ paternalheight: Named num [1:61] 66 70 72 69.5 64 68 73 72 69 67 ...
##   ..- attr(*, "names")= chr [1:61] "66" "70" "72" "69.5" ...
```

```r
# If the heights are stored as factors or characters, we need to convert them to numeric
heights_data$studentheight <- as.numeric(as.character(heights_data$studentheight))
heights_data$maternalheight <- as.numeric(as.character(heights_data$maternalheight))
heights_data$paternalheight <- as.numeric(as.character(heights_data$paternalheight))

# Now let's check the structure again to confirm the conversion
str(heights_data)
```

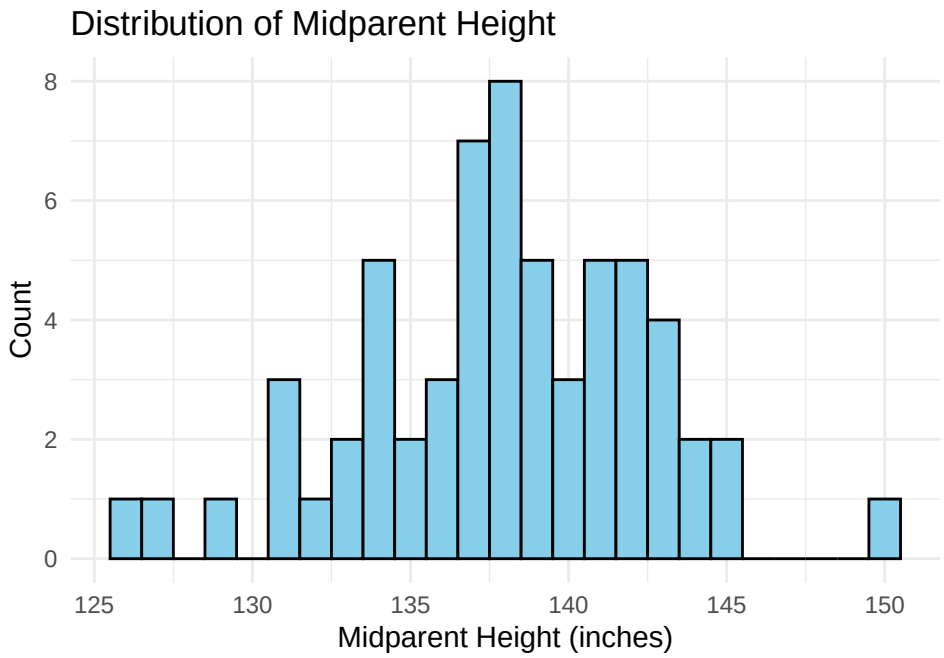
```
## tibble [61 x 3] (S3: tbl_df/tbl/data.frame)
##  $ studentheight : num [1:61] 61 67.5 70.5 66.5 61 70 68.5 69.5 71.5 71 ...
##  $ maternalheight: num [1:61] 63 62 61 69 60 64 64 62 63 60 ...
##  $ paternalheight: num [1:61] 66 70 72 69.5 64 68 73 72 69 67 ...
```

```r
# If the conversion was successful, we can proceed with creating the new variables
heights_data$midparentheight <- heights_data$paternalheight + 1.08 * heights_data$maternalheigh
heights_data$tallparents <- as.integer(heights_data$midparentheight >= median(heights_data$midp

# Summary of new variables
summary(heights_data[c("midparentheight", "tallparents")])
```

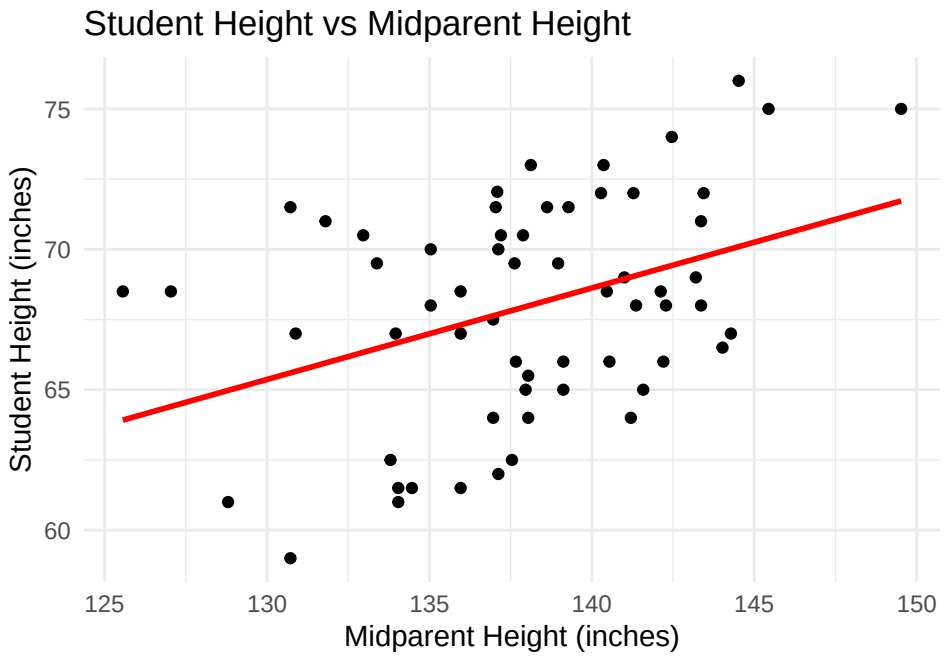
```
##  midparentheight  tallparents
##  Min.   :125.6   Min.   :0.0000
##  1st Qu.:135.0   1st Qu.:0.0000
##  Median :138.0   Median :1.0000
##  Mean   :138.0   Mean   :0.5082
##  3rd Qu.:141.3   3rd Qu.:1.0000
##  Max.   :149.5   Max.   :1.0000
```

```r
# Visualize midparentheight distribution
ggplot(heights_data, aes(x = midparentheight)) +
  geom_histogram(binwidth = 1, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Midparent Height", x = "Midparent Height (inches)", y = "Count
  theme_minimal()
```

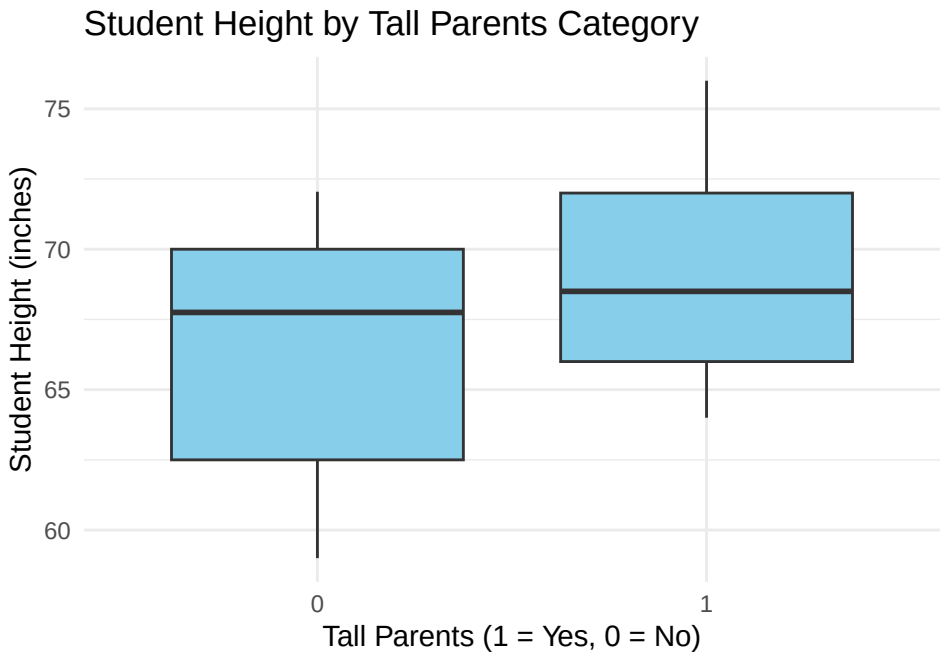



```
# Visualize relationship between midparentheight and studentheight
ggplot(heights_data, aes(x = midparentheight, y = studentheight)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Student Height vs Midparent Height",
       x = "Midparent Height (inches)",
       y = "Student Height (inches)") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



```r
# Compare student heights for tall vs not tall parents
ggplot(heights_data, aes(x = factor(tallparents), y = studentheight)) +
  geom_boxplot(fill = "skyblue") +
  labs(title = "Student Height by Tall Parents Category",
       x = "Tall Parents (1 = Yes, 0 = No)",
       y = "Student Height (inches)") +
  theme_minimal()
```



(c) Fit a simple linear regression in R, with `studentheight` as your dependent variable, and `tallparents` as your independent variable. Interpret the slope coefficient from your model.

```r
# Fit the linear regression model
model <- lm(studentheight ~ tallparents, data = heights_data)

# Display the summary of the model
summary(model)
```

```
##
## Call:
## lm(formula = studentheight ~ tallparents, data = heights_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7016 -3.1774  0.2984  2.8226  6.8226
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  66.7016     0.6842   97.49   <2e-16 ***
## tallparents   2.4758     0.9598    2.58   0.0124 *
## ---
```

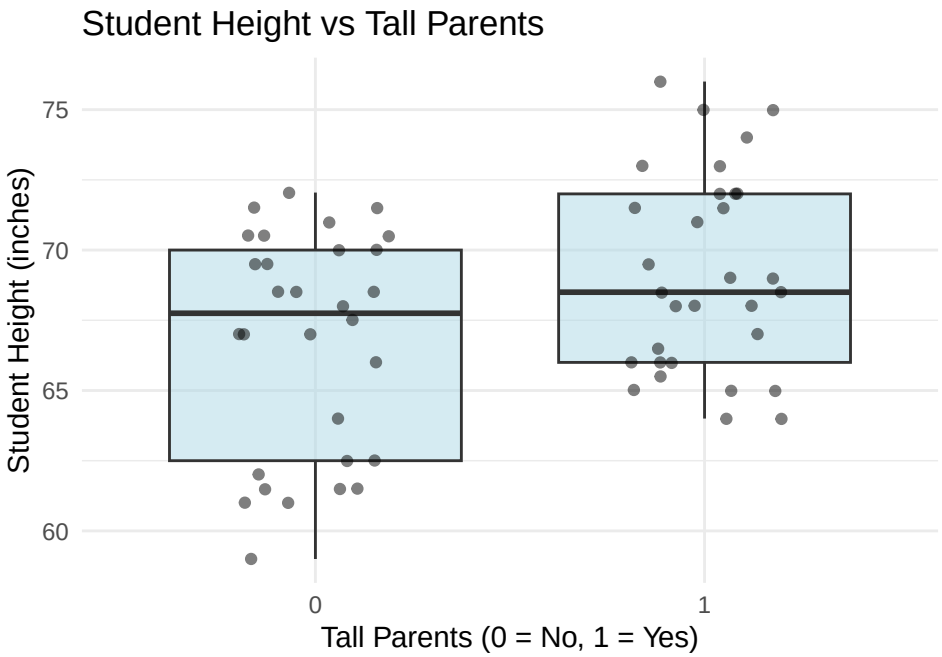
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.747 on 59 degrees of freedom
## Multiple R-squared:  0.1014, Adjusted R-squared:  0.08613
## F-statistic: 6.655 on 1 and 59 DF,  p-value: 0.0124
```

```
# Calculate the confidence interval for the coefficients
confint(model)
```

```
##                  2.5 %    97.5 %
## (Intercept) 65.3325039 68.070646
## tallparents  0.5553636  4.396325
```

```
# Plot the regression line
ggplot(heights_data, aes(x = factor(tallparents), y = studentheight)) +
  geom_boxplot(fill = "lightblue", alpha = 0.5) +
  geom_jitter(width = 0.2, alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Student Height vs Tall Parents",
       x = "Tall Parents (0 = No, 1 = Yes)",
       y = "Student Height (inches)") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Interpretation: For each unit increase in the "tallparents" variable, we expect an average increase of 2.4758 units in student height. Thus, if the student has tall parents, we can expect them to be 2.4758 inches taller on average. However, this should not be confused with causality.

However, R^2 is relatively low for the data (.10), which indicates that lots of the behavior of the data is not explained by the model.

(d) Test whether there is sufficient evidence to indicate a true mean difference in heights of the children of taller versus shorter parents, using a two sample t-test in R at the $\alpha = 0.05$ level of significance. Make sure to formally state your hypothesis, report your test statistic and interpret the associated p-value.

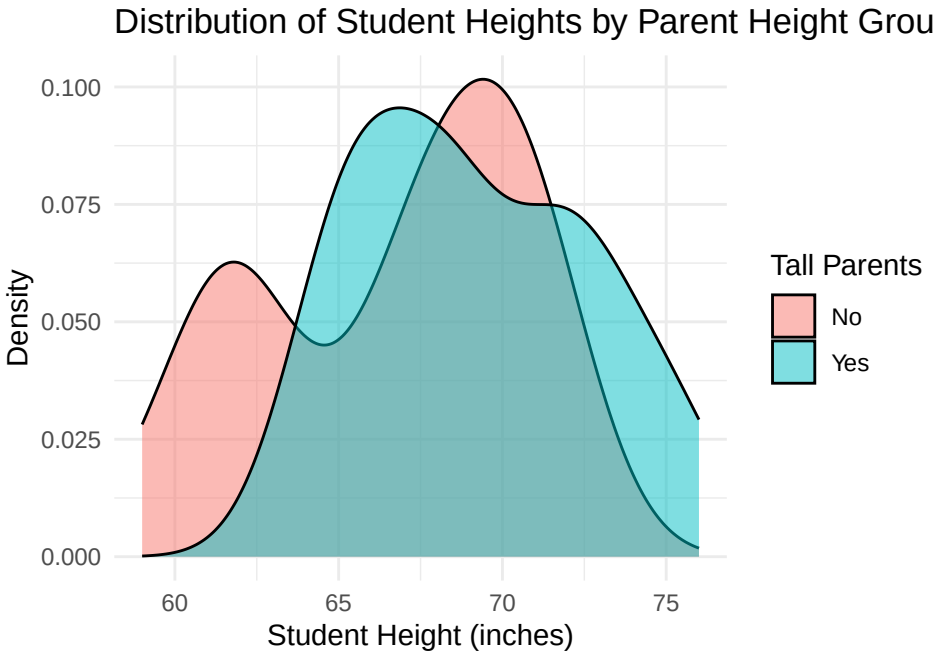
Statements: - Null Hypothesis (H_0): There is no difference in the mean heights of children with taller parents and children with shorter parents. - $\mu_1 = \mu_2$, where μ_1 is the mean height of children with taller parents and μ_2 is the mean height of children with shorter parents. - Alternative Hypothesis (H_1): There is a difference in the mean heights of children with taller parents and children with shorter parents. $\mu_1 \neq \mu_2$

```
# Perform the two-sample t-test
t_test_result <- t.test(studentheight ~ tallparents, data = heights_data)

# Display the results
print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  studentheight by tallparents
## t = -2.5745, df = 57.66, p-value = 0.01263
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
## 95 percent confidence interval:
##  -4.4010873 -0.5506018
## sample estimates:
## mean in group 0 mean in group 1
##        66.70157        69.17742
```

```
# Visualize the distribution of heights for both groups
ggplot(heights_data, aes(x = studentheight, fill = factor(tallparents))) +
  geom_density(alpha = 0.5) +
  labs(title = "Distribution of Student Heights by Parent Height Group",
       x = "Student Height (inches)",
       y = "Density",
       fill = "Tall Parents") +
  scale_fill_discrete(labels = c("No", "Yes")) +
  theme_minimal()
```



Data: t = -2.5745, df = 57.66, p-value = 0.01263 95 percent confidence interval: -4.4010873 -0.5506018

Interpretation: The p-value associated with our test statistic is 0.012, which is much smaller than our significance level of $\alpha = 0.05$. Given this small p-value ($p < 0.05$), we reject the null hypothesis. There is strong evidence to suggest that there is a significant difference in the mean heights of children with taller parents compared to children with shorter parents. THe 95% confidence intercal does not contain 0, which also gives us reasoning for why we should reject the null hypothesis.

(e) Comment on the consistencies and/or inconsistencies between the output of `lm()` in part (c) and the output of `t.test()` in part (d). Is there a connection between the estimate from `lm()` and the test statistic from `t.test()`? What about the inference from the two approaches? You don't need to derive anything here, just comment on the outputs.

Consistencies

- The p value from both isare very very similar to each other.
- the ttest staitisc is close to the value of the slope coefficient.
  - linear regression coefficient: 2.4758
  - ttest differnce in means: $69.17742 - 66.70157 = 2.47585$
- the confidence intervals are very similar (negated)
  - Linear regression CI for tallparents: $(0.5553636, 4.396325)$
  - T-test CI for the difference in means: $(-4.4010873, -0.5506018)$

Inconsistencies

- The degrees of freedom are different
  - 59 vs 57.66, respectively. I do not know why this difference occurs, but my guess is that the ttest is using a more sophisticated method of calculating DF than just $N - 1$. Pehaps it is taking into account variance in the data.

Conclusion:

Both the linear regression and t-test provide consistent results, supporting the conclusion that there is a significant relationship between tall parents and student height.