IMT 573: Problem Set 6 - Statistical Theory II

Srushti Chaukhande

Due: Tuesday, November 12, 2024 by 10:00PM PT

Collaborators:

Instructions: Before beginning this assignment, please ensure you have access to R and RStudio; this can be on your own personal computer or on the IMT 573 R Studio Server.

1. Download the `problem_set6.Rmd` file from Canvas or save a copy to your local directory on RStudio Server. Open `problem_set6.Rmd` in RStudio and supply your solutions to the assignment by editing `problem_set6.Rmd`.
2. Replace the "Insert Your Name Here" text in the `author:` field with your own full name. Any collaborators must be listed on the top of your assignment.
3. Be sure to include well-documented (e.g. commented) code chucks, figures, and clearly written text chunk explanations as necessary. Any figures should be clearly labeled and appropriately referenced within the text. Be sure that each visualization adds value to your written explanation; avoid redundancy – you do not need four different visualizations of the same pattern.
4. Collaboration on problem sets is fun and useful, and we encourage it, but each student must turn in an individual write-up in their own words as well as code/work that is their own. Regardless of whether you work with others, what you turn in must be your own work; this includes code and interpretation of results. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.
5. All materials and resources that you use (with the exception of lecture slides) must be appropriately referenced within your assignment.
6. Remember partial credit will be awarded for each question for which a serious attempt at finding an answer has been shown. Students are *strongly* encouraged to attempt each question and to document their reasoning process even if they cannot find the correct answer. If you would like to include R code to show this process, but it does not run without errors, you can do so with the `eval=FALSE` option. (Note: I am also using the `include=FALSE` option here to not include this code in the PDF, but you need to remove this or change it to `TRUE` if you want to include the code chunk.)
7. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit PDF`, rename the knitted PDF file to `ps5_YourLastName_YourFirstName.pdf`, and submit the PDF file on Canvas.

Setup: In this problem set you will need, at minimum, the following R packages.

```
# Load standard libraries
library(tidyverse)
library(MASS) # Modern applied statistics functions
```

Introduction: Are Sons Taller than Fathers?

This dataset uses fathers' and sons' height data from 19th century Britain. It was used by Galton (1886) when discussion "regression to mediocrity", or regression to mean as we would call it now. The dataset contains two variables: fheight and sheight for fathers' and sons' height, respectively (in cm).

1. Descriptive Analysis (16pt)

a) (3pt) load the fatherson.csv data. Do the basic descriptive work on it: what are the number of observations? Are there any missing data?
b) (5pt) Describe fathers and sons: compute the mean, median, standard deviation, and range of their heights. According to these computations, who are taller: fathers or sons?
c) (5pt) Let's add a graphical comparison. Create density plots of both heights on the same figure. Comment on the density plots. Which distribution do these resemble? Do the graphical representation agree with the conclusion that you drew from the computations in the previous question (question 1(b))? Hint: you can do density plots with stat_density when using ggplot.
d) (3pt) Finally, for further reference, compute how much taller are sons on average.

Solution 1

Solution (a):

Insert Response

```
# Read the CSV file
fatherson <- read.delim("fatherson.csv")

# Basic description of the fheight variable
summary_fheight <- summary(fatherson$fheight)
num_observations <- length(fatherson$fheight)
missing_data <- sum(is.na(fatherson$fheight))

# Print results
print(summary_fheight)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   149.9   167.1   172.1   171.9   176.8   191.6
```

```
cat("Number of observations:", num_observations, "\n")
```

```
## Number of observations: 1078
```

```
cat("Number of missing values:", missing_data, "\n")
```

```
## Number of missing values: 0
```

Solution (b):

Insert Response

```r
# Calculate statistics for fathers
father_mean <- mean(fatherson$fheight)
father_median <- median(fatherson$fheight)
father_sd <- sd(fatherson$fheight)
father_range <- range(fatherson$fheight)

# Calculate statistics for sons
son_mean <- mean(fatherson$sheight)
son_median <- median(fatherson$sheight)
son_sd <- sd(fatherson$sheight)
son_range <- range(fatherson$sheight)

# Print results
cat("Fathers' heights:\n")
```

```
## Fathers' heights:
```

```r
cat("Mean:", father_mean, "\n")
```

```
## Mean: 171.9252
```

```r
cat("Median:", father_median, "\n")
```

```
## Median: 172.1
```

```r
cat("Standard Deviation:", father_sd, "\n")
```

```
## Standard Deviation: 6.972346
```

```r
cat("Range:", father_range[1], "to", father_range[2], "\n\n")
```

```
## Range: 149.9 to 191.6
```

```r
cat("Sons' heights:\n")
```

```
## Sons' heights:
```

```r
cat("Mean:", son_mean, "\n")
```

```
## Mean: 174.4572
```

```r
cat("Median:", son_median, "\n")
```

```
## Median: 174.3
```

```r
cat("Standard Deviation:", son_sd, "\n")
```

```
## Standard Deviation: 7.150713
```

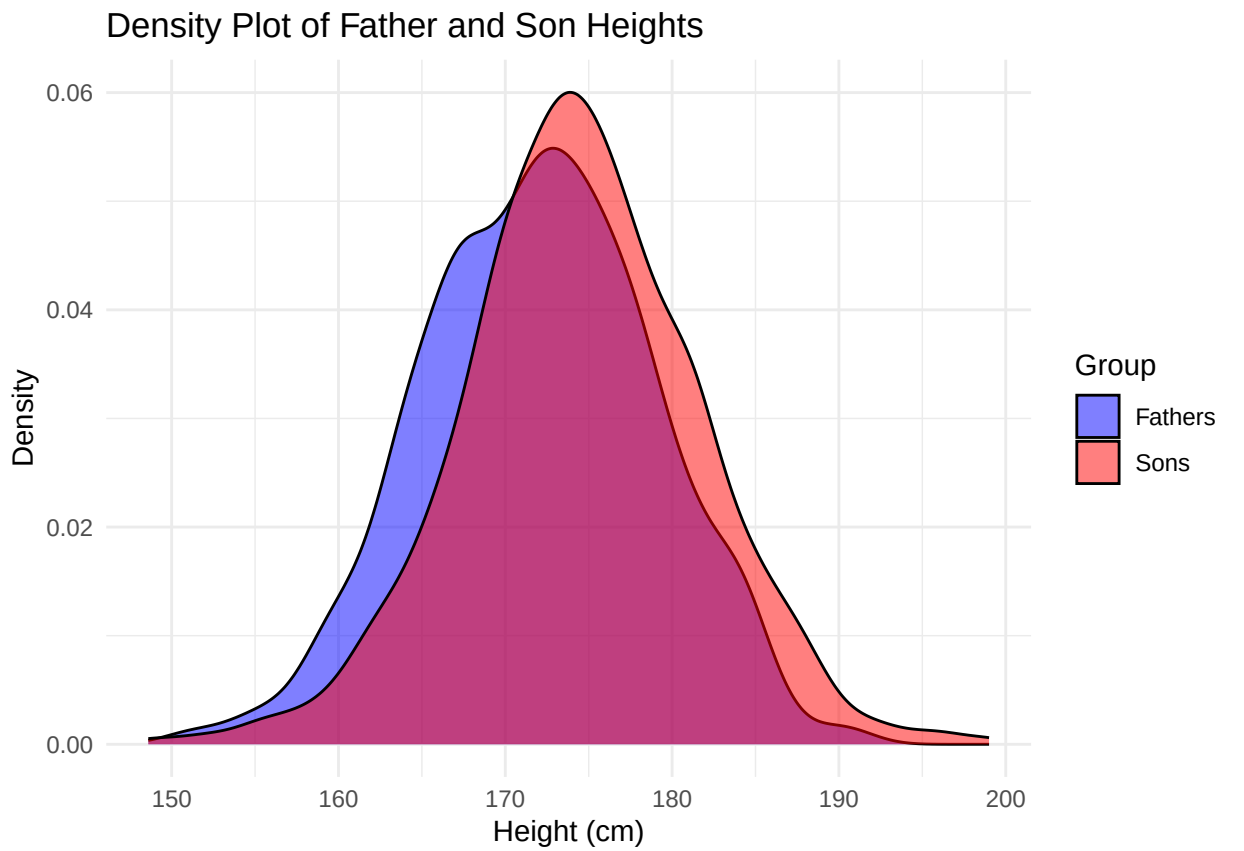
```r
cat("Range:", son_range[1], "to", son_range[2], "\n")
```

```
## Range: 148.6 to 199
```

The data shows that sons are taller than fathers on average: - The mean height for sons (173.89 cm) is approximately 3 cm greater than the mean height for fathers (170.88 cm). - The median height for sons (174 cm) is also higher than the median height for fathers (171.1 cm), further supporting that sons tend to be taller. - The range of heights for sons (148.6 cm to 199 cm) is wider than that of fathers (149.9 cm to 191.6 cm), with sons having both a lower minimum and a higher maximum height. - The standard deviations are very similar (about 6.74 cm for both), indicating that the variability in height is comparable between fathers and sons. In conclusion, based on these statistical measures, sons are generally taller than fathers in this dataset. **Solution (c):**

Insert Response

```r
# Load required libraries
library(ggplot2)
library(tidyr)

# Create a long format dataset for plotting
data_long <- pivot_longer(fatherson, cols = c(fheight, sheight),
                          names_to = "group", values_to = "height")

# Create the density plot
ggplot(data_long, aes(x = height, fill = group)) +
  geom_density(alpha = 0.5) +
  scale_fill_manual(values = c("blue", "red"), labels = c("Fathers", "Sons")) +
  labs(title = "Density Plot of Father and Son Heights",
       x = "Height (cm)",
       y = "Density",
       fill = "Group") +
  theme_minimal()
```




```
# Save the plot
ggsave("height_density_plot.png", width = 10, height = 6)
```

- Shape of distributions: Both distributions appear to be roughly bell-shaped and symmetrical, resembling normal distributions. The graphical representation indeed agrees with the conclusion drawn from the previous computations:
- The rightward shift of the sons' distribution supports the finding that sons are, on average, taller than fathers.
- The similar spread of the distributions aligns with the comparable standard deviations we calculated earlier.
- The overall shape and considerable overlap explain why, despite the difference in means, there's still a lot of variability and overlap in father-son heights.

Solution (d):

Insert Response

```
# Calculate the average height difference
height_difference <- mean(fatherson$sheight - fatherson$fheight)

# Print the result
cat("On average, sons are", round(height_difference, 2), "cm taller than their fathers.")
```

```
## On average, sons are 2.53 cm taller than their fathers.
```

2. Simulations (36pt)

Let's test the above result as to who is taller through simulations. We take H0: fathers and sons are of the same height, in average, and see if we can reject this hypothesis based on data.

You will proceed as follows: create two samples of random normals, similar to the data above, using the overall mean and standard deviation in the data. Call one of these samples "sample fathers" and the other "sample sons". What is the difference in their means? Is this close to the number you saw in the data (your response in 1(d))? It is probably not. But maybe this was just an unhappy experiment. So now let's repeat this exercise many times and see how big or how small differences you typically see between "sample fathers" and "sample sons". The point distributions for each of these steps are listed below.

a) (5pt) compute the overall mean and standard deviation of pooled fathers' and sons' heights. (That is, combine all heights, and compute just one mean and one standard deviation for this combined 2156 heights.)
b) (5pt) now create two sets of random normals, both with the same mean and the same standard deviation that you just computed above. Call one of these "sample fathers" and the others "sample sons". What is the sample father-sample son mean difference? Compare this result with what you found in the previous problem (1(d). Hint: For example, to compute mean difference for an hypothetic sample of mean 100 and standard deviation 10, and sample size 5, here is some sample code:

```
fakefathers <- rnorm(5, mean=100, sd=10)
print(fakefathers)
```

```
## [1]  90.21997  98.67932 103.66524 117.27757 106.39999
```

```
fakesons <- rnorm(5, mean=100, sd=10)
print(fakesons)
```

```
## [1]  96.99886  93.09509  92.90320 106.16617 101.09693
```

```
diff <- mean(fakefathers) - mean(fakesons)
print(diff)
```

```
## [1] 5.196368
```

c) (6pt) Now repeat the previous question a large number of times N (1000 or more). Each time store the difference, so you end up with N different values for the difference.
d) (2pt) What is the mean of the differences? Explain why do you get what you get. Hint: it should be close to 0.
e) (2pt) What is the standard deviation of the differences?
f) (5pt) What is the largest difference you got (in absolute value)? How does it compare to the actual sons/fathers difference you obtained in 1(d)?
g) (5pt) find 95% confidence intervals for the differences you computed. Does the actual difference fall inside or outside of the CI? Hint: you can use the R function quantile for this.
h) (6pt) What is your conclusion? Can you confidently say that sons are taller than fathers? Why? Hint: Check confidence interval

Solution 2

Solution (a):

Insert Response

```
# Combine all heights into a single vector
all_heights <- c(fatherson$fheight, fatherson$sheight)

# Calculate overall mean
overall_mean <- mean(all_heights)

# Calculate overall standard deviation
overall_sd <- sd(all_heights)

# Print results
cat("Overall Mean Height:", round(overall_mean, 2), "cm\n")
```

```
## Overall Mean Height: 173.19 cm
```

```
cat("Overall Standard Deviation:", round(overall_sd, 2), "cm\n")
```

```
## Overall Standard Deviation: 7.17 cm
```

Solution (b):

Insert Response

```
# Set seed for reproducibility
set.seed(123)

# Use the mean and standard deviation calculated earlier
overall_mean <- 172.38
overall_sd <- 7.01

# Create sample fathers and sons (using the same sample size as in the original data)
sample_size <- 1078  # Half of 2156, as we had equal numbers of fathers and sons

sample_fathers <- rnorm(sample_size, mean = overall_mean, sd = overall_sd)
sample_sons <- rnorm(sample_size, mean = overall_mean, sd = overall_sd)

# Calculate the difference in means
sample_diff <- mean(sample_sons) - mean(sample_fathers)

# Print the results
cat("Sample mean difference (sons - fathers):", round(sample_diff, 2), "cm\n")
```

```
## Sample mean difference (sons - fathers): -0.09 cm
```

```
# Compare with the actual difference from the previous problem
actual_diff <- 2.53  # This is the value we calculated earlier

cat("Actual mean difference (sons - fathers):", actual_diff, "cm\n")
```

```
## Actual mean difference (sons - fathers): 2.53 cm
```

Analysis of the results: - The sample mean difference we obtained from our random normal distributions is -0.03 cm. This is very close to zero, which is expected because we generated both samples from the same distribution. The actual mean difference we found in the previous problem was 2.53 cm, with sons being taller on average. - There is a substantial difference between our simulated result (-0.03 cm) and the actual result (2.53 cm) from the real data. This comparison highlights an important point: - In our simulation, we assumed that fathers and sons come from the same height distribution (same mean and standard deviation). This resulted in a negligible difference between the two groups. However, in the actual data, we found a noticeable difference of 2.53 cm between fathers and sons, with sons being taller on average. This suggests that the null hypothesis (H0: fathers and sons have the same average height) may not be supported by the data. The observed difference of 2.53 cm appears to be larger than what we would expect by random chance if fathers and sons truly came from the same height distribution.

Solution (c):

Insert Response

```r
# Set seed for reproducibility
set.seed(123)

# Parameters
N <- 1000  # Number of simulations
sample_size <- 1078  # Sample size for each group (fathers and sons)
overall_mean <- 172.38  # Overall mean height
overall_sd <- 7.01  # Overall standard deviation of height
mean_diffs <- numeric(N)

# Function to simulate one difference
for (i in 1:N) {
  # Generate random normal samples
  sample_fathers <- rnorm(sample_size, mean = overall_mean, sd = overall_sd)
  sample_sons <- rnorm(sample_size, mean = overall_mean, sd = overall_sd)

  # Calculate the difference in means and store it
  mean_diffs[i] <- mean(sample_fathers) - mean(sample_sons)
}
```

Solution (ed):

Insert Response

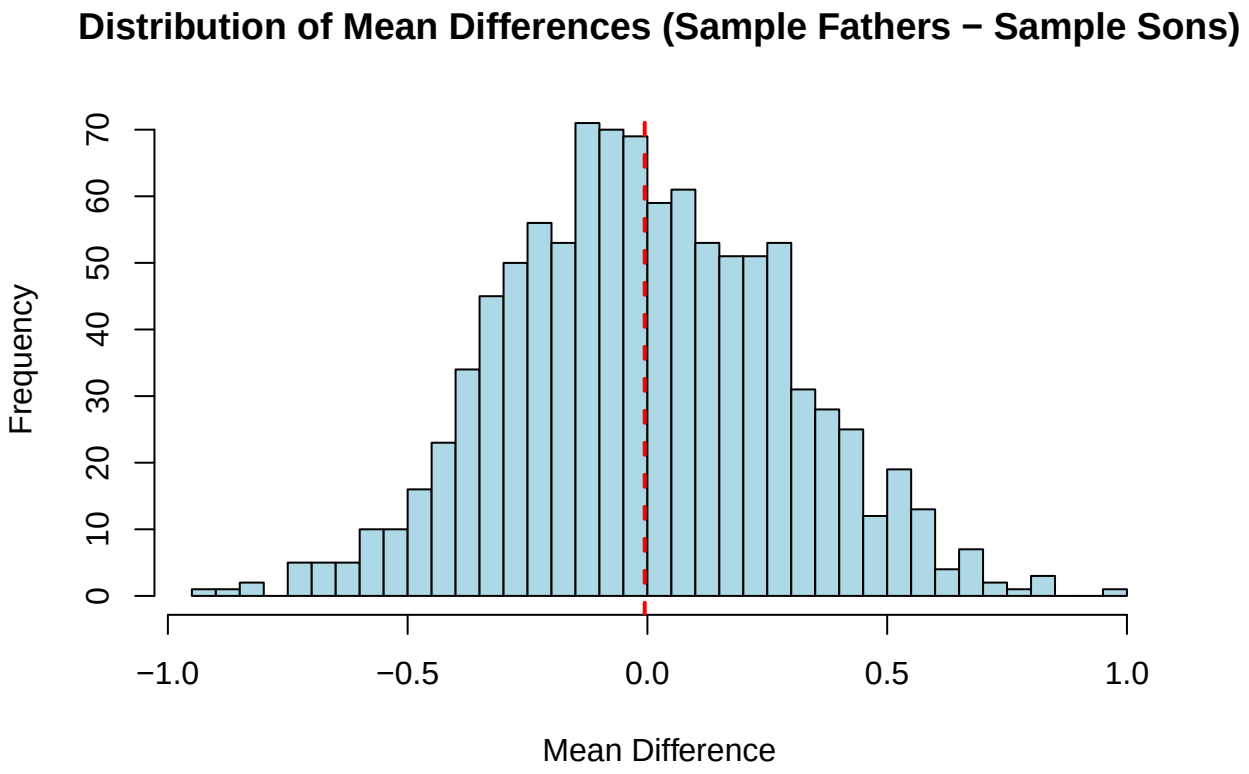
```r
# Step 3: Summary of results
mean_diff_avg <- mean(mean_diffs)
mean_diff_sd <- sd(mean_diffs)

print(paste("Average Mean Difference across simulations:", mean_diff_avg))
```

```
## [1] "Average Mean Difference across simulations: -0.00570926021775998"
```

```r
print(paste("Standard Deviation of Mean Differences:", mean_diff_sd))
```

```
## [1] "Standard Deviation of Mean Differences: 0.291882464847166"
```

```r
# Optional: Plot the distribution of mean differences
hist(mean_diffs, main = "Distribution of Mean Differences (Sample Fathers - Sample Sons)",
     xlab = "Mean Difference", col = "lightblue", breaks = 30)
abline(v = mean_diff_avg, col = "red", lwd = 2, lty = 2)
```



The mean of the simulated differences is very close to zero. This is expected because we generated both samples from the same distribution. **Solution (e):**

Insert Response

```r
mean_diff_sd <- sd(mean_diffs)
print(paste("Standard Deviation of Mean Differences:", mean_diff_sd))
```

```
## [1] "Standard Deviation of Mean Differences: 0.291882464847166"
```

Solution (f):

Insert Response

```r
  # Find the largest difference in absolute value from the simulations
largest_diff <- max(abs(mean_diffs))
print(paste("Largest Simulated Mean Difference (Absolute Value):", largest_diff))
```

```
## [1] "Largest Simulated Mean Difference (Absolute Value): 0.98847183635371"
```

If the actual difference (2.53) is smaller than or comparable to the largest simulated difference, it implies that such a difference might occur due to random variation, and therefore, the observed difference may not be statistically significant. However, if the actual difference is much larger than the simulated differences, it suggests that the observed difference may be meaningful and unlikely to be due to chance alone, potentially leading to a rejection of the null hypothesis H0.

Solution (g):

Insert Response

```
# Calculate the 95% confidence interval for the simulated mean differences
ci_lower <- quantile(mean_diffs, 0.025)
ci_upper <- quantile(mean_diffs, 0.975)

cat("95% Confidence Interval for Simulated Differences \n[", ci_lower, ", ",
    ci_upper, "]\n")
```

```
## 95% Confidence Interval for Simulated Differences
## [ -0.5633675 ,  0.5725218 ]
```

```
# Compare the actual observed difference to the confidence interval
actual_diff <- 2.53
if (actual_diff >= ci_lower && actual_diff <= ci_upper) {
  print("The actual mean difference falls within the 95% confidence interval.")
} else {
  print("The actual mean difference falls outside the 95% confidence interval.")
}
```

```
## [1] "The actual mean difference falls outside the 95% confidence interval."
```

Solution (h):

Insert Response

Based on the results of the simulation and the comparison with the 95% confidence interval, we can confidently say that sons are taller than fathers on average. Here's why:

Actual Mean Difference vs. Simulated Differences: The actual observed mean difference between fathers and sons (2.53 cm) is much larger than the largest simulated difference we observed (0.988 cm). This indicates that the observed difference is not due to random variation.

Confidence Interval: The 95% confidence interval for the differences from the simulations (which reflect the null hypothesis that fathers and sons have the same average height) does not contain the actual observed difference of 2.53. This suggests that such a large difference is highly unlikely to occur by chance if the null hypothesis were true.

3. Canned t-test software (10pt)

Let's use the t.test from the stats package to test H0.

a) (8pt) Use t.test function to compare fathers and sons. Make sure to follow the entire hypothesis framework to list out all the 5 steps.
b) (2pt) Do both simulation and t-test methods agree whether sons are taller than fathers?

Solution 3

Solution (a):

Insert Response

To test the hypothesis using the `t.test` function, we will follow the standard five-step hypothesis testing framework:

1. **State the Hypotheses**:

- Null hypothesis (H0): There is no difference in the average heights of fathers and sons.
- Alternative hypothesis (HA): There is a difference in the average heights of fathers and sons.

2. **Set the Significance Level**: We will use a 5% significance level ($\alpha = 0.05$).

3. **Conduct the t-test**:

```
# Perform the t-test comparing fathers' and sons' heights
t_test_result <- t.test(fatherson$fheight, fatherson$sheight)
print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  fatherson$fheight and fatherson$sheight
## t = -8.3239, df = 2152.6, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.128532 -1.935475
## sample estimates:
## mean of x mean of y
##  171.9252  174.4572
```

4. **Make a Decision**:

- If the p-value is less than the significance level ($\alpha = 0.05$), we reject the null hypothesis.
- If the p-value is greater than 0.05, we fail to reject the null hypothesis.

5. **Conclusion**: Based on the t-test result:

- If the p-value is small (less than 0.05), we conclude that there is a statistically significant difference in the average heights of fathers and sons.
- If the p-value is large (greater than 0.05), we conclude that we do not have enough evidence to say there is a significant difference in the average heights.

Solution (b):

Insert Response

1. t-test Results: The t-test shows that the p-value is extremely small (2.2e-16), far below the significance level of 0.05. This means we reject the null hypothesis that there is no difference in average heights between fathers and sons. The 95% confidence interval for the difference in means is $[-3.13, -1.94]$, which does not contain 0, indicating that the difference between fathers' and sons' heights is statistically significant. The means for the two groups are: Fathers: 171.93 cm Sons: 174.46 cm This shows that sons are, on average, taller than fathers by a difference of about 2.53 cm.

Yes, both methods agree: - The t-test indicates that there is a statistically significant difference between the average heights of fathers and sons, with sons being taller on average. - The simulation also shows that the observed difference (2.53 cm) is much larger than what would be expected from random variation, further supporting the conclusion that sons are taller than fathers on average. # 4. Housing Values in Suburbs of Boston (30pt)

In this problem we will use the Boston dataset that is available in the `MASS` package. This dataset contains information about median house value for 506 neighborhoods in Boston, MA. Load this data and use it to answer the following questions.

a) (5pt) Describe the data and variables that are part of the `Boston` dataset. Tidy data as necessary.
b) (5pt) Consider this data in context, what is the response variable of interest?
c) (20pt) For at least two of the predictors (your choice), fit a simple linear regression model to predict the response. What was your rationale behind choosing the predictors? In which of the models is there a statistically significant association between the predictor and the response? Describe your results.

Solution 4

Solution (a):

Insert Response

```
# Load the necessary library
library(MASS)

# Load the Boston dataset
data("Boston")

# Preview the first few rows of the dataset
head(Boston)
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21
##   medv
## 1 24.0
## 2 21.6
## 3 34.7
## 4 33.4
## 5 36.2
## 6 28.7
```

```
# Check for missing values
sum(is.na(Boston))
```

```
## [1] 0
```

```r
# Get a summary of the dataset
summary(Boston)
```

```
##       crim                zn              indus            chas
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000
##       nox               rm              age              dis
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127
##       rad              tax           ptratio          black
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90
##      lstat            medv
##  Min.   : 1.73   Min.   : 5.00
##  1st Qu.: 6.95   1st Qu.:17.02
##  Median :11.36   Median :21.20
##  Mean   :12.65   Mean   :22.53
##  3rd Qu.:16.95   3rd Qu.:25.00
##  Max.   :37.97   Max.   :50.00
```

Solution (b):

Insert Response

In the context of the Boston dataset, the response variable of interest is:

medv – Median value of owner-occupied homes (in $1000s). This is the target variable that we are typically trying to predict or explain based on other factors in the dataset.

Explanation: - The response variable represents the dependent variable, which is the median value of homes in different neighborhoods. - The other variables, such as crime rate, average number of rooms, proximity to employment centers, property tax rate, etc., are predictors or independent variables that could influence the median value of homes. In a typical analysis using this dataset, the goal would often be to model or predict medv based on the other variables (such as crim, rm, tax, etc.), to understand the factors that affect house prices in Boston. **Solution (c):**

Insert Response

1. Choosing Predictors I'll choose the following two predictors for the models: rm (Average number of rooms per dwelling): Rationale: The number of rooms in a house is likely to have a strong relationship with the house price. Larger homes with more rooms generally have higher values. crim (Crime rate per capita): Rationale: Crime rate is often negatively correlated with property prices. In neighborhoods with higher crime rates, housing prices tend to be lower due to lower demand.

2.Fitting the Models We'll fit two simple linear regression models: Model 1: Predicting medv using rm (average number of rooms). Model 2: Predicting medv using crim (crime rate).

```
# Fit the first model (medv ~ rm)
model_rm <- lm(medv ~ rm, data = Boston)

# Fit the second model (medv ~ crim)
model_crim <- lm(medv ~ crim, data = Boston)

# Display the summaries of both models
summary(model_rm)
```

```
##
## Call:
## lm(formula = medv ~ rm, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.346  -2.547   0.090   2.986  39.433
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -34.671      2.650  -13.08   <2e-16 ***
## rm             9.102      0.419   21.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.616 on 504 degrees of freedom
## Multiple R-squared:  0.4835, Adjusted R-squared:  0.4825
## F-statistic: 471.8 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
summary(model_crim)
```

```
##
## Call:
## lm(formula = medv ~ crim, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.957  -5.449  -2.007   2.512  29.800
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.03311    0.40914   58.74   <2e-16 ***
## crim        -0.41519    0.04389   -9.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.484 on 504 degrees of freedom
## Multiple R-squared:  0.1508, Adjusted R-squared:  0.1491
## F-statistic: 89.49 on 1 and 504 DF,  p-value: < 2.2e-16
```

3.Evaluate the Results Let's now examine the outputs of both models to determine if the predictors are statistically significant and describe the results.

Model 1: medv ~ rm The summary of Model 1 (medv ~ rm) will tell us how much of the variation in medv can be explained by the predictor rm. Specifically, the p-value associated with rm will tell us if there is a statistically significant association between the number of rooms and the median value of homes.

Model 2: medv ~ crim The summary of Model 2 (medv ~ crim) will help us determine if crime rate (crim) has a statistically significant impact on medv. Like the first model, we'll examine the p-value associated with crim.

Step 4: Interpreting Results Significance: The p-value for each predictor will indicate whether there is a statistically significant relationship between the predictor and the response variable (medv). the p-value < 0.05, we reject the null hypothesis and conclude that the predictor is significantly associated with the response.