ST 518 Homework 1

Eric Warren

September 1, 2023

Contents

1 Problem 1

Read in the data to use for the problem

```
library(tidyverse)
heights <- read_table("~/ST-518/heights-tall.txt")
heights
```

```
## # A tibble: 928 x 2
##    parent   son
##     <dbl> <dbl>
##  1   73    72.2
##  2   73    73.2
##  3   73    73.2
##  4   73    73.2
##  5   72.5  68.2
##  6   72.5  69.2
##  7   72.5  69.2
##  8   72.5  70.2
##  9   72.5  71.2
## 10   72.5  71.2
## # i 918 more rows
```

We can see this data has 928 rows with two columns. The rows represent the observation of the parent's height represents the first column and the son's height represents the second column. Height seems to be in inches.

1.1 Part A

Consider the population from which this simple random sample of adult males was drawn. Let μ denote the mean of the population. Use statistical software to compute $\bar{y}$ as an estimate of μ. Obtain a 95% confidence interval for μ. If another person is sampled at random, would you expect this interval to capture this person's height with high confidence?

Since we are trying to look at y variable, we are going to find the mean of the `son` column and we are going to make a 95% confidence interval of this variable.

```
library(stats)
# Find the mean of adult heights
y_bar <- mean(heights$son)
y_bar
```

```
## [1] 68.08847
```

```
# Find the confidence interval
confidence_adult <- t.test(heights$son, conf.level = 0.95)
confidence_adult$conf.int[1]  # Lower confidence value
```

```
## [1] 67.92626
```

```
confidence_adult$conf.int[2]  # Upper confidence value
```

```
## [1] 68.25068
```

Here we can see that the $\bar{y} = 68.0884698$ inches. We are also 95% confident that the true average height of adult men is between 67.9262563 and 68.2506834 inches. Despite obtaining this interval, this does not mean that adding another observation will be within this amount will occur 95% of the time. We are saying that we believe that 95 out of 100 samples taken the same way as this one will occur the true population mean. Let us look at this sample for example. We can see that 120 out of 928 observations are within the confidence interval. This just gives us proof that we should not be overly confident that sample one additional observation will necessarily be within a confidence interval.

1.2 Part B

Here we are going to find some important summary statistics.

1.2.1 Part i

Finding $\bar{y} = \frac{1}{n} * \sum y_i$. This is the son average height.

```
y_bar <- mean(heights$son)
y_bar
```

```
## [1] 68.08847
```

Here we can see that the $\bar{y} = 68.0884698$ inches.

1.2.2 Part ii

Finding $\bar{x} = \frac{1}{n} * \sum x_i$. This is the midparent average height.

```
x_bar <- mean(heights$parent)
x_bar
```

```
## [1] 68.30819
```

Here we can see that the $\bar{x} = 68.3081897$ inches.

1.2.3 Part iii

Finding $S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y})$.

```
S_xy <- sum((heights$parent - x_bar) * (heights$son - y_bar))
S_xy
```

```
## [1] 1913.898
```

Here we can see that S_{xy} is 1913.8976293.

1.2.4 Part iv

Finding $s_{xy} = \frac{1}{n-1} * \sum(x_i - \bar{x})(y_i - \bar{y})$.

```
s_xy <- (1/(nrow(heights) - 1)) * sum((heights$parent - x_bar) * (heights$son -
    y_bar))
s_xy
```

```
## [1] 2.064614
```

Here we can see that s_{xy} is 2.0646145.

1.2.5 Part v

Finding $\frac{1}{n-1} \sum (x_i - \bar{x}) y_i$

```
sum_v <- (1/(nrow(heights) - 1)) * sum((heights$parent - x_bar) * heights$son)
sum_v
```

```
## [1] 2.064614
```

We can find this value to be 2.0646145.

1.2.6 Part vi

Finding $\frac{1}{n-1} \sum x_i (y_i - \bar{y})$

```
sum_vi <- (1/(nrow(heights) - 1)) * sum((heights$son - y_bar) * heights$parent)
sum_vi
```

```
## [1] 2.064614
```

We can find this value to be 2.0646145.

1.2.7 Part vii

Finding $S_{xx} = \sum(x_i - \bar{x})^2$

```
S_xx <- sum((heights$parent - x_bar)^2)
S_xx
```

```
## [1] 2961.358
```

Here we can see that S_{xx} is 2961.3577586.

1.2.8 Part viii

Finding $s_x^2 = \frac{1}{n-1} * \sum(x_i - \bar{x})^2$

```
s_x2 <- (1/(nrow(heights) - 1)) * sum((heights$parent - x_bar)^2)
s_x2
```

```
## [1] 3.194561
```

Here we can see that s_x^2 is 3.1945607.

1.2.9 Part ix

Finding $S_{yy} = \sum(y_i - \bar{y})^2$

```
S_yy <- sum((heights$son - y_bar)^2)
S_yy
```

```
## [1] 5877.207
```

Here we can see that S_{yy} is 5877.2066272.

1.2.10 Part x

Finding $s_y^2 = \frac{1}{n-1} * \sum(y_i - \bar{y})^2$

```
s_y2 <- (1/(nrow(heights) - 1)) * sum((heights$son - y_bar)^2)
s_y2
```

```
## [1] 6.340029
```

Here we can see that s_y^2 is 6.3400287.

1.2.11 Part xi

Finding r. In this case, the equation we are using to find r is $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{xy}}}$

```
r <- S_xy/(sqrt(S_xx * S_xy))
r
```

```
## [1] 0.803922
```

Here we can see the r value is 0.803922

1.3 Part C

We are going to find the slope using r, s_x, and s_y. We will call the slope, b. The equation to find b is $b = r * \frac{s_y}{s_x}$. Note we have s_x^2 and s_y^2 so we will have to take the square roots of these values to use them in the equation.

```r
slope <- r * (sqrt(s_y2)/sqrt(s_x2))
slope
```

```
## [1] 1.132541
```

Here we can see that with our x variable being midparent's height and our y variable is son's height that our slope is 1.1325411.

2 Problem 2

When solving parts we are going to use the following information:

- Pre-treatment is the x variable and post-treatment is the y variable
- $r = 0.7$
- $\bar{x} = 10.7$
- $\bar{y} = 7.9$
- $s_x = 4.8$
- $s_y = 6.7$
- $n = 30$

2.1 Part A

Obtain a 95% confidence interval for the mean of the population from which the sample was drawn, post-treatment. The formula we will use is confidence interval $= \bar{y} \pm t_{\alpha/2,n-2} * SE_y = \bar{y} \pm t_{\alpha/2,n-2} * \frac{s_y}{\sqrt{n}}$. We use t instead of z for our test-statistic since the population standard deviation is unknown (σ is not known).

```r
r_treatment <- 0.7
x_bar_treatment <- 10.7
y_bar_treatment <- 7.9
s_x_treatment <- 4.8
s_y_treatment <- 6.7
n_treatment <- 30

# This is the t critical value with a 95% confidence interval with
# n-2 (or 30-2=28) degrees of freedom
t_treatment <- qt(p = 0.05/2, df = n_treatment - 2, lower.tail = FALSE)

# Find lower and upper confidence bounds
post_treat_confidence_interval_lower <- y_bar_treatment - (t_treatment *
    (s_y_treatment/sqrt(n_treatment)))  #Lower confidence bound
post_treat_confidence_interval_lower
```

```
## [1] 5.394292
```

```r
post_treat_confidence_interval_upper <- y_bar_treatment + (t_treatment *
    (s_y_treatment/sqrt(n_treatment)))  #Upper confidence bound
post_treat_confidence_interval_upper
```

```
## [1] 10.40571
```

We are also 95% confident that the mean count of the bacteria from post-treatment is between 5.394292 and 10.405708.

2.2 Part B

Obtain the least squares estimates of the slope and intercept in a linear regression model in which the mean of Y is linear in x. Please note that we can find the slope of a regression line by $b = r * \frac{s_y}{s_x}$. The intercept can be found by $a = \bar{y} - b * \bar{x}$

```
slope_treatment <- r_treatment * (s_y_treatment/s_x_treatment)
slope_treatment
```

```
## [1] 0.9770833
```

```
intercept_treatment <- y_bar_treatment - (slope_treatment * x_bar_treatment)
intercept_treatment
```

```
## [1] -2.554792
```

We get a regression line that we analyze the pre-treatment and post-treatment bacteria counts by saying that the predicted bacteria count post_treatment = -2.5547917 + 0.9770833 * pre-treatment bacteria count. Also known as $\hat{y}$ = -2.5547917 + 0.9770833 * x.

2.3 Part C

Use the regression line to estimate mean post-treatment score among those with average pre-treatment levels of the bacteria (x = $\bar{x}$).

```
average_post <- intercept_treatment + slope_treatment * x_bar_treatment
average_post
```

```
## [1] 7.9
```

Using our regression line, we would expect the post-treatment bacteria count to be 7.9 if the pre-treatment bacteria count is the average amount of 10.7 (when x = $\bar{x}$).

2.4 Part D

Estimate the standard error of the estimate in part(c). Since we are predicting the y values from x, we should use the formula: Standard Error = $\sqrt{\frac{MSE}{n}}$. Now MSE = $\frac{SSE}{n-2}$ and SSE = $(\sum(y_i - \bar{y})^2) * (1 - r^2) = ((s_y^2) * (n-1)) * (1 - r^2)$. We are showing all of this below.

```
sum_of_squares_error_treatment <- (s_y_treatment^2) * (n_treatment - 1) *
    (1 - (r_treatment^2))
mean_square_error_treatment <- sum_of_squares_error_treatment/(n_treatment -
    2)
standard_error_treatment <- sqrt(mean_square_error_treatment/n_treatment)
standard_error_treatment
```

```
## [1] 0.8890358
```

We can see here that the standard error of our estimate is 0.8890358.

2.5 Part E

Estimate the standard deviation of the estimate in part(c). Since we are predicting the y values from x, we should use the formula: Standard Deviation = standard error * $\sqrt{n}$.

```
standard_deviation_treatment <- standard_error_treatment * sqrt(n_treatment)
standard_deviation_treatment
```

```
## [1] 4.86945
```

We can see here that the standard deviation of our estimate is 4.8694496.

2.6 Part F

Here we can see that there is a difference between the standard error and the standard deviation. The standard error is equal to the standard deviation divided by the square root of n (or known as the sample size). Standard deviation tends to describe a single sample's variation between data points, while standard error describes multiple samples variability that come from the population. We can also say that standard deviation assesses how far a data point likely falls from the mean, while standard error assesses how far a sample statistic likely falls from a population parameter.

2.7 Part G

Use the regression line to obtain 95% confidence limits for the mean post-treatment score among those with average pre-treatment levels of the bacteria (x = $\bar{x}$). Remember, the formula we will use is confidence interval = $\bar{y} \pm t_{\alpha/2,n-2} * SE_y$. We use t instead of z for our test-statistic since the population standard deviation is unknown (σ is not known). Compare the width of this interval with that from part (a). Explain, briefly, how and why it is different.

```
lower_new_conf_interval <- average_post - (t_treatment * standard_error_treatment)
lower_new_conf_interval  # Lower confidence bound
```

```
## [1] 6.078893
```

```
upper_new_conf_interval <- average_post + (t_treatment * standard_error_treatment)
upper_new_conf_interval  # Upper confidence bound
```

```
## [1] 9.721107
```

We are 95% confident that the mean count of the bacteria from post-treatment is between 6.0788927 and 9.7211073, using the least squares regression model to try to predict this. From part (a), we were 95% confident that the mean count of the bacteria from post-treatment is between 5.394292 and 10.405708. We can see that by using regression, we have a narrower confidence interval and thus we now have a more precise estimate on the population parameter, which is the mean number of bacteria post-treatment. This is a good thing, since our goal is trying our best to estimate the population parameter and a narrower confidence interval (or smaller standard error) is what we want when making these estimates.