Quantitative_Methods_Assignment

Firstname Lastname

T-tests and linear regression: A comparison

T-test Recap

One Sample T-test

Remember that, for a random variable, X, the standard error of $\bar{x}$ for a sample of size n is $SE(\bar{x}) = \frac{\sigma}{\sqrt{n}}$; therefore, for a single sample of size n, we can estimate the standard error of $\bar{x}$ using $SE(\bar{x}) = \frac{s}{\sqrt{n}}$ where s is the sample standard deviation.

This can be used to derive a T statistic, which can then be compared to a t distribution with $n-1$ degrees of freedom.

Two (Independent) Sample T-test

Remember that we can estimate the standard error of $\bar{x_1} - \bar{x_2}$ using the following formula (derived in the math review packet) :

$$SE(\bar{x_1} - \bar{x_2}) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

In this formula, s_1 and s_2 represent the sample standard deviations for group 1 and group 2, respectively. If we believe that the two populations have unequal variances, then we can estimate s_1 and s_2 separately. If we believe that the two populations have equal variances, then we can say $s_1 = s_2 = s_{pooled}$ where:

$$s_{pooled} = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}$$

We can then use the estimated standard error to calculate a test statistic, which we can then compare to a t distribution with degrees of freedom as follows:
In general, we use the smaller of $n_1 - 1$ and $n_2 - 1$ as the degrees of freedom for the t distribution.
If assuming equal variances and using the s_{pooled}, we can use $n_1 + n_2 - 2$ degrees of freedom.

Linear Regression with a Dummy Variable

First, let's read in some data. In order to read in the file, you will need to download height_sex.csv to your computer and include the full file path in the function call below. For example, I could put the file in my "Downloads" folder and call:
height <- read.csv("/Users/sophiesommer/Downloads/height_sex.csv")

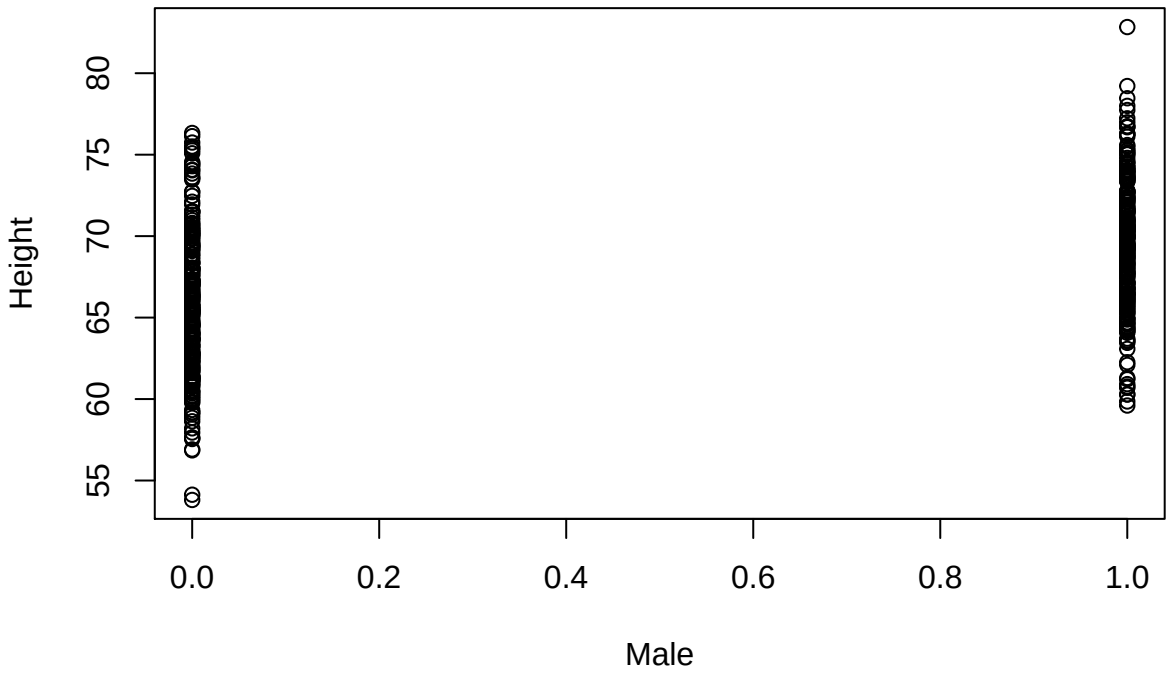
```
#Read in data
height <- read.csv("height_sex.csv")

#Inspect the first few rows
height[1:5,]
```

```
##   X id Gender Birth.Order   Height    Weight Male
## 1 1  1      M           1 65.44441 166.14611    1
## 2 2  2      F           2 65.44836  76.74407    0
## 3 3  3      M           2 68.39149 140.14253    1
## 4 4  4      M           1 73.92210 165.30409    1
## 5 5  5      F           1 70.98392 117.61932    0
```

For this assignment, we are going to focus on the "Height" and "Male" variables. "Height" is given in inches and "Male" is an indicator variable (equal to 1 if the person is male and equal to 0 if the person is female). If we plot height against the gender indicator variable, we get:

```
plot(height$Male, height$Height, xlab="Male", ylab="Height")
```



Suppose for a moment that you only had one set of heights, say, $60, 63, 63, 70, 72$ and you wanted to find some value, M, such that the sum of the differences between M and each value in the set is minimized. It turns out that the value of M which satisfies this minimization problem is the mean of the dataset (you can prove this with some relatively simple calculus. We leave the proof to the interested reader, but hopefully this claim already feels somewhat intuitive).

Now, suppose we try to fit an OLS regression line to this data. This would involve finding values of β_0 and β_1 such that the vertical distance between each point (in the above graph) and the line $Height = \beta_0 + \beta_1 * Male$ is minimized.

There are only two possible values of "Male" in this dataset: 0 and 1. Therefore, one way to think about minimizing the vertical distances between each data point and the line is this: Find the mean height of women (i.e., everyone with Male=0), find the mean height of men (i.e., everyone with Male=1), and connect the two points $(0, \text{mean womens height})$ and $(1, \text{mean mens height})$.

Now we can try to write the equation for this line in the form $Height = \beta_0 + \beta_1 * Male$. Using the strategy described above, we know that $(0, \text{mean womens height})$ is the y-intercept of the line. Therefore $\beta_0 =$ mean womens height. We also know that β_1 is the slope of the line. Since we know two points on the line, we can calculate the slope as $\frac{y_2 - y_1}{x_2 - x_1} = \frac{(\text{mean mens height} - \text{mean womens height})}{(1-0)} =$ mean mens height $-$ mean womens height. So, $\beta_1 =$ mean mens height $-$ mean womens height.

Assignment

1. The mean height of men is calculated below and saved as MM. Fill in the rest of the code to a) save the mean height of women as MW and b) print the mean height of women

```
# Save mean height of men as MM
MM <- mean(height$Height[height$Male==1])

# Print mean height of men
MM
```

```
## [1] 69.55706
```

```
# Save mean height of women as MM

# Print mean height of women
```

2. Calculate the mean difference for the two groups and plot the regression line

a) Save and print: MeanDiff = MM - MW

b) Add (0, MW) and (1,MM) to the plot from above (in yellow)

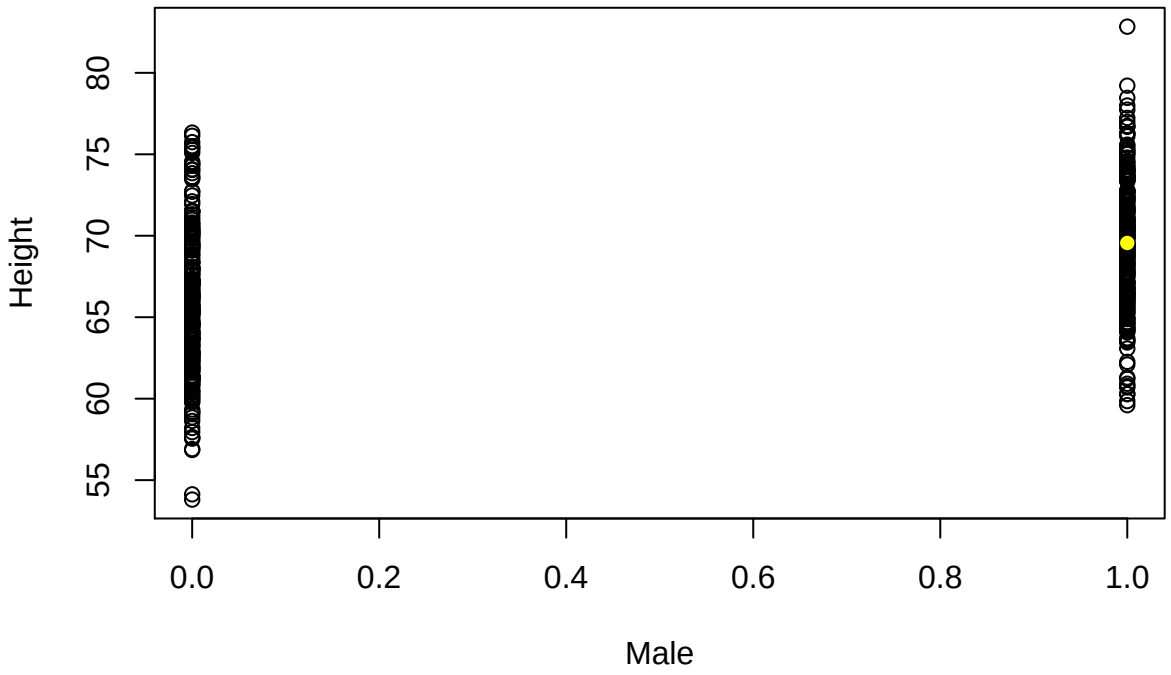
c) Draw a line connecting (0, MW) and (1,MM) using the abline() function

```
# Save the difference in means (MM-MW) as the variable MeanDiff:

# Print MeanDiff

# Re-plot the data
plot(height$Male, height$Height, xlab="Male", ylab="Height")

# Add (0, MW) to the plot (in yellow) (note: (1,MM) has been added for you)
# Note: pch=16 makes the dot filled in, which makes it easier to see
points(1, MM, col=7, pch=16)
```




```
# Draw the line Height=b_0 + b_1*Male where b_0=MW and b_1=MeanDiff
# Use the abline() function which takes parameters a and b where a=y-intercept and b=slope
# Syntax is: abline(a,b)
```

3. Estimate the standard error for an independent sample T-test by hand (note: assume pooled standard deviation)

a) Calculate nM and nW, the number of men and women in the sample, respectively

b) Calculate sM and sW, the sample standard deviations for the men's heights and women's heights, respectively

c) Calculate s_{pooled} for the two groups using $s_{pooled} = \sqrt{\frac{s_1^2(n_1-1)+s_2^2(n_2-1)}{n_1+n_2-2}}$ and save the result as spool

d) Use the following equation: $SE(\bar{x_1} - \bar{x_2}) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ to estimate $SE(\bar{x_1} - \bar{x_2})$ where $s_1 = s_2 = s_{pooled}$ and $n_1 = nW$, $n_2 = nM$. Save the result as SEMeanDiff and print the result.

```
# Add code to calculate the number of women and save the result as nW
nM <- length(height$Height[height$Male==1])

# Add code to calculate sW, the sample standard deviation for women's heights
sM <- sd(height$Height[height$Male==1])

# Calculate pooled standard deviation and save as spool
# (note: this is done for you; all you need to do is uncomment the line below)
# spool <- sqrt((sW^2 * (nW - 1) + sM^2 * (nM - 1))/(nM + nW - 2))

# Print spool

# Estimate the standard error of the difference in means and save it as SEMeanDiff

# Print SEMeanDiff
```

4. Calculate a T-Statistic by hand to compare the null hypothesis that $MM - MW = 0$ versus the alternative hypothesis that $MM - MW \neq 0$

a) Calculate $T = \frac{(\text{MeanDiff})-0}{\text{SEMeanDiff}}$ and save the result as Ttest

b) Use the pt function to calculate a p-value for Ttest (degrees freedom= nW+nM-2). Hint: You will want to use the parameter lower.tail=FALSE because Ttest should be > 0 for this example, then multiply by 2 to get a 2-sided test

```
# Calculate a test statistic and save it as Ttest

# Print Ttest

# Use pt to calculate a p-value
```

5. Do the same thing using the t.test() function in R. Compare the value of t and p-value computed by the t.test() function to the values you calculated above. Note that setting var.equal=TRUE causes t.test() to use pooled variance. Do you get the same value (to two decimal places) for the T statistic?

```
# Use the t.test function to compare the men's heights and female's heights
# Note: the code has already been written for you (just uncomment the line below to run it)

# t.test(height$Height[height$Male==1],height$Height[height$Male==0], var.equal = TRUE)
```

6. Use the lm() function to implement ordinary least squares regression to estimate β_0 and β_1 in the equation $Height = \beta_0 + \beta_1 * Male$.

```
# Use lm() to estimate b_0 and b_1 in the equation Height=b_0+b_1*Male; save the result as linmod
# Note: this has been done for you
linmod <- lm(Height~Male, data=height)

#Inspect a summary of the results
summary(linmod)
```

```
##
## Call:
## lm(formula = Height ~ Male, data = height)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.919  -2.957  -0.076   2.607  13.275
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  65.7260     0.2576  255.14   <2e-16 ***
## Male          3.8311     0.3695   10.37   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.13 on 498 degrees of freedom
## Multiple R-squared:  0.1775, Adjusted R-squared:  0.1759
## F-statistic: 107.5 on 1 and 498 DF,  p-value: < 2.2e-16
```

7. Reprint the following values that you saved in previous problems: Mean women's height (MW), difference in mean height for men and women (MeanDiff), standard error of this mean difference (SEMeanDiff), the T statistic for the difference in means (Ttest). Which of these values can you find in the lm() output above?

```
# Re-print MW

# Re-print MeanDiff

# Re-print SEMeandiff

# Re-print Ttest
```

8. Note that the T-test for the intercept in the lm() output above tests the null hypothesis: $\beta_0 = 0$ against the alternative hypothesis: $\beta_0 \neq 0$. In the space below, please show how you could calculate the Std. Error (.2576) and the t value (255.14) of the intercept without using lm() or t.test(). Hint: lm() uses s_{pooled} to estimate the standard deviation of heights for both groups.

```
# Estimate standard error of b_0
# (hint: remember that b_0=MW and you can estimate the SE of the mean using the central limit theorem)

# Use the standard error of b_0 to calculate the t value
```