Classwork 7: Hypothesis Testing (Part 2)

In chapter 6 of the workbook, we determined that (given some assumptions): $\hat{\beta}_1$ is distributed $N(\beta_1, \frac{\sigma_u^2}{\sum_i (x_i - \bar{x})^2})$. So the variance of $\hat{\beta}_1$ is $\frac{\sigma_u^2}{\sum_i (x_i - \bar{x})^2}$.

Let the mean squared deviance of X be: $MSD(x) = \frac{1}{n} \sum_i (x_i - \bar{x})^2$. This is very much related to the estimate of the variance of X ($\hat{\sigma}_X^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$). Then another formula for the variance of $\hat{\beta}_1$ is:

$$\frac{\sigma_u^2}{n\text{MSD}(x)}$$

1) When we fit a model, we should prefer more precise estimates of model parameters. That is, if we can easily take steps to decrease the variance of $\hat{\beta}_1$, we should take those steps because then we would get a more precise estimate of the relationship between X and Y. Would we be more certain about our estimate of β_1 if we increased the sample size n? Why/why not? Draw a picture to demonstrate this idea. (Hint: reference the formula $\text{Var}(\hat{\beta}_1) = \frac{\sigma_u^2}{n\text{MSD}(x)}$)

2) Would we be more certain about our estimate of β_1 if the explanatory variable X was more spread out, and why/why not? Draw a picture to demonstrate this idea.

3) Would we be more certain about our estimate of β_1 if the unobservable variable U was more spread out, and why/why not? Draw a picture to demonstrate this idea.

For the next few questions, consider this dataset and model:

```r
library(tidyverse)

sample_data <- tibble(
  x = 1:10,
  y = c(-8, 0, -8, -1, 4, 3, 1, 8, 8, 6)
  )

sample_data %>%
  lm(y ~ x, data = .)
```

```
##
## Call:
## lm(formula = y ~ x, data = .)
##
## Coefficients:
## (Intercept)            x
##      -7.600        1.618
```

Recall that the simple linear regression estimate $\hat{\beta}_1$ is equal to the covariance of x and y divided by the variance. Here I use dplyr verbs to get that value:

```r
sample_data %>%
  summarize(cov = cov(x, y), var = var(x)) %>%
  mutate(b1 = cov / var)
```

```
## # A tibble: 1 x 3
##     cov   var    b1
##   <dbl> <dbl> <dbl>
## 1  14.8  9.17  1.62
```

You can also use dplyr verbs to find the standard error for $\hat{\beta}_1$ like this:

```
sample_data %>%
  mutate(e = residuals(lm(y ~ x, data = sample_data))) %>%
  summarize(se = sqrt(var(e) / (8 * var(x))))
```

```
## # A tibble: 1 x 1
##      se
##   <dbl>
## 1 0.361
```

4) Your task is to explain where the numbers come from in the result below, especially

the statistic = 4.48, the p.value = .00205, the conf.low = .785, and the conf.high = 2.45. Some hints: the null hypothesis for regression parameters is always $\beta_1 = 0$, or that x does not actually effect y and the observed correlation between the two variables can be chalked up to sampling error. Use the alternative hypothesis $\beta_1 \neq 0$. Another hint: the degrees of freedom here is $n - 2 = 8$ because we lose a degree of freedom when we use residuals `e` as an estimate for `u` and we lose another degree of freedom when we use the sample variance of `e` as an estimate for the population variance of `e`.

```
sample_data %>%
  lm(y ~ x, data = .) %>%
  broom::tidy(conf.int = T)
```

```
## # A tibble: 2 x 7
##   term        estimate std.error statistic p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
## 1 (Intercept)    -7.6      2.24      -3.39 0.00948  -12.8       -2.43
## 2 x               1.62     0.361      4.48 0.00205    0.785      2.45
```

5) Should we reject the null hypothesis in the example above for $\hat{\beta}_1$ at the .05 significance level, or fail to reject it?