Section 5 Activity

Main

The goal of this question is to practice simulating data and running regressions in R. It is taken from your textbook.

1. Set the RNG seed to be 1
2. Create a vector `x` that has 30 observations randomly drawn from a standard normal distribution. (Hint: use the `rnorm()` function).
3. Create a second vector `eps` that has 30 observations randomly drawn from a normal distribution with a mean of 0 and a standard deviation of 0.25.
4. Using `x` and `eps` create a vector `y` according to the following data generating process

$$Y = -1 + 0.5X + \epsilon$$

5. What is the length of y? What are the values of β in the DGP?
6. Create a data frame called `dgp` with the variables created in 2-4.
7. Using `ggplot2` create a scatterplot of the relationship between `x` and `y`.
8. Run the regression of `y` on `x` and report the summary of the model. Call this `m1`. Comment on why you expect the result. (Hint: consider the discussion of the Conditional Expectation Function from lecture)
9. Using `ggplot2` add the least squares line to your previous plot. Give it a color other than black. Draw the population regression line on the plot in a different color.
10. Create a second model `m2` that adds a squared term x^2 to the model. Is there evidence that the term improves the model fit? Which model is "correct"?
11. For both models, manually predict the result of `y` when `x = 4`. Would you trust either prediction?

Bonus

1. Add a new variable `z` to the `dgp` data frame that has 30 observations randomly drawn from a Poisson distribution. Set `lambda=3`.
2. Update the `y` variable in the `dgp` data frame so that Y is now drawn from the following data generating process.

$$-1 + 0.5X + .25Z + .75(XZ) + \epsilon$$

3. Run a new model called `m3` that would perfectly estimate the CEF in expectation. Report the summary of this model.