Lab 3: Parameter inference in multiple linear regression

To better understand the sampling distribution of our the fitted β regression coefficients, you will perform several small simulation studies. The basic code for the simulation study can be found in the sampling-distribution-simulation.R file. Open this code file and read it carefully. In all, this code is running a small simulation study to generate a "sampling distribution" of the regression coefficients for a simple linear regression model $\mathbb{E}[y|x] = \beta_0 + \beta_1 \cdot x$ where both x and y are simulated data. Try to understand what each line of code is doing before you run the file, and then run it all the way through.

> **Exercise 1** In a few sentences, describe in your own words the procedure followed by the simulation study.

In a simulation study like this, we might be interested in several different characteristics of our fitted models. We may be interested in how often we reject the null hypothesis that β_1=0; in how wide the confidence intervals are on average; how accurate the $\hat{\beta}_1$ are on average; how much variability do we see in the $\hat{\beta}_1$ estimates; etc...

> **Exercise 2** Compute the 95% confidence interval coverage for β_1. That is, calculate the 95% confidence interval for each estimated β_1 and compute the percentage of times that the confidence interval covered the true value of β_1.

> **Exercise 3** Also, compute the mean squared error of your estimates of β_1. This can be computed as $MSE = \sum_{i=1}^{M}(\hat{\beta}_{1,i} - \beta_1)^2/M$, where M is the number of simulations that you ran.

> **Exercise 4** Given the parameters defined at the top of the file, determine what the sampling distribution for β_1 should be. That is, determine analytically what $\mathbb{E}[\hat{\beta}_1]$ and $Var[\hat{\beta}_1]$ are. Hint: use the formulae provided in the slides.

> **Exercise 5** Using the estimated values of the $\hat{\beta}_1$, calculate summary metrics and or use appropriate visualizations to determine whether your simulated distribution of $\hat{\beta}_1$ match up with with the theoretical distribution (i.e. with $\mathbb{E}[\hat{\beta}_1]$ and $Var[\hat{\beta}_1]$).

The top rows of the code file define a set of parameters that are used by the simulation study. Experiment with modifying them, and observe how the results change.

> **Exercise 6** For three new combinations of parameters, calculate the 95% CI coverage and the MSE. Report the results for these three new sets of parameters and the original set of parameters in a table, with one row for each parameter set, and a column for each metric, including the expected value and variance of $\hat{\beta}_1$. Hint: it may be easier to make vectors of the parameters, and then lop through each combination and for each set of parameters, run the full simulation M times.

Adapt the simulation to simulate data for two covariates, x_1 and x_2, using `mvrnorm()`. To start, assume that both x_1 and x_2 have mean 0 and variance 1. However, define x_1 and x_2 so that you may modify the degree of correlation between them, using a correlation parameter ρ.

> **Exercise 7** Run the original simulation again for two scenarios, one with low and one with high correlation. Add a row to your table above reporting the results from this simulation

For this lab, you should submit a modified version of the original code file, embedded into an RMarkdown document, with any figures used and a single table showing your results.