STA302H1F/STA1001HF: Mini Project 1
Due on 24th May, 2020 11:59 PM Sharp in Quercus

The purpose of these mini projects is to deepen your understanding of linear regression properties and develop your data analysis skills (which will be useful for the final project and future courses). Also, the emphasis will be on `R` coding, which you learned during lecture2-lecture5. The purpose of these mini projects is to develop your data analysis skills which will be useful for the final project and future courses. This will be a simulation based study. Recall that: for the mini projects you will be asked to do at least one of the following:

- To submit your results, you will be required to prepare a 5 minute presentation that you will need to record (using your computer, phone, etc.). You will be required to display your T-card alongside your face at the beginning of your video to verify your identity.
- You will be required to submit the R codes that you have created in a separate file. This helps us to check the reproducibility of your codes.
- You will need to display the results of your project in a logical way using slides (e.g. PowerPoint, latex, R Markdown or other) and record yourself discussing these results, with a focus on why you chose to do certain things and interpretation of your results for non-statisticians.
- Presentations should be submitted on time (i.e. by the deadline). Late submissions will receive a 20% penalty for each day that the project is late.
- In general, extensions will not be given unless a valid reason is provided. In such cases, the instructor may decide to grant an extension of up to 5 days.
- **There are no make-up mini projects**. A missed mini project will be given a grade of 0.

Project Description: In this course we have come across many properties of least squares estimates and variance estimates. For example the least squares estimates are unbiased and follow normal distribution. The mean residual sum of squares (MRS) is an unbiased estimator for the error variance σ^2. We have seen the mathematical proofs of these properties. Furthermore, we have made some assumptions about the linear regression model.

Task 1: In the first task you need to show and explain the following steps:

1. Simulate the paramters using the following codes,

```
## Simulation ##
set.seed(1002656486)
beta0 <- rnorm(1, mean = 0, sd = 1) ## The population beta0
beta1 <- runif(n = 1, min = 1, max = 3)  ## The population beta1
sig2 <- rchisq(n = 1, df = 25) ## The error variance sigma^2

## Multiple simulation may require loops ##
nsample <- 5  ## Sample size
n.sim <- 100 ## The number of simulations
sigX <- 0.2 ## The variances of X
```

```
## Simulate the predictor variable ##
X <- rnorm(nsample, mean = 0, sd = sqrt(sigX))
```

The provided seed is my student ID. Please change the seed to your student ID

2. Fix the sample size `nsample = 5` . Execute 100 simulations (i.e., `n.sim = 100`). For each simulation estimate the regression coefficients. Calculate the mean of the estimates from the different simulations. Comment on your observations.

3. Plot the histogram of each of the regression parameters. Explain the pattern of the distribution.

4. Obtain the variance of the regression parameters for each simulation. This is the variance obtained from the outputs of the `lm` function. Calculate their means. How do these means compare to the true variance of the regression parameters? Explain.

5. Construct the 95% t and z confidence intervals for β_0 and β_1 during every simulation. What is the proportion of the intervals for each method containing the true value of the parameters? Is this consistent with the definition of confidence interval? What differences do you observe in the t and z confidence intervals?

6. For steps 2-4 the sample size was fixed at 5. Start increasing the sample size (e.g., 10, 25, 50, 100) and run steps 2-4. Explain what happens to the mean, variance and distribution of the estimates as sample size increases.

7. Choose the largest sample size you have used in step 5. Fix the sample size to that and start changing the error variance (`sig2`). You can increase and decrease the value of the error variance. For each value of error variance execute steps 2-4. Explain what happens to the mean, variance and distribution of the estimates as the error variance changes.

Note: For steps 5, 6 and 6 you can present the results according to your convenience. For example you can add further plots and tables which you think are going to be useful. **The presentation video needs to be between 4-5 minutes**. You are going to lose important points if your presentations exceeds 5 minutes or is under 4 minutes.