STA302H1F: Assignment 1
Due on 3rd June, 2022 11:59 PM Sharp in Crowdmark

The purpose of the assignment 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 the lectures. This will be a simulation based study.

Assignment 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, under the Gauss Markov assumptions. The mean residual sum of squares (MRS) is an unbiased estimator for the error variance σ^2. We have seen (or will see) the mathematical proofs of these properties. Furthermore, we have made some assumptions about the linear regression model.

Task 1:

Assume the following simple regression model,

$$Y = \beta_0 + \beta_1 X + \epsilon$$
$$\epsilon \sim N(0, \sigma^2)$$

Now run the following code to generate values of $\sigma^2 =$ `sig2`, $\beta_1 =$ `beta1` and $\beta_0 =$ `beta0`. In the first task you need to show and explain the following steps:

(a) Simulate the parameters using the following codes,

```
## Simulation ##
set.seed("INSERT YOUR STUDENT ID")
beta0 <- rnorm(1, mean = 0, sd = 1) ## The true beta0
beta1 <- runif(n = 1, min = 1, max = 3)  ## The true beta1
sig2 <- rchisq(n = 1, df = 25) ## The true value of the error variance sigma^2

## Multiple simulation will 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))
```

Please change the seed to your student ID. The seed is used to generate random numbers. Since every student will set the seed to their own simulations every student will have unique datasets. **If you don't set your seed to your student ID then you will receive a 0 for this assignment**.

(b) Fix the sample size `nsample = 5` . Here, the values of X are fixed. You just need to generate ϵ and Y. Execute 100 simulations (i.e., `n.sim = 100`). For each simulation estimate the regression coefficients (β_0, β_1) and the error variance (σ^2). Calculate the mean of the estimates from the different simulations. Comment on your observations. What did you expect the mean to be?

(c) Plot the histogram of each of the regression parameter estimates from (b). Explain the pattern of the distributions.

(d) Obtain the variance of the regression parameter estimator (i.e., β_0 and β_1) from the simulations. That is calculate the sample variances of the regression parameter estimates from the 100 simulations. Is this variance approximately equal to the true variances of the regression parameter estimates? Explain.

(e) 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? Does it help if you increase the number of simulations from 100?

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

(g) Choose the largest sample size you have used in step (f). 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 (a) - (d). Explain what happens to the mean, variance and distribution of the estimates as the error variance changes.

Note: For steps (e), (f) and (g) 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.

Task 2

Assume the following multiple linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$
$$\epsilon \sim N(0, \sigma^2)$$

In the first task you need to simulate a dataset for multiple linear regression. The dataset will consist of one outcome variable (Y) and three predictor variables ($\mathbf{X} = (X_1, X_2, X_3)$). The $\mathbf{X}$ has to be simulated from a multivariate normal distribution. You can use the following simulation codes (Note: these are just initial codes)

```
library(MASS)

## Simulation for correlated predictors ##
set.seed("INSERT YOUR STUDENT ID")

nsample <- 10; nsim <- 100
sig2 <- rchisq(1, df = 1) ## The true error variance
bet <- c(rnorm(3, 0, 1), 0) ## 4 values of beta that is beta0, beta1, beta2, beta3 = 0
muvec <- rnorm(3, 0, 1)
sigmat <- diag(rchisq(3, df = 4))
X <- mvrnorm(nsample, mu = muvec, Sigma = sigmat)
Xmat <- cbind(1, X)

## Simulate the response ##
bets <- matrix(NA, ncol = length(bet), nrow = nsim)
for(i in 1:nsim){
Y <- Xmat%*%bet + rnorm(nsample, 0, sqrt(sig2))
model1 <- lm(Y ~ X)
bets[i,] <- coef(model1)
}
```

Please change the seed to your student ID

There are few things to be noticed from simulations. The $\boldsymbol{\beta}$ has four values, $\beta_0, \beta_1, \beta_2$ and β_3. You can see that $\beta_3 = 0$, i.e., the third predictor is not linearly related with the response. Here `sigmat` is the variance-covariance matrix for $\mathbf{X}$ the independent predictors, where the diagonal elements are variances (not standard deviations) and the off diagonals are covariances.

(a) First assume that the correlation between the three predictors are zero, i.e., the off diagonals of `sigmat` are zero, like the codes provided above. Set the number of simulations `nsim = 100` and sample size for each simulation to 10. Generate Y for each simulation. Then run simple linear regression for each of the three variables separately. Obtain the regression parameter estimates and their variances from the coefficients tables obtained from the `lm` function. Comment on whether the estimators are unbiased. That is calculate the mean of all regression parameter estimates and check it the values are approximately equal to the true values.

(b) Now fit a multiple linear regression and obtain the regression parameter estimates along with their variances from each simulation. Again check the unbiasedness and the variances. Compare the results with step (a). Remember in step (a) you fitted wrong models and in step (b) you are fitting the correct model.

(c) Now assume X_1 and X_2 are correlated. You can select a value for correlation (e.g., $r_{12} = 0.2$). Then add the following covariance terms in the `sigmat` matrix,

```
## The correlation ##
r12 <- 0.2
sigmat[1,2] <- sigmat[2,1] <- r12*sqrt(sigmat[1,1])*sqrt(sigmat[2,2])
## Simulation for Categorical Variables with Interaction ##
set.seed(1002656486)
X <- mvrnorm(nsample, mu = muvec, Sigma = sigmat); cor(X[,1], X[,2])
Xmat <- cbind(1, X)
```

Again run simple linear regressions on each of the predictors and also a multiple linear regression. Compare the results with step (a) and (b) and comment on the differences/similarities between the results. Start increasing the value of the correlation coefficient r_{12}, (e.g., 0.5, 0.7, 0.8 etc.) and again perform step (a) and (b). How do the estimated values and standard error of $\hat{\beta}_1$ and $\hat{\beta}_2$ change for simple and multiple linear regressions as the correlation changes?

(d) Now assume X_1 and X_2 are uncorrelated, i.e., $r_{12} = 0$ and `sigmat[1,2]` = `sigmat[2,1]` = 0. Instead X_1 and X_3 are correlated. Select a value for r_{13} arbitrarily (e.g., $r_{13} = 0.5$). Now change the values of `sigma[1,3]` and `sigmat[3,1]` using similar codes as the previous step. You can select a high value for correlation (e.g., $r_{13} > 0.5$). Recall, that the true $\beta_3 = 0$. Again perform step (a) and (b). Compare the results with the results obtained from step (c) and comment on the differences/similarities. Start increasing the value of the correlation coefficient r_{13}, (e.g., 0.6, 0.7, 0.8, 0.9, 0.95 etc.). How do the estimated values and standard error of $\hat{\beta}_1$ and $\hat{\beta}_2$ and $\hat{\beta}_3$ change for simple and multiple linear regression as the correlation changes?

Note: The answers to the tasks are open ended. You don't necessarily need to show every result. You just need to show the summary statistics or plots from the 100 simulations. How you present your results is up to you. These are very subjective choices. You will be marked based on your presentation of the results through your plots, tables and interpretations.