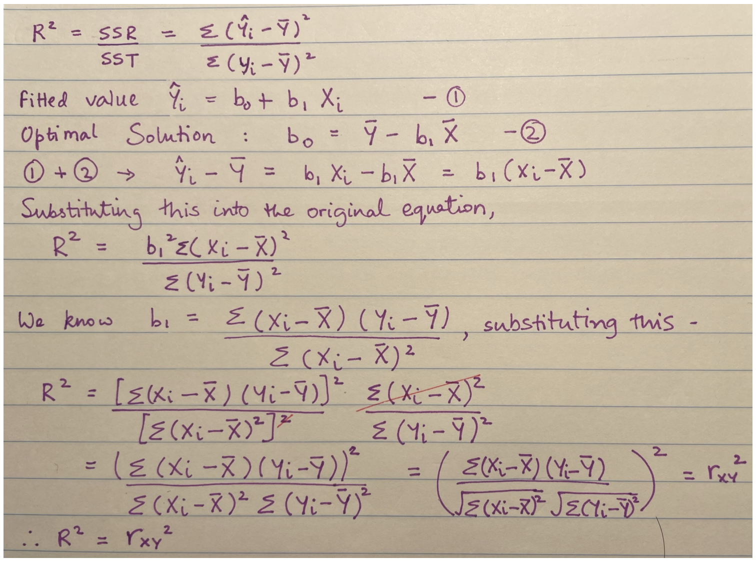
Homework 2
MSBA 400: Statistical Foundations for Data Analytics
UID 106082225, Sarvari Pidaparty

Question 1

Q1, part A

In the class notes, we introduced the concept of R^2. Show that the formula $R^2 = \frac{SSR}{SST}$ implies that R^2 is the square of the sample correlation coefficient between X and Y, r_{XY}. Hint: recall from the notes how the fitted regression line can be expressed in terms of deviations from the mean.



$$R^2 = \frac{SSR}{SST} = \frac{\Sigma(\hat{Y}_i - \bar{Y})^2}{\Sigma(y_i - \bar{Y})^2}$$

Fitted value $\hat{Y}_i = b_0 + b_1 X_i$ — ①

Optimal Solution : $b_0 = \bar{Y} - b_1 \bar{X}$ — ②

① + ② → $\hat{Y}_i - \bar{Y} = b_1 X_i - b_1 \bar{X} = b_1 (X_i - \bar{X})$

Substituting this into the original equation,

$$R^2 = \frac{b_1^2 \Sigma(X_i - \bar{X})^2}{\Sigma(Y_i - \bar{Y})^2}$$

We know $b_1 = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2}$, substituting this -

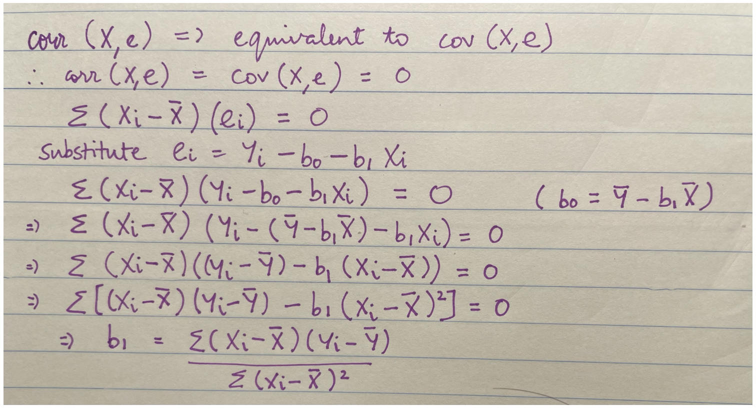
$$R^2 = \frac{[\Sigma(X_i - \bar{X})(Y_i - \bar{Y})]^2}{[\Sigma(X_i - \bar{X})^2]^{\cancel{2}}} \cdot \frac{\cancel{\Sigma(X_i - \bar{X})^2}}{\Sigma(Y_i - \bar{Y})^2}$$

$$= \frac{(\Sigma(X_i - \bar{X})(Y_i - \bar{Y}))^2}{\Sigma(X_i - \bar{X})^2 \, \Sigma(Y_i - \bar{Y})^2} = \left(\frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\Sigma(X_i - \bar{X})^2}\sqrt{\Sigma(Y_i - \bar{Y})^2}}\right)^2 = r_{XY}^2$$

$\therefore R^2 = r_{XY}^2$

Q1, part B

In the class notes, we used intuition to argue the regression property, $corr(X, e) = 0$. Show this directly results from the formula for b_1. Hint: substitute, $e_i = Y_i - b_0 - b_1 X_i$ into $corr(X, e)$.



corr (X, e) => equivalent to cov (X, e)

$\therefore$ corr (X, e) = cov $(X, e) = 0$

$$\Sigma(X_i - \bar{X})(e_i) = 0$$

Substitute $e_i = Y_i - b_0 - b_1 X_i$

$$\Sigma(X_i - \bar{X})(Y_i - b_0 - b_1 X_i) = 0 \qquad (b_0 = \bar{Y} - b_1 \bar{X})$$

$$\Rightarrow \Sigma(X_i - \bar{X})(Y_i - (\bar{Y} - b_1 \bar{X}) - b_1 X_i) = 0$$

$$\Rightarrow \Sigma(X_i - \bar{X})((Y_i - \bar{Y}) - b_1(X_i - \bar{X})) = 0$$

$$\Rightarrow \Sigma[(X_i - \bar{X})(Y_i - \bar{Y}) - b_1(X_i - \bar{X})^2] = 0$$

$$\Rightarrow b_1 = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2}$$

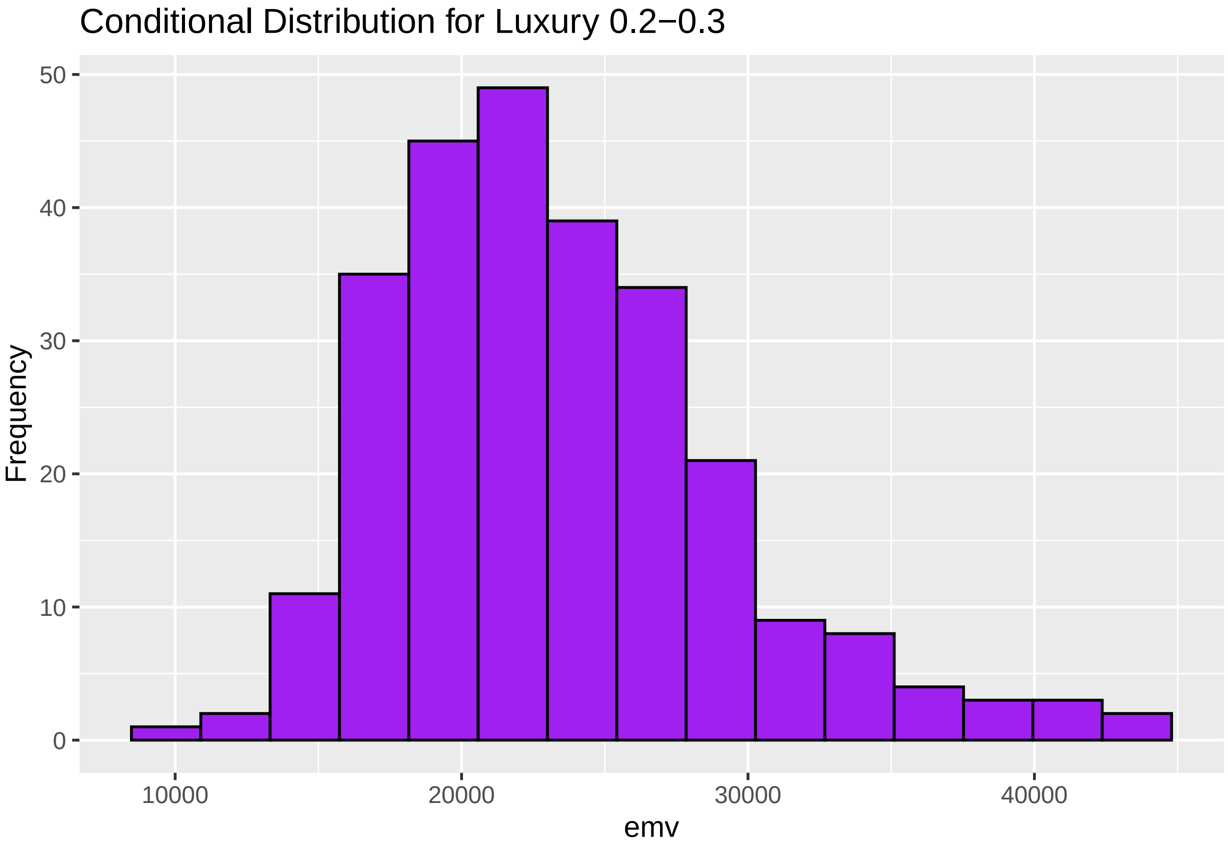
Question 2 : More on Nearest Neighbor Approaches

The simplest view of nearest neighbor methods is to "slice" into the data for only a small interval of X values. This distribution is called the conditional distribution of Y given X.

Q 2, part A

Display a histogram of the conditional distribution of emv given that luxury is in the interval (.2, .3) in the cars dataset (from problem set 1).

```
library(DataAnalytics)
library(ggplot2)
data(mvehicles)
cars=mvehicles[mvehicles$bodytype != "Truck",]
carsm1=cars[cars$luxury<=0.3 & cars$luxury>=0.2,]
qplot(emv,data=carsm1, geom="histogram",col=I("black"),fill=I("purple"), bins=15, main="Conditional Dis
```



Q 2, part B

Compute the mean of the conditional distribution in part A and compute a prediction interval that takes up 95% of the data (an interval that stretches from the .025 quantile (2.5 percentile) to the .975 (97.5 percentile)). Use the `quantile()` command.

```
mean(carsm1$emv)
```

```
## [1] 23376.5
```

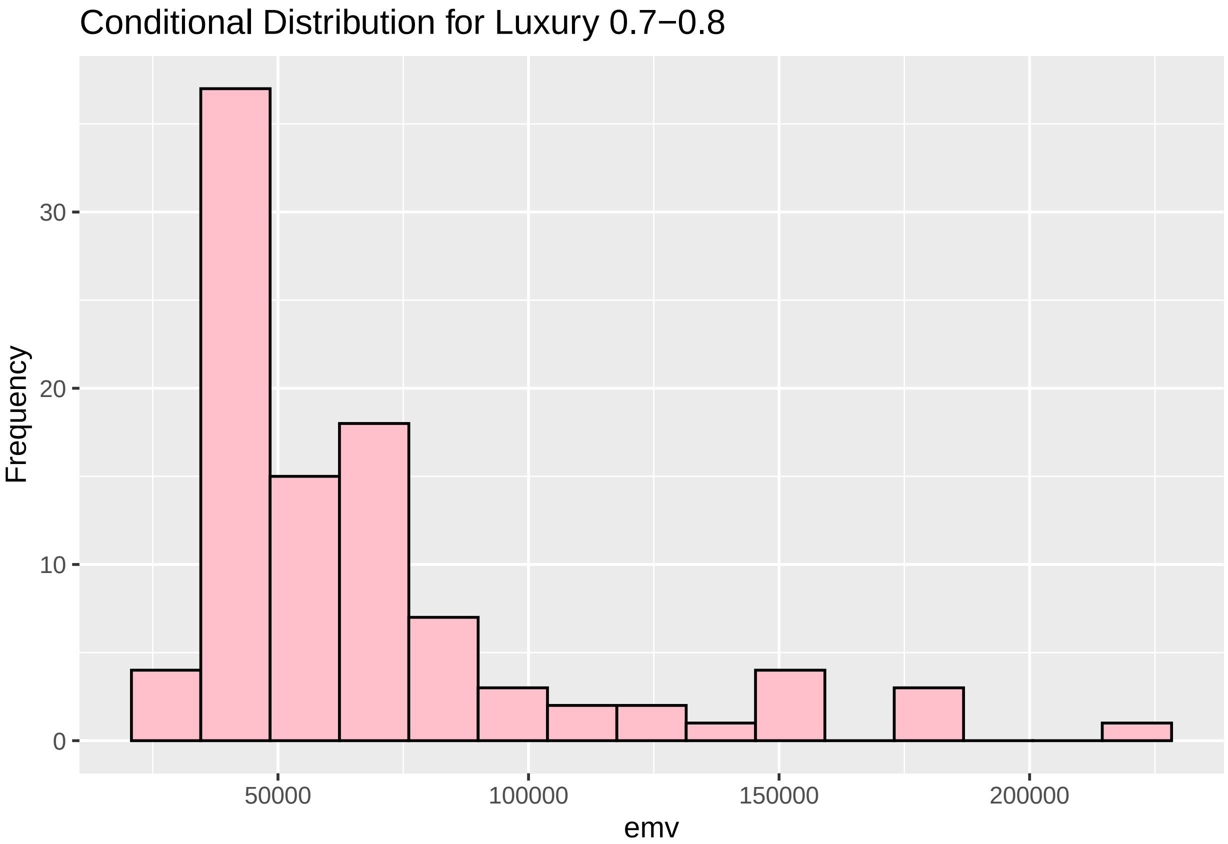
```
quantile(carsm1$emv,c(0.025,0.975))
```

```
##     2.5%    97.5%
## 14699.09 38632.34
```

Q 2, part C

Repeat part A and B for a much higher level of luxury, namely the interval (.7,.8). Describe the difference between these two conditional distributions.

```
carsm2=cars[cars$luxury<=0.8 & cars$luxury>=0.7,]
qplot(emv,data=carsm2, geom="histogram",col=I("black"), fill=I("pink"), bins=15, main="Conditional Dist
```



```
mean(carsm2$emv)
```

```
## [1] 68101.87
```

```
quantile(carsm2$emv,c(0.025,0.975))
```

```
##      2.5%     97.5%
##  34254.33 179179.22
```

- It is clear that the mean of emv is higher, where the level of luxury is higher.
- For the lower level of luxury, the distribution is more evenly spread out than the higher level of luxury. The data is highly concentrated around ~ 50000 for the 0.7-0.8 luxury case.
- The range of the 95% confidence interval for case 2 (higher level of luxury) is also much greater than that of case 1.

Q 2, part D

Explain why the results of part B and C show that luxury is probably (by itself) not sufficiently informative to give highly accurate predictions of `emv`.

Ans. The 95% confidence interval, especially for case 2, ranges from $34,254 to $179,179, which means we can estimate with 95% confidence that the price of a vehicle in the luxury level 0.7-0.8 would be in between

$34,254 and $179,179. This is a large range and standard error is high so we cannot accurately predict just with the feature 'luxury', how much 'emv' would be.

Question 3 : Optimal Pricing and Elasticities

Q 3, part A

Use the `detergent` dataset to determine the price elasticity of demand for 128 oz Tide. Compute the 90 percent confidence interval for this elasticity.

```
data(detergent)

lm_log=lm(log(q_tide128)~log(p_tide128),data=detergent)
print(elasticity_est<-coef(lm_log)[2])
```

```
## log(p_tide128)
##      -4.412049
```

```
confint(lm_log,level=.90)  # 90 percent C.I.
```

```
##                       5 %      95 %
## (Intercept)    13.072601 13.522963
## log(p_tide128) -4.518186 -4.305912
```

Q3, part B

One simple rule of pricing is the "inverse elasticity" rule that the optimal gross margin should be equal to the reciprocal of the absolute value of the price elasticity, i.e. Gross Margin $= \frac{1}{|elasticity|}$. For example, suppose we estimate that the price elasticity is -2 (a 1 per cent increase in price will reduce sales (in units) by 2 per cent. Then the optimal gross margin is 50 percent.

Suppose this retailer is earning a 25 per cent gross margin on 128 oz Tide. Perform appropriate hypothesis test to check if the retailer is pricing optimally at the 90 per cent confidence level?

Hints:
i. use the inverse elasticity rule to determine what elasticity is consistent with a 25 per cent gross margin.
ii. Use the confidence interval!

```
margin = 0.25
cat("Margin:", margin,'\n')
```

```
## Margin: 0.25
```

```
margin_est = 1/(elasticity_est)
cat("Estimated Margin", margin_est,'\n')
```

```
## Estimated Margin -0.2266521
```

```
elasticity = -1/margin
cat("Elasticity", elasticity,'\n')
```

```
## Elasticity -4
```

The ideal elasticity we need is -4, which does not fall in our estimated 90% confidence interval (-4.528 to -4.305). This implies the retailer is not pricing optimally.

Questions 4-5 explore the sampling properties of least squares

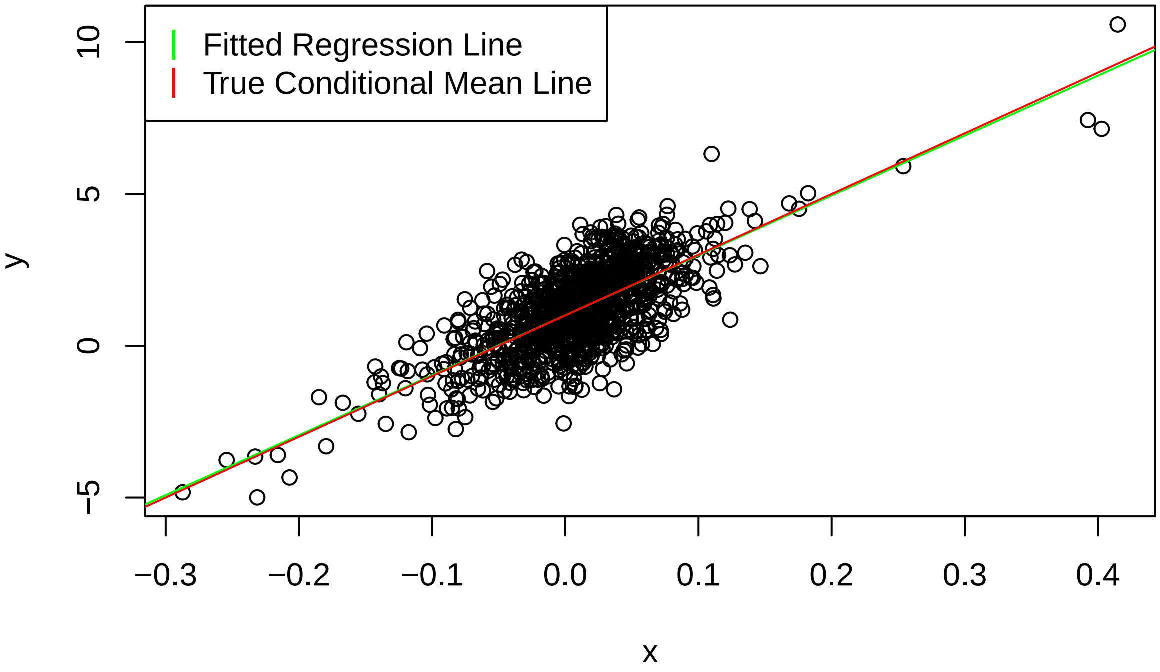
Question 4

a. Write your own function in `R` (using `function()`) to simulate from a simple regression model. This function should accept as inputs: b_0 (intercept), b_1 (slope), X (a vector of values), and σ (error standard deviation). You will need to use `rnorm()` to simulate errors from the normal distribution. The function should return a vector of Y values.

```
library(DataAnalytics)
library(ggplot2)
library(reshape2)

sim_reg=function(beta0,beta1,sigma,x){
  y=beta0+beta1*x+rnorm(length(x),sd=sigma)
}
```

b. Simulate Y values from your function and make a scatterplot of X versus simulated Y. When simulating, use the `vwretd` data from the `marketRf` dataset as the X vector, and choose $b_0 = 1$, $b_1 = 20$, and $\sigma = 1$. Then add the fitted regression line to the plot as well as the true conditional mean line (the function `abline()` may be helpful).

```
data(marketRf)

x=marketRf$vwretd
beta0=1
beta1=20
sigma=1
y=sim_reg(beta0,beta1,sigma,x)

plot(x, y)
abline(lm(y~x, data = marketRf$coef), col="green")
abline(a=beta0,b=beta1, col="red")
legend("topleft", legend=c("Fitted Regression Line","True Conditional Mean Line"), pch="|",col=c("green
```

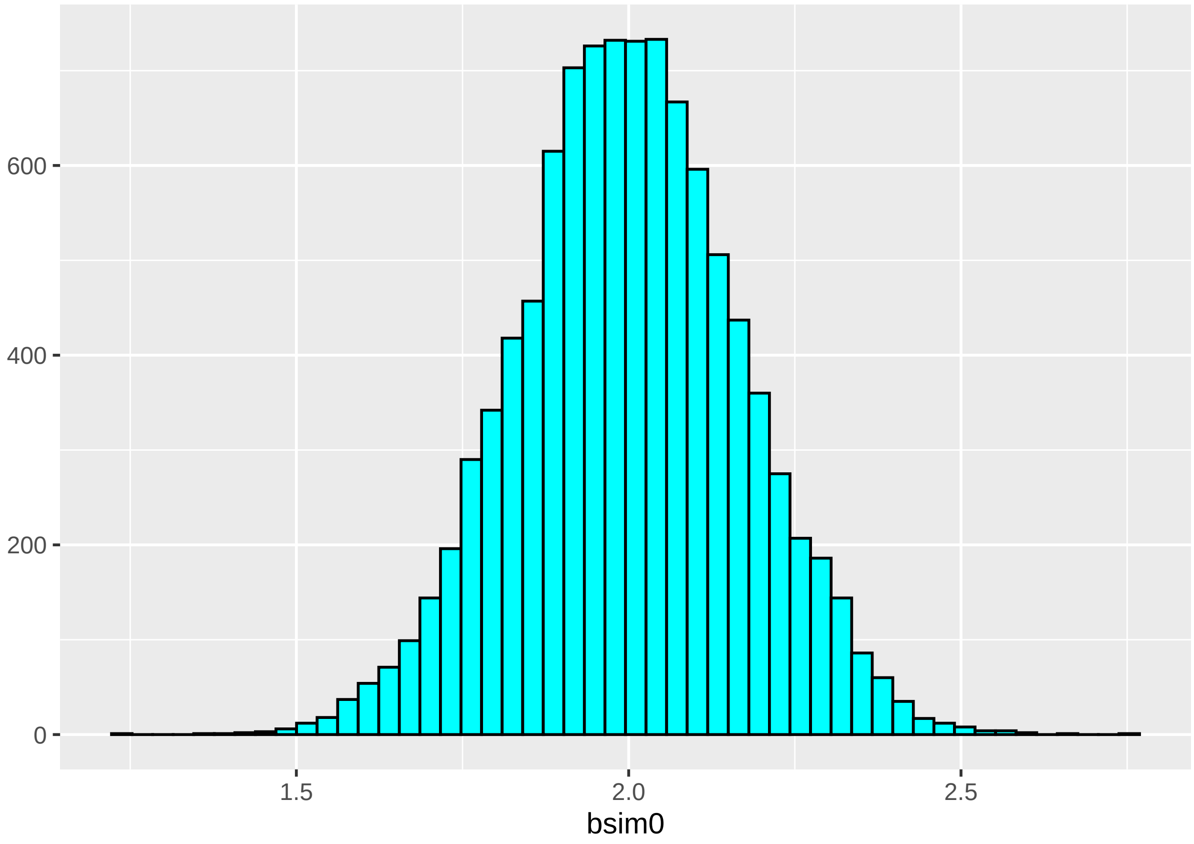


Question 5

Assume $Y = \beta_0 + \beta_1 X + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Let $\beta_0 = 2$, $\beta_1 = 0.6$, and $\sigma^2 = 2$. You can make X whatever you like, for example you could simulate X from a uniform distribution.

a. Use your `R` function from question 4 to simulate the sampling distribution of the slope. Use a sample size of $N = 300$ and calculate b_0 & b_1 for 10,000 samples. Plot a histogram of the sampling distribution of b_0.

```r
n=300
x1=runif(n)
beta0=2
beta1=0.6
sigma=sqrt(2)
nsample=10000
bsim0=double(nsample)
bsim1=double(nsample)
for(i in 1:nsample){
  y=sim_reg(beta0,beta1,sigma,x1)
  bsim0[i]=lm(y~x1)$coef[1]
  bsim1[i]=lm(y~x1)$coef[2]
}
qplot(bsim0, col=I("black"), fill=I("cyan"), bins=50)
```



b. Calculate the empirical value for $\mathbb{E}[b_1]$ from your simulation and provide the theoretical value for $\mathbb{E}[b_1]$. Compare the simulated and theoretical values.

```r
cat("Empirical value of b1:", mean(bsim1),'\n')
```

```
## Empirical value of b1: 0.5999602
```

```
cat("Theoretical value of b1:", beta1)
```

```
## Theoretical value of b1: 0.6
```

The simulated and theoretical values are close, the simulated value is slightly higher than the theoretical value.

c. Calculate the empirical value for $\text{Var}(b_1)$ from your simulation and provide the theoretical value for $\text{Var}(b_1)$. Compare the simulated and theoretical values.

```
cat("Empirical variance of b1:", var(bsim1),'\n')
```

```
## Empirical variance of b1: 0.0873859
```

```
cat("Theoretical variance of b1:", sigma**2/((n-1)*var(x1)))
```

```
## Theoretical variance of b1: 0.08581949
```

The variance values as well are really close, so we can say our simulation is pretty accurate.

Question 6

Standard errors and p-values.

a. What is a standard error (of a sample statistic or an estimator)? How is a standard error different from a standard deviation?

Ans. Standard error is the estimate of the standard deviation of the sample population's statistic (or estimator) that is selected from an entire population. Standard error indicates how accurate a sample statistic that we calculate is with respect to the population statistic. For example, standard error of mean would tell you how accurate the sample mean is in comparison to the population mean. Standard deviation is a measure of the dispersion of a set of values from the mean. It measures the variability of among values in a sample population, while standard error looks at how close the sample mean is to the real population mean.

$SD = \sqrt(variance)$
$StandardError = SD/\sqrt(N)$

b. What is sampling error? How does the standard error capture sampling error?

Ans. Sampling errors are errors in statistics that arise from the fact that the chosen sample is not representative of the actual population. We estimate population mean, variance and so on from the sample mean, variance and so on, but there might be a difference in these values. This is the sampling error. Standard error measures the accuracy of a sample statistic in comparison to population statistics, so it basically captures the sampling errors.

c. Your friend Steven is working as a data analyst and comes to you with some output from some statistical method that you've never heard of. Steven tells you that the output has both parameter estimates and standard errors. He then asks, "how do I interpret and use the standard errors?' ' What do you say to Steven to help him even though you don't know what model is involved?

Ans. The standard errors can be used to estimate how far the actual population values can be from the sample statistic values. Using the parameter estimates and standard errors, the confidence values or range of values within which the parameters fall within can be estimated. Steven can use the standard error to say whether the sample population is representative of the real population.

d. Your friend Xingua works with Steven. She also needs help with her statistical model. Her output reports a test statistic and the p-value. Xingua has a Null Hypothesis and a significance level in mind, but she asks “how do I interpret and use this output?’ ’ What do you say to Xingua to help her even though you don’t know what model is involved?

Ans. The p-value can be used to say if a hypothesis can be considered to be true, or it should be rejected. A small p-value, lesser than the chosen significance level would prompt you to reject the null hypothesis, and a large p-value, greater than or equal to the significance level would prompt you to accept the null hypothesis.