Assignment-2 S for R

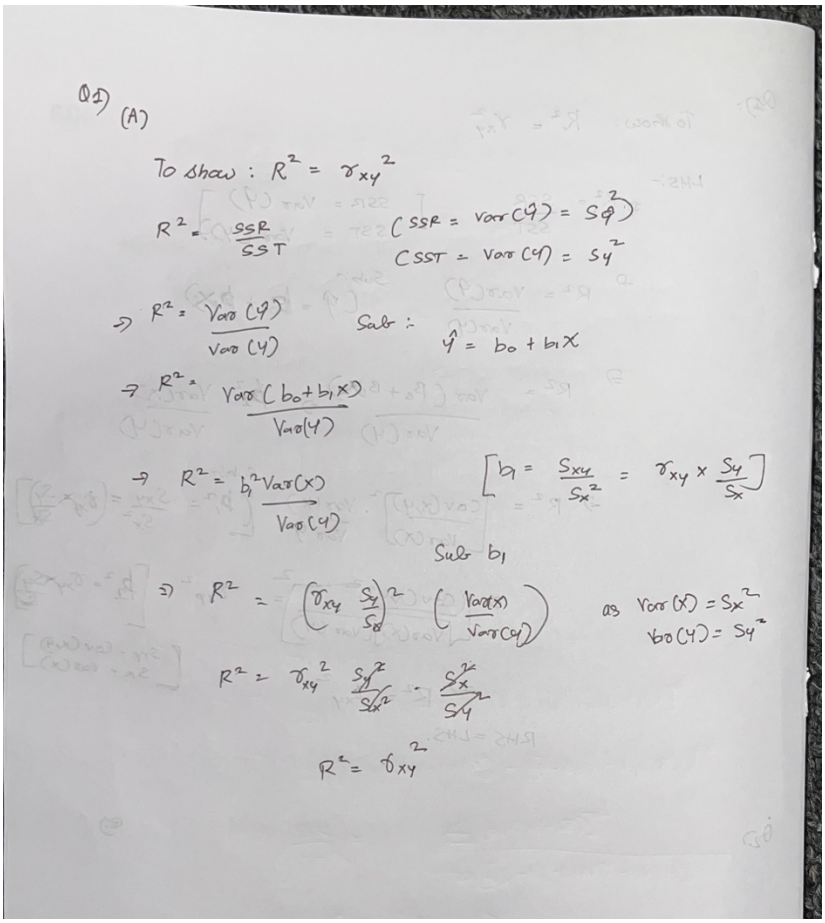
Vishanth

2023-10-28

Question 1

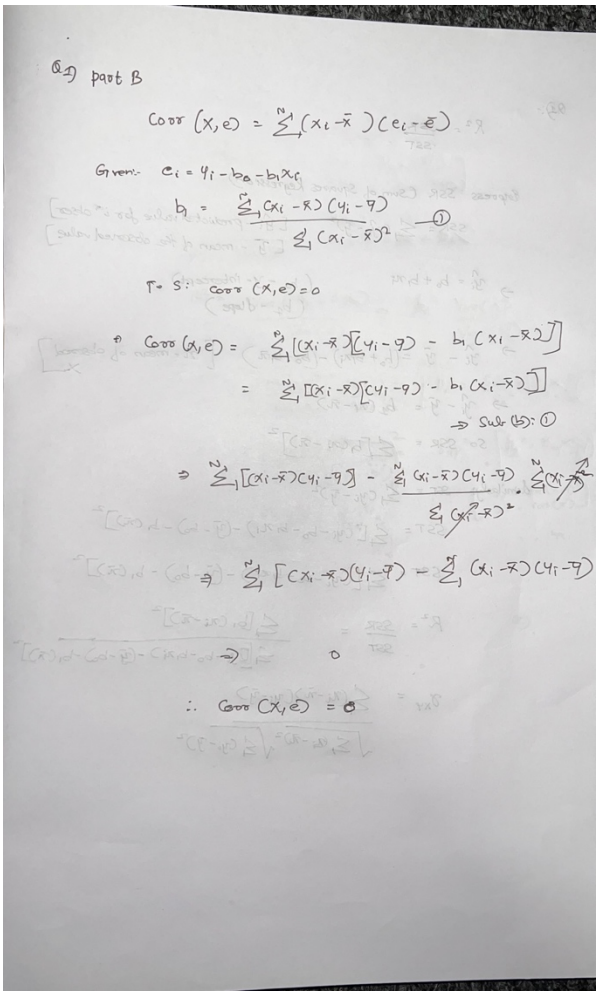
Q1, part A

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.



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 $corr(X, e)$.



Question 2 : More on Nearest Neighbor Approaches

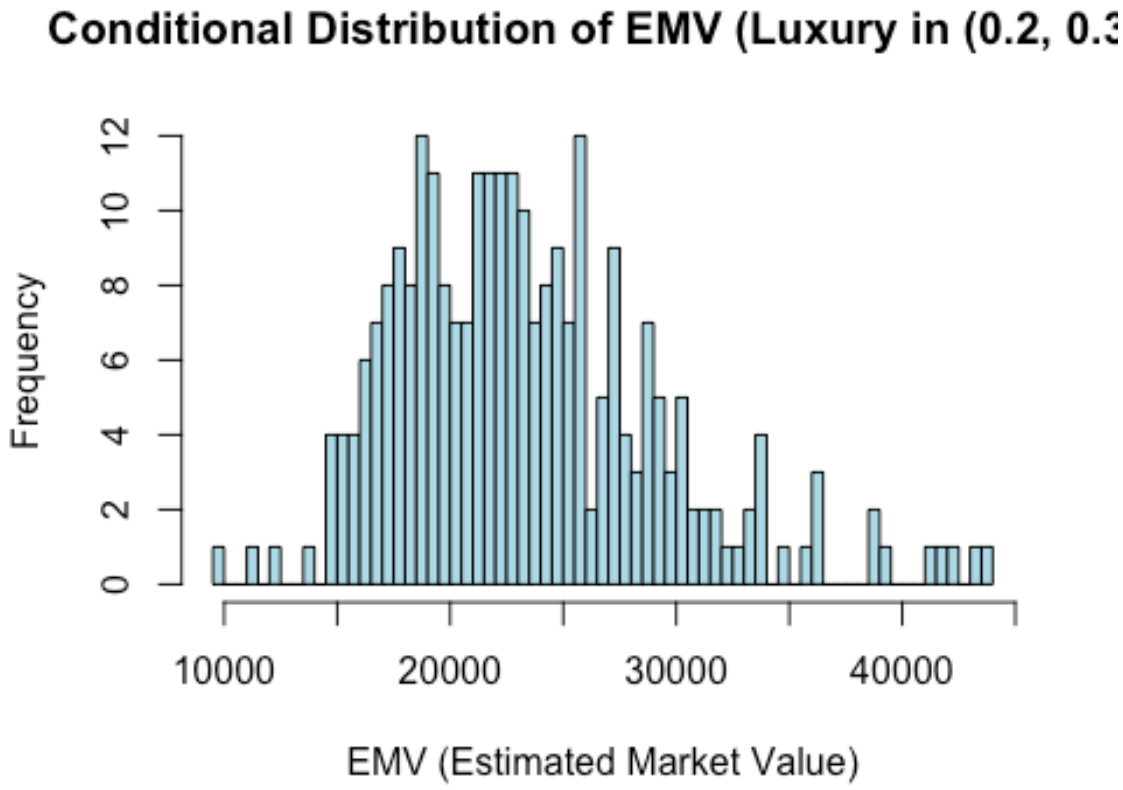
The simplest view of nearest neighbor methods is to “slice” into the data for only a small interval of X values. This distribution is called the conditional distribution of Y given X.

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)
data(mvehicles)

cars = mvehicles[mvehicles$bodytype != "Truck",]

# Filter the dataset for rows where Luxury is in the interval (0.2, 0.3)
filtered_data = cars[cars$luxury > 0.2 & cars$luxury < 0.3,]

# Create a histogram of the EMV for the filtered data
hist(filtered_data$emv,
```

```
     main = "Conditional Distribution of EMV (Luxury in (0.2, 0.3))",
     xlab = "EMV (Estimated Market Value)",
     ylab = "Frequency",
     col = "lightblue", # Adjust the color if needed
     breaks = 50)
```



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.

```
# Calculate the mean of the conditional distribution
mean_conditional = mean(filtered_data$emv)

# Calculate the quantiles for the prediction interval
lower_quantile = quantile(filtered_data$emv, 0.025)
upper_quantile = quantile(filtered_data$emv, 0.975)

# Print the mean and the prediction interval
cat("Mean of the conditional distribution:", mean_conditional, "\n")

## Mean of the conditional distribution: 23376.5
```

```r
cat("95% Prediction Interval (2.5% quantile to 97.5% quantile): [",
lower_quantile, " , ", upper_quantile, "]\n")
```

```
## 95% Prediction Interval (2.5% quantile to 97.5% quantile): [ 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.

```r
# Filter the dataset for rows where Luxury is in the interval (0.7, 0.8)
filtered_data_high_luxury = cars[cars$luxury > 0.7 & cars$luxury < 0.8, ]

# Calculate the mean of the conditional distribution for high luxury
mean_conditional_high_luxury = mean(filtered_data_high_luxury$emv)

# Calculate the quantiles for the prediction interval for high luxury
lower_quantile_high_luxury = quantile(filtered_data_high_luxury$emv, 0.025)
upper_quantile_high_luxury = quantile(filtered_data_high_luxury$emv, 0.975)

# Print the results for high luxury
cat("For Luxury in (0.7, 0.8):\n")
```

```
## For Luxury in (0.7, 0.8):
```

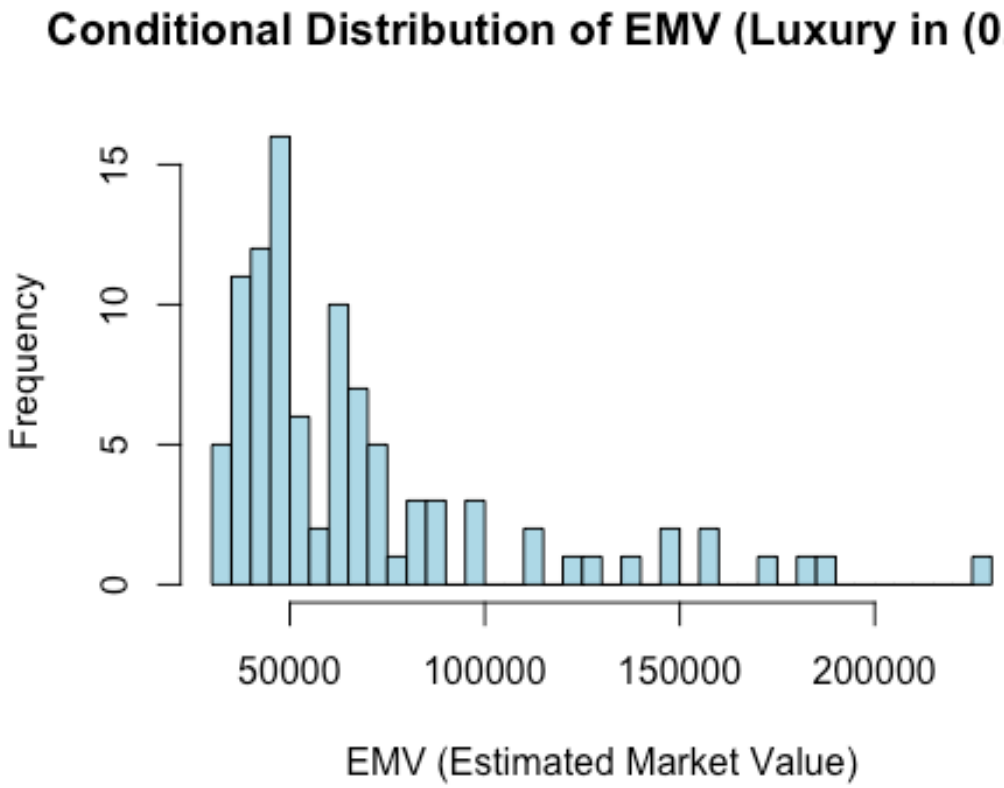
```r
cat("Mean of the conditional distribution:", mean_conditional_high_luxury,
"\n")
```

```
## Mean of the conditional distribution: 68101.87
```

```r
cat("95% Prediction Interval (2.5% quantile to 97.5% quantile): [",
lower_quantile_high_luxury, " , ", upper_quantile_high_luxury, "]\n")
```

```
## 95% Prediction Interval (2.5% quantile to 97.5% quantile): [ 34254.33  ,
179179.2 ]
```

```r
hist(filtered_data_high_luxury$emv,
     main = "Conditional Distribution of EMV (Luxury in (0.7, 0.8))",
     xlab = "EMV (Estimated Market Value)",
     ylab = "Frequency",
     col = "lightblue", # Adjust the color if needed
     breaks = 50)
```



The difference in conditional distribution, there is discretion in plotting, and it is shifted more toward right side and as the luxury index increases the range, the emv is in decreasing trend. 95% prediction interval is wide as compared to previous graph.

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`.

ANSWER:

1. ***Variability within Luxury Intervals:*** *Both in Part B and Part C, you calculated prediction intervals for the conditional distributions of EMV based on different luxury intervals (0.2-0.3 and 0.7-0.8). These intervals were relatively wide, indicating a significant spread of EMV values within each luxury range. This wide variability suggests that luxury alone does not provide precise information about the exact EMV of a car.*

2. ***Overlap in Luxury Ranges:*** *It's important to note that there can be significant overlap in luxury values between different car models or brands. Two cars with similar luxury values may have very different EMVs due to other factors such as make, model, year, technology, and so on. Therefore, luxury, by itself, cannot capture all the nuanced factors that affect EMV.*

3. ***Multifactorial Nature of EMV:*** *EMV is influenced by a multitude of factors, including make, model, year, features, performance, technology, and market demand. While luxury is one aspect, it is just one of many factors that collectively determine EMV. Ignoring these other factors can lead to incomplete predictions.*

4. ***Sample Size and Distribution:*** *The accuracy of predictions is also influenced by the size and distribution of the dataset. The more data points available within specific luxury intervals, the more accurate the predictions are likely to 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.

```
#
# let's run the price elasticity and un-tranformed regressions
#
data("detergent")
lm_log=lm(log(q_tide128)~log(p_tide128),data=detergent)
confint(lm_log,level=.90)  # use 90 percent
```

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

```
# Extract the coefficient for p_tide128
price_coefficient = coef(lm_log)["log(p_tide128)"]

# Calculate the anti-log of the coefficient to get the price elasticity
price_elasticity = exp(price_coefficient)

#cat("price elasticity:-", price_elasticity, "/n")
cat("Price elasticity with log :- ", price_coefficient)
```

```
## Price elasticity with log :-  -4.412049
```

Q 3, 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!

Questions 4-5 explore the sampling properties of least squares

```
# Elasticity consistent with a 25% gross margin
desired_elasticity = -(1 / 0.25)

# Confidence interval of the estimated elasticity
confidence_interval = confint(lm_log, level = 0.90)

# Check if the desired elasticity falls within the confidence interval
is_optimal_pricing = (desired_elasticity >= confidence_interval[1]) &&
(desired_elasticity <= confidence_interval[2])

# Print the results
cat("Desired Elasticity for 25% Gross Margin:", desired_elasticity, "\n")

## Desired Elasticity for 25% Gross Margin: -4

cat("90% Confidence Interval for Elasticity:", confidence_interval, "\n")

## 90% Confidence Interval for Elasticity: 13.0726 -4.518186 13.52296 -
4.305912

if (is_optimal_pricing) {
  cat("The retailer is pricing optimally at the 90% confidence level.\n")
} else {
  cat("The retailer may not be pricing optimally at the 90% confidence
level.\n")
}

## The retailer may not be pricing optimally at the 90% confidence level.

cat("90% Confidence Interval for Elasticity:", confidence_interval, "\n")

## 90% Confidence Interval for Elasticity: 13.0726 -4.518186 13.52296 -
4.305912

if (is_optimal_pricing) {
  cat("The retailer is pricing optimally at the 90% confidence level.\n")
} else {
  cat("The retailer may not be pricing optimally at the 90% confidence
level.\n")
}

## The retailer may not be pricing optimally at the 90% confidence level.
```

As you can identify from code, the output below, 90% confident interval is [-4.518, -4.30], it doesn't match with desired elasticity -4, therefore hypothesis Ho is rejected, below is the output of the code

```
Desired Elasticity for 25% Gross Margin: -4
90% Confidence Interval for Elasticity: 13.0726 -4.518186 13.52296 -4.305912
The retailer may not be pricing optimally at the 90% confidence level.
```

Question 4

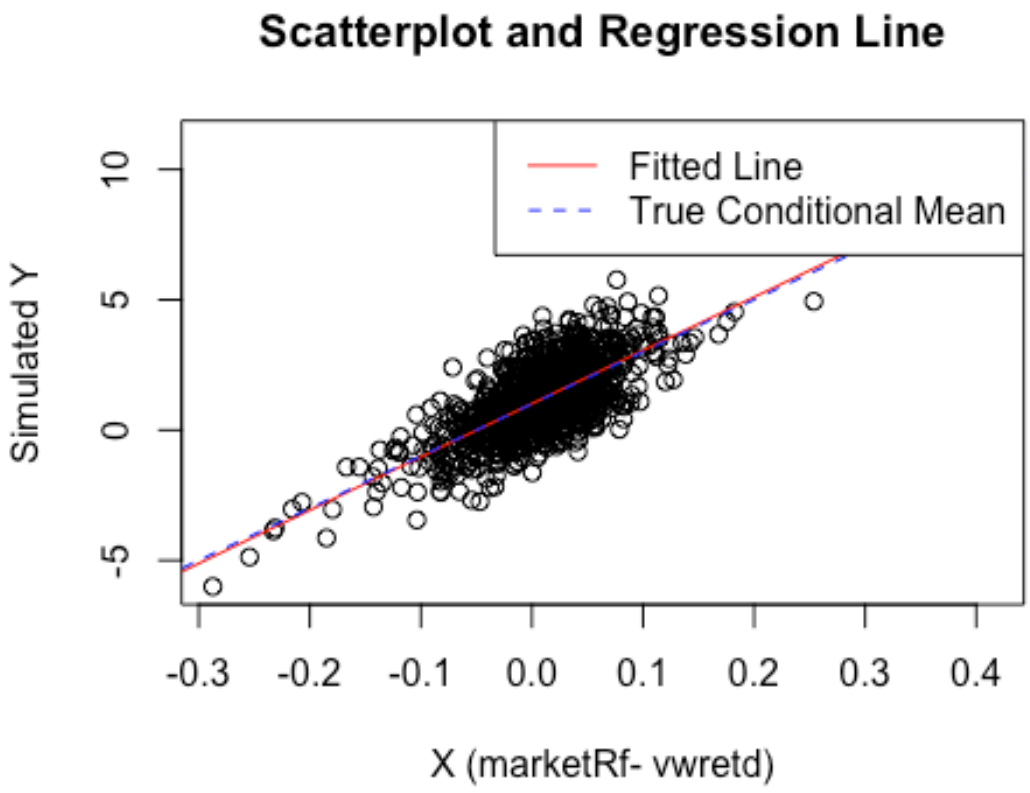
a. Write your own function in R (using `function()`) to simulate from a simple regression model. This function should accept as inputs: b_0 (intercept), b_1 (slope), X (a vector of values), and σ (error standard deviation). You will need to use `rnorm()` to simulate errors from the normal distribution.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

```r
simulate_simple_regression = function(b0, b1, X, sigma) {
  # Generate random errors from a normal distribution
  errors = rnorm(length(X), mean = 0, sd = sigma)

  # Calculate the corresponding Y values
  Y = b0 + b1 * X + errors

  return(Y)
}


set.seed(123)  # For reproducibility
b0 = 2.5
b1 = 1.8
X = seq(1, 100, by = 1)
sigma = 0.5
simulated_Y = simulate_simple_regression(b0, b1, X, sigma)
cat("Simulated Value Y", simulated_Y)
```

```
## Simulated Value Y 4.019762 5.984911 8.679354 9.735254 11.56464
14.15753 15.33046 16.26747 18.35657 20.27717 22.91204 24.27991 26.10039
27.75534 29.22208 32.19346 33.34893 33.91669 37.05068 38.2636 39.76609
41.99101 43.387 45.33555 47.18748 48.45665 51.51889 52.97669 54.13093
57.12691 58.51323 59.95246 62.34756 64.13907 65.91079 67.64432 69.37696
70.86904 72.54702 74.30976 75.95265 77.99604 79.2673 82.78448 84.10398
84.73845 86.89856 88.66667 91.08998 92.45832 94.42666 96.08573 97.87856
100.3843 101.3871 104.0582 104.3256 107.1923 108.7619 110.608 112.4898
113.8488 115.7334 117.1907 118.9641 121.4518 123.3241 124.9265 127.1611
129.525 130.0545 130.9454 134.4029 135.3454 137.156 139.8128 140.9576
142.2896 144.7907 146.4306 148.3029 150.2926 151.7147 154.0222 155.3898
157.4659 159.6484 161.1176 162.537 165.0744 166.7968 168.3742 170.0194
171.386 174.1803 174.9999 178.1937 179.6663 180.5821 181.9868
```

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")
# Simulate Y values using the custom function
b0 = 1
b1 = 20
sigma = 1
X = marketRf$vwretd
simulated_Y = simulate_simple_regression(b0, b1, X, sigma)

# Create a scatterplot of X versus simulated Y
plot(X, simulated_Y, xlab = "X (marketRf- vwretd)", ylab = "Simulated
Y", main = "Scatterplot and Regression Line")

# Add the fitted regression line
fit = lm(simulated_Y ~ X)
abline(fit, col = "red")

# Add the true conditional mean line
true_mean_line = b0 + b1 * X
abline(a = b0, b = b1, col = "blue", lty = 2)

# Add a legend
legend("topright", legend = c("Fitted Line", "True Conditional Mean"),
col = c("red", "blue"), lty = c(1, 2))
```



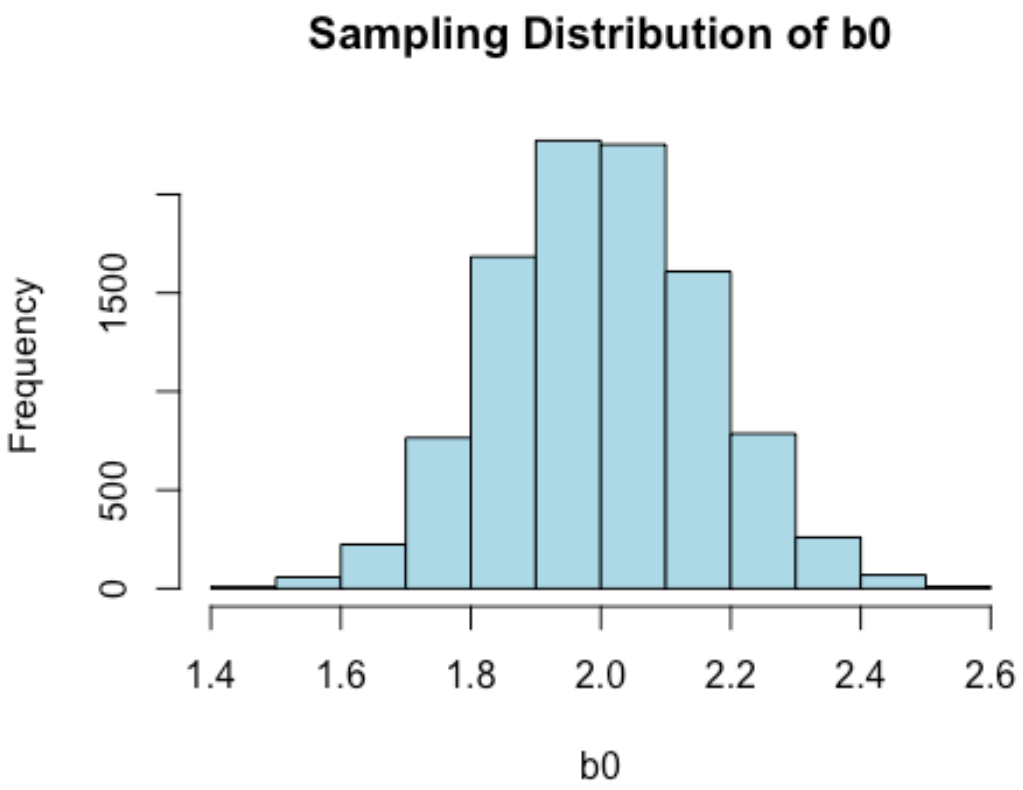
Question 5

Assume $Y = beta_0 + beta_1 X + \varepsilon$ with $\varepsilon \sim N(0, \sigma^2)$. 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.

```
# Define the parameters


set.seed(123)  # For reproducibility

beta_0 = 2
beta_1 = 0.6
sigma_squared = 2

# Specify the sample size and number of samples
sample_size = 300
num_samples = 10000
```

```r
# Initialize vectors to store b0 values
b0_values = numeric(num_samples)
b1_values = numeric(num_samples)

# Simulate the sampling distribution
for (i in 1:num_samples) {
  # Simulate X from a uniform distribution
  X = runif(sample_size, min = 0, max = 10)

  # Simulate Y using the custom function
  Y = simulate_simple_regression(beta_0, beta_1, X,
sqrt(sigma_squared))

  # Fit a linear regression model to the data
  model = lm(Y ~ X)

  # Extract the coefficients
  b0_values[i] = coef(model)[1]  # Intercept (b0)
  b1_values[i] = coef(model)[2]  #slope
}

#As the output values are to many to print,
#cat("b0 values", b0_values)
#cat("b1 values", b1_values)

# Plot a histogram of the sampling distribution of b0
hist(b0_values, main = "Sampling Distribution of b0", xlab = "b0", col
= "lightblue")
```



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

ANSWER:

Theoretical value $E[b1]$: In the linear regression model, the theoretical value for $E[b1]$ is equal to the true parameter $b1$. In this case, $b1 = 0.6$

Simulated value for $E[b1]$: You have already simulated the sampling distribution in code. Now, calculate the mean

```
# Calculate the empirical value for E[b1] from the simulation
empirical_mean_b1 = mean(b1_values)

# Theoretical value for E[b1]
theoretical_mean_b1 = beta_1

# Compare the simulated and theoretical values
cat("Simulated E[b1]:", empirical_mean_b1, "\n")

## Simulated E[b1]: 0.5996354

cat("Theoretical E[b1]:", theoretical_mean_b1, "\n")
```

```
## Theoretical E[b1]: 0.6
```

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

ANSWER:

In a simple linear regression model $Y = beta_0 + beta_1 X + \varepsilon$, *the theoretical value for* $\mathrm{Var}(b_1)$ *can be calculated as:* $\mathrm{Var}(b_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$ *where* σ^2 *is the error variance,* n *is the sample size, and* $\bar{X}$ *is the mean of the* X *values. In this case,* $\sigma^2 = 2$, *and* n *is the sample size.*

Simulated can be calculated from the code

```
# Calculate the empirical value for Var(b1) from the simulation
empirical_var_b1 = var(b1_values)

# Theoretical value for Var(b1)
n = sample_size
mean_X = mean(X)
theoretical_var_b1 = sigma_squared / sum((X - mean_X)^2)

# Compare the simulated and theoretical values
cat("Simulated Var(b1):", empirical_var_b1, "\n")

## Simulated Var(b1): 0.0007961139

cat("Theoretical Var(b1):", theoretical_var_b1, "\n")

## Theoretical Var(b1): 0.0008839872
```

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?

ANSWER:

A standard error is a measure of the variation or uncertainty in an estimate or sample statistic. It provides a way to quantify how much the estimate or sample statistic is likely to vary from one sample to another, assuming random sampling from the same population. Standard errors are commonly used in statistics to assess the precision and reliability of estimates.

The key difference is that the standard error relates to the precision of an estimate or statistic when taking multiple samples, while the standard deviation describes the variation within a single dataset. Standard errors are especially important when

making inferences about population parameters based on sample statistics, as they provide a measure of how confident we can be in the estimate's accuracy.

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

ANSWER:

Sampling error *refers to the variation or discrepancy between a sample statistic (such as the sample mean, sample proportion, or sample regression coefficient) and the true population parameter it is intended to estimate. It is the result of random sampling, where different samples from the same population may yield different values for the sample statistic. Sampling error is a fundamental concept in inferential statistics because it quantifies the uncertainty or variability associated with sample-based estimates.*

The ***standard error*** *(SE) captures sampling error by providing a quantitative measure of the expected variation in the sample statistic. The standard error is calculated based on the sample data and reflects the degree of uncertainty in the estimate due to random sampling. It is a key component in constructing confidence intervals and performing hypothesis tests.*

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?

ANSWER: *We can say smaller standard errors indicate greater precision in the parameter estimates. A smaller standard error implies that the sample estimate is likely to be closer to the true population value.*

standard errors represent a measure of the uncertainty or variability associated with the estimated parameters. They indicate how much we expect the parameter estimates to vary if we were to obtain different samples from the same population.

standard errors are often used to construct confidence intervals. A confidence interval is a range of values around the point estimate (parameter estimate) within which we are reasonably confident the true population parameter lies. Common confidence levels are 95% or 99%, for example.

Standard errors play a crucial role in hypothesis testing. They help determine whether the parameter estimate is statistically different from a null hypothesis value. If the point estimate is far from the null hypothesis value in terms of standard errors, it suggests statistical significance.

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?

ANSWER:

I would suggest the following points and tell the below key inference from the output

- *The test statistic measures how far your sample statistic (e.g., the difference between means, regression coefficient, etc.) is from what is expected under the null hypothesis.*
- *A larger absolute value of the test statistic suggests a greater departure from the null hypothesis.*
- *The p-value is a measure of evidence against the null hypothesis.*
- *It quantifies the probability of observing a test statistic as extreme as, or more extreme than, the one calculated from your data, assuming the null hypothesis is true.*
- *A smaller p-value indicates stronger evidence against the null hypothesis.*
- *To make an inference, you compare the observed test statistic to what is expected under the null hypothesis.*
- *The significance level (often denoted as alpha, is the threshold below which you consider the results statistically significant.*
- *Common significance levels are 0.05 (5%) and 0.01 (1%). If the p-value is less than the chosen significance level, you reject the null hypothesis.*
- *If the p-value is less than or equal to the chosen significance level, you have evidence to reject the null hypothesis. This suggests that your data provides support for the alternative hypothesis, indicating a significant effect, difference, or relationship.*
- *If the p-value is greater than the chosen significance level, you fail to reject the null hypothesis. This means that your data does not provide strong evidence against the null hypothesis.*
- *The decision to reject or fail to reject the null hypothesis is based on comparing the p-value to the significance level.*