Homework\_3 (Group 3)

Sohail Shaikh, Aishwarya Jawalkar, Samarth Sathe, Tanmay Khairnar, Ashutosh Lonkar

2023-11-07

# Question 1

Alumni Donation Data (Multiple Linear Regression). Continue with the same data from homework 1 and fit a multiple linear regression model to the data, where the alumni giving rate is the response variable (Y), and the percentage of classes with fewer than 20 students (X1) and Student/Faculty Ratio (X2) as the predictors.

df = read.csv("E:\\Linder\_college\\Linear Regression\\dataset\\alumni.csv")  
  
X1 = df$percent\_of\_classes\_under\_20  
X2 = df$student\_faculty\_ratio  
Y = df$alumni\_giving\_rate  
  
head(df)

## school percent\_of\_classes\_under\_20  
## 1 Boston College 39  
## 2 Brandeis University 68  
## 3 Brown University 60  
## 4 California Institute of Technology 65  
## 5 Carnegie Mellon University 67  
## 6 Case Western Reserve Univ. 52  
## student\_faculty\_ratio alumni\_giving\_rate private  
## 1 13 25 1  
## 2 8 33 1  
## 3 8 40 1  
## 4 3 46 1  
## 5 10 28 1  
## 6 8 31 1

model = lm(Y ~ X1 + X2, data = df)  
  
summary(model)

##   
## Call:  
## lm(formula = Y ~ X1 + X2, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.00 -6.57 -1.95 4.42 24.56   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 39.6556 13.5076 2.936 0.005225 \*\*   
## X1 0.1662 0.1626 1.022 0.312128   
## X2 -1.7021 0.4421 -3.850 0.000371 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 9.098 on 45 degrees of freedom  
## Multiple R-squared: 0.5613, Adjusted R-squared: 0.5418   
## F-statistic: 28.79 on 2 and 45 DF, p-value: 8.869e-09

1. What is your final estimated model?

The estimated model is Y = 39.6556 + 0.1662X1 - 1.7021X2

1. What is the predicted alumni giving rate for an observation with (X1=50,X2=10) ?

new\_data <- data.frame(X1 = c(50), X2 = c(10))  
predictions <- predict(model, newdata = new\_data)  
  
predictions

## 1   
## 30.94291

For observation X1 = 50 and X2 = 10, the model predicts the alumni giving rate of 30.94 %

1. Using the model summary, test the statistical significance of each regression coefficient using t-tests; use α=0.05. Obtain the t-statistics and p-values, interpret the results, make a conclusion (i.e. reject or not reject), and explain why. Note: please explain what the null hypothesis is.

summary(model)

##   
## Call:  
## lm(formula = Y ~ X1 + X2, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.00 -6.57 -1.95 4.42 24.56   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 39.6556 13.5076 2.936 0.005225 \*\*   
## X1 0.1662 0.1626 1.022 0.312128   
## X2 -1.7021 0.4421 -3.850 0.000371 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 9.098 on 45 degrees of freedom  
## Multiple R-squared: 0.5613, Adjusted R-squared: 0.5418   
## F-statistic: 28.79 on 2 and 45 DF, p-value: 8.869e-09

Null Hypothessis: The coefficient B1(Beta 1) is zero and coefficient B2 (Beta2) is zero. Alternate hypothesis: The coefficient B1(Beta 1) and coefficient B2 (Beta2) is not zero.

For coefficient Beta 1 (B1) we do not reject the Null hypothesis as its p value is 0.312 which is greater than 0.05. Hence there is insufficient evidence to conclude that the predictor variable has a statistically significant relationship with the response variable.

For coefficient Bets 2 (B2) we reject the null hypothesis as the p value is 0.000371 which is less than 0.05.

Since the p value is less than 0.05 there is strong evidence to suggest that the X2 coefficient is not 0 at alpha = 0.05 significance level.

1. What is the F statistic of the model? Is it significant? Clearly write out the null hypothesis, F-statistic, and p-value and interpret the test results.

summary(model)

##   
## Call:  
## lm(formula = Y ~ X1 + X2, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.00 -6.57 -1.95 4.42 24.56   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 39.6556 13.5076 2.936 0.005225 \*\*   
## X1 0.1662 0.1626 1.022 0.312128   
## X2 -1.7021 0.4421 -3.850 0.000371 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 9.098 on 45 degrees of freedom  
## Multiple R-squared: 0.5613, Adjusted R-squared: 0.5418   
## F-statistic: 28.79 on 2 and 45 DF, p-value: 8.869e-09

NUll Hypothesis: All the coefficent B1 (Beta 1) and B2 (Beta 2) are zero

Alternate hypothesis: Atleast one of the coefficient is not zero

We get the F-statistic value as 28.79 on 2 and 45 DF and p value = 8.869e-09

Since the p value is less than 0.05 we reject the null hypothesis at significance level alpha = 0.05. We conclude that at least one of the predictor variables in the model has a significant relationship with the response variable.

1. What is the value of the coefficient of determination? Please interpret.

r\_squared <- summary(model)$r.squared  
  
cat("R-squared:", r\_squared, "\n")

## R-squared: 0.5613406

We have the coeeficient of determination to be 0.5613. So approximately 56.13% of variation in the dependant variable (Y) is explained by the independant variables (X1, X2) in our present model.

1. What is the correlation coefficient r1 between X1 and Y and the correlation coefficient r2 between X2 and Y? Do you see any relationship between r1, r2, and R2 ?

r1 <- cor(X1, Y)  
r2 <- cor(X2, Y)  
  
# Calculate R-squared (R²) from your linear regression model (assuming you've already fitted the model)  
  
r\_squared <- summary(model)$r.squared  
  
# Print the correlation coefficients and R-squared  
cat("Correlation coefficient (r1) between X1 and Y:", r1, "\n")

## Correlation coefficient (r1) between X1 and Y: 0.6456504

cat("Correlation coefficient (r2) between X2 and Y:", r2, "\n")

## Correlation coefficient (r2) between X2 and Y: -0.7423975

cat("R-squared (R²) from the linear regression model:", r\_squared, "\n")

## R-squared (R²) from the linear regression model: 0.5613406

X1 and Y show a moderate positive correlation (r1 = 0.6456), indicating they move in the same direction, while X2 and Y exhibit a moderate negative correlation (r2 = -0.742), indicating opposite movement. The overall fit of the model, as represented by the R-squared value (0.5613), suggests that both X1 and X2 collectively explain a significant proportion of the variance in Y.

Although there is no direct mathematical link between individual correlations (r1 and r2) and R-squared, it is clear that the model effectively captures a substantial portion of Y’s variance, taking both X1 and X2 into account.

##Q.2) Simulation Study (Multiple Linear Regression). Assume mean function E(Y|X)=10+5X1−2X2

**a. Generate data with X1∼N(μ=2,σ=0.1), X2∼N(μ=0,σ=0.4), sample size n=100, and error term ϵ∼N(μ=0,σ=0.5).**

n <- 100  
x1 <- rnorm(n, mean = 2, sd = 0.1)  
x2 <- rnorm(n, mean = 0, sd = 0.4)  
  
error <- rnorm(n, mean = 0, sd = 0.5) # e ~ N(0, sigma = 0.5)  
  
  
y1 <- 10 + (5\*x1) - (2\*x2) + error # equivalent

**b. Fit a simple linear regression to the simulated data from part a. What is the estimated prediction equation? Report the estimated coefficients and their standard errors. Are they significant? Clearly write out the null and alternative hypotheses, observed t-statistic(s), p-value(s), and interpret the estimates and test results. What is fitted model’s MSE?**

# Part b: Fit Simple Linear Regression  
fit <- lm(y1 ~ x1 + x2)  
  
# Print summary of the linear regression model  
summary(fit)

##   
## Call:  
## lm(formula = y1 ~ x1 + x2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.26922 -0.38961 -0.04784 0.39209 1.35817   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 10.7712 0.9432 11.420 < 0.0000000000000002 \*\*\*  
## x1 4.5925 0.4668 9.837 0.000000000000000301 \*\*\*  
## x2 -2.0431 0.1296 -15.767 < 0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5271 on 97 degrees of freedom  
## Multiple R-squared: 0.7801, Adjusted R-squared: 0.7756   
## F-statistic: 172.1 on 2 and 97 DF, p-value: < 0.00000000000000022

# Extract coefficients and standard errors  
coefficients <- coef(fit)  
std\_errors <- summary(fit)$coef[, "Std. Error"]

### Null and Alternative Hypotheses:

Null Hypothesis: There is no relation between y and x1, x2.

Alternate Hypothesis: There is a relation between y and x1, x2

Since the p-value for both the estimates is less than alpha = 0.05, we reject the null hypothesis.

# Observed t-statistic and p-value for the slope (β1)  
t\_stat\_x1 <- as.numeric(coefficients["x1"] / std\_errors["x1"])  
p\_value\_x1 <- 2 \* (1 - pt(abs(t\_stat\_x1), df = n - 2))  
  
t\_stat\_x2 <- as.numeric(coefficients["x2"] / std\_errors["x2"])  
p\_value\_x2 <- 2 \* (1 - pt(abs(t\_stat\_x2), df = n - 2))  
  
  
# Interpretation of the estimates and test results  
cat("Estimated Prediction Equation: Y = ", round(coefficients["(Intercept)"], 2), " + ", round(coefficients["x1"], 2), " \* X1", " + ", round(coefficients["x2"], 2), " \* X2\n")

## Estimated Prediction Equation: Y = 10.77 + 4.59 \* X1 + -2.04 \* X2

cat("Estimated Coefficients:\n", coefficients, "\n")

## Estimated Coefficients:  
## 10.77124 4.592451 -2.043116

cat("Standard Errors:\n", std\_errors, "\n")

## Standard Errors:  
## 0.9431502 0.4668498 0.1295843

cat("t-statistic X1 (β1): ", round(t\_stat\_x1, 2), "\n")

## t-statistic X1 (β1): 9.84

cat("t-statistic X2 (β2): ", round(t\_stat\_x2, 2), "\n")

## t-statistic X2 (β2): -15.77

cat("p-value X1 (β1): ", format(p\_value\_x1, scientific = TRUE, digits = 2), "\n")

## p-value X1 (β1): 2.2e-16

cat("p-value X2 (β2): ", format(p\_value\_x2, scientific = TRUE, digits = 2), "\n")

## p-value X2 (β2): 0e+00

# Calculate Mean Squared Error (MSE)  
mse1 <- sum(fit$residuals^2) / (n - 2)  
cat("Fitted Model's MSE (n=100, σ=0.5): ", round(mse1, 2), "\n\n")

## Fitted Model's MSE (n=100, σ=0.5): 0.28

**c. Repeat part b), but re-simulate the data and change the error term to ϵ∼N(0,σ=1)**

n <- 100  
x1 <- rnorm(n, mean = 2, sd = 0.1)  
x2 <- rnorm(n, mean = 0, sd = 0.4)  
  
error2 <- rnorm(n, mean = 0, sd = 1) # e ~ N(0, sigma = 1)  
  
  
y2 <- 10 + (5\*x1) - (2\*x2) + error2 # equivalent  
  
fit2 <- lm(y2 ~ x1+x2)  
  
# Print summary of the linear regression model  
summary(fit2)

##   
## Call:  
## lm(formula = y2 ~ x1 + x2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.70759 -0.68049 0.04353 0.63624 2.98898   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.7750 1.7955 5.444 0.0000003923006 \*\*\*  
## x1 5.1768 0.8972 5.770 0.0000000947393 \*\*\*  
## x2 -1.9589 0.2692 -7.277 0.0000000000893 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9313 on 97 degrees of freedom  
## Multiple R-squared: 0.4798, Adjusted R-squared: 0.469   
## F-statistic: 44.73 on 2 and 97 DF, p-value: 0.00000000000001722

# Extract coefficients and standard errors  
coefficients2 <- coef(fit2)  
std\_errors2 <- summary(fit2)$coef[, "Std. Error"]

# Observed t-statistic and p-value for the slope (β1)  
t\_stat\_x3 <- as.numeric(coefficients2["x1"] / std\_errors2["x1"])  
p\_value\_x3 <- 2 \* (1 - pt(abs(t\_stat\_x3), df = n - 2))  
  
t\_stat\_x4 <- as.numeric(coefficients2["x2"] / std\_errors2["x2"])  
p\_value\_x4 <- 2 \* (1 - pt(abs(t\_stat\_x4), df = n - 2))  
  
  
# Interpretation of the estimates and test results  
cat("Estimated Prediction Equation: Y = ", round(coefficients2["(Intercept)"], 2), " + ", round(coefficients2["x1"], 2), " \* X1", " + ", round(coefficients2["x2"], 2), " \* X2\n")

## Estimated Prediction Equation: Y = 9.77 + 5.18 \* X1 + -1.96 \* X2

cat("Estimated Coefficients:\n", coefficients2, "\n")

## Estimated Coefficients:  
## 9.774987 5.176841 -1.958886

cat("Standard Errors:\n", std\_errors2, "\n")

## Standard Errors:  
## 1.795544 0.8972001 0.2692037

cat("t-statistic X1 (β1): ", round(t\_stat\_x3, 2), "\n")

## t-statistic X1 (β1): 5.77

cat("t-statistic X2 (β2): ", round(t\_stat\_x4, 2), "\n")

## t-statistic X2 (β2): -7.28

cat("p-value X1 (β1): ", format(p\_value\_x3, scientific = TRUE, digits = 2), "\n")

## p-value X1 (β1): 9.3e-08

cat("p-value X2 (β2): ", format(p\_value\_x4, scientific = TRUE, digits = 2), "\n")

## p-value X2 (β2): 8.6e-11

# Calculate Mean Squared Error (MSE)  
mse2 <- sum(fit2$residuals^2) / (n - 2)  
cat("Fitted Model's MSE (n=100, σ=1): ", round(mse2, 2), "\n\n")

## Fitted Model's MSE (n=100, σ=1): 0.86

Null Hypothesis: There is no relation between y and x1, x2.

Alternate Hypothesis: There is a relation between y and x1, x2

Since the p-value for both the estimates is less than alpha = 0.05, we reject the null hypothesis.

**d. Repeat parts a)–c) using n=400. What do you conclude? What is the effect to the model parameter estimates when the error variance gets smaller? What is the effect when the sample size gets bigger?**

n <- 400  
x3 <- rnorm(n, mean = 2, sd = 0.1)  
x4 <- rnorm(n, mean = 0, sd = 0.4)  
  
error3 <- rnorm(n, mean = 0, sd = 0.5) # e ~ N(0, sigma = 0.5)  
  
  
y3 <- 10 + (5\*x3) - (2\*x4) + error3 # equivalent  
  
fit3 <- lm(y3 ~ x3+x4)  
  
# Print summary of the linear regression model  
summary(fit3)

##   
## Call:  
## lm(formula = y3 ~ x3 + x4)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.70776 -0.33201 0.00989 0.33920 1.31426   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.15352 0.51211 17.87 <0.0000000000000002 \*\*\*  
## x3 5.40843 0.25526 21.19 <0.0000000000000002 \*\*\*  
## x4 -1.93572 0.06408 -30.21 <0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5029 on 397 degrees of freedom  
## Multiple R-squared: 0.7728, Adjusted R-squared: 0.7716   
## F-statistic: 675 on 2 and 397 DF, p-value: < 0.00000000000000022

# Extract coefficients and standard errors  
coefficients3 <- coef(fit3)  
std\_errors3 <- summary(fit3)$coef[, "Std. Error"]

# Observed t-statistic and p-value for the slope (β1)  
t\_stat\_x5 <- as.numeric(coefficients3["x3"] / std\_errors3["x3"])  
p\_value\_x5 <- 2 \* (1 - pt(abs(t\_stat\_x5), df = n - 2))  
  
t\_stat\_x6 <- as.numeric(coefficients3["x4"] / std\_errors3["x4"])  
p\_value\_x6 <- 2 \* (1 - pt(abs(t\_stat\_x6), df = n - 2))  
  
  
# Interpretation of the estimates and test results  
cat("Estimated Prediction Equation: Y = ", round(coefficients3["(Intercept)"], 2), " + ", round(coefficients3["x3"], 2), " \* X1", " + ", round(coefficients3["x4"], 2), " \* X2\n")

## Estimated Prediction Equation: Y = 9.15 + 5.41 \* X1 + -1.94 \* X2

cat("Estimated Coefficients:\n", coefficients3, "\n")

## Estimated Coefficients:  
## 9.153521 5.40843 -1.935725

cat("Standard Errors:\n", std\_errors3, "\n")

## Standard Errors:  
## 0.512105 0.2552635 0.06407933

cat("t-statistic X1 (β1): ", round(t\_stat\_x5, 2), "\n")

## t-statistic X1 (β1): 21.19

cat("t-statistic X2 (β2): ", round(t\_stat\_x6, 2), "\n")

## t-statistic X2 (β2): -30.21

cat("p-value X1 (β1): ", format(p\_value\_x5, scientific = TRUE, digits = 2), "\n")

## p-value X1 (β1): 0e+00

cat("p-value X2 (β2): ", format(p\_value\_x6, scientific = TRUE, digits = 2), "\n")

## p-value X2 (β2): 0e+00

# Calculate Mean Squared Error (MSE)  
mse3 <- sum(fit3$residuals^2) / (n - 2)  
cat("Fitted Model's MSE (n=100, σ=0.5): ", round(mse3, 2), "\n\n")

## Fitted Model's MSE (n=100, σ=0.5): 0.25

Null Hypothesis: There is no relation between y and x1, x2.

Alternate Hypothesis: There is a relation between y and x1, x2

Since the p-value for both the estimates is less than alpha = 0.05, we reject the null hypothesis.

n <- 400  
x3 <- rnorm(n, mean = 2, sd = 0.1)  
x4 <- rnorm(n, mean = 0, sd = 0.4)  
  
error4 <- rnorm(n, mean = 0, sd = 1) # e ~ N(0, sigma = 0.5)  
  
  
y4 <- 10 + (5\*x3) - (2\*x4) + error4 # equivalent  
  
fit4 <- lm(y4 ~ x3+x4)  
  
# Print summary of the linear regression model  
summary(fit4)

##   
## Call:  
## lm(formula = y4 ~ x3 + x4)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.3070 -0.6650 -0.0088 0.6515 2.7999   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.3073 0.9671 8.59 <0.0000000000000002 \*\*\*  
## x3 5.8339 0.4838 12.06 <0.0000000000000002 \*\*\*  
## x4 -1.9666 0.1238 -15.88 <0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9855 on 397 degrees of freedom  
## Multiple R-squared: 0.5088, Adjusted R-squared: 0.5063   
## F-statistic: 205.6 on 2 and 397 DF, p-value: < 0.00000000000000022

# Extract coefficients and standard errors  
coefficients4 <- coef(fit4)  
std\_errors4 <- summary(fit4)$coef[, "Std. Error"]

# Observed t-statistic and p-value for the slope (β1)  
t\_stat\_x7 <- as.numeric(coefficients4["x3"] / std\_errors4["x3"])  
p\_value\_x7 <- 2 \* (1 - pt(abs(t\_stat\_x7), df = n - 2))  
  
t\_stat\_x8 <- as.numeric(coefficients4["x4"] / std\_errors4["x4"])  
p\_value\_x8 <- 2 \* (1 - pt(abs(t\_stat\_x8), df = n - 2))  
  
  
# Interpretation of the estimates and test results  
cat("Estimated Prediction Equation: Y = ", round(coefficients4["(Intercept)"], 2), " + ", round(coefficients4["x3"], 2), " \* X1", " + ", round(coefficients4["x4"], 2), " \* X2\n")

## Estimated Prediction Equation: Y = 8.31 + 5.83 \* X1 + -1.97 \* X2

cat("Estimated Coefficients:\n", coefficients4, "\n")

## Estimated Coefficients:  
## 8.30734 5.83386 -1.966578

cat("Standard Errors:\n", std\_errors4, "\n")

## Standard Errors:  
## 0.9670778 0.4837535 0.1238191

cat("t-statistic X1 (β1): ", round(t\_stat\_x7, 2), "\n")

## t-statistic X1 (β1): 12.06

cat("t-statistic X2 (β2): ", round(t\_stat\_x8, 2), "\n")

## t-statistic X2 (β2): -15.88

cat("p-value X1 (β1): ", format(p\_value\_x7, scientific = TRUE, digits = 2), "\n")

## p-value X1 (β1): 0e+00

cat("p-value X2 (β2): ", format(p\_value\_x8, scientific = TRUE, digits = 2), "\n")

## p-value X2 (β2): 0e+00

# Calculate Mean Squared Error (MSE)  
mse4 <- sum(fit4$residuals^2) / (n - 2)  
cat("Fitted Model's MSE (n=100, σ=0.5): ", round(mse3, 2), "\n\n")

## Fitted Model's MSE (n=100, σ=0.5): 0.25

Null Hypothesis: There is no relation between y and x1, x2.

Alternate Hypothesis: There is a relation between y and x1, x2

Since the p-value for both the estimates is less than alpha = 0.05, we reject the null hypothesis. As the sample size increases the parameter estimates (including MSE) are closer to the true values (same is observed when the error variance decreases).

**e. What about the MSE from each model?**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **B2** | **Std error B2** | **B1** | **Std error B1** | **B0** | **Std Error B0** | **MSE** |
| **N = 100, sd = 0.5** | -2.04 | 0.1296 | 4.59 | 0.4668 | 10.77 | 0.9432 | 0.28 |
| **N = 100, sd = 1** | -1.95 | 0.2692 | 5.17 | 0.8872 | 9.77 | 1.7955 | 0.86 |
| **N = 400, sd = 0.5** | -1.93 | 0.06408 | 5.40 | 0.25526 | 9.15 | 0.51211 | 0.25 |
| **N = 400, sd = 1** | -1.966 | 0.1238 | 5.83 | 0.4838 | 8.30 | 0.9671 | 0.25 |

# We observe that when we use N = 100 and variance = 0.5 the MSE is the least 0.28. This is because less number of points are scattered across the space that is not very widely spread.

# On the other hand, when we use N = 400 and variance of 0.5, the MSE is 0.25. The MSE has increased probably due to the increase in the number of points. There are more points now that are farther from the fitted line as compared to N = 100.

# The MSE further increases when we increase the variance to 1 with N = 100, this is because the points are now more widespread. The MSE in this case is 0.86. However, for N = 400 and variance = 1, the value of MSE has decreased to 0.95 which is due to the increase in the number of points and more points spread in the periphery of our variance value.

# Question 3





