SCMA 288

# NAME and ID

**Student name 1**:

**Student ID 1**:

**Role:**

**Student name 2**:

**Student ID 2**:

**Role:**

**Student name 3**:

**Student ID 3**:

**Role:**

**Student name 4**:

**Student ID 4**:

**Role:**

**Student name 5**:

**Student ID 5**:

**Role:**

# **set.seed**

**set.seed(6505769)**

# 2020 project

## Scenario 1

1. **Answer:**

1. **Answer:**



**Comment**: The data seem to be normal distribution.

1. **Answer:** 
   1. The predicted mortality rate of a person aged 50 is 0.003858708.
   2. The prediction interval at 93% is .
2. **Answer:**
   1. The number of positive signs is 27.
   2. 0 because we counted 27 positive signs and 27 negative signs in our test, making them equal. This means that the middle value of our data is likely zero. Since we're testing at a 5% significance level and found no strong evidence against this, we can't say the median isn't zero. So, our model passes the test.
3. **Answer:**
4. **Answer:**

|  |  |  |
| --- | --- | --- |
|  | **SLR** | **MLR** |
|  | 0.9954 | 0.9911 |
|  | 0.9953 | 0.9908 |
|  |  |  |
|  |  |  |
|  |  |  |

**7. Answer:**

**A screenshot of a computer code

Description automatically generated**

**a. F-statistic: 2854 on 2 and 51 DF**

**b.** 1 since given that the p-value of the F-test is less than which approximates to zero, it indicates that the model is appropriate and successfully passes the test.

**8. Answer**



SLR Fitted line (Blue line):

MLR Fitted line (Green line):

**9. Answer**

The fitted line in the Simple Linear Regression (SLR) model's plot appears to fit the data better than that of the Multiple Linear Regression (MLR) model. Furthermore, the adjusted R-squared value for the SLR model is 0.9953, which is higher than the MLR model's adjusted R-squared value of 0.9908. Therefore, we can conclude that the SLR model is more suitable for this dataset.

## Scenario 2

1. **Answer:** 
   1. Baht.
2. **Answer:** 
   1. The explanatory variables x\_1 and x\_2 in the multiple linear regression (MLR) model demonstrate a substantial issue with multicollinearity. This is evident from their high Variance Inflation Factor (VIF) values and correlation coefficient. Specifically, the VIF values for both x\_1 and x\_2 are 37.86251, significantly exceeding the recommended threshold of 10. Additionally, the correlation coefficient between x\_1 and x\_2 stands at 0.986706, well above the advisable limit of 0.8. These figures clearly indicate that the MLR model is severely affected by multicollinearity.
   2. To fix the issue with x\_1 and x\_2 in our model, we could try two things. First, we could use x\_1 and x\_2 in their own simple linear regression (SLR) models instead of putting them together. This means we treat x\_1 and x\_2 separately. Another option is to keep our multiple linear regression (MLR) model but find different variables to use instead of x\_1 or x\_2. This way, we can still have a more complex model without the problems caused by using x\_1 and x\_2 together.
3. **Answer:** 
   1. The adjusted R-squared value of -0.00365 suggests that the model lacks predictive power, and the independent variable x3 does not significantly contribute to it. The high P-value of 0.9529, being greater than the threshold of 0.05, further indicates that the model is statistically insignificant. Additionally, the presence of leverage and outliers is evident from the analysis of hat values and studentized residuals. Leverage is confirmed by a hat value of 0.9920557, surpassing the threshold of (2(k+1))/n, which is 0.014545. Furthermore, the model has 11 outliers, identified by 11 studentized residuals exceeding the value of 2.



* 1. It's advisable to remove all leverage points and outliers from the SLR model.

Model before adjusted

Fitted line:

A screenshot of a computer code

Description automatically generated

Adjusted model

Fitted line:

A screenshot of a computer code

Description automatically generated

Since the p-value of the adjusted model from the F-test is less than , which is much smaller than the standard significance level of 0.05. In simpler terms, this means there's strong evidence that our model accurately represents the observed data.

## R code

**Please include your R code here. Make sure there are NO ERRORS.**

################Q1

SLR <- lm(Obs ~ AGE)

SLR$coef

#Q2

Z.x <- rstandard(SLR)

qqnorm(Z.x, main = "Q-Q Plot of the Standardized residuals")

lines(Z.x, Z.x, col = "red")

#Q3

NewData <- data.frame(AGE=50)

Predict <- predict(SLR, NewData, interval="prediction", level=.93)

Predict

exp(Predict[1])

exp(Predict[2])

exp(Predict[3])

#Q4

Sign(Z.x)

#Q5

AGE3 = AGE ^ 3

AGE5 = AGE ^ 5

MLR <- lm(Obs~AGE3+AGE5)

MLR$coef

#Q6 Q7

summary(SLR)

summary(MLR)

# Q8

plot(AGE, Obs,xlab = "Age",ylab = "log(mortality)", main = "Scatter plot between logarithmic mortality rate and age")

lines(AGE, SLR$fit,col = "blue")

lines(AGE, MLR$fit, col = "green" )

SLR$coef

MLR$coef

#Q9

summary(SLR)

summary(MLR)

#---------------------------------------------------------------------------------------

# Scenario 2

## Q1

MLR2 <- lm(Benefit ~ x1+x2, data=Future.Life)

summary(MLR2)

predict\_x1 = 3500

predict\_x2 = 5000

NewData2 <- data.frame(x1=predict\_x1, x2=predict\_x2)

Predict\_val <- predict(MLR2, NewData2, interval="prediction")

Predict\_val

#Q2

vif(MLR2,)

cor(Future.Life[,c(X1,X2)], use = "complete.obs")

#Q3

#a

SLR2 <- lm(Benefit ~ x3, data=Future.Life)

summary(SLR2)

SLR2$coef

#b

plot(x3,Benefit, xlab = "x3", ylab = "Benefit")

lines(x3,SLR2$coef[1]+SLR2$coef[2]\*x3, col = "red")

summary(SLR2)

#c

#outlier

r\_t <- rstudent(SLR2)

outlier\_indices <- which(abs(r\_t) > 2)

length(outlier\_indices)

# Leverages

hatval <- hatvalues(SLR2)

average\_leverage <- 2\*(length(coef(SLR2)))/nrow(Future.Life)

high\_leverage\_indices <- which(hatval > average\_leverage)

length(high\_leverage\_indices)

#Newmodel

indices\_to\_remove <- unique(c(outlier\_indices, high\_leverage\_indices))

cleaned\_data <- Future.Life[-indices\_to\_remove, ]

x3\_cleaned <- x3[-indices\_to\_remove]

cleaned\_data$x3 <- x3\_cleaned

SLR2\_cleaned <- lm(Benefit ~ x3, data=cleaned\_data)

summary(SLR2\_cleaned)