**STAT 46700/CS 5900 Topics in Data Science Spring 2025**

**Lab 4  
[Vaishak Balachandra]**

**Q.N. 1)** An economist studied 10 mutual firms and 10 stock firms. Let y= Number of months elapsed, X1= Size of the firm and x2= Type of firm. Below is the data

|  |  |  |
| --- | --- | --- |
| y | X1 | X2 |
| 17 | 151 | Mutual |
| 26 | 92 | Mutual |
| 21 | 175 | Mutual |
| 30 | 31 | Mutual |
| 22 | 104 | Mutual |
| 0 | 277 | Mutual |
| 12 | 210 | Mutual |
| 19 | 120 | Mutual |
| 4 | 290 | Mutual |
| 16 | 238 | Mutual |
| 28 | 164 | Stock |
| 15 | 272 | Stock |
| 11 | 295 | Stock |
| 38 | 68 | Stock |
| 31 | 85 | Stock |
| 21 | 224 | Stock |
| 20 | 166 | Stock |
| 13 | 305 | Stock |
| 30 | 124 | Stock |
| 14 | 246 | Stock |

1. Draw a scatter plot of Size of the firm vs. Number of months elapsed. Also choose different colors to display Type of the firm.
2. Fit a regression model with indicator variable and write out the regression models.

> #################### Q1

>

> install.packages("readxl")

> library(readxl)

> # file.choose()

> Q1 <- read\_excel("C:/Users/PNW\_checkout/Downloads/sem2/0. Coursework/Data science/Lab/Lab 4/Q1.xlsx")

> head(Q1,5)

# A tibble: 5 × 3

y X1 X2

*<dbl>* *<dbl>* *<chr>*

1 17 151 Mutual

2 26 92 Mutual

3 21 175 Mutual

4 30 31 Mutual

5 22 104 Mutual

> names(Q1)

[1] "y" "X1" "X2"

> dim(Q1)

[1] 20 3

> attach(Q1)

>

> # a

> plot(X1,y,pch = 17, main = "Size of the firm vs Number of months elapsed", col = ifelse(X2 == "Mutual", "red", "green"))

> legend(225,35,fill = c("red","green"), c("Mutual","Stock"))

>

> # b

> model = lm(y~X1+X2)

> model

Call:

lm(formula = y ~ X1 + X2)

Coefficients:

(Intercept) X1 X2Stock

33.8741 -0.1017 8.0555

> cat("Fitted Model Equation:

+

+ for Mutual:

+ y = 33.8741 -0.1017\*Size

+

+ for Stock:

+ y = 41.9296 -0.1017\*Size")

Fitted Model Equation:

for Mutual:

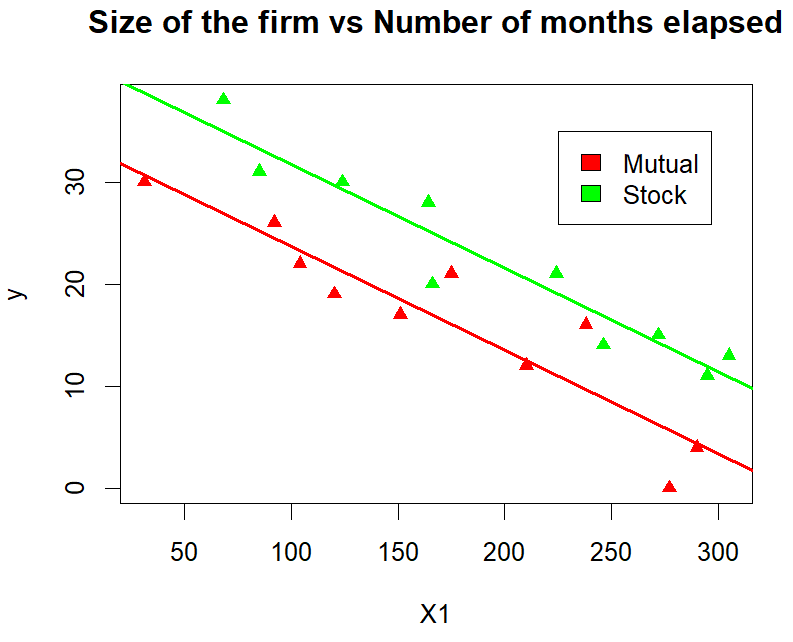
y = 33.8741 -0.1017\*Size

for Stock:

y = 41.9296 -0.1017\*Size

> abline(33.8741, -0.1017, lwd = 2, col = "red") # for Mutual

> abline(41.9296, -0.1017, lwd = 2, col = "green") # for Stock



**Q. N. 2)** The dataset *mtcars* in R was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

1. Import the data in R and extract the variables included in the dataset.
2. Fit a multiple linear regression model to model mpg using disp, hp, wt and qsec as the predictor variables.
3. Preform marginal t-test to check the significance of each predictor variables.
4. Determine the coefficient of determination.

> #################### Q2

>

> # a

> data("mtcars")

> head(mtcars)

mpg cyl disp hp drat wt qsec vs am gear carb

Mazda RX4 21.0 6 160 110 3.90 2.620 16.46 0 1 4 4

Mazda RX4 Wag 21.0 6 160 110 3.90 2.875 17.02 0 1 4 4

Datsun 710 22.8 4 108 93 3.85 2.320 18.61 1 1 4 1

Hornet 4 Drive 21.4 6 258 110 3.08 3.215 19.44 1 0 3 1

Hornet Sportabout 18.7 8 360 175 3.15 3.440 17.02 0 0 3 2

Valiant 18.1 6 225 105 2.76 3.460 20.22 1 0 3 1

> colnames(mtcars)

[1] "mpg" "cyl" "disp" "hp" "drat" "wt" "qsec" "vs" "am" "gear" "carb"

> dim(mtcars)

[1] 32 11

> attach(mtcars)

>

> # b

> model2 <- lm(mpg~disp+hp+wt+qsec)

> model2

Call:

lm(formula = mpg ~ disp + hp + wt + qsec)

Coefficients:

(Intercept) disp hp wt qsec

27.329638 0.002666 -0.018666 -4.609123 0.544160

> cat("Fitted Model Equation:

+ mpg = 27.329638 + 0.002666\*disp - 0.018666\*hp - 4.609123\*wt + 0.544160\*qsec")

Fitted Model Equation:

mpg = 27.329638 + 0.002666\*disp - 0.018666\*hp - 4.609123\*wt + 0.544160\*qsec

>

> # c

> summary(model2)

Call:

lm(formula = mpg ~ disp + hp + wt + qsec)

Residuals:

Min 1Q Median 3Q Max

-3.8664 -1.5819 -0.3788 1.1712 5.6468

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 27.329638 8.639032 3.164 0.00383 \*\*

disp 0.002666 0.010738 0.248 0.80576

hp -0.018666 0.015613 -1.196 0.24227

wt -4.609123 1.265851 -3.641 0.00113 \*\*

qsec 0.544160 0.466493 1.166 0.25362

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.622 on 27 degrees of freedom

Multiple R-squared: 0.8351, Adjusted R-squared: 0.8107

F-statistic: 34.19 on 4 and 27 DF, p-value: 3.311e-10

> cat("From the model summary, using the pvalue = 0.00113(\*\*),

+ 'wt' is the only variable that is significant, while other variables have pvalue > 0.05, that's makes them insignificant!!")

From the model summary, using the pvalue = 0.00113(\*\*),

'wt' is the only variable that is significant, while other variables have pvalue > 0.05, that's makes them insignificant!!

>

>

> # d

> # anova(model2)

> summary(model2)

Call:

lm(formula = mpg ~ disp + hp + wt + qsec)

Residuals:

Min 1Q Median 3Q Max

-3.8664 -1.5819 -0.3788 1.1712 5.6468

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 27.329638 8.639032 3.164 0.00383 \*\*

disp 0.002666 0.010738 0.248 0.80576

hp -0.018666 0.015613 -1.196 0.24227

wt -4.609123 1.265851 -3.641 0.00113 \*\*

qsec 0.544160 0.466493 1.166 0.25362

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.622 on 27 degrees of freedom

Multiple R-squared: 0.8351, Adjusted R-squared: 0.8107

F-statistic: 34.19 on 4 and 27 DF, p-value: 3.311e-10

> cat("Coefficient of Determination:

+ Multiple R^2 squared value: 83.51%

+ Adjusted R^2 squared value: 81.07%")

Coefficient of Determination:

Mulitple R^2 squared value: 83.51%

Adjusted R^2 squared value: 81.07%