**STAT 40001/ STAT 50001 Statistical Computing Fall 2024**

**Lab -16**

1. An article in the Journal of the Environmental Engineering Division [Least Squares Estimates of BOD Parameters (1980, Vol. 106, pp. 1197–1202)] took a sample from the Holston River below Kingport, Tennessee, during August 1977. The biochemical oxygen demand (BOD) test is conducted over a period of time in days. The resulting data follow:

Time (days): 1 2 4 6 8 10 12 14 16 18 20

BOD(mg/liter): 0.6 0.7 1.5 1.9 2.1 2.6 2.9 3.7 3.5 3.7 3.8

1. Assuming that a linear regression model is appropriate, fit the regression model relating BOD (y) to the time (x).

b) What is the estimated value of?

c) Calculate the expected BOD level when the time is 15 days? Provide a 90% confidence interval and prediction interval

d) What change in mean BOD is expected when the time changes by three days?

> #Q1

> time = scan()

1: 1 2 4 6 8 10 12 14 16 18 20

12:

Read 11 items

> # 1 2 4 6 8 10 12 14 16 18 20

> bod = scan()

1: 0.6 0.7 1.5 1.9 2.1 2.6 2.9 3.7 3.5 3.7 3.8

12:

Read 11 items

> # 0.6 0.7 1.5 1.9 2.1 2.6 2.9 3.7 3.5 3.7 3.8

> time;bod

[1] 1 2 4 6 8 10 12 14 16 18 20

[1] 0.6 0.7 1.5 1.9 2.1 2.6 2.9 3.7 3.5 3.7 3.8

> plot(time, bod, pch = 17, col= 2, main = "BOD vs Time")

> model1 <- lm(bod~time)

> model1

Call:

lm(formula = bod ~ time)

Coefficients:

(Intercept) time

0.6578 0.1781

> cat("Fitted model: bod = 0.6578 + 0.1781\*time")

Fitted model: bod = 0.6578 + 0.1781\*time

> abline(model1, col = "green",lwd = 2)

> anova(model1)

Analysis of Variance Table

Response: bod

Df Sum Sq Mean Sq F value Pr(>F)

time 1 13.3445 13.3445 161.69 4.694e-07 \*\*\*

Residuals 9 0.7428 0.0825

---

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

> cat("Estimated sigma square = 0.0825") # using mean square value under residuals in anova

Estimated sigma square = 0.0825

> #or

> summary(model1)

Call:

lm(formula = bod ~ time)

Residuals:

Min 1Q Median 3Q Max

-0.41892 -0.19933 0.01775 0.14581 0.54942

Coefficients:

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

(Intercept) 0.6578 0.1657 3.969 0.00326 \*\*

time 0.1781 0.0140 12.716 4.69e-07 \*\*\*

---

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

Residual standard error: 0.2873 on 9 degrees of freedom

Multiple R-squared: 0.9473, Adjusted R-squared: 0.9414

F-statistic: 161.7 on 1 and 9 DF, p-value: 4.694e-07

> predict(model1, data.frame(time = 15))

1

3.328639

> cat("For time = 15 Days -> bod -> 3.328639 units")

For time = 15 Days -> bod -> 3.328639 units

> # confidence interval

> predict(model1, data.frame(time = 15), interval = "conf", level = 0.9)

fit lwr upr

1 3.328639 3.125933 3.531346

> cat("Predicted value for 15 days time for 90% conf int of (3.12, 3.53)")

Predicted value for 15 days time for 90% conf int of (3.12, 3.53)

> # prediction interval

> predict(model1, data.frame(time = 15), interval = "pred", level = 0.9)

fit lwr upr

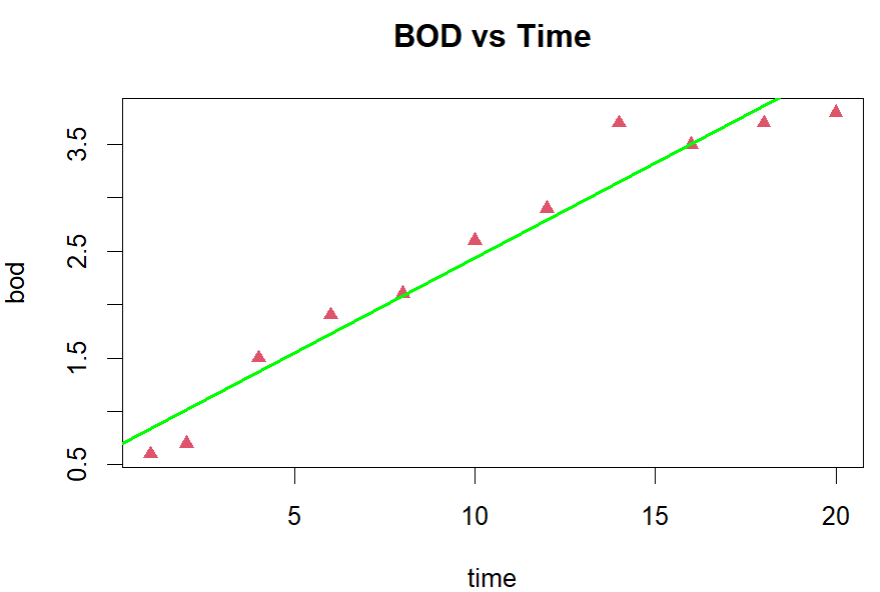
1 3.328639 2.764355 3.892924

> cat("Predicted value for 15 days time for 90% conf int of (2.76, 3.89)")

Predicted value for 15 days time for 90% conf int of (2.76, 3.89)

> cat("Change in mean BOD is expected when the time changes by three days: ", 3\*0.1781)

Change in mean BOD is expected when the time changes by three days: 0.5343



1. Air Pollution is currently one of the most serious public health worries worldwide. Many epidemiological studies have proved that some chemical compounds such as sulphur dioxide (SO2), nitrogen dioxide (NO2), ozone (O3) or other air-borne dust particles can have on our health. missMDA library in R contains ozone dataset containing 112 observations recorded during summer 2001 in Rennes (France).
2. Import the dataset in R.
3. Generate the list of variables included in the data using R.
4. Create the subset of the data containing only the first 11 variables.
5. Fit a multiple linear regression model for maxO3 as a response variable and all the remaining 10 variables as regressor variables. Use summary option to identify which of these variables appear to be significant.

> #Q2

> install.packages("missMDA")

> library(missMDA)

> data(ozone)

> head(ozone)

maxO3 T9 T12 T15 Ne9 Ne12 Ne15 Vx9 Vx12 Vx15 maxO3v vent

20010601 87 15.6 18.5 18.4 4 4 8 0.6946 -1.7101 -0.6946 84 Nord

20010602 NA 17.0 18.4 17.7 5 5 7 -4.3301 -4.0000 -3.0000 87 Nord

20010603 92 15.3 17.6 19.5 2 5 4 2.9544 1.8794 0.5209 82 Est

20010604 114 16.2 19.7 22.5 1 NA 0 0.9848 NA NA 92 <NA>

20010605 94 17.4 20.5 20.4 8 8 7 -0.5000 -2.9544 -4.3301 114 Ouest

20010606 80 17.7 19.8 18.3 6 6 7 -5.6382 -5.0000 -6.0000 NA Ouest

pluie

20010601 Sec

20010602 Sec

20010603 <NA>

20010604 Sec

20010605 Sec

20010606 Pluie

> dim(ozone)

[1] 112 13

> names(ozone)

[1] "maxO3" "T9" "T12" "T15" "Ne9" "Ne12" "Ne15" "Vx9"

[9] "Vx12" "Vx15" "maxO3v" "vent" "pluie"

> newdata <- ozone[,c(-12,-13)]

> dim(newdata)

[1] 112 11

> names(newdata)

[1] "maxO3" "T9" "T12" "T15" "Ne9" "Ne12" "Ne15" "Vx9"

[9] "Vx12" "Vx15" "maxO3v"

> attach(newdata)

> model = lm(maxO3~T9+T12+T15+Ne9+Ne12+Ne15+Vx9+Vx12+Vx15+maxO3v, data = newdata)

> summary(model)

Call:

lm(formula = maxO3 ~ T9 + T12 + T15 + Ne9 + Ne12 + Ne15 + Vx9 +

Vx12 + Vx15 + maxO3v, data = newdata)

Residuals:

Min 1Q Median 3Q Max

-34.283 -9.348 -1.107 8.886 24.448

Coefficients:

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

(Intercept) 2.9941 25.8868 0.116 0.9089

T9 6.0162 3.7342 1.611 0.1208

T12 -4.2879 3.7180 -1.153 0.2606

T15 2.2675 3.4019 0.667 0.5117

Ne9 -1.2472 2.1251 -0.587 0.5630

Ne12 0.1917 2.4544 0.078 0.9384

Ne15 -0.2909 2.3791 -0.122 0.9037

Vx9 3.1040 2.5347 1.225 0.2331

Vx12 -0.8650 2.2006 -0.393 0.6979

Vx15 -0.3501 2.0442 -0.171 0.8655

maxO3v 0.3023 0.1336 2.263 0.0334 \*

---

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

Residual standard error: 15.83 on 23 degrees of freedom

(78 observations deleted due to missingness)

Multiple R-squared: 0.7757, Adjusted R-squared: 0.6781

F-statistic: 7.952 on 10 and 23 DF, p-value: 2.156e-05

> cat("Significant Variables is(are): maxO3v")

Significant Variables is(are): maxO3v

> install.packages("car")

Error in install.packages : Updating loaded packages

Restarting R session...

> install.packages("car")

> library(car)

> vif(model)

T9 T12 T15 Ne9 Ne12 Ne15 Vx9 Vx12

21.748581 35.112292 35.467451 4.690343 4.666454 4.025955 6.582401 6.054002

Vx15 maxO3v

5.432559 1.953959

> cat("Also, VIF of maxO3v is much lesser when compared to others! -> good!!")

Also, VIF of maxO3v is much lesser when compared to others! -> good!!

1. Is a baby's birth weight related to the mother's smoking during pregnancy? Researchers (Daniel, 1999) interested in answering the above research question collected the following data (birthsmokers.txt) on a random sample of n = 32 births. The data are provided in the link below

<https://online.stat.psu.edu/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/birthsmokers/index.txt>

1. Import the data in R
2. Display the data using a scatter plot. Please be sure to address all three variables used in the dataset.
3. Fit a multiple linear regression model using gestation period and smoking status as the predictor variables to model the weight.
4. Display the models based on the smoking status.

> #Q3

> Q3 <- read.table("https://online.stat.psu.edu/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/birthsmokers/index.txt",header = T)

> head(Q3)

Wgt Gest Smoke

1 2940 38 yes

2 3130 38 no

3 2420 36 yes

4 2450 34 no

5 2760 39 yes

6 2440 35 yes

> dim(Q3)

[1] 32 3

> names(Q3)

[1] "Wgt" "Gest" "Smoke"

> attach(Q3)

> plot(Gest, Wgt, pch = ifelse(Smoke == "yes", 17, 16), col = ifelse(Smoke == "yes", "red", "green"), main = "Wgt vs Gest", col.main = "maroon",col.lab = "orange")

> model3 <- lm(Wgt~Gest+Smoke)

> model3

Call:

lm(formula = Wgt ~ Gest + Smoke)

Coefficients:

(Intercept) Gest Smokeyes

-2389.6 143.1 -244.5

> summary(model3)

Call:

lm(formula = Wgt ~ Gest + Smoke)

Residuals:

Min 1Q Median 3Q Max

-223.693 -92.063 -9.365 79.663 197.507

Coefficients:

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

(Intercept) -2389.573 349.206 -6.843 1.63e-07 \*\*\*

Gest 143.100 9.128 15.677 1.07e-15 \*\*\*

Smokeyes -244.544 41.982 -5.825 2.58e-06 \*\*\*

---

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

Residual standard error: 115.5 on 29 degrees of freedom

Multiple R-squared: 0.8964, Adjusted R-squared: 0.8892

F-statistic: 125.4 on 2 and 29 DF, p-value: 5.289e-15

> cat("If Smoking:- "); cat("Fitted Model: Wgt = -2,634.1 + 143.1\*Gest")

If Smoking:- Fitted Model: Wgt = -2,634.1 + 143.1\*Gest

> cat("If Not Smoking:- "); cat("Fitted Model: Wgt = -2389.6 + 143.1\*Gest")

If Not Smoking:- Fitted Model: Wgt = -2389.6 + 143.1\*Gest

> abline(-2634.1, 143.1, lwd = 2, col = "red")

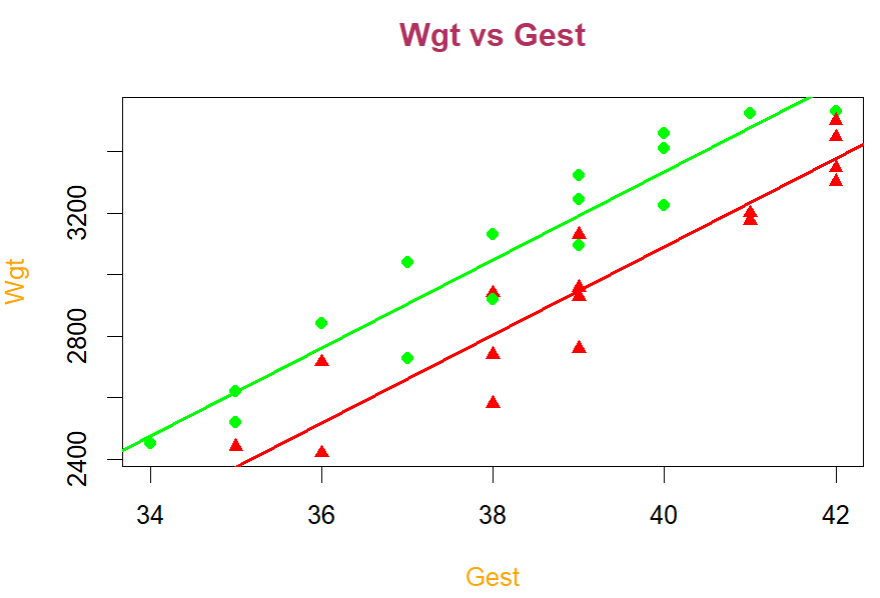
> abline(-2389.6, 143.1, lwd = 2, col = "green")

> # predict(model3, data.frame(Gest =35, Smoke = "yes"))

> predict(model3, data.frame(Gest =35, Smoke = "no"))

1

2618.937



4) 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 month elapsed. Also choose different colors to display Type of the firm.
2. Fit a regression model with indicator variable and write out the regression model.

> #Q4

> library(readxl)

> Q4 <- read\_excel("Q4.xlsx")

> head(Q4)

# A tibble: 6 × 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

6 0 277 Mutual

> dim(Q4)

[1] 20 3

> attach(Q4)

> names(Q4)

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

> plot(y,X1, pch = ifelse(X2 == "Mutual", "m","s"), col = ifelse(X2 == "Mutual", "green","red"), main = "Size of the firm vs. Number of month elapsed", xlab= "Size of the firm ",ylab = "Number of month elapsed",col.main = "maroon",col.lab = "orange")

> model4 <- lm(y~X1+X2)

> summary(model4)

Call:

lm(formula = y ~ X1 + X2)

Residuals:

Min 1Q Median 3Q Max

-5.6915 -1.7036 -0.4385 1.9210 6.3406

Coefficients:

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

(Intercept) 33.874069 1.813858 18.675 9.15e-13 \*\*\*

X1 -0.101742 0.008891 -11.443 2.07e-09 \*\*\*

X2Stock 8.055469 1.459106 5.521 3.74e-05 \*\*\*

---

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

Residual standard error: 3.221 on 17 degrees of freedom

Multiple R-squared: 0.8951, Adjusted R-squared: 0.8827

F-statistic: 72.5 on 2 and 17 DF, p-value: 4.765e-09

> model4

Call:

lm(formula = y ~ X1 + X2)

Coefficients:

(Intercept) X1 X2Stock

33.8741 -0.1017 8.0555

> cat("If Stock:- "); cat("Fitted Model: y = 41.9296 - 0.1017\*X1")

If Stock:- Fitted Model: y = 41.9296 - 0.1017\*X1

> cat("If Mutual:- "); cat("Fitted Model: y = 33.8741 - 0.1017\*X1")

If Mutual:- Fitted Model: y = 33.8741 - 0.1017\*X1

> abline(41.9296 ,-0.1017, lwd = 2, col = "red")

> abline(33.8741 ,-0.1017, lwd = 2, col = "green")

