De thi hoc phan

Le Dang Tan – 11236199 – ĐỀ 1

10/12/2024

# Câu 2:

data1 <- read.table("data1.txt", header = TRUE)  
filter <- subset(data1, tinh == '17' | tinh == '10' & M1 > 6)  
result2 <- filter[sample(nrow(filter), 8), ]  
print(result2)

## khoi tinh M1 M2 M3  
## 8326 D1 17 5.50 4.50 6.50  
## 4284 A 17 6.75 6.50 7.75  
## 4298 A 17 6.00 7.00 5.50  
## 6545 A1 17 7.75 8.00 8.50  
## 8606 D1 17 4.50 5.25 7.50  
## 8954 D1 17 5.50 3.25 5.00  
## 121 A 10 6.25 6.50 7.25  
## 113 A 10 6.75 5.75 4.75

# Câu 3:

library(quantmod)

## Loading required package: xts

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

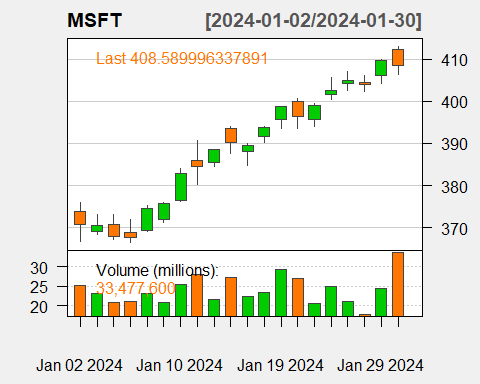
## Loading required package: TTR

## Registered S3 method overwritten by 'quantmod':  
## method from  
## as.zoo.data.frame zoo

getSymbols(Symbols = "MSFT", src = "yahoo", from = "2024-01-01", to = "2024-01-31")

## [1] "MSFT"

chartSeries(MSFT, theme = "white")



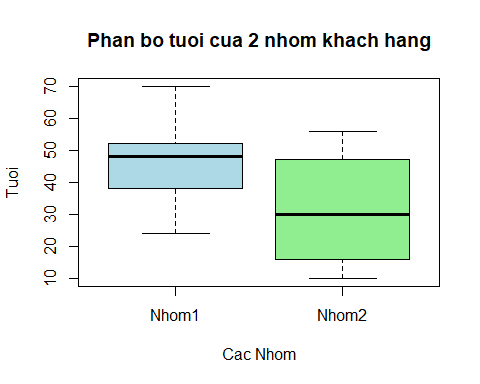
# Câu 4:

data31 <- read.table("data31.txt", header = TRUE)  
counter <- table(data31$Group)  
print(counter)

##   
## Nhom1 Nhom2   
## 30 30

Cả 2 nhóm gồm Nhóm 1 và 2 đều có 30 người

data31 <- read.table("data31.txt", header = TRUE)  
boxplot(Data ~ Group, data = data31,  
 main = "Phan bo tuoi cua 2 nhom khach hang",  
 xlab = "Cac Nhom",  
 ylab = "Tuoi",  
 col = c("lightblue", "lightgreen"))



- Nhóm 1: Tuổi tập trung, trung vị gần 50, phạm vi từ 20 đến 70, ít phân tán.  
- Nhóm 2: Tuổi phân tán hơn, trung vị gần 40, phạm vi từ 10 đến 70.  
- So sánh: Nhóm 1 đồng nhất hơn, trung vị cao hơn. Nhóm 2 đa dạng hơn, phạm vi rộng hơn.

# Câu 5:

data41 <- read.table("data41.txt", header = TRUE)  
model <- lm(data41[, 1] ~ ., data = data41)  
result5 <- step(model, direction = "both")

## Start: AIC=-22023.76  
## data41[, 1] ~ y + a1 + a2 + a3 + a4

## Warning: attempting model selection on an essentially perfect fit is nonsense

## Df Sum of Sq RSS AIC  
## - a3 1 0 0 -22026  
## - a1 1 0 0 -22026  
## - a4 1 0 0 -22025  
## - a2 1 0 0 -22024  
## <none> 0 -22024  
## - y 1 7912509 7912509 3967  
##   
## Step: AIC=-22025.67  
## data41[, 1] ~ y + a1 + a2 + a4

## Warning: attempting model selection on an essentially perfect fit is nonsense  
## Warning: attempting model selection on an essentially perfect fit is nonsense

## Df Sum of Sq RSS AIC  
## - a1 1 0 0 -22027.2  
## - a4 1 0 0 -22027.1  
## <none> 0 -22025.7  
## - a2 1 0 0 -22024.7  
## + a3 1 0 0 -22023.8  
## - y 1 184750110 184750110 5225.2  
##   
## Step: AIC=-22027.17  
## data41[, 1] ~ y + a2 + a4

## Warning: attempting model selection on an essentially perfect fit is nonsense  
## Warning: attempting model selection on an essentially perfect fit is nonsense

## Df Sum of Sq RSS AIC  
## - a4 1 0 0 -22028.5  
## <none> 0 -22027.2  
## - a2 1 0 0 -22026.5  
## + a1 1 0 0 -22025.7  
## + a3 1 0 0 -22025.5  
## - y 1 263201675 263201675 5364.8  
##   
## Step: AIC=-22028.54  
## data41[, 1] ~ y + a2

## Warning: attempting model selection on an essentially perfect fit is nonsense  
## Warning: attempting model selection on an essentially perfect fit is nonsense

## Df Sum of Sq RSS AIC  
## <none> 0 -22028.5  
## - a2 1 0 0 -22027.8  
## + a4 1 0 0 -22027.2  
## + a1 1 0 0 -22027.1  
## + a3 1 0 0 -22026.8  
## - y 1 266876243 266876243 5368.3

summary(result5)

## Warning in summary.lm(result5): essentially perfect fit: summary may be  
## unreliable

##   
## Call:  
## lm(formula = data41[, 1] ~ y + a2, data = data41)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.937e-13 -1.320e-13 -5.650e-14 3.790e-14 2.168e-11   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.746e-11 1.125e-12 -1.552e+01 <2e-16 \*\*\*  
## y 1.000e+00 6.709e-17 1.491e+16 <2e-16 \*\*\*  
## a2 4.634e-16 2.791e-16 1.661e+00 0.0976 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.096e-12 on 397 degrees of freedom  
## Multiple R-squared: 1, Adjusted R-squared: 1   
## F-statistic: 1.219e+32 on 2 and 397 DF, p-value: < 2.2e-16

Phương trình hồi quy tối ưu là:

data41[, 1] = -1.746 ^{-11} + y + 4.634 ^{-16} a2

A number with black numbers

Description automatically generated with medium confidence

- Biến y có ảnh hưởng mạnh nhất, với hệ số 1.000, nghĩa là khi y tăng 1 đơn vị, data41[, 1] tăng 1 đơn vị. - Biến a2 không có ý nghĩa thống kê (do p-value = 0.0976)

# Câu 6:

data51 <- read.table("DATA51.txt", header = TRUE)  
model <- glm(Thangnay ~ Tienmua + Thangqua, data = data51, family = 'binomial')  
summary(model)

##   
## Call:  
## glm(formula = Thangnay ~ Tienmua + Thangqua, family = "binomial",   
## data = data51)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.817e+00 5.231e-01 -3.474 0.000512 \*\*\*  
## Tienmua 1.746e-05 7.675e-06 2.274 0.022942 \*   
## Thangqua 7.710e-01 4.401e-01 1.752 0.079792 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 134.60 on 99 degrees of freedom  
## Residual deviance: 124.22 on 97 degrees of freedom  
## AIC: 130.22  
##   
## Number of Fisher Scoring iterations: 4

Hàm hồi quy (P) cho trạng thái mua hàng trong tháng hiện tại là:

data51 <- read.table("DATA51.txt", header = TRUE)  
new\_data <- data.frame(Tienmua = 50000, Thangqua = 0)  
predict(model, new\_data, type = 'response')

## 1   
## 0.2799635

new\_data <- data.frame(Tienmua = c(40000, 40000), Thangqua = c(1, 0))  
predict(model, new\_data, type = 'response')

## 1 2   
## 0.4138232 0.2461579

- Với hàm hồi quy đã tính được (P), ta thay số vào phương trình với: - Tienmua = 50000 - Thangqua = 0 Ta tính được P = -0.944 Với số đó ta tính được xác suất từ công thức đã học thu được xác suất mua hàng dự báo là 28%

- Khả năng mua hàng của hai khách hàng (1: mua, 2: không mua) là:

Khách 1:

- Tienmua=40000

- Thangqua=1

Khách 2:

- Tienmua=40000

- Thangqua=0

Từ output ta có khả năng mua hàng của khách 1 là 41,38% và khách 2 là 24,62%