Giữa kì

Bùi Khánh Duy - 20001898

2023-03-30

Câu 1.

Cho bộ dữ liệu nghiên cứu về sự biến động của tiền thuê đất nông nghiệp để trồng cỏ linh lăng - thức ăn cho bò sữa vào năm 1977 tại 67 hạt của Minnesota.

(i) Nhập dữ liệu vào R/RStudio.

```
data = read.csv("ThueDatTrongCo.csv")
```

(ii) Tìm trung bình mẫu, ma trận hiệp phương sai mẫu và ma trận tương quan mẫu.

```
cm = colMeans(data) # Trung binh mau
data.cov = cov(data) # Ma tran hiep phuong sai
data.cor = cor(data) # Ma tran tuong quan mau
cm
```

```
##       Rent    AllRent       Cows    Pasture
## 42.1661194 43.6398507 20.5632836  0.1697015
```

```
data.cov
```

```
##               Rent    AllRent       Cows    Pasture
## Rent    510.154060 418.372136 106.886387 -1.0553906
## AllRent 418.372136 447.344177  15.802796 -1.5275258
## Cows    106.886387  15.802796 235.197416  1.1580722
## Pasture  -1.055391  -1.527526   1.158072  0.0208787
```

```
data.cor
```

```
##               Rent     AllRent       Cows    Pasture
## Rent     1.0000000  0.87577226 0.30857124 -0.3233783
## AllRent  0.8757723  1.00000000 0.04871882 -0.4998227
## Cows     0.3085712  0.04871882 1.00000000  0.5225979
## Pasture -0.3233783 -0.49982268 0.52259791  1.0000000
```

(iii) Tìm giá trị riêng, vectơ riêng của ma trận hiệp phương sai mẫu.

```
data.eigen = eigen(data.cov)
data.eigen$values # Giá trị riêng
```

```
## [1] 9.101420e+02 2.433629e+02 3.920241e+01 9.274382e-03
```

```
data.eigen$vectors # vectơ riêng
```

```
##              [,1]         [,2]          [,3]         [,4]
## [1,]  0.733105729 -0.118574101  0.6696974204  0.001240119
## [2,]  0.667235780  0.316077012 -0.6744521565  0.002454512
## [3,]  0.131716192 -0.941275788 -0.3108356237 -0.005652451
## [4,] -0.001802393 -0.005949401 -0.0009320555  0.999980243
```

(iv) Từng biến ngẫu nhiên có phân bố chuẩn 1—chiều không?

Với biến `Rent`

```
shapiro.test(data$Rent)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$Rent
## W = 0.95406, p-value = 0.01471
```

Vì p-value = 0.01471 $<$ 0.05 => với khoảng tin cậy 95%, không thể kết luận biến `Rent` tuân theo phân bố chuẩn 1-chiều

Với biến `AllRent`

```
shapiro.test(data$AllRent)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$AllRent
## W = 0.9552, p-value = 0.01686
```

Vì p-value = 0.01686 $<$ 0.05 => với khoảng tin cậy 95%, không thể kết luận biến `AllRent` tuân theo phân bố chuẩn 1-chiều

Với biến `Cows`

```
shapiro.test(data$Cows)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$Cows
## W = 0.90801, p-value = 0.0001149
```

Vì p-value = 0.0001149 < 0.05 => với khoảng tin cậy 95%, không thể kết luận biến `Cows` tuân theo phân bố chuẩn 1-chiều

Với biến `Pasture`

```
shapiro.test(data$Pasture)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$Pasture
## W = 0.81827, p-value = 1.229e-07
```

Vì p-value = 1.229e-07 < 0.05 => với khoảng tin cậy 95%, không thể kết luận biến `Pasture` tuân theo phân bố chuẩn 1-chiều

(v) X có phân bố chuẩn 4−chiều không?

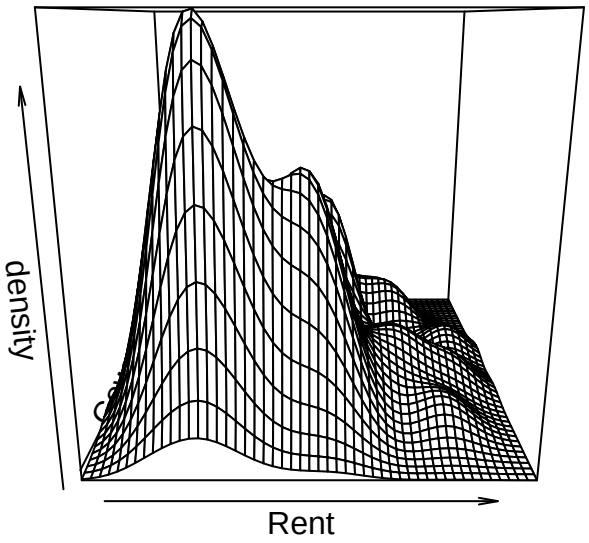
Từ kết quả ở câu `(iv)` ta có thể kết luận X không phân bố chuẩn 4-chiều

(vi) Vẽ biểu đồ thể hiện rõ hàm mật độ hai chiều của hai biến Rent và Cows.

```
# install.packages("KernSmooth")
library(KernSmooth)
```

```
## KernSmooth 2.23 loaded
## Copyright M. P. Wand 1997-2009
```

```
x <- data[,c("Rent", "Cows")]
do_thi <- bkde2D(x, bandwidth =c(dpik(data$Rent), dpik(data$Cows)))
persp(x = do_thi$x1, y = do_thi$x2, z = do_thi$fhat , xlab="Rent", ylab="Cows", zlab="density")
```



(vii) Giải thuật từng bước step backward để tìm mô hình biểu diễn Rent theo các biến còn lại "phù hợp nhất". Viết phương trình hồi quy tuyến tính.

```
# Backward = all to only
# install.packages(stats)
library(stats)
only = lm(Rent ~ 1, data = data)
all = lm(Rent ~ ., data = data)

backward = step(object = all, scope = formula(only), direction = "backward", trace = 0)
summary(backward)
```

```
##
## Call:
## lm(formula = Rent ~ AllRent + Cows, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.4827  -5.8720   0.3321   4.3855  28.6007
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.11433    2.96123  -2.065    0.043 *
## AllRent      0.92137    0.05382  17.121  < 2e-16 ***
```

```
## Cows         0.39255    0.07422   5.289 1.59e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.236 on 64 degrees of freedom
## Multiple R-squared:  0.8379, Adjusted R-squared:  0.8328
## F-statistic: 165.3 on 2 and 64 DF,  p-value: < 2.2e-16
```

```
backward$coefficients
```

```
## (Intercept)     AllRent        Cows
##  -6.1143282   0.9213684   0.3925476
```

Từ kết quả này ta có phương trình HQTT:

$$\text{Rent} = -6.1143282 + 0.9213684 \times \text{AllRent} + 0.3925476 \times \text{Cows}$$

(viii) Kiểm định xem phần dư trong mô hình có tuân theo phân phối chuẩn với giá trị trung bình bằng 0 không?

Dùng shapiro test để kiểm định xem phần dư trong mô hình có tuân theo phân phối chuẩn hay không.

```
shapiro.test(backward$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  backward$residuals
## W = 0.98075, p-value = 0.3855
```

Vì p-value = 0.3855 > 0.05

=> Với khoảng tin cậy 95%, có thể kết luận phần dư của mô hình tuân theo phân phối chuẩn

Bài toán: $H0 : EX(\varepsilon) = EX(\varepsilon)$; $H1 : EX(\varepsilon) \neq 0$

Dùng wilcox test để kiểm định phần dư có giá trị trung bình bằng 0 hay không

```
wilcox.test(backward$residuals)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  backward$residuals
## V = 1117, p-value = 0.8932
## alternative hypothesis: true location is not equal to 0
```

Vì p-value = 0.8932 > 0.05 nên chấp nhận H0

Kết luận: Với khoảng tin cậy 95%, không đủ cơ sở để bác bỏ H0, vậy phần dư có giá trị trung bình bằng 0.

Câu 2

Sinh ngẫu nhiên 1000 giá trị của biến X biết $X \sim N(160, 16.5)$ và sinh ngẫu nhiên 1000 giá trị của biến Y biết $Y \sim N(170, 22.5)$. Khi đó, vectơ ngẫu nhiên $(X, Y)^T$ có phân bố chuẩn 2−chiều không?

Chuẩn bị dữ liệu, (X, Y) được lưu dưới dạng ma trận

```
set.seed(216)
X <- rnorm(1000, 160, 16.5)
Y <- rnorm(1000, 170, 22.5)
head(X)
```

```
## [1] 186.7862 177.5148 156.4747 150.9186 161.6321 144.9999
```

```
head(Y)
```

```
## [1] 184.5723 202.0601 159.0995 191.3525 228.7761 162.1760
```

```
matrix_xy = matrix(c(X, Y), ncol=2)
```

Bài toán: H0: $(X, Y)^T$ có pb chuẩn 2-chiều; H1: $(X, Y)^T$ có không có pb chuẩn 2-chiều

Sử dụng hàm `mshapiro.test` của thư viện `mvnormtest` để kiểm định cho $(X, Y)^T$ có phân bố chuẩn 2-chiều hay không?

```
# install.packages("mvnormtest")
mvnormtest::mshapiro.test(t(matrix_xy))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Z
## W = 0.99787, p-value = 0.2306
```

Vì p-value $= 0.2306 > 0.05 \Rightarrow$ Chấp nhận H0

Vậy với khoảng tin cậy 95%, có thể kết luận rằng $(X, Y)^T$ tuân theo phân bố chuẩn 2-chiều.