Bài tập phân tích thành phần chính (tiếp)

Bùi Khánh Duy

2023-04-19

Cho bộ dữ liệu Sleep in Mammals về một số nhân tố liên quan đến giấc ngủ của 62 loài động vật có vú.

Biến:

- **BodyWt** - trọng lượng cơ thể (kg)
- **BrainWt** - trọng lượng não (g)
- **NonDreaming** - số giờ ngủ không mơ (giờ/ngày)
- **Dreaming** - số giờ ngủ có mơ (giờ/ngày)
- **TotalSleep** - tổng số giờ ngủ (giờ/ngày)
- **LifeSpan** - tuổi thọ (năm)
- **Gestation** - thời gian mang thai (ngày)
- **Predation** - chỉ số bị săn mồi (1-5): 1= ít khả năng bị săn mồi nhất; 5 = rất có thể là con mồi
- **Exposure** - chỉ số tiếp xúc khi ngủ (1-5): 1 = ít tiếp xúc nhất (ngủ trong hang được bảo vệ tốt); 5 = tiếp xúc nhiều nhất
- **Danger** - chỉ số gặp nguy hiểm (1-5): 1 = ít gặp nguy hiểm nhất; 5 = dễ gặp nguy hiểm nhất,

Thực hiện phân tích hồi quy tuyến tính tổng số giờ ngủ theo trọng lượng cơ thể, trọng lượng não, tuổi thọ, thời gian mang thai, chỉ số săn mồi, chỉ số tiếp xúc khi ngủ và chỉ số nguy hiểm.

1) Có nên thực hiện phân tích hồi quy nói trên trực tiếp không? Vì sao?

```
data = read.csv("Sleep in Mammals.csv")
dt = data[-1]
x = dt[, c(1,2,5,6,7,8,9,10)]
x
```

```
##      BodyWt BrainWt TotalSleep LifeSpan Gestation Predation Exposure Danger
## 1  6654.000 5712.00        3.3     38.6     645.0         3        5      3
## 2     1.000    6.60        8.3      4.5      42.0         3        1      3
## 3     3.385   44.50       12.5     14.0      60.0         1        1      1
## 4     0.920    5.70       16.5      7.0      25.0         5        2      3
```

```
## 5  2547.000 4603.00       3.9     69.0     624.0        3        5      4
## 6    10.550  179.50       9.8     27.0     180.0        4        4      4
## 7     0.023    0.30      19.7     19.0      35.0        1        1      1
## 8   160.000  169.00       6.2     30.4     392.0        4        5      4
## 9     3.300   25.60      14.5     28.0      63.0        1        2      1
## 10   52.160  440.00       9.7     50.0     230.0        1        1      1
## 11    0.425    6.40      12.5      7.0     112.0        5        4      4
## 12  465.000  423.00       3.9     30.0     281.0        5        5      5
## 13    0.550    2.40      10.3     13.0      58.0        2        1      2
## 14  187.100  419.00       3.1     40.0     365.0        5        5      5
## 15    0.075    1.20       8.4      3.5      42.0        1        1      1
## 16    3.000   25.00       8.6     50.0      28.0        2        2      2
## 17    0.785    3.50      10.7      6.0      42.0        2        2      2
## 18    0.200    5.00      10.7     10.4     120.0        2        2      2
## 19    1.410   17.50       6.1     34.0     621.0        1        2      1
## 20   60.000   81.00      18.1      7.0      26.0        1        1      1
## 21  529.000  680.00       5.4     28.0     400.0        5        5      5
## 22   27.660  115.00       3.8     20.0     148.0        5        5      5
## 23    0.120    1.00      14.4      3.9      16.0        3        1      2
## 24  207.000  406.00      12.0     39.3     252.0        1        4      1
## 25   85.000  325.00       6.2     41.0     310.0        1        3      1
## 26   36.330  119.50      13.0     16.2      63.0        1        1      1
## 27    0.101    4.00      13.8      9.0      28.0        5        1      3
## 28    1.040    5.50       8.2      7.6      68.0        5        3      4
## 29  521.000  655.00       2.9     46.0     336.0        5        5      5
## 30  100.000  157.00      10.8     22.4     100.0        1        1      1
## 31   35.000   56.00      12.2     16.3      33.0        3        5      4
## 32    0.005    0.14       9.1      2.6      21.5        5        2      4
## 33    0.010    0.25      19.9     24.0      50.0        1        1      1
## 34   62.000 1320.00       8.0    100.0     267.0        1        1      1
## 35    0.122    3.00      10.6      8.3      30.0        2        1      1
## 36    1.350    8.10      11.2      8.1      45.0        3        1      3
## 37    0.023    0.40      13.2      3.2      19.0        4        1      3
## 38    0.048    0.33      12.8      2.0      30.0        4        1      3
## 39    1.700    6.30      19.4      5.0      12.0        2        1      1
## 40    3.500   10.80      17.4      6.5     120.0        2        1      1
## 41  250.000  490.00      11.0     23.6     440.0        5        5      5
## 42    0.480   15.50      17.0     12.0     140.0        2        2      2
## 43   10.000  115.00      10.9     20.2     170.0        4        4      4
## 44    1.620   11.40      13.7     13.0      17.0        2        1      2
## 45  192.000  180.00       8.4     27.0     115.0        4        4      4
## 46    2.500   12.10       8.4     18.0      31.0        5        5      5
## 47    4.288   39.20      12.5     13.7      63.0        2        2      2
## 48    0.280    1.90      13.2      4.7      21.0        3        1      3
## 49    4.235   50.40       9.8      9.8      52.0        1        1      1
## 50    6.800  179.00       9.6     29.0     164.0        2        3      2
## 51    0.750   12.30       6.6      7.0     225.0        2        2      2
## 52    3.600   21.00       5.4      6.0     225.0        3        2      3
## 53   14.830   98.20       2.6     17.0     150.0        5        5      5
## 54   55.500  175.00       3.8     20.0     151.0        5        5      5
## 55    1.400   12.50      11.0     12.7      90.0        2        2      2
## 56    0.060    1.00      10.3      3.5      18.0        3        1      2
## 57    0.900    2.60      13.3      4.5      60.0        2        1      2
## 58    2.000   12.30       5.4      7.5     200.0        3        1      3
```

```
## 59    0.104    2.50       15.8      2.3      46.0         3        2      2
## 60    4.190   58.00       10.3     24.0     210.0         4        3      4
## 61    3.500    3.90       19.4      3.0      14.0         2        1      1
## 62    4.050   17.00        5.9     13.0      38.0         3        1      1
```

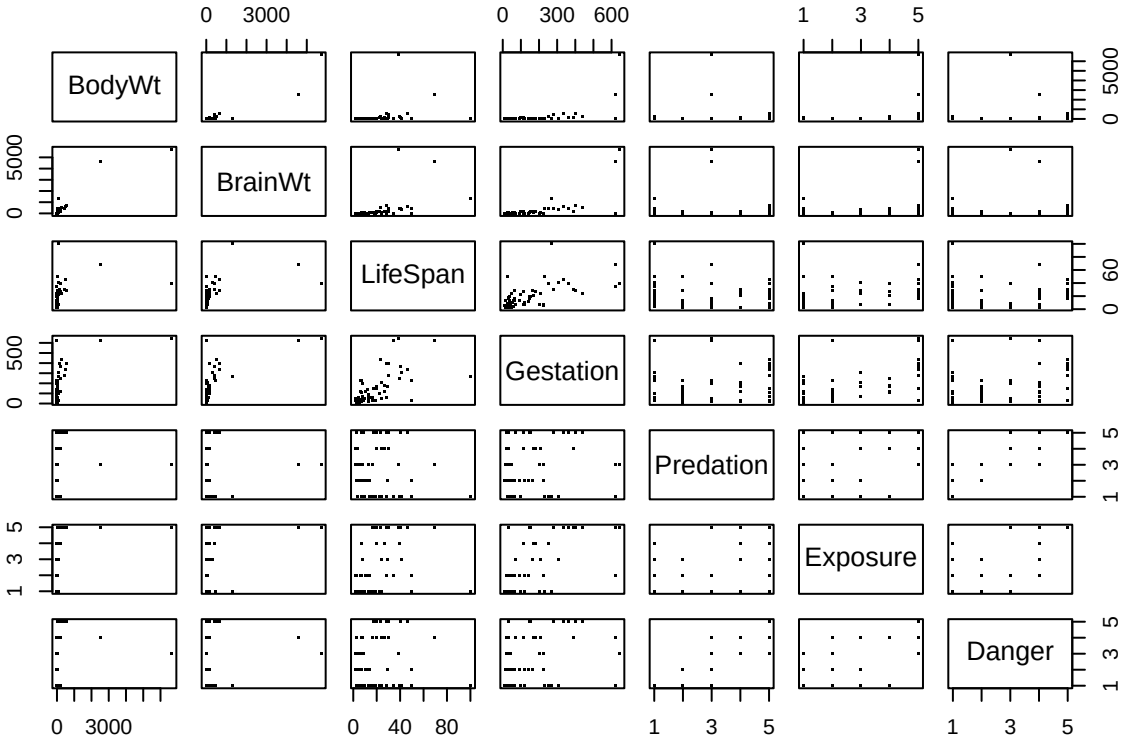
```
cor = cor(x[,-3])
cor
```

```
##               BodyWt    BrainWt    LifeSpan Gestation   Predation  Exposure
## BodyWt    1.00000000 0.93416384  0.30709665 0.5853735  0.05949472 0.3382737
## BrainWt   0.93416384 1.00000000  0.51349541 0.6708432  0.03385548 0.3678004
## LifeSpan  0.30709665 0.51349541  1.00000000 0.6216972 -0.10700759 0.3762348
## Gestation 0.58537352 0.67084318  0.62169720 1.0000000  0.13218594 0.5836509
## Predation 0.05949472 0.03385548 -0.10700759 0.1321859  1.00000000 0.6182460
## Exposure  0.33827367 0.36780037  0.37623476 0.5836509  0.61824597 1.0000000
## Danger    0.13358123 0.14587888  0.06876557 0.3015836  0.91604245 0.7872031
##                Danger
## BodyWt    0.13358123
## BrainWt   0.14587888
## LifeSpan  0.06876557
## Gestation 0.30158361
## Predation 0.91604245
## Exposure  0.78720311
## Danger    1.00000000
```

```
usair_sleep <- princomp(x[,-3], cor=T)
usair_sleep
```

```
## Call:
## princomp(x = x[, -3], cor = T)
##
## Standard deviations:
##    Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6    Comp.7
## 1.8651616 1.4444424 0.9182644 0.5471894 0.4541718 0.2295744 0.1820886
##
##  7  variables and  62 observations.
```

```
pairs(x[,-3], pch=".", cex=1.5)
```



Do **BodyWt** và **BrainWt** có tương quan mạnh => không thực hiện hồi quy tuyến tính trên 7 biến

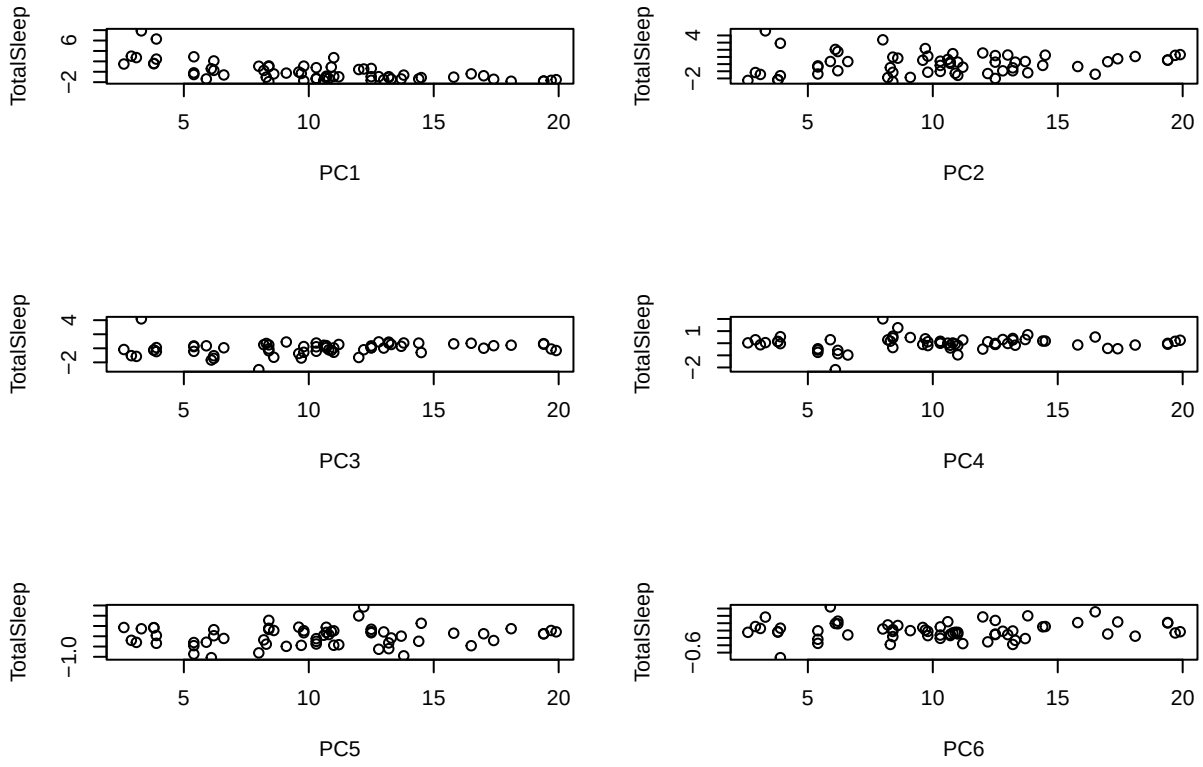
2) Vẽ ma trận biểu đồ tán xạ giữa các biến trọng lượng cơ thể, trọng lượng não, tuổi thọ, thời gian mang thai, chỉ số săn mồi, chỉ số tiếp xúc khi ngủ và chỉ số nguy hiểm.

```
cor(usair_sleep$scores)
```

```
##               Comp.1        Comp.2        Comp.3        Comp.4        Comp.5
## Comp.1  1.000000e+00  1.395793e-16 -7.825753e-16  3.044135e-16 -2.039902e-16
## Comp.2  1.395793e-16  1.000000e+00  7.675047e-16 -3.251117e-16 -1.941424e-16
## Comp.3 -7.825753e-16  7.675047e-16  1.000000e+00 -1.741806e-16  1.824818e-16
## Comp.4  3.044135e-16 -3.251117e-16 -1.741806e-16  1.000000e+00 -3.329815e-15
## Comp.5 -2.039902e-16 -1.941424e-16  1.824818e-16 -3.329815e-15  1.000000e+00
## Comp.6 -1.467864e-15 -1.047603e-15  5.160284e-16  1.571578e-15 -1.657304e-15
## Comp.7 -7.908801e-17 -8.263517e-16 -8.875476e-16  1.004190e-15  2.138222e-15
##               Comp.6        Comp.7
## Comp.1 -1.467864e-15 -7.908801e-17
## Comp.2 -1.047603e-15 -8.263517e-16
## Comp.3  5.160284e-16 -8.875476e-16
## Comp.4  1.571578e-15  1.004190e-15
## Comp.5 -1.657304e-15  2.138222e-15
## Comp.6  1.000000e+00  3.533999e-16
## Comp.7  3.533999e-16  1.000000e+00
```

```
par(mfrow = c(3,2))
out <- sapply(1:6, function(i){
  plot(x$TotalSleep, usair_sleep$scores[,i],
       xlab = paste("PC", i, sep=""),
```

```
       ylab = "TotalSleep")
})
```



3) Thực hiện phân tích thành phần chính đối với ma trận tương quan giữa các biến trong ý 2

```
summary(usair_sleep,loadings = TRUE)
```

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
## Standard deviation     1.8651616 1.4444424 0.9182644 0.54718939 0.45417181
## Proportion of Variance 0.4969754 0.2980591 0.1204585 0.04277375 0.02946743
## Cumulative Proportion  0.4969754 0.7950345 0.9154930 0.95826676 0.98773419
##                           Comp.6      Comp.7
## Standard deviation     0.2295744 0.182088562
## Proportion of Variance 0.0075292 0.004736606
## Cumulative Proportion  0.9952634 1.000000000
##
## Loadings:
##           Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
## BodyWt     0.386  0.311  0.554         0.133  0.258  0.598
## BrainWt    0.419  0.350  0.336  0.255        -0.313 -0.652
## LifeSpan   0.311  0.295 -0.678  0.554         0.113  0.175
## Gestation  0.442  0.196 -0.232 -0.708 -0.458
## Predation  0.263 -0.564  0.163  0.240 -0.431  0.550 -0.198
## Exposure   0.441 -0.271 -0.200 -0.222  0.756  0.195 -0.183
## Danger     0.345 -0.514         0.119        -0.694  0.337
```

4) Đưa ra tỉ lệ biến sai tổng cộng của từng thành phần chính

Proportion of Variance: PC1 = 0.4969754, PC2 = 0.2980591, PC3 = 0.1204585, PC4 = 0.04277375, PC5 = 0.02946743, PC6 = 0.0075292, PC7 = 0.004736606

5) Biểu diễn thành phần chính thứ nhất và thứ hai theo các biến ban đầu.

```
usair_sleep$loadings[,1]
```

```
##    BodyWt   BrainWt  LifeSpan Gestation Predation  Exposure    Danger
## 0.3860353 0.4189993 0.3108115 0.4424461 0.2628181 0.4413774 0.3452051
```

```
usair_sleep$loadings[,2]
```

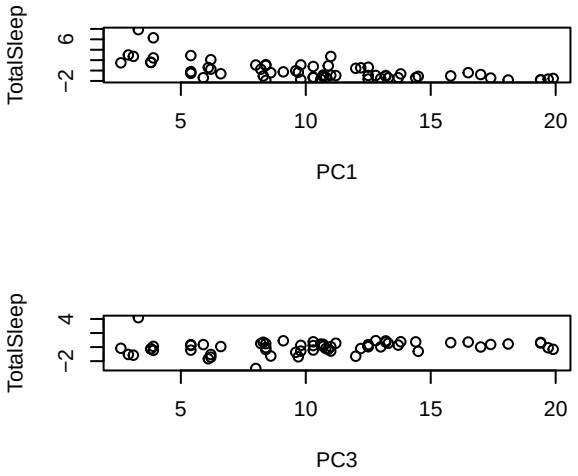
```
##     BodyWt    BrainWt   LifeSpan  Gestation  Predation   Exposure     Danger
##  0.3113421  0.3498188  0.2947869  0.1958503 -0.5642979 -0.2705969 -0.5135966
```

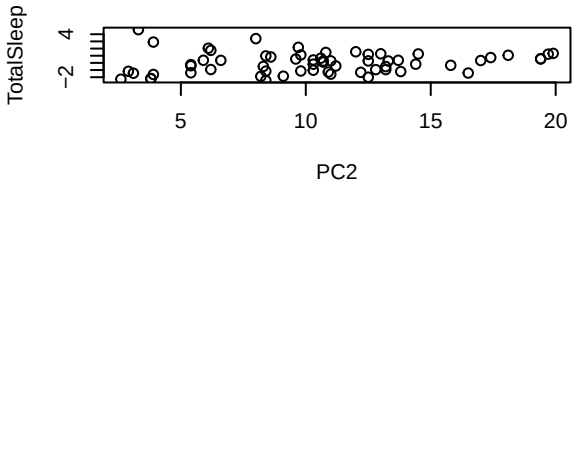
6) Cần mấy thành phần chính để thu được 90% thông tin về tập dữ liệu ban đầu.

Chỉ cần 3 thành phần chính để biểu diễn 90% thông tin về tập dữ liệu bởi vì Cumulative Proportion của 3 thành phần chính ban đầu đã đạt 0.9154930 = 91.55% > 90 %.

7) Vẽ các biểu đồ tán xạ giữa biến tổng số giờ ngủ theo m thành phần chính vừa tìm được ở ý 6

```
par(mfrow = c(3,2))
out <- sapply(1:3, function(i){
  plot(x$TotalSleep, usair_sleep$scores[,i],
       xlab = paste("PC", i, sep = ""),
       ylab = "TotalSleep")
})
```





8) Đưa ra phương trình hồi quy tuyến tính của tổng số giờ ngủ theo m thành phần chính vừa tìm được ở ý 6.

```
usair_reg <- lm(TotalSleep ~ usair_sleep$scores[,c(1,2,3)], data=x)
summary(usair_reg)
```

```
##
## Call:
## lm(formula = TotalSleep ~ usair_sleep$scores[, c(1, 2, 3)], data = x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1517 -2.0387 -0.1761  2.1974  6.8197
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           10.4097     0.4299  24.216  < 2e-16 ***
## usair_sleep$scores[, c(1, 2, 3)]Comp.1  -1.5601     0.2305  -6.769 7.15e-09 ***
## usair_sleep$scores[, c(1, 2, 3)]Comp.2   0.4455     0.2976   1.497   0.1398
## usair_sleep$scores[, c(1, 2, 3)]Comp.3   0.9220     0.4681   1.970   0.0537 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.385 on 58 degrees of freedom
## Multiple R-squared:  0.4724, Adjusted R-squared:  0.4452
## F-statistic: 17.31 on 3 and 58 DF,  p-value: 3.803e-08
```

9) Mô hình này có tốt không

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

```
##
##  Shapiro-Wilk normality test
##
## data:  usair_reg$residuals
## W = 0.98478, p-value = 0.6377
```

$p - value = 0.6377 > 0.05$ -> Chấp nhận giả thiết, độ lệch tuân theo phân bố chuẩn.

```
t.test(usair_reg$residuals)
```

```
##
##  One Sample t-test
##
## data:  usair_reg$residuals
## t = 2.0506e-16, df = 61, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.8381629  0.8381629
```

```
## sample estimates:
##    mean of x
## 8.595275e-17
```

$p - value = 1 > 0.05$ -> giá trị trung bình của phần dư $= 0$ => Mô hình tốt