ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KIỂM TRA CUỐI KỲ II
NĂM HỌC 2022 - 2023

Môn thi: PHÂN TÍCH HỒI QUY VÀ ỨNG DỤNG

Mã môn học: **MAT3379** Số tín chỉ: **3** Đề số: **1**
Dành cho sinh viên hệ: Chính quy *Ngành:* Khoa học dữ liệu
Thời gian làm bài: **90 phút** *(không kể thời gian phát đề)*
(Đề thi gồm 06 trang.)

Câu 1. Giả sử có mẫu cỡ n của biến ngẫu nhiên X là $\{x_1, x_2, ..., x_n\}$. Hãy trình bày phương pháp để kiểm định xem X có tuân theo phân phối chuẩn không? Giải thích.

Câu 2. Cho bộ dữ liệu **data** về giới tính và 11 chỉ số sức khỏe của 202 vận động viên tại Viện Thể thao Úc được thu thập bởi Richard Telford và Ross Cunningham. Ký hiệu:

- Sex - Giới tính, có giá trị là nữ hoặc nam;
- RCC - Số lượng hồng cầu (triệu tế bào/cm^3);
- WCC - Số lượng bạch cầu (triệu tế bào/cm^3);
- Hc - Chỉ số các tế bào hồng cầu trong máu (%);
- Hg - Nồng độ huyết sắc tố trong các tế bào hồng cầu (mg/dL);
- Ferr - Nồng độ ferritin huyết tương (mg/dL);
- BMI - Chỉ số thể trọng (kg/m^2);
- SSF - Tổng số nếp gấp da;
- XBfat - Tỷ lệ mỡ cơ thể (%);
- LBM - Khối lượng nạc (kg);
- Ht - Chiều cao (cm);
- Wt - Cân nặng (kg).

(Nguồn: http://www.statsci.org/data/oz/ais.html)
Cho **dat** là bộ dữ liệu con của **data** sau khi loại bỏ đi biến *Sex*.
Dữ liệu **data** và **dat** được phân tích bằng phần mềm RStudio.

(A) Chia ngẫu nhiên bộ dữ liệu **data** thành tập học *train* và tập thử *test*. Thực hiện phân tích hồi quy logistic của biến *Sex* theo một số biến (Hình 1). Sử dụng biến *Dir_pred* là biến đưa ra dự đoán trên tập thử *test* với ngưỡng 0.5 (Hình 2). Trả lời các câu hỏi sau.

(i) Biểu diễn biến *Sex* theo mô hình hồi quy logistic.

(ii) Nhận xét về các biến tham gia mô hình. Mô hình có cần cải tiến không?

(iii) Tìm độ nhạy, độ chính xác và độ đặc hiệu của mô hình.

Hình 1. *Phân tích hồi quy logistic*

```
Call:
glm(formula = Sex ~ ., family = binomial, data = data[, 1:8],
    subset = train)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -48.49644   14.36035  -3.377 0.000733 ***
RCC           0.58535    2.29334   0.255 0.798539
WCC           0.37147    0.32534   1.142 0.253542
HC            0.11122    0.43066   0.258 0.796214
Hg            1.91112    1.26475   1.511 0.130771
Ferr          0.02109    0.01346   1.568 0.116952
BMI           0.75371    0.25505   2.955 0.003126 **
SSF          -0.13232    0.03352  -3.947 7.91e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 175.82  on 139  degrees of freedom
Residual deviance:  38.97  on 132  degrees of freedom
AIC: 54.97

Number of Fisher Scoring iterations: 8
```

Hình 2

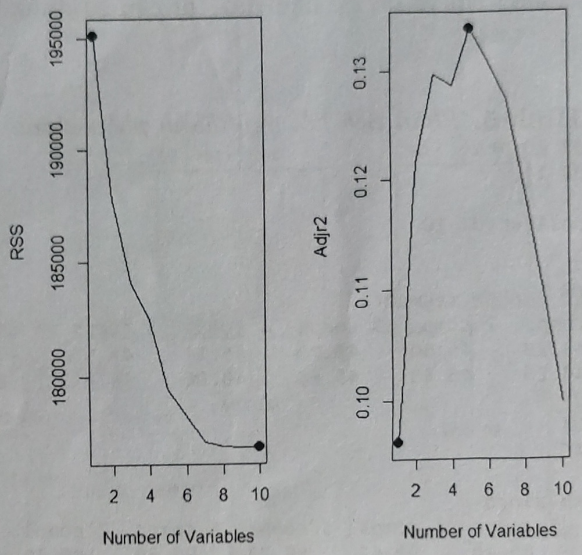
```
> table(test$Dir_pred , test$Sex)

        female male
  female      3    1
  male        2   56
```

(B) Chia ngẫu nhiên bộ dữ liệu **dat** thành tập học *train* và tập thử *test*. Gọi **regfit.best** là mô hình hồi quy tuyến tính biểu diễn *Ferr* theo các biến còn lại có tập con phù hợp nhất thu được từ tập học *train*. Hình 3 mô tả giá trị RSS và R^2-hiệu chỉnh (Adjr2) của mô hình **regfit.best**. Hình 4 là kết quả về các hệ số trong một số mô hình hồi quy của **regfit.best**. Trả lời các câu hỏi sau.

(iv) Theo tiêu chuẩn RSS, mô hình nào là mô hình phù hợp nhất? Viết phương trình hồi quy tuyến tính tương ứng.

(v) Theo tiêu chuẩn R^2-hiệu chỉnh, mô hình nào là mô hình phù hợp nhất? Viết phương trình hồi quy tuyến tính tương ứng.

Hình 3. *Mối liên hệ giữa các tiêu chuẩn và số biến trong mô hình hồi quy tuyến tính*



Hình 4

```
> coef(regfit.best, 1)
(Intercept)          Hg
  -93.46936    11.56560
> coef(regfit.best, 3)
(Intercept)         WCC         BMI       XBfat
 -29.129013    3.440760    4.974954   -2.421667
> coef(regfit.best, 5)
(Intercept)         WCC         BMI         SSF       XBfat         LBM
 -6.3709622   3.1313108   7.1931783   0.7483204  -7.0690242  -0.9323607
> coef(regfit.best, 8)
(Intercept)         RCC         WCC          Hc          Hg         BMI         SSF
-42.4243162  -7.1686044   2.9660309  -2.1755916  12.0138161   6.0405241   0.7069856
      XBfat         LBM
 -6.1893204  -0.8257502
> coef(regfit.best, 10)
(Intercept)         RCC         WCC          Hc          Hg         BMI         SSF
 11.2692853  -6.9195118   2.9549913  -2.1422255  11.7927861   4.9761185   0.6738485
      XBfat         LBM          Ht          Wt
 -6.8946262  -1.9631634  -0.2304010   1.3129806
```

(C) Trên bộ dữ liệu **dat**, thực hiện phân tích hồi quy thành phần chính (sử dụng kiểm chứng chéo) biểu diễn *Ferr* theo các biến còn lại (Hình 5). Trả lời các câu hỏi sau.

(vi) Mô hình hồi quy thành phần chính phù hợp nhất có bao nhiêu biến (thành phần chính) và MSE tương ứng bằng bao nhiêu?

(vii) Để thu được 99% thông tin về bộ dữ liệu **dat** thì cần sử dụng bao nhiêu thành phần chính?

Hình 5. *Phân tích hồi quy thành phần chính*

```
Data:   X dimension: 202 10
        Y dimension: 202 1
Fit method: svdpc
Number of components considered: 10

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
CV           47.62    45.29    45.50    45.73    46.14    44.93    45.08    45.06    44.92
adjCV        47.62    45.26    45.45    45.66    46.06    44.85    44.99    44.95    44.81
       9 comps  10 comps
CV       45.28     45.54
adjCV    45.15     45.40

TRAINING: % variance explained
      1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps  10 comps
X       48.56    74.11    85.46    93.41    98.23    99.29    99.70    99.94    99.99    100.00
Ferr    11.06    11.26    11.80    11.80    16.14    16.16    17.59    18.26    18.26     18.26
```

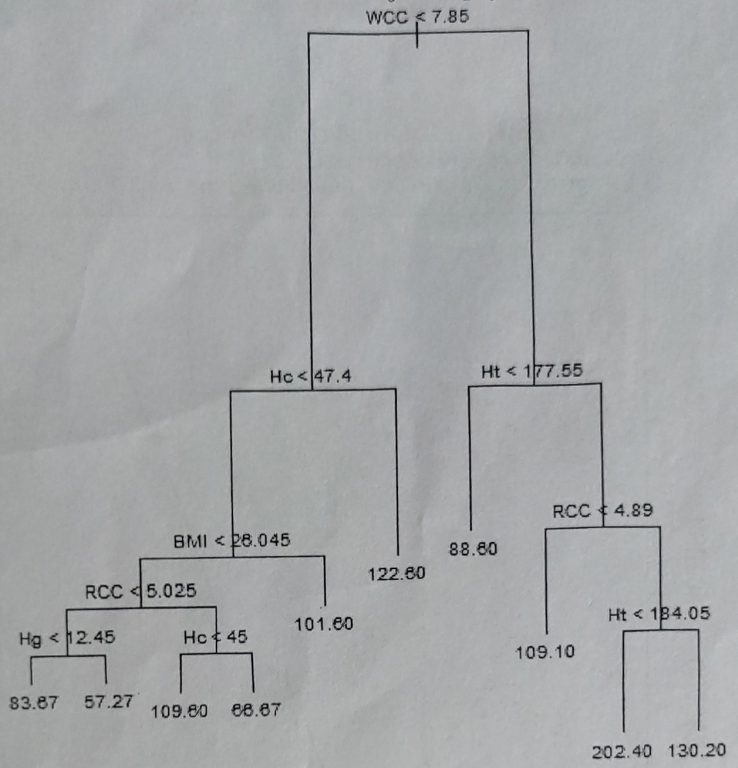
(D) Sử dụng bộ dữ liệu **dat**, gọi **tree.dat** là việc thực hiện hồi quy cây dự đoán **Ferr** theo các biến còn lại (Hình 6).

(viii) Cây thu được không xuất hiện những biến nào?

(ix) Những vận động viên có nồng độ **Ferr** cao nhất thường có các thông số sức khỏe khác như thế nào?

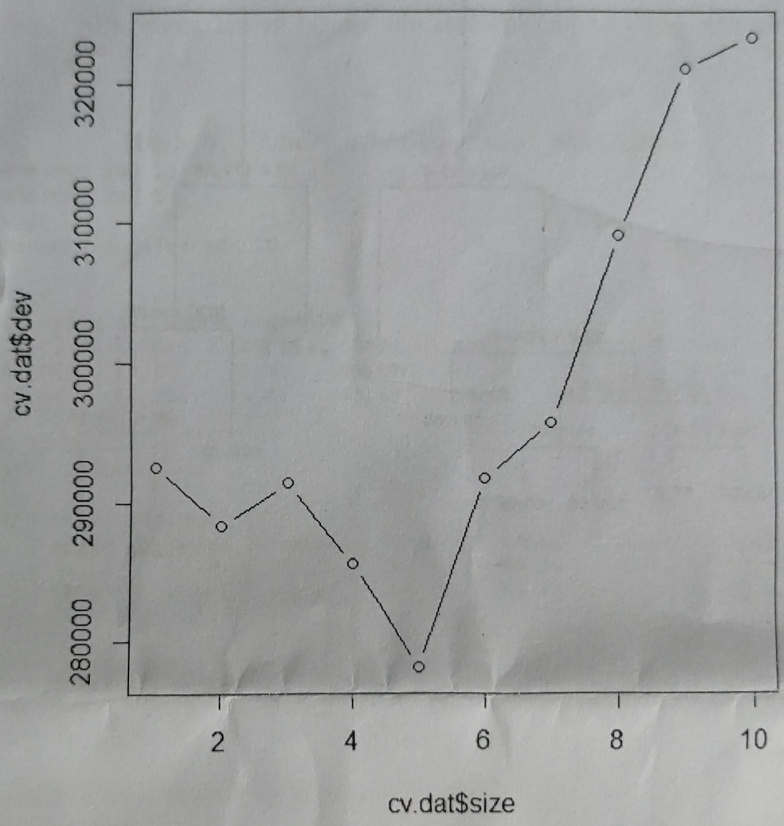
(x) Hình 7 là kết quả chạy hàm **cv.tree** để cắt tỉa cây trong Hình 6 theo tiêu chuẩn về tỷ lệ sai số xác thực chéo. Việc cắt tỉa cây có cần thiết không? Để thu được cây tối ưu, cây ban đầu cần cắt bao nhiêu lá?

Hình 6. *Cây hồi quy*



Hình 7. *Điều kiện cắt tỉa cây*

```
> cv.dat = cv.tree(tree.dat)
> plot(cv.dat$size, cv.dat$dev, type = "b")
```



Hết

Ghi chú: *Sinh viên không được dùng tài liệu.*