BÀI TẬP PHÂN TÍCH HỒI QUY VÀ ỨNG DỤNG

TUẦN 7: ÔN TẬP HỒI QUY LOGISTIC

Cho bộ dữ liệu **Australian Institute of Sport** 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)

1. Nhập bộ dữ liệu vào R và đặt tên **data**.
2. Bộ dữ liệu có bao nhiêu quan sát chứa dữ liệu trống? Loại bỏ các quan sát đó.
3. Dùng hàm *set.seed(1)* và *sample* để sinh ngẫu nhiên vectơ **sample** gồm 202 phần tử, trong đó, 70% phần tử có giá trị *TRUE*, còn lại mang giá trị *FALSE*.
4. Chia bộ dữ liệu **data** thành tập học *train* và tập thử *test*. Tập học gồm các quan sát ở vị trí có giá trị *TRUE* trong vectơ **sample**. Tập thử gồm các quan sát còn lại.
5. Xây dựng mô hình hồi quy logistic của biến *Sex* theo các biến *RCC*, *Ferr*, *BMI* và *SSF*. Biểu diễn biến *Sex* theo mô hình được xây dựng.
6. Nhận xét về các biến tham gia mô hình. Mô hình có cần cải tiến không?
7. Đưa ra dự đoán về giới tính trên tập thử *test* với ngưỡng 0.5. Tìm độ chính xác của mô hình.
8. Đưa ra dự đoán về giới tính của một vận động viên có các chỉ số sức khỏe như sau: $RCC = 4.2$, $Ferr = 68$, $BMI = 24$ và $SSF = 114$.

9. Vẽ biểu đồ nomogram để đưa ra dự đoán về xác suất một vận động viên là nam. Giải thích cách sử dụng biểu đồ. Đưa ra dự đoán về xác suất vận động viên có các chỉ số trong ý (8) là nam. So sánh với kết quả sử dụng công thức tính toán trong mô hình được xây dựng ở ý (5).

10. Vẽ đồ thị đường cong ROC, đưa ra giá trị ngưỡng và diện tích dưới đường cong ROC (AUC) trong đồ thị. Giá trị của AUC thể hiện điều gì? Tính độ nhạy, độ đặc hiệu của mô hình.