**Bài tập thực hành số 3**

**Hồi quy Logistic dự đoán nguy cơ Trầm cảm ở thanh niên**

**Mục tiêu:**

* Xây dựng mô hình hồi quy Logistic đơn biến và đa biến.
* Hiểu và diễn giải các khái niệm liên quan đến hồi quy Logistic (Odds Ratio, Log-odds).
* Kiểm tra các giả định liên quan đến mô hình hồi quy Logistic (tính tuyến tính của log-odds, không có đa cộng tuyến).
* Kiểm tra và xử lý vấn đề đa cộng tuyến.
* Diễn giải các hệ số, p-value và khoảng tin cậy trong mô hình Logistic.

**1. Bộ số liệu mẫu**

Bộ dữ liệu có 350 đối tượng ở độ tuổi 18-25 với 10 biến độc lập và 1 biến phụ thuộc nhị phân.

**Biến phụ thuộc:**

* TramCam: Tình trạng trầm cảm (0 = Không trầm cảm, 1 = Có trầm cảm )

**Biến độc lập:**

1. Tuoi : Tuổi (năm, liên tục)
2. GioHoc : Số giờ học/làm việc/tuần (liên tục)
3. GioNgu : Số giờ ngủ trung bình/ngày (liên tục)
4. MoiQuanHeXaHoi : Điểm số mối quan hệ xã hội (thang điểm 1-10, liên tục)
5. TienSuTramCamGiaDinh : Tiền sử trầm cảm trong gia đình (0 = Không, 1 = Có - biến nhị phân)
6. ApLucHocTap : Mức độ áp lực học tập (thang điểm 1-10, liên tục)
7. HoatDongTheChat :Mức độ hoạt động thể chất (0 = Ít, 1 = Trung bình, 2 = Nhiều - biến thứ bậc)
8. SuDungThucPham : Mức độ tiêu thụ thực phẩm không lành mạnh (0 = Ít, 1 = Trung bình, 2 = Nhiều - biến thứ bậc)
9. ThuNhapGiaDinh : Thu nhập gia đình (triệu VND/tháng, liên tục)
10. SuDungMangXaHoi : Số giờ sử dụng mạng xã hội/ngày (liên tục)

**Nhiệm vụ**

1. Xây dựng mô hình hồi quy Logistic để dự đoán TramCam chỉ dựa vào biến ApLucHocTap.

* Hệ số hồi quy (Estimate) của ApLucHocTap là bao nhiêu? Nó có ý nghĩa thống kê không (dựa vào p-value)?
* Làm thế nào để chuyển hệ số hồi quy này thành Tỷ số Chênh (Odds Ratio - OR)? Diễn giải ý nghĩa của OR. (Gợi ý: exp(estimate)).
* Giá trị AIC (Akaike Information Criterion) của mô hình là bao nhiêu? (AIC thấp hơn thường được ưu tiên).

1. Xây dựng mô hình hồi quy Logistic để dự đoán TramCam dựa trên *tất cả* các biến độc lập đã cho

* Những biến độc lập nào có ý nghĩa thống kê trong mô hình này (p-value < 0.05)?
* So sánh AIC của mô hình đơn biến và đa biến. Mô hình nào "tốt hơn" dựa trên tiêu chí này?
* Giải thích ý nghĩa của Odds Ratio cho biến TienSuTramCamGiaDinh và GioNgu (nếu có ý nghĩa thống kê).

1. Kiểm tra xem mối quan hệ giữa các biến độc lập liên tục và log-odds của biến phụ thuộc có tuyến tính hay không

* Dựa trên kiểm tra bậc hai và/hoặc biểu đồ LOESS, mối quan hệ giữa biến độc lập liên tục (ví dụ ApLucHocTap) và log-odds có vẻ tuyến tính không?
* Nếu không tuyến tính, bạn sẽ đề xuất cách xử lý nào (ví dụ: biến đổi biến, thêm thuật ngữ bậc hai/bậc cao hơn)?

1. Kiểm tra các điểm dữ liệu có ảnh hưởng lớn đến mô hình

* Có điểm dữ liệu nào có Cook's Distance cao bất thường không?
* Nếu có, bạn sẽ làm gì với những điểm này trong một phân tích thực tế?

1. Kiểm tra mức độ đa cộng tuyến giữa các biến độc lập và đưa ra phương án xử lý nếu cần

* Giá trị VIF của từng biến là bao nhiêu?
* Có biến nào có VIF cao (thường > 5 hoặc > 10) không? Nếu có, đó là biến nào?
* Nếu có biến có VIF cao, bạn đề xuất cách xử lý nào? (Ví dụ: loại bỏ một biến có VIF cao nhất, kết hợp các biến, v.v.).
* Hãy thử loại bỏ biến có VIF cao nhất (nếu có) và chạy lại mô hình. VIF của các biến còn lại có thay đổi không?

1. Kết quả của mô hình Logistic đa biến cuối cùng (sau khi đã xử lý đa cộng tuyến và các giả định khác nếu cần).