Bài tập 4 : Chọn mô hình

A. Chọn hai "đề tài"tự do để tìm hiểu "nghiên cứu", vận dụng những kiến thức học được từ môn này.

1. Dữ liệu 1: mô hình hồi quy đa biến.

2. Dữ liệu 2: hồi quy cho các thành phần chính.

- Tên "đề tài"
- Nguồn gốc của dữ liệu, giới thiệu các biến, ...
- Mô hình chọn được
- Phân tích kết quả
- Đưa ra những phương pháp/phân tích khác có thể giúp cho kết quả tốt hơn nếu có thể.
- Kết luận.

B. Chọn mô hình với dữ liệu cho trước

1. Chọn mô hình phù hợp nhất giải thích biến phụ thuộc với từng bộ dữ liệu sau. Phương pháp chọn và tiêu chuẩn chọn mô hình cho mỗi bộ dữ liệu là không trùng nhau.

2. Nêu rõ phương pháp chọn mô hình và lý do chọn phương pháp đó.

3. Nói rõ ý nghĩa của mô hình đã chọn.

- **data1** : Những thông tin về các giám đốc điều hành các tập đoàn Hoa Kỳ. Chúng ta quan tâm đến biến phụ thuộc "salary"(hàng năm và tính theo đơn vị nghìn $) hoặc/và "lsalary". Các biến được mô tả như sau :

1. salary : 1990 compensation, $1000s$
2. age : in years
3. college : =1 if attended college
4. grad : =1 if attended graduate school
5. comten : years with company
6. ceoten : years as ceo with company
7. sales : 1990 firm sales, millions
8. profits : 1990 profits, millions
9. mktval : market value, end 1990, mills

10. lsalary : log(salary)

11. lsales : log(sales)

12. lmktval : log(mktval)

13. comtensq : $comten^2$

14. ceotensq : $ceoten^2$

15. profmarg : profits as % of sales.

- **data 2** : Chúng ta quan tâm đến bộ dữ liệu ghi lại lịch sử về những ngôi nhà được bán từ 5/2014 đến 5/2015 ở quận King thuộc bang Washington, Hoa Kỳ. Bộ dữ liệu bao gồm 21613 quan trắc và gồm 21 biến sau:

1. id: id của một ngôi nhà;

2. date: ngày nhà được bán;

3. Price: giá nhà ;

4. bebrooms: số phòng ngủ;

5. bathrooms: số phòng tắm;

6. sqftliving: diện tích của ngôi nhà;

7. sqrt-lot: diện tích của lô đất;

8. floor: tổng số tầng trong nhà;

9. waterfront: hương nhà ra bờ sông;

10. view: số lượt xem mà công ty bất động sản thu được;

11. condition: đánh giá tình trạng của nhà (1 cho biết bất động sản cũ và 5 là tuyệt vời);

12. grade: điểm tổng thể cho nhà ở (theo phân loại của Quận King :1: kém, 13 xuất sắc);

13. sqftabove: diện tích của ngôi nhà ngoài tầng hầm;

14. sqftbasement: diện tích của tầng hầm;

15. yrbuilt: năm xây dựng;

16. yrrenovated: năm ngôi nhà được cải tạo;

17. zipcode: mã vùng;

18. lat: tọa độ vĩ đạo;

19. long: tọa độ kinh độ;

20. sqftliving15: diện tích ngôi nhà năm 2015;

21. sqftlot15: diện tích lô đất khu vực vào năm 2015.

- **data 3** : Chúng ta quan tâm đến tỷ lệ tai nạn với bộ dữ liệu gồm 39 quan trắc được thực hiện trên vài đoạn đường cao tốc ở tiểu bang Minnesota vùng Trung Tây của Hoa Kỳ, gồm các biến sau :

– X1 : chiều dài đoạn đường (dặm) ;

– X2 : lượng giao thông trung bình hàng ngày (nghìn xe) ;

– X3 : tỷ lệ % xe tải trên tổng số ;

– X4 :tốc đệ giới hạn cho phép (dặm/giờ) ;

– X5 : chiều rộng làn đường (bước chân) ;

– X6 : chiều rộng làn đường khẩn cấp (bước chân) ;

– X7 : số làn đường thay đổi tự do trên đoạn đường cao tốc ;

– X8 : số làn đường thay đổi (báo hiệu) trên đoạn đường cao tốc ;

– X9 : số cửa vào đoạn đường cao tốc ;

– X10 : tổng số làn đường (trên hai chiều của đường cao tốc) ;

– X11 : 1 nếu là tuyến đường liên thông xa lộ và cao tốc , 0 nếu ngược lại ;

– X12 : 1 nếu là tuyến đường lớn của cao tốc , 0 nếu ngược lại ;

– X13 : 1 nếu là tuyến đường cao tốc chính, 0 nếu ngược lại.

- **data 4** : Chúng ta muốn tìm hiểu những yếu tố ảnh hưởng đến mức lương ($ giờ) của người lao động ở Anh năm 1976. Các biến được mô tả như sau:

1. wage average hourly earnings

2. educ years of education

3. exper years potential experience

4. tenure years with current employer

5. nonwhite =1 if nonwhite

6. female =1 if female

7. married =1 if married

8. numdep number of dependents

9. smsa =1 if live in SMSA

tenure years with current employer

10. northcen =1 if live in north central U.S

11. south =1 if live in southern region

12. west =1 if live in western region

13. construc =1 if work in construc. indus.

14. ndurman =1 if in nondur. manuf. indus.

15. trcommpu =1 if in trans, commun, pub ut

16. trade =1 if in wholesale or retail

17. services =1 if in services indus.

18. profserv =1 if in prof. serv. indus.

19. profocc =1 if in profess. occupation

20. clerocc =1 if in clerical occupation

21. servocc =1 if in service occupation

22. lwage log(wage)

23. expersq exper^2

24. tenursq tenure^2