Môn học: Xác suất thống kê
Bài tập lớn số 2 (1 TC)
Học kì: 201

Bộ môn Toán Ứng Dụng
Khoa Khoa Học Ứng Dụng, trường Đại Học Bách Khoa TPHCM

Ngày 11 tháng 10 năm 2020

Yêu cầu:

- Tất cả sinh viên **không thuộc** 2 khoa sau: Khoa Điện – Điện tử, Khoa Kỹ Thuật Địa Chất và Dầu Khí, đều **bắt buộc** làm bài tập lớn số 2.
- Trong bài tập lớn số 2, sinh viên được yêu cầu dùng phần mền `Rstudio` để xử lý các số liệu thống kê.
- Tất cả các nhóm đều bắt buộc làm cả 2 phần: phần chung và phần riêng. Trong phần chung giáo viên sẽ chọn ngẫu nhiên 1 bài để phân cho mõi nhóm.

1 Phần chung: (4 điểm)

Phần chung bao gồm 2 chủ đề: Hồi quy tuyến tính bội và Anova. Sinh viên được yêu cầu làm 2 nhiệm vụ:

- Trình bày lý thuyết của chủ đề mình được phân công.
- Làm bài tập xử lý số liệu được phân công.

A. Hồi quy tuyến tính bội

Bài tập 1. Tập tin "**gia_nha.csv**" chứa thông tin về giá bán ra thị trường (đơn vị đô la) của 21613 ngôi nhà ở quân King nước Mỹ trong khoảng thời gian từ tháng 5/2014 đến 5/2015. Bên cạnh giá nhà, dữ liệu còn bao gồm các thuộc tính mô tả chất lượng ngôi nhà. Dữ liệu gốc được cung cấp tại: `https://www.kaggle.com/harlfoxem/housesalesprediction`.

Các biến chính trong bộ dữ liệu:

- **price**: Giá nhà được bán ra.
- **sqft_living15**: Diện tích trung bình của 15 ngôi nhà gần nhất trong khu dân cư.
- **floors**: Số tầng của ngôi nhà được phân loại từ 1-3.5.
- **condition**: Điều kiện kiến trúc của ngôi nhà từ $1-5$, 1: rất tệ và 5: rất tốt.
- **sqft_above**: Diện tích ngôi nhà.
- **sqft_living**: Diện tích khuôn viên nhà.

Câu hỏi:

1. **Đọc dữ liệu (Import data):**
   Hãy dùng lênh `read.csv()` để đọc tệp tin.
2. **Làm sạch dữ liệu (Data cleaning):**
   (a) Hãy trích ra một dữ liệu con đặt tên là **new_DF** chỉ bao gồm các biến chính mà ta quan tâm như đã trình bày trong phần giới thiệu dữ liệu. Từ câu hỏi này về sau, mọi yêu cầu xử lý đều dựa trên tập dữ liệu con **new_DF** này.
   (b) Kiểm tra các dữ liệu bị khuyết trong tập tin. (*Các câu lênh tham khảo:* `is.na(), which(), apply()`). Nếu có dữ liệu bị khuyết, hãy đề xuất phương pháp thay thế cho những dữ liệu bị khuyết này.
3. **Làm rõ dữ liệu (Data visualization):**
   (a) Chuyển đổi các biến **price**, **sqft_living15**, **sqft_above**, **sqft_living** lần lượt thành $\log(price)$, $\log(sqft\_living15)$, $\log(sqft\_above)$, và $\log(sqft\_living)$. Từ đây mọi sự tính toán với các biến trên được hiểu là đã qua đổi biến dạng log.
   (b) Đối với các biến liên tục, hãy tính các giá trị thống kê mô tả bao gồm: trung bình, trung vị, độ lệch chuẩn, giá trị lớn nhất và giá trị nhỏ nhất. Xuất kết quả dưới dạng bảng. (Hàm gợi ý: `mean(), median(), sd(), min(), max() , apply(), as.data.frame(), rownames()`)
   (c) Đối với các biến phân loại, hãy lập một bảng thống kê số lượng cho từng chủng loại (*Hàm gợi ý:* `table()`).
   (d) Hãy dùng hàm `hist()` để vẽ đồ thị phân phối của biến **price**.
   (e) Hãy dùng hàm `boxplot()` vẽ phân phối của biến **price** cho từng nhóm phân loại của biến **floors** và biến **condition**.
   (f) Dùng lệnh `pairs()` vẽ các phân phối của biến **price** lần lượt theo các biến **sqft_living15**, **sqft_above**, và **sqft_living**
4. **Xây dựng các mô hình hồi quy tuyến tính (Fitting linear regression models):**
   Chúng ta muốn khám phá rằng có những nhân tố nào và tác động như thế nào đến giá nhà ở quận King.

(a) Xét mô hình hồi quy tuyến tính bao gồm biến **price** là một biến phụ thuộc, và tất cả các biến còn lại đều là biến độc lập. Hãy dùng lệnh `lm()` để thực thi mô hình hồi quy tuyến tính bội.

(b) Dựa vào kết quả của mô hình hồi quy tuyến tính trên, những biến nào bạn sẽ loại khỏi mô hình tương ứng với mức tin cậy 5%?

(c) Xét 2 mô hình tuyến tính cùng bao gồm biến **price** là biến phụ thuộc nhưng:

- mô hình M1 chứa tất cả các biến còn lại là biến độc lập
- mô hình M2 là loại bỏ biến **condition** từ mô hình M1.

Hãy dùng lệnhh `anova()` để đề xuất mô hình hồi quy hợp lý hơn.

(d) Chọn mô hình hợp lý hơn từ câu (c) hãy suy luận sự tác động của các biến lên giá nhà.

(e) Từ mô hình hồi quy mà bạn chọn ở câu (c) hãy dùng lệnh `plot()` để vẽ đồ thị biểu thị sai số hồi quy (residuals) và giá trị dự báo (fitted values). Nêu ý nghĩa và nhận xét đồ thị.

5. **Dự báo (Predictions:)**

(a) Từ mô hình bạn chọn trong câu (c), hãy dùng lệnh `predict()` để dự báo giá nhà tại 2 thuộc tính như sau:
x1: sqft_living15 = mean(sqft_living15), sqft_above = mean(sqft_above), sqft_living = mean(sqft_living), floor = 2, condition = 3
x2: sqft_living15 = max(sqft_living15), sqft_above = max(sqft_above), sqft_living = max(sqft_living), floor = 2, condition = 3.
So sánh khoảng tin cậy cho 2 giá trị dự báo này.

Bài tập 2. Tập tin "**diem_so.csv**" chứa thông tin về điểm toán của các em học sinh trung học thuộc hai trường học ở Bồ Đào Nha. Các thuộc tính dữ liệu bao gồm điểm học sinh, nơi cư trú, và một số hoạt động xã hội khác. Dữ liệu được thu thập bằng cách sử dụng báo cáo của các trường và các kết quả khảo sát sinh viên. Dữ liệu gốc được cung cấp tại: `https://archive.ics.uci.edu/ml/datasets/student+performance`.

Các biến chính trong bộ dữ liệu:

- **G1:** Điểm thi học kì 1.
- **G2:** Điểm thi học kì 2.
- **G3:** Điểm cuối khoá.
- **studytime:** Thời gian tự học trên tuần (1 - ít hơn 2 giờ, 2 - từ 2 đê 5 giờ, 3 - từ 5-10 giờ, or 4 - lớn hơn 10 giờ).
- **failures:** số lần không qua môn (1,2,3, hoặc 4 chỉ nhiều hơn hoặc bằng 4 lần).
- **absences:** số lần nghỉ học.

- **higher:** Có muốn học cao hơn hay không (yes: có, no: không).
- **age:** Tuổi của học sinh.

Câu hỏi:

1. **Đọc dữ liệu:**
   Hãy dùng lênh `read.csv()` để đọc tệp tin.
2. **Làm sạch dữ liệu (Data cleaning):**
   (a) Hãy trích ra một dữ liệu con đặt tên là **new_DF** chỉ bao gồm các biến chính mà ta quan tâm như đã trình bày trong phần giới thiệu dữ liệu. Từ câu hỏi này về sau, mọi yêu cầu xử lý đều dựa trên tập dữ liệu con **new_DF** này.
   (b) Kiểm tra các dữ liệu bị khuyết trong tập tin. *(Các câu lênh tham khảo:* `is.na()`, `which()`, `apply()`*)*. Nếu có dữ liệu bị khuyết, hãy đề xuất phương pháp thay thế cho những dữ liệu bị khuyết này.
3. **Làm rõ dữ liệu (Data visualization):**
   (a) Đối với các biến liên tục, hãy tính các giá trị thống kê mô tả bao gồm: trung bình, trung vị, độ lệch chuẩn, giá trị lớn nhất và giá trị nhỏ nhất. Xuất kết quả dưới dạng bảng. (Hàm gợi ý: `mean()`, `median()`, `sd()`, `min()`, `max()` , `apply()`, `as.data.frame()`, `rownames()`)
   (b) Đối với các biến phân loại, hãy lập một bảng thống kê số lượng cho từng chủng loại.
   (c) Hãy dùng hàm `hist()` để vẽ đồ thị phân phối của biến **G3**.
   (d) Hãy dùng hàm `boxplot()` vẽ phân phối của biến **G3** cho từng nhóm phân loại của biến **studytime**, **failures**, và biến **higher**.
   (e) Dùng lệnh `pairs()` vẽ các phân phối của biến **G3** lần lượt theo các biến **G2**, **G1**, **age**, và **absences**.
4. **Xây dựng các mô hình hồi quy tuyến tính (Fitting linear regression models):**
   Chúng ta muốn khám phá rằng có những nhân tố nào và tác động như thế nào đến điểm cuối khoá môn Toán của các em học sinh.
   (a) Xét mô hình hồi quy tuyến tính bao gồm biến **G3** là một biến phụ thuộc, và tất cả các biến còn lại đều là biến độc lập. Hãy dùng lệnh `lm()` để thực thi mô hình hồi quy tuyến tính bội.
   (b) Dựa vào kết quả của mô hình hồi quy tuyến tính trên, những biến nào bạn sẽ loại khỏi mô hình tương ứng với các mức tin cậy 5% và 1%?
   (c) Xét 3 mô hình tuyến tính cùng bao gồm biến **G3** là biến phụ thuộc nhưng:
   - Mô hình M1 chứa tất cả các biến còn lại là biến độc lập
   - Mô hình M2 là loại bỏ biến **higher** từ M1,

- Mô hình M3 là loại bỏ biến **failure** từ M2.

Hãy dùng lệnhh `anova()` để đề xuất mô hình hồi quy hợp lý hơn.

(d) Từ mô hình hồi quy hợp lý nhất từ câu (c) hãy suy luận sự tác động của các biến điểm thi cuối kì.

(e) Từ mô hình hồi quy hợp lý nhất từ câu (c) hãy dùng lệnh `plot()` để vẽ đồ thị biểu thị sai số hồi quy và giá trị dự báo. Nêu ý nghĩa và nhận xét.

5. **Dự báo (Predictions:)**

(a) Trong dữ liệu của bạn, hãy tạo thêm biến đặt tên là **evaluate**, biến này biểu diễn tỷ lệ đạt ($G3 >= 10$) hoặc không đạt ($G3 < 10$) của sinh viên trong điểm thi cuối kì. Hãy thống kê tỷ lệ đạt/không đạt *(Hàm gợi ý:* `cbind()` *).*

(b) Xét mô hình hồi quy hợp lý nhất mà bạn đã chọn trong câu 4(c). Hãy lập một bảng số liệu mới đặt tên là **new_X** bao gồm toàn bộ các biến độc lập trong mô hình này, và dùng lênh `predict()` để đưa ra số liệu dự báo cho biến **G3** phụ thuộc vào **new_X**. Gọi kết quả dự báo này là biến **pred_G3**.

(c) Khảo sát độ chính xác trong kết quả dự báo của câu trên bằng cách lập một bảng so sánh kết quả dự báo **pred_G3** với kết quả thực tế của biến **G3**.

	Đạt	Không đạt
Quan sát		
Dự báo		

B. Anova

Bài tập 3. Tập tin **Diet.csv** (cung cấp bởi Đại học Sheffield, Anh) chứa thông tin về một thử nghiệm về hiệu quả của các chế độ ăn kiêng trong việc giảm cân nặng đối với những người trưởng thành. Một người tham gia sẽ được áp dụng một trong ba chế độ ăn kiêng khác nhau trong vòng 6 tuần lễ. Cân nặng của người tham gia sẽ được ghi nhận trước và sau khi kết thúc thử nghiệm để đánh giá hiệu quả của từng chế độ ăn kiêng. Chi tiết về bộ dữ liệu như sau:

- Tổng số người tham gia: 78.
- Tổng số biến 7.
- Mô tả các biến:
  1. $Person$ = số thứ tự của người tham gia thử nghiệm
  2. $gender$ = giới tính của người tham gia (1 = nam, 0 = nữ)
  3. Age = tuổi (năm)
  4. $Height$ = chiều cao (cm)
  5. $pre.weight$ = cân nặng trước khi áp dụng chế độ ăn kiêng (kg)
  6. $Diet$ = chế độ ăn kiêng (3 chế độ khác nhau)
  7. $weight6weeks$ = cân nặng sau 6 tuần ăn kiêng

Câu hỏi:

1. **Đọc file dữ liệu, thực hiện thống kê mô tả và kiểm định**

   (a) Đọc dữ liệu vào **R** và tính toán các giá trị thống kê mô tả cho các biến *gender*, *Age*, *Height*, *pre.weight* và *weight6weeks* theo từng nhóm chế độ ăn kiêng tương ứng.

   (b) Biến *gender* có chứa hai giá trị khuyết (NA = Not Available) của người tham gia thứ 25 và 26. Hãy đề xuất một phương pháp để thay thế hai giá trị khuyết này.

   (c) Tạo biến $weight.loss = pre.weight - weight6weeks$. Hãy vẽ biểu đồ boxplot cho biến *weight.loss* tương ứng theo 3 chế độ ăn kiêng. Dựa trên các biểu đồ boxplot vừa vẽ, đưa ra nhận xét về 3 chế độ ăn kiêng.

   (d) Dựa trên hai biến *pre.weight* và *weight6weeks*, hãy thực hiện một kiểm định t theo cặp (paired t-test) để đánh giá xem liệu chế độ ăn kiêng (nói chung) có làm giảm cân nặng?

2. **Phân tích phương sai một nhân tố (one way ANOVA)**

   (a) Trình bày mô hình phân tích phương sai một nhân tố, phát biểu các giả thuyết và đối thuyết và nêu các giả định của mô hình cần kiểm tra.

   (b) Thực hiện kiểm tra các giả định của mô hình (giả định về phân phối chuẩn, tính đồng nhất của các phương sai). *Gợi ý: ta có thể sử dụng phân tích thặng dư kết hợp với việc sử dụng đồ thị QQ-plot, kiểm định Shapiro-Wilk để kiểm tra giả định về phân phối chuẩn, kiểm định Levene hay Bartlett để kiểm tra giả định về tính đồng nhất của các phương sai.*

   (c) Thực hiện phân tích ANOVA một nhân tố. Trình bày bảng phân tích phương sai trong báo cáo. Cho kết luận về hiệu quả của các phương pháp ăn kiêng đối với việc giảm cân.

   (d) Thực hiện các so sánh bội (multiple comparisons) sau phân tích phương sai. Phương pháp ăn kiêng nào có hiệu quả tốt nhất trong việc giảm cân?

3. **phân tích phương sai hai nhân tố (two way ANOVA)**

   (a) Thực hiện phân tích phương sai hai nhân tố để xét xem liệu **chế độ ăn kiêng** và **giới tính** ảnh hưởng như thế nào đến sự giảm cân?

   (b) Phân tích sự tương tác giữa **chế độ ăn kiêng** và **giới tính** đến sự giảm cân.

Bài tập 4. Tập tin **flights.rda** cung cấp thông tin về 162049 chuyến bay đã khởi hành từ hai sân bay lớn của vùng Tây bắc Thái Bình Dương của Mỹ, SEA ở Seattle và PDX ở Portland trong năm 2014. Dữ liệu cung cấp bởi Văn phòng Thống kê Vận tải, Mỹ (https://www.transtats.bts.gov/). Dữ liệu này được dùng để phân tích các nguyên nhân gây ra sự khởi hành trễ hoặc hoãn các chuyến bay. Chi tiết về bộ dữ liệu như sau:

- Tổng chuyến bay được thống kê: 162049.
- Tổng số biến 16.

- Mô tả các biến chính:

    1. *year*, *month*, *day*: ngày khởi hành của mỗi chuyến bay
    2. *carrier*: tên của hãng hàng không, được mã hóa bằng 2 chữ cái in hoa. Ví dụ: UA = United Air Lines, AA = American Airlines, DL = Delta Airlines, v.v.
    3. *origin* và *dest*: tên sân bay đi và đến. Đối với sân bay đi, ta chỉ có hai giá trị SEA (Seattle) và PDX (Portland)
    4. *dep_time* và *arr_time*: thời gian cất cánh và hạ cánh (theo lịch dự kiến)
    5. *dep_delay* và *arr_time*: chênh lệch (phút) giữa thời gian cất cánh/hạ cánh thực tế với thời gian cất cánh/hạ cánh in trong vé
    6. *distance*: khoảng cách giữa hai sân bay (dặm)

Câu hỏi:

1. **Nhập và làm sạch dữ liệu, thực hiện các thống kê mô tả**

    (a) Trong **R**, hãy sử dụng lệnh `read.table` để đọc dữ liệu từ tập tin **flights.rda**. Chú ý rằng hàng đầu tiên dùng để đặt tên biến và dấu ngăn cách giữa các cột là dấu "," thay vì khoảng trắng như mặc định.

    (b) Hãy tạo một data.frame mới, đặt tên là **newFlights**, chỉ chứa các biến chúng ta cần quan tâm là: *carrier*, *origin*, *dep_time*, *arr_time*, *dep_delay* và *arr_time*. Từ câu hỏi này về sau, mọi yêu cầu xử lý đều được thực hiện trên data.frame **newFlights** này.

    (c) Trong các biến đang xét, có một số biến chứa nhiều giá trị khuyết (NA - Not Available). Hãy in bảng thống kê tỷ lệ giá trị khuyết đối với từng biến. Hãy đề xuất một phương pháp để xử lý những giá trị khuyết này.

    (d) Tính các giá trị thống kê mô tả (cỡ mẫu, trung bình, độ lệch chuẩn, min, max, các điểm tứ phân vị) của thời gian khởi hành trễ (biến *dep_delay*) của từng hãng hàng không (*carrier*). Xuất kết quả ra dưới dạng bảng.

    (e) Vẽ đồ thị boxplot cho thời gian khởi hành trễ *dep_delay* tương ứng với từng hãng hàng không *carrier*.

    (f) Ta sẽ quan sát thấy rằng có rất nhiều điểm outliers trên các đồ thị boxplot vừa vẽ (đối với biến *dep_delay*). Hãy sử dụng khoảng tứ phân vị (interquartile range) để loại bỏ các điểm outlier này và vẽ lại các đồ thị boxplot cho *dep_delay*. Dựa trên đồ thị boxplot, cho nhận xét về thời gian khởi hành trễ của từng hãng hàng không.

2. **Phân tích phương sai một nhân tố (one way ANOVA)**

    Ta quan tâm đến việc kiểm định rằng liệu có sự khác biệt về thời gian khởi hành trễ trung bình giữa các hãng hàng không đối với các chuyến bay khởi hành từ Portland trong năm 2014 hay không?

(a) Hãy giải thích tại sao ta cần dùng phân tích phương sai để trả lời cho câu hỏi trên. Xác định biến phụ thuộc và các nhân tố (hay các biến độc lập).

(b) Phát biểu các giả thuyết và đối thuyết bằng lời và công thức toán. Nêu các giả định cần kiểm tra của mô hình.

(c) Thực hiện kiểm tra các giả định của mô hình (giả định về phân phối chuẩn, tính đồng nhất của các phương sai). *Gợi ý: ta có thể sử dụng phân tích thặng dư kết hợp với việc sử dụng đồ thị QQ-plot, kiểm định Shapiro-Wilk để kiểm tra giả định về phân phối chuẩn, kiểm định Levene hay Bartlett để kiểm tra giả định về tính đồng nhất của các phương sai.*

(d) Thực hiện phân tích ANOVA một nhân tố. Trình bày bảng phân tích phương sai trong báo cáo. Cho kết luận.

Bài tập 5. Chăn nuôi gà là một ngành công nghiệp trị giá nhiều tỷ đô ở Mỹ. Bất kỳ phương pháp nào có thể làm tăng tốc độ tăng trưởng của gà con đều giúp giảm chi phí trong chăn nuôi và làm tăng lợi nhuận của công ty, có thể trị giá đến hàng triệu đô-la. Một thí nghiệm đã được thực hiện để đo lường và so sánh hiệu quả của các loại thức ăn khác nhau đối với tốc độ tăng trưởng của gà con. Thí nghiệm được thực hiện như sau: người ta chia ngẫu nhiên những con gà con mới nở vào sáu nhóm và mỗi nhóm được cung cấp một loại thức ăn khác nhau. Sáu loại thức ăn được thử nghiệm là *casein*, đậu răng ngựa (horsebean), hạt lanh (linseed), thịt xay (meatmeal), đậu tương (soybean) và hoa hướng dương (sunflower). Kết quả của thí nghiệm được cung cấp trong tập tin **chicken_feed.csv**, gồm hai biến: *weight* là trọng lượng của gà con sau thời gian dài được ăn loại thức ăn thử nghiệm, $feed$ là biến nhân tố với các giá trị là tên 6 loại thức ăn được thử nghiệm.

Câu hỏi:

1. **Đọc file dữ liệu, thực hiện thống kê mô tả và kiểm định**

   (a) Đọc dữ liệu vào **R** và tính toán các giá trị thống kê mô tả cho biến *weight* theo từng loại thức ăn ($feed$) tương ứng.

   (b) Biến *weight* có chứa hai giá trị khuyết (NA = Not Available). Hãy đề xuất một phương pháp để thay thế hai giá trị khuyết này.

   (c) Vẽ biểu đồ boxplot cho trọng lượng của gà con *weight* theo từng loại thức ăn tương ứng. Dựa trên biểu đồ boxplot vừa vẽ, hãy cho nhận xét về ảnh hưởng của từng loại thức ăn lên sự tăng trưởng của gà con.

2. **Phân tích phương sai một nhân tố (one way ANOVA)**

   Ta quan tâm đến câu hỏi rằng liệu có sự khác biệt về ảnh hưởng của các loại thức ăn lên tốc độ tăng trưởng của gà con hay không?

   (a) Hãy giải thích tại sao ta cần dùng phân tích phương sai để trả lời cho câu hỏi trên. Xác định biến phụ thuộc và các nhân tố (hay các biến độc lập).

   (b) Phát biểu các giả thuyết và đối thuyết bằng lời và công thức toán. Nêu các giả định cần kiểm tra của mô hình.

(c) Thực hiện kiểm tra các giả định của mô hình (giả định về phân phối chuẩn, tính đồng nhất của các phương sai). *Gợi ý: ta có thể sử dụng phân tích thặng dư kết hợp với việc sử dụng đồ thị QQ-plot, kiểm định Shapiro-Wilk để kiểm tra giả định về phân phối chuẩn, kiểm định Levene hay Bartlett để kiểm tra giả định về tính đồng nhất của các phương sai.*

(d) Thực hiện phân tích ANOVA một nhân tố. Trình bày bảng phân tích phương sai trong báo cáo. Cho kết luận.

(e) Thực hiện các so sánh bội (multiple comparisons) sau phân tích phương sai. Loại thức ăn nào có tác động tốt nhất lên sự tăng trưởng của gà con?

2 Phần riêng (6 điểm)

- Mỗi nhóm **bắt buộc** tự tìm một bộ dữ liệu **thuộc về chuyên ngành** của mình. Khuyến khích sinh viên sử dụng dữ liệu thực tế sẵn có từ các thí nghiệm, khảo sát, dự án ... trong chuyên ngành của mình. Ngoài ra sinh viên có thể tự tìm kiếm dữ liệu từ những nguồn khác hoặc tham khảo trong kho dữ liệu cung cấp trong tập tin "kho_du_lieu_BTL_xstk.xlsx".

- Các nhóm được yêu cầu xử lý số liệu mà mình đã chọn. Sinh viên được tự do chọn phương pháp lý thuyết phù hợp để áp dụng phân tích dữ liệu của mình, nhưng phải đảm bảo 2 phần: Làm rõ dữ liệu (data visualization) và mô hình dữ liệu (model fitting).

Tài liệu

[1] Douglas C. Montgomery, George C. Runger. Hoboken. *Applied Statistics and Probability for Engineers.* NJ: Wiley, (2007).

[2] Peter Dalgaard *Introductory Statistics with R.* Springer, (2008).