

BUILDING A MODEL TO ANALYZE CARDIOVASCULAR DISEASE DATA USING R SOFTWARE

Do Thi Phuong Quynh

TNU – University of Medicine and Pharmacy

ARTICLE INFO	ABSTRACT
Received: 04/4/2022	Having a data analysis model has been a matter that many statisticians interest and specially desing to build an analytical model for a specific object. Therefore, through the use of R software (an open source software with many advanced features, updated daily, and commonly employed by many researchers in Vietnam as well as around the world), the author of this study wanted to build a data analysis model of cardiovascular disease basing on the cardiovascular disease data set used in teaching, learning and researching at the Institute for Advanced Study in Mathematics (VIASM). Approaching by statistical method, the author biult an Analytical Model consisting of 2 main steps: descriptive statistics and inferential statistics. After processing data with R software with clear and sharp models and especially using logistic regression, the author came to some conclusions for medical research such as: exercise angina wast the early symptoms of heart failure; The maximum heart rate achieved by the patients gradually decreased according to the degree of heart failure from I to IV; and 2 models using logistic regression correlation.
Revised: 31/5/2022	
Published: 31/5/2022	
KEYWORDS	
Analysis	
Heart-related diseases	
R software	
Descriptive statistics	
Inferential statistics	

XÂY DỰNG MÔ HÌNH PHÂN TÍCH DỮ LIỆU BỆNH TIM MẠCH BẰNG PHẦN MỀM R

Đỗ Thị Phương Quỳnh

Trường Đại học Y Dược – ĐH Thái Nguyên

THÔNG TIN BÀI BÁO	TÓM TẮT
Ngày nhận bài: 04/4/2022	Việc có một mô hình phân tích dữ liệu cũng đang là vấn đề được nhiều nhà thống kê học quan tâm và thiết kế đặc biệt xây dựng mô hình phân tích cho một đối tượng cụ thể. Chính vì vậy thông qua sử dụng phần mềm R (một phần mềm mã nguồn mở với nhiều tính năng ưu việt luôn được cập nhật mới hàng ngày đang là phần mềm được nhiều nhà nghiên cứu trên thế giới cũng như tại Việt Nam sử dụng) tác giả muốn xây dựng mô hình phân tích dữ liệu về bệnh tim mạch dựa trên bộ dữ liệu về bệnh tim mạch được sử dụng trong giảng dạy học tập và nghiên cứu tại Viện Nghiên Cứu Cao Cấp về Toán (VIASM). Tiếp cận bằng phương pháp thống kê học, tác giả sẽ xây dựng Mô hình phân tích gồm 2 bước chính là thống kê mô tả và thống kê suy diễn. Sau khi xử lý dữ liệu bằng phần mềm R với những mô hình rõ ràng sắc nét và đặc biệt sử dụng phần hồi quy logistic, tác giả đã đưa ra được một số kết luận phục vụ trong nghiên cứu y học như: Một số triệu chứng đau thắt ngực khi tập thể dục là những triệu chứng ban đầu cho suy tim; Nhịp tim tối đa của các bệnh nhân đạt được giảm dần theo mức độ suy tim từ I đến IV; và 2 mô hình sử dụng tương quan hồi quy logistic.
Ngày hoàn thiện: 31/5/2022	
Ngày đăng: 31/5/2022	
TỪ KHÓA	
Phân tích	
Bệnh tim mạch	
Phần mềm R	
Thống kê mô tả	
Thống kê suy diễn	

DOI: <https://doi.org/10.34238/tnu-jst.5809>

Email: dothiphuongquynh@tump.edu.vn

<http://jst.tnu.edu.vn>

534

Email: jst@tnu.edu.vn

1. Giới thiệu

Việc có một mô hình phân tích dữ liệu cũng đang là vấn đề được nhiều nhà thống kê học quan tâm và xây dựng, đặc biệt xây dựng mô hình phân tích cho một đối tượng cụ thể. Chính vì vậy thông qua sử dụng phần mềm R [1]-[4], (một phần mềm mã nguồn mở [5] với nhiều tính năng ưu việt luôn được cập nhật mới hàng ngày đang là phần mềm được nhiều nhà nghiên cứu trên thế giới cũng như tại Việt Nam sử dụng [6]) tác giả sẽ xây dựng mô hình phân tích dữ liệu về bệnh tim mạch dựa trên bộ dữ liệu về bệnh tim mạch được sử dụng trong giảng dạy học tập và nghiên cứu tại Viện Nghiên cứu cao cấp về toán (VIASM).

Thông thường thống kê học thường nghiên cứu vấn đề hay xử lý dữ liệu theo hai bước sau:

Bước 1: Thống kê mô tả.

Bước 2: Thống kê suy diễn [7], [8].

Trong bước thống kê mô tả, chúng ta thường xử lý dữ liệu từ dạng “thô”, từ bảng số liệu ban đầu điều tra được phải đưa ra một số thông tin tổng quát như trung bình, phương sai, một số biểu đồ thể hiện được mối quan hệ giữa các thuộc tính nghiên cứu. Sau đó sang bước thứ 2 dùng chính những thông tin có được ở bước 1 để đưa ra các nhận định dự đoán, đồng thời kiểm định lại xem những nhận định dự đoán đó có thể chấp nhận được ở mức ý nghĩa bao nhiêu phần trăm.

Việc xây dựng mô hình phân tích dữ liệu cho bệnh tim mạch là quá trình khai thác dữ liệu cụ thể [3], tập trung vào mô hình thống kê để khám phá thông tin hữu ích khi đưa ra kết luận trong việc hỗ trợ điều trị bệnh nhân mắc bệnh tim mạch.

2. Phương pháp nghiên cứu

Nghiên cứu định tính: Khai thác bằng phần mềm R từ bộ dữ liệu bệnh tim mạch đưa ra các suy đoán hỗ trợ.

Phương pháp thống kê phân tích dùng để đưa ra các mô hình thể hiện mối quan hệ thích hợp giữa các yếu tố tác động đến bệnh tim mạch, đưa ra các kết luận hỗ trợ điều trị trong điều trị bệnh tim mạch.

Phạm vi nghiên cứu: Từ bộ dữ liệu cụ thể xây dựng mô hình phân tích dữ liệu bệnh tim mạch, với mô hình này khi thay bộ dữ liệu khác (đầu vào khác) ta cũng được những kết luận (đầu ra) phù hợp hỗ trợ trong điều trị bệnh tim mạch.

3. Kết quả luận bàn

3.1. Thống kê mô tả

Dùng các câu lệnh chính sau để có cái nhìn tổng quan về bộ dữ liệu ban đầu. Nhập dữ liệu từ một file trên excel bằng lệnh:

```
> dl1 <- read.csv("D:/de tai nckh/2021/processed.cleveland.data")
```

Đặt tên cho các biến:

```
> names(dl1) <- c("age", "sex", "cp", "trestbps", "chol", "fbs", "restecg", "thalach", "exang", "oldpeak", "slope", "ca", "thal", "num")
```

Xem tổng quát bộ dữ liệu:

```
> view(dl1)
```

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
67	1	4	160	286	0	2	108	1	1.5	2	3.0	3.0	2
67	1	4	120	229	0	2	129	1	2.6	2	2.0	7.0	1
37	1	3	130	250	0	0	187	0	3.5	3	0.0	3.0	0
41	0	2	130	204	0	2	172	0	1.4	1	0.0	3.0	0
56	1	2	120	236	0	0	178	0	0.8	1	0.0	3.0	0
62	0	4	140	268	0	2	160	0	3.6	3	2.0	3.0	3
57	0	4	120	354	0	0	163	1	0.6	1	0.0	3.0	0

Hình 1. Tổng quan dữ liệu bệnh tim mạch

Bộ dữ liệu có 14 cột và 302 hàng, cột 1 (Age): tuổi; cột 2 (Sex): giới tính, 1 là nam 2 là nữ; cột 3 (cp): kiểu đau ngực, gồm: Giá trị 1: cơn đau thắt ngực ổn định, giá trị 2: đau thắt ngực không ổn định, giá trị 3: không đau thắt ngực, giá trị 4: không có triệu chứng; Cột 4 (trestbps): huyết áp

khí nhập viện (mm/Hg); Cột 5 (*chol*): cholestoral trong huyết thanh (đơn vị đo mg/dl); Cột 6 (*lbs*): (đường huyết lúc đói >120 mg/dl) (1: true; 0: false); Cột 7 (*restecg*): kết quả điện tâm đồ lúc nghỉ gồm: Giá trị 0: bình thường, giá trị 1: có bất thường sóng ST (đảo ngược sóng ST chênh lên hoặc xuống lớn hơn 0,05 mV), giá trị 2: hiện thị phì đại thất trái có thể xảy ra hoặc xác định theo tiêu chuẩn của Romhilt-Estes. Cột 8 (*thalach*): nhịp tim tối đa đạt được; Cột 9 (*exang*): đau thắt ngực do tập thể dục, giá trị 1 là có, giá trị 0 là không; Cột 10 (*oldpeak*): trầm cảm gây ra do tập thể dục liên quan đến nghỉ ngơi; Cột 11 (*Slope*): độ dốc của đoạn tập luyện đỉnh cao, giá trị 1: độ dốc cao; giá trị 2: bằng phẳng; giá trị 3: độ dốc thấp; Cột 12 (*ca*): Số lượng các mạch chính (0-3) khi sử dụng nội soi huỳnh quang; Cột 13 (*thal*): bệnh tan máu bẩm sinh, giá trị 3: trạng thái bình thường; giá trị 6: trạng thái khuyết tật cố định, giá trị 7: trạng thái khiếm khuyết có thể hồi phục; Cột 14 (*Num*): cột chỉ trường mục tiêu, theo mức độ rối loạn chức năng: Hội Tim mạch New York chia suy tim thành 4 mức độ I, II, III, IV dựa theo triệu chứng cơ năng, với các giá trị 1,2,3,4 là các mức độ suy tim, còn giá trị 0 là không bị tim. Trên đây là 14 thuộc tính cơ bản trong bộ 76 thuộc tính nghiên cứu, mục tiêu của bài báo là từ bộ dữ liệu này bằng các công cụ chuyên biệt của phần mềm R có thể phân tích và đưa ra được các mối quan hệ giữa các yếu tố để từ đó góp phần đưa ra các thông tin hữu ích cho các bệnh nhân tim mạch từ bộ dữ liệu đã khai thác được.

➤ Câu lệnh: `summary(dl1)`

```
> summary(dl1)
      age      sex      cp      trestbps      chol      fbs      restecg      thalach
Min.   :29.00 Min.   :0.0000 Min.   :1.000 Min.   : 94.0 Min.   :126.0 Min.   :0.0000 Min.   :0.0000 Min.   : 71.0
1st Qu.:48.00 1st Qu.:0.0000 1st Qu.:3.000 1st Qu.:120.0 1st Qu.:211.0 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:133.2
Median :55.50 Median :1.0000 Median :3.000 Median :130.0 Median :241.5 Median :0.0000 Median :0.5000 Median :153.0
Mean   :54.41 Mean   :0.6788 Mean   :3.166 Mean   :131.6 Mean   :246.7 Mean   :0.1457 Mean   :0.9868 Mean   :149.6
3rd Qu.:61.00 3rd Qu.:1.0000 3rd Qu.:4.000 3rd Qu.:140.0 3rd Qu.:275.0 3rd Qu.:0.0000 3rd Qu.:2.0000 3rd Qu.:166.0
Max.   :77.00 Max.   :1.0000 Max.   :4.000 Max.   :200.0 Max.   :564.0 Max.   :1.0000 Max.   :2.0000 Max.   :202.0

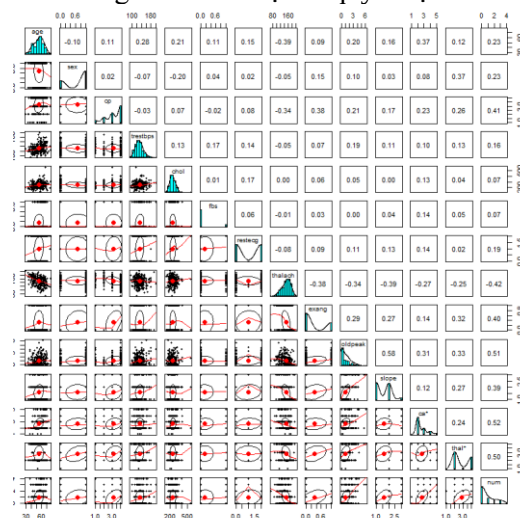
      exang      oldpeak      slope      ca      thal      num
Min.   :0.0000 Min.   :0.000 Min.   :1.000 Length:302 Length:302 Min.   :0.0000
1st Qu.:0.0000 1st Qu.:0.000 1st Qu.:1.000 Class :character Class :character 1st Qu.:0.0000
Median :0.0000 Median :0.800 Median :2.000 Mode :character Mode :character Median :0.0000
Mean   :0.3278 Mean   :1.035 Mean   :1.596                   Mean   :0.9404
3rd Qu.:1.0000 3rd Qu.:1.600 3rd Qu.:2.000                   3rd Qu.:2.0000
Max.   :1.0000 Max.   :6.200 Max.   :3.000                   Max.   :4.0000
```

Hình 2. Tổng quan thuộc tính và tham số của bộ dữ liệu về bệnh tim mạch.

Với câu lệnh này cho cái nhìn tương đối tổng quan về dữ liệu: Min, Max là giá trị nhỏ nhất và giá trị lớn nhất của thuộc tính, 1 st Qu: cho biết 25% đối tượng nghiên cứu có giá trị nhỏ hơn giá trị Min. 3rd Qu: cho biết 75% đối tượng nghiên cứu có giá trị thấp hơn giá trị Max [8].

➤ Với câu lệnh `pairs.panels(dl1)`

Lệnh này giúp chúng ta có cái nhìn tổng quan về dữ liệu cũng như mối liên quan giữa các biến (thuộc tính) với nhau để từ đó chúng ta đi đến một số quyết định cho bước suy diễn tiếp theo.

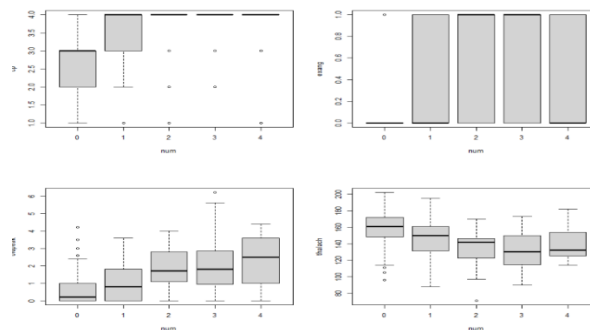


Hình 3. Biểu đồ tương quan đa biến

Nhìn hình 3 chúng ta thấy, biểu đồ trên thực chất là một ma trận gồm 2 phần. Phần dưới của ma trận là biểu đồ và đường biểu diễn cho từng biến. Phần trên của ma trận là hệ số tương quan,

với con số chỉ mức độ tương quan (có thể hiểu trong y học quy ước giá trị tuyệt đối của hệ số tương quan bé hơn 0,3 là tương quan lỏng lẻo và lớn hơn 0,7 là tương quan mạnh. Các ô trong đường chéo vẽ phân bố của từng biến. Đây là một biểu đồ cung cấp khá nhiều thông tin có ích. Tuy nhiên, khi số biến nhiều lên thì biểu đồ rất khó đọc. Với số lượng 14 thuộc tính (biến) mà chúng ta đang nghiên cứu thì biểu đồ này vẫn rất phù hợp. Với các biến tuổi (*age*), biến huyết áp lúc nghỉ (*trestbps*), cholestoral trong huyết thanh (*chol*), nhịp tim tối đa đạt được (*thalach*) là 4 biến có hình chuông cân đối, vậy ta dự đoán được các biến này tuân theo quy luật phân phối chuẩn. Các biến còn lại có dáng không giống hình chuông nên có thể không tuân theo quy luật phân phối chuẩn. Để khẳng định chính xác các biến này có tuân theo quy luật phân phối chuẩn hay không ta có thể dùng lệnh `shapiro.test(x)` để kiểm tra, với biến cần kiểm tra, lệnh này cho giá trị $p\text{-value} > 0,05$ thì biến đó được coi là có phân phối chuẩn [4], [5], [9].

Qua biểu đồ tương quan đa biến chúng ta thấy có các đám mây dữ liệu gần giống hình chữ nhật dẹt, cũng có nghĩa là các biến tương ứng sẽ có mối tương quan tuyến tính với nhau. Nhìn phần hệ số tương quan bên tay trái, ta nhận thấy với các con số lớn hơn 0,3 chứng tỏ giữa 2 thuộc tính đó có mối quan hệ tuyến tính với nhau cần thiết lập. Vậy có 7 hệ số có giá trị tuyệt đối lớn hơn 0,3 chỉ hệ số tương quan giữa 7 cặp sau: (*cp*; *num*), (*exang*; *num*), (*oldpeak*; *num*), (*slope*; *num*), (*ca*; *num*); (*thal*; *num*), (*thalach*; *num*) và đây cũng chính là 7 cặp thuộc tính mà chúng ta cần quan tâm. Trong thuộc tính *thalach* là thuộc tính có mối tương quan với *exang*, *oldpeak*, *slope*, *ca*, *num* nhưng chủ yếu là tương quan nghịch biến. Bên cạnh đó, ta có thể sử dụng lệnh `boxplot` để vẽ được biểu đồ hình hộp của một số cặp trên như sau:

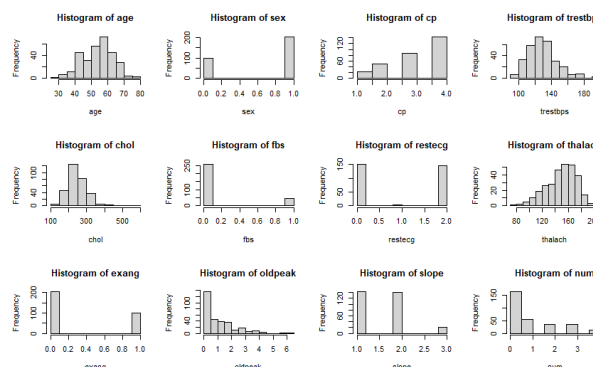


Hình 4. Biểu đồ hình hộp chữ nhật

Qua Hình 4 nhận thấy các đối tượng có mức độ suy tim bậc II, III, IV thuộc kiểu đau ngực không có triệu chứng, và các bệnh nhân suy tim đều cảm thấy đau thắt ngực do tập thể dục. Nhịp tim tối đa đạt được của các bệnh nhân cũng giảm dần theo mức độ suy tim.

Để xem các phân phối của các thuộc tính cùng một lúc để đưa ra nhận định về phân phối chuẩn của chúng, chúng ta dùng hệ thống lệnh:

➤ `op<-par(mfrow=c(2,3));hist(age); hist(sex);hist(cp);hist(trestbps); hist(chol);hist(fbs); hist(restecg); hist(thalach); hist(exang); hist(oldpeak);hist(slope); hist(ca);hist(thal); hist(num)`



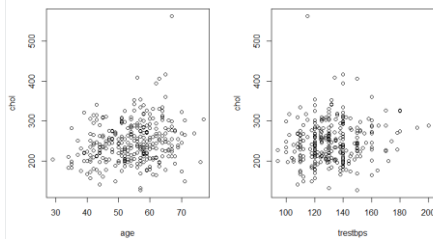
Hình 5. Hệ thống biểu đồ của 12 thuộc tính nghiên cứu

Thông qua hình 5 cho ta thấy, chỉ có 4 thuộc tính: age; trestbps; chol; thalach có phân bố gần với phân bố chuẩn. Từ điều nghi ngờ về phân bố chuẩn đó để khẳng định chúng ta có thể dùng lệnh shapiro.test cho 4 thuộc tính này để kiểm tra. Như vậy, kết quả của hình 3 và hình 5 cũng cho kết quả dự đoán tương đối giống nhau.

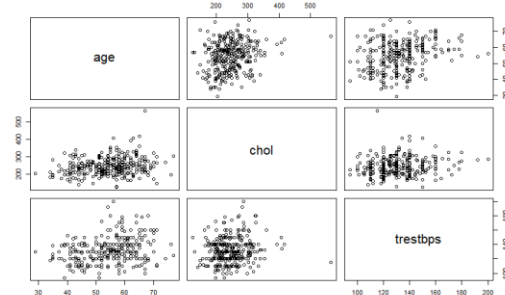
3.2. Thống kê suy diễn

Sau khi đã có cái nhìn tương đối tổng quan về dữ liệu, tiếp đến chúng ta sẽ dùng đến phân tích hồi quy tuyến tính để tiếp tục phân tích bộ dữ liệu trên. Trước tiên chúng ta sẽ nghiên cứu mối quan hệ của 3 thuộc tính chính: age, chol, trestbps.

```
> data<-data.frame(age,chol,trestbps)
> op<-par(mfrow=c(1,2))
> plot(chol~age)
> plot(chol~trestbps)
```



Hình 6. Biểu đồ giá trị tiên đoán của age và plot với độ tin cậy 95%.



Hình 7. Biểu đồ giá trị tiên đoán với độ tin cậy 95% của 3 thuộc tính chol, trestbps, age.

Biểu đồ cho thấy một đám mây không có hình chữ nhật dẹt, vậy dự đoán được các biến này không có mối tương quan tuyến tính với nhau. Để khẳng định lại một lần nữa mối tương quan giữa 3 thuộc tính này có phải là mối tương quan tuyến tính hay không, chúng ta dùng lệnh:

```
> pairs(data)
```

Nhìn biểu đồ tổng thể trên với hi vọng thuộc tính chol và trestbps là các thuộc tính có mối quan hệ tuyến tính với nhau, để kiểm chứng được điều đó chúng ta sử dụng lệnh: summary(lm(chol~trestbps)).

```
call:
lm(formula = chol ~ trestbps)

Residuals:
    Min       1Q   Median       3Q      Max
-127.81  -33.24   -5.24   28.70  323.68

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 195.9904    22.3833   8.756  <2e-16 ***
trestbps     0.3855     0.1685   2.287  0.0229 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 51.5 on 300 degrees of freedom
Multiple R-squared:  0.01714,    Adjusted R-squared:  0.01386
F-statistic: 5.232 on 1 and 300 DF,  p-value: 0.02287
```

Hình 8. Mối quan hệ giữa 2 thuộc tính chol và trestbps

```
> logistic <-glm( exang ~ num, family="binomial")
> summary(logistic)

Call:
glm(formula = exang ~ num, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7666  -0.6485  -0.6485   0.9239   1.8235

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.4522     0.1802  -8.060 7.64e-16 ***
num           0.6942     0.1088   6.381 1.76e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 382.10 on 301 degrees of freedom
Residual deviance: 335.67 on 300 degrees of freedom
AIC: 339.67

Number of Fisher Scoring iterations: 4
```

Hình 9. Mối quan hệ giữa 2 thuộc tính exang và num

Với kết quả này cho thấy, thuộc tính trestbps giải thích khoảng 1,7% độ dao động phụ thuộc tính chol giữa các cá nhân, độ phụ thuộc này rất thấp gần như không có ý nghĩa trong y học. Hai biến tiếp theo mà chúng ta đang hướng đến là exang (đau thắt ngực do tập thể dục) và num số lượng người bị bệnh tim phân theo từng mức độ. Sử dụng hàm lệnh:

```
> tapply(exang,num,mean)
      0      1      2      3      4
0.1411043 0.4545455 0.6111111 0.6571429 0.4615385
```

Kết quả trên cho thấy hiện tượng đau thắt ngực khi tập thể dục của mỗi nhóm được thể hiện qua giá trị trung bình.

Đối với 2 biến *exang* và biến *fbs* là 2 biến không liên tục và mang tính đo lường nhị phân nên chúng ta có thể sử dụng hồi quy logistic để tìm được mô hình tối ưu giữa các biến độc lập và phụ thuộc.

Đối với lệnh này, chúng ta đang yêu cầu R phân tích theo mô hình *exang* là một hàm số với biến số là *num*. Trong hàm lệnh *glm* có nhiều luật phân phối, mà trong đó phân phối nhị thức (binomial) là một luật phân phối chuẩn cho hồi quy logistic. Kết quả ở dòng thứ hai phản ánh độ khác biệt về giá trị quan sát và ước tính của $\text{logit}(p)$. Phần tiếp theo cung cấp các hệ số trong phương trình hồi quy và sai số chuẩn. Qua kết quả này, chúng ta có $\alpha = -1,4522$ và $\beta = 0,6942$, hai số này trái dấu vậy mỗi liên hệ giữa độ suy tim và nguy cơ đau thắt ngực do tập thể dục là mỗi liên hệ nghịch đảo. Vậy chúng ta mới chỉ phân tích được mối quan hệ giữa 2 biến *chol* và *trestbps*, nhưng chúng ta cần tìm được một mối quan hệ thích hợp cho nhiều biến, điều này mới thể hiện được sự liên hệ chặt chẽ giữa các yếu tố với nhau. Vì vậy, sử dụng hồi quy logistic đa biến là một câu trả lời hoàn hảo cho yêu cầu trên. Tuy nhiên, một trong những vấn đề khó khăn khá nan giải trong việc sử dụng hồi quy logistic đa biến là chọn một mô hình để có thể mô tả đầy đủ dữ liệu hay mô hình tối ưu (Đơn giản, đầy đủ, có ý nghĩa thực tế [4], [7], [8]). Để tìm được mô hình tương quan tối ưu giữa *exang* và các yếu tố còn lại ta sử dụng lệnh:

```
> temp <- glm(exang ~ ., family = "binomial", data=dl1)
search <- step(temp)
```

Với lệnh này cho ta kết quả của 9 bước tìm kiếm mô hình hồi quy logistic như sau:

Mô hình	AIC
exang ~ age + sex + cp + trestbps + chol + fbs + restecg + thalach + oldpeak + slope + ca + thal + num	319.55
exang ~ age + sex + cp + trestbps + chol + fbs + restecg + thalach + oldpeak + slope + ca + num	315.82
exang ~ age + sex + cp + trestbps + chol + fbs + thalach + oldpeak + slope + ca + num	313.82
exang ~ age + sex + cp + trestbps + chol + thalach + oldpeak + slope + ca + num	312.2
exang ~ age + sex + cp + trestbps + chol + thalach + oldpeak + slope + num	310.65
exang ~ age + sex + cp + trestbps + chol + thalach + slope + num	309
exang ~ age + sex + cp + chol + thalach + slope + num	307.81
exang ~ sex + cp + chol + thalach + slope + num	306.79
exang ~ sex + cp + thalach + slope + num	306.11

Hình 10. Mô hình hồi quy logistic của thuộc tính *exang*

Qua 9 bước tìm kiếm mô hình với hàm *exang* ta thấy thuộc tính này phụ thuộc chủ yếu vào 5 thuộc tính *sex + cp + thalach + slope + num* và đây cũng là mô hình cho thông số AIC là bé nhất [5], [3], [10]. Căn cứ vào thông tin này có thể thấy 6 thuộc tính trên liên quan mật thiết đến nhau. Đây cũng chính là mô hình đáp ứng 3 tiêu chuẩn: Đơn giản, đầy đủ, có ý nghĩa thực tế. Sở dĩ mô hình này đơn giản vì chỉ có 5 biến phụ thuộc trên tổng số 14 biến và mô hình này đã tiên đoán gần với giá trị thực tế quan sát của biến phụ thuộc *exang*. Đây cũng là cơ sở cho các suy đoán và nghiên cứu tiếp theo trong y học. Tiếp tục có thể sử dụng hồi quy logistic đối với biến *fbs*, đây là biến nhị thức, mục tiêu tiếp theo để tìm một mô hình đơn giản, đầy đủ có ý nghĩa thực tế nhất. Sử dụng hàm lệnh sau:

```
> temp1 <- glm(fbs ~ ., family = "binomial", data=dl1)
```

search <- step(temp1). Với câu lệnh này ta được mô hình 13 bước tìm kiếm như sau:

Mô hình	AIC
fbs ~ age + sex + cp + trestbps + chol + restecg + thalach + exang + oldpeak + slope + ca + thal + num	265.94
fbs ~ age + sex + cp + trestbps + chol + restecg + thalach + exang + oldpeak + slope + ca + num	262.58
fbs ~ age + sex + cp + trestbps + restecg + thalach + exang + oldpeak + slope + ca + num	260.69
fbs ~ age + sex + cp + trestbps + restecg + thalach + exang + oldpeak + slope + ca	258.88
fbs ~ age + sex + cp + trestbps + thalach + exang + oldpeak + slope + ca	257.05

$\text{fbs} \sim \text{age} + \text{sex} + \text{cp} + \text{trestbps} + \text{thalach} + \text{oldpeak} + \text{slope} + \text{ca}$	255.29
$\text{fbs} \sim \text{age} + \text{sex} + \text{trestbps} + \text{thalach} + \text{oldpeak} + \text{slope} + \text{ca}$	253.5
$\text{fbs} \sim \text{age} + \text{sex} + \text{trestbps} + \text{thalach} + \text{oldpeak} + \text{slope}$	251.84
$\text{fbs} \sim \text{age} + \text{sex} + \text{trestbps} + \text{oldpeak} + \text{slope}$	250.18
$\text{fbs} \sim \text{age} + \text{sex} + \text{trestbps} + \text{oldpeak}$	248.87
$\text{fbs} \sim \text{age} + \text{sex} + \text{trestbps}$	247.68
$\text{fbs} \sim \text{age} + \text{trestbps}$	246.96
$\text{fbs} \sim \text{trestbps}$	246.42

Hình 11. Mô hình hồi quy logistic của thuộc tính *fbs*

Đối với thuộc tính *fbs* (đường huyết lúc đói $> 120\text{mg/dl}$) để thấy thuộc tính này chỉ phụ thuộc vào *trestbps* (huyết áp khi nhập viện). Như vậy, chúng ta cũng sẽ loại bỏ thuộc tính này khi nghiên cứu đến mức độ suy tim. Và một lần nữa cho chúng ta nghi ngờ về mối quan hệ giữa cao huyết áp và đường huyết lúc đói; chính vì vậy cũng có thể đưa ra một giả thuyết mới trong y học, để từ đó ta có thể có những nghiên cứu tiếp theo.

Qua 2 mô hình trên khi sử dụng hồi quy logistic giúp chúng ta có cái nhìn khá tổng quan về mối tương quan giữa các thuộc tính với nhau, tuy nhiên 1 hạn chế khi sử dụng hồi quy logistic chỉ nghiên cứu trên các hàm nhận 2 giá trị 0 và 1 hay các biến nhị thức.

4. Kết luận

Thông qua mô hình phân tích dữ liệu cho bệnh tim mạch ta có thể đưa ra một số kết luận tương đối quan trọng phục vụ cho y học:

Một số triệu chứng đau thắt ngực khi tập thể dục là những triệu chứng ban đầu cho suy tim.

Nhịp tim tối đa của các bệnh nhân đạt được giảm dần theo mức độ suy tim từ I đến IV.

Mô hình các thuộc tính liên quan đến việc đau thắt ngực khi tập thể dục của các bệnh nhân suy tim có liên quan đến 5 thuộc tính: *sex* + *cp* + *thalach* + *slope* + *num*.

Hoàn toàn loại trừ được thuộc tính đường huyết lúc đói và bệnh suy tim, đồng thời cho chúng ta một hướng nghiên cứu mới là mối liên quan mật thiết giữa đường huyết lúc đói và huyết áp khi nhập viện.

Như vậy, việc xây dựng được mô hình phân tích dữ liệu đã được xây dựng gồm 2 bước chính: Bước 1: Thống kê mô tả; Bước 2: Thống kê suy diễn đã cho ta được những kết luận khá thú vị về bệnh tim mạch. Với những kết luận này chính là gợi ý cho những phương án điều trị bệnh lý hay những hướng mới cho nghiên cứu về bệnh tim mạch.

TÀI LIỆU THAM KHẢO/ REFERENCES

- [1] D. T. P. Quynh, "Application of software R in teaching estimation and hypothesis testing in medicine and pharmacy at university of medicine and pharmacy TNU," *TNU Journal of Science and Technology*, vol. 225, no. 04, pp. 107-112, 2019.
- [2] V. T. Nguyen, *Data analyze with question and answers*, Publishing company Ho Chi Minh city, 2018.
- [3] V. T. Nguyen, *Data analysis and chart R*, Garvan Institute of Medical Research Sydney, Australia, 2013.
- [4] T. H. Dang, *Statistics for social sciences and life sciences with R software*. Publishing company Ha Noi University, 2019.
- [5] J. Veani, "Simple R-Using R for Introductory Statistics," 2001. [Online]. Available: <https://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf>. [Accessed Jan. 2020].
- [6] E. Paradis, "R for Beginners," 2005. [Online]. Available: https://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf. [Accessed Jan. 2020].
- [7] J. H. Maindonald, "Using R for Data Analysis and Graphics," Australian National University, 2008. [Online]. Available: <https://cran.r-project.org/doc/contrib/usingR.pdf>. [Accessed Feb. 2020].
- [8] D. H. Nguyen, "GGEbiplot and R Language," *J. Sci. & Devel*, vol. 11, no. 8, pp. 1164-1169, 2013.
- [9] M. Bodin, *A trustworthy mechanize formalization of R*, Boston, MA, USA. ACM, New York, NY, USA, 2018, p. 12.
- [10] John. Fox, "R and the Journal of Statistical Software," *Journal of Statistical Software*, vol. 73, no. 2, pp. 1-13, 2016, doi:10.18637/jss.v073.i02.