A picture containing rectangle, screenshot, text, frame

Description automatically generated

**Giảng viên lý thuyết:** TS. Nguyễn Đình Thuân

**Giảng viên thực hành**: Nguyễn Minh Nhật

**Nhóm sinh viên thực hiện**: Nhóm 20

1. Cao Hoài Sang MSSV: 21522541

2. Thi Thành Công MSSV: 21521897

3. Nguyễn Trần Gia Kiệt MSSV: 21522258

4. Nguyễn Hoàng Đăng Khoa MSSV: 21520999

5. Cù Ngọc Hoàng MSSV: 21552086

**Môn học: PHÂN TÍCH DỮ LIỆU KINH DOANH**

ĐẠI HỌC QUỐC GIA TP.HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHOA HỆ THỐNG THÔNG TIN

BÁO CÁO LAB03

**BẢNG PHÂN CÔNG CÔNG VIỆC**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Cao Hoài Sang | Thi Thành Công | Nguyễn Trần Gia Kiệt | Cù Ngọc Hoàng | Nguyễn Hoàng Đăng Khoa |
| Câu 1 |  |  |  |  | x |
| Câu 2 | x |  | x |  |  |
| Câu 3 | x | x |  |  |  |
| Câu 4 | x |  |  |  | x |
| File báo cáo | x | x | x | x | x |

**DANH SÁCH HÌNH ẢNH**

**MỤC LỤC**

[**1.** **Phần 1** 2](#_Toc163376958)

[**2.** **Phần 2** 9](#_Toc163376959)

[2.1. Câu a 9](#_Toc163376960)

[2.1.1. Yêu cầu 9](#_Toc163376961)

[2.1.2. Bài giải 9](#_Toc163376962)

[**CÁC TÀI NGUYÊN KHÁC** 28](#_Toc163376963)

# **Phần 1**

* 1. **Câu a:**
     1. **Yêu cầu:**
* Hồi quy tuyến tính đa biến (Multivariable Linear Regression) nghĩa là gì? Cho ví dụ minh họa.
  + 1. **Bài giải:**
* Định nghĩa:
  + Hồi quy tuyến tính đa biến (Multiple Variable Linear Regression) là phương pháp thống kê được sử dụng để phân tích mối quan hệ giữa một biến phụ thuộc và nhiều biến độc lập. Trong hồi quy tuyến tính đa biến, mục tiêu chính là xác định và dự đoán mối quan hệ tuyến tính giữa biến phụ thuộc (dependent variables hay thường gọi là biến y) và một hoặc nhiều biến độc lập (independent variable hay thường gọi là biến x) bằng cách tìm ra các hệ số hồi quy phù hợp. [1]
  + Cách thức tính toán hồi quy tuyến tính đa biến thường dựa trên phương pháp của bình phương nhỏ nhất (least squares), trong đó chúng ta cố gắng tìm ra một đường thẳng hoặc một siêu mặt phẳng sao cho tổng bình phương sai số giữa các giá trị dự đoán và giá trị thực tế là nhỏ nhất.
  + Mô hình hồi qui tuyến tính đa biến là phương trình mô tả mối quan hệ giữa biến phụ thuộc y với các biến độc lập xn và số hạng sai số Ꜫ:

* + Với:
    - β là các tham số.
    - Ꜫ là biến ngẫy nhiên gọi là số hạng sai số.
  + **Kiểm định F**: được dùng để xác định có tồn tại mối liên hệ có ý nghĩa giữa biến phụ thuộc và toàn bộ các biến độc lập => Kiểm định tổng thể của toàn bộ mô hình.
    - Giả thuyết: H0: β1 = β2 = ... = βp = 0; Ha: Có ít nhất 1 tham số βi khác 0.
    - Trị kiểm định: F = MSR/MSE
    - Quy tắc bác bỏ: Bác bỏ H0 nếu p-value < a hay nếu F > Fa, với Fa lấy từ bảng phân phối F .Bậc tự do trên tử số là p và bậc tự do dưới mẫu số là n - p - 1.
  + Kiểm định t: được dùng để xác định xem từng biến độc lập riêng có ý nghĩa hay không ở trong mô hình.
    - Giả thuyết: H0: βi = 0; Ha: βi khác 0.
    - Trị kiểm định: t = β/SE(βi) (β là hệ số hồi quy và SE(β) là sai số chuẩn)
    - Quy tắc bác bỏ: Bác bỏ H0 nếu p-value < a hay nếu t < -ta/2 hay t > ta/2 với ta/2 được lấy từ bảng phân phối t với bậc tự do là n - p - 1.
* Ví dụ: Khảo sát lương lập trình viên một công ty phần mềm thu thập dữ liệu của một mẫu gồm 20 lập trình viên. Người ta đề nghị sử dụng phân tích hồi qui để xác định xem lương có mối liên hệ với số năm kinh nghiệm và điểm thi năng khiếu về lập trình do công ty tổ chức hay không? Số năm kinh nghiệm, điểm thi năng khiếu Và mức lương hàng năm ($1000s) của 20 lập trình viên được trình bày ở bảng sau:

|  |  |  |
| --- | --- | --- |
| Experience | Score | Salary |
| 4 | 78 | 24.0 |
| 7 | 100 | 43.0 |
| 1 | 86 | 23.7 |
| 5 | 82 | 34.3 |
| 8 | 86 | 35.8 |
| 10 | 84 | 38.0 |
| 0 | 75 | 22.2 |
| 1 | 80 | 23.1 |
| 6 | 83 | 30.0 |
| 6 | 91 | 33.0 |
| 9 | 88 | 38.0 |
| 2 | 73 | 26.6 |
| 10 | 75 | 36.2 |
| 5 | 81 | 31.6 |
| 6 | 74 | 29.0 |
| 8 | 87 | 34.0 |
| 4 | 79 | 30.1 |
| 6 | 94 | 33.9 |
| 3 | 70 | 28.2 |
| 3 | 89 | 30.0 |

* Nhập vào Excel ta có:

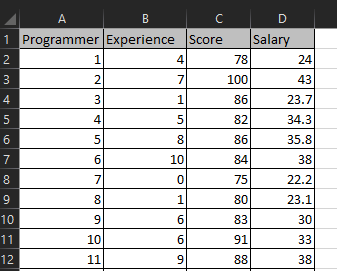


Figure 1. Khảo sát lập trình viên nhập vào Excel

* Kết quả thu được trong Excel sau khi sử dụng công cụ Data Analysis:

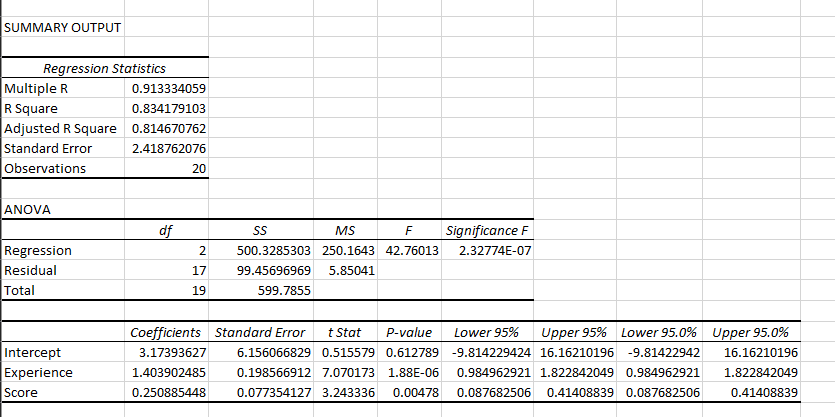


Figure 2. Kết quả thu được Excel

* Giải thích:
  + Với β1 = 1.404, ta kết luận lương được kỳ vọng tăng $1.404 đối với mỗi 1 năm kinh nghiệm tăng thêm (Khi điểm năng khiếu được giữ không đổi).
  + Với β2 = 0.251, ta kết luận lương được kỳ vọng tăng $251 đối với điểm năng tăng thêm (Khi số năm kinh nghiệm được giữ không đổi).
* Kiểm định F:
  + **Giả thuyết**: H0: β1 = β2 = 0; Ha: Có ít nhất 1 tham số βi khác 0.
  + **Quy tắc bác bỏ**: Với a = 5% và bậc tự do là 17. Tra bảng F.05 = 3.59 => Bác bỏ H0 nếu p-value < .05 hay F > 3.59

Dựa vào bảng kết quả ANOVA ta có MSR = 250.1643, MSE = 5.85041, F = MSR/MSE = 42.76 và p-value = 2.32774E-07:

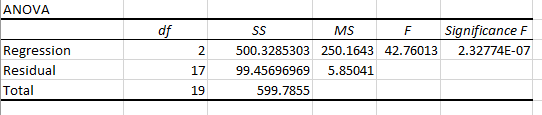


Figure 3. Kết quả bảng tính ANOVA

* + **Kết luận**: Do p-value < .05 (hay F = 42.76 > 3.59) vì vậy có thể bác bỏ H0.

Kiểm định t:

* **Giả thuyết**: H0: βi = 0; Ha: βi khác 0.
* **Quy tắc bác bỏ**: Với a = .05 và bậc tự do = 17, tra bảng t.025 = 2.11. Bác bỏ H0 nếu p-value < .05 hay t > 2.11.

Dựa vào bảng kết quả ta có được, βExperience = 1.403902485 và SE(βExperience) = 0.198566912; , βScore = 0.250885448 và SE(βScore) = 0.077354127. Ta tính được tExperience = 7.070173 và tScore = 3.243336.

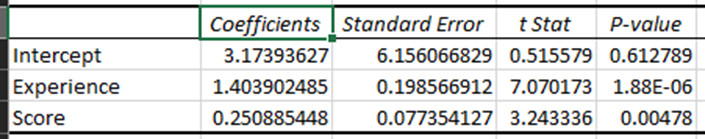
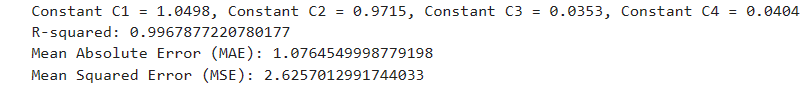


Figure 4. Bảng kết quả Coeficients, Standard Error, giá trị t và giá trị P

* + - **Kết luận**: Bác bỏ cả H0: β1 = 0 và H0: β2 = 0 => Cả hai biến độc lập đều có ý nghĩa.
  1. **Câu b:**
     1. **Yêu cầu:**
* Hồi quy phi tuyến tính (Multivariable Nonlinear Regression) là gì? Cho ví dụ minh họa.
  + 1. **Bài giải**
* Định nghĩa: Hồi quy phi tuyến tính đa biến là một phương pháp thống kê được sử dụng để tìm ra mối quan hệ giữa biến phụ thuộc và nhiều biến độc lập khi mối quan hệ giữa chúng không phải là tuyến tính. Mô hình hồi quy phi tuyến tính có tính linh hoạt hơn so với mô hình hồi qui tuyến tính. Ngoài ra trong thực tế, dữ liệu không phải lúc nào cũng có mối quan hệ tuyến tính, nên việc sử dụng mô hình hồi qui phi tuyến tính sẽ cho độ chính xác cao hơn
* Mô hình hồi qui phi tuyến tính đa biến tổng quát:
  + Trong đó:
  + X: là vector các biến độc lập
  + β: là vector các tham số
  + f: hàm số phi tuyến tính
  + ε: hệ số lỗi
* Tổng bình phương sai số
* Trong đó:
* : là giá trị quan sát thứ i
* : giá trị dự đoán thứ i theo mô hình
* Chúng ta sẽ ước lượng tham số để hàm SSE đạt giá trị nhỏ nhất.
* Phương pháp Gauss-Newton: là một kỹ thuật tối ưu hóa được sử dụng để ước lượng tham số trong các hàm phi tuyến tính
* Đặt ,ta được:
* SSE =
* Trong đó: i = 1,2,…m
* Ta có:
* = 1,2,…n
* Với J là ma trận jacobian
* Và
* Ma trận Hessian
* Từ đó suy ra :
* Điều kiện dừng:
* Số lần lặp chỉ định
* Khoảng cách bằng ngưỡng mong muốn
* Ví dụ: Dự đoán giá dầu (OIL) từ (WTI), (HH), (MB)
* Trong đó: WTI: West Texas Intermediate price
* WTI: Henry Hub gas price
* MB: Mont Belvieu propane spot price
* Tập dữ liệu
* A screenshot of a computer screen

  Description automatically generated
* Giả sử mô hình dự đoán
* Sử dụng python để tìm tham số
* 
* Vậy dự đoán
* Đánh giá:  = 0.996 => mô hình phù hợp mạnh mẽ
* MAE = = 1.076
* MSE = = 2.626
  1. **Câu c:**
     1. **Yêu cầu:**
* Hồi quy Logistic (Logistic Regression) nghĩa là gì? Cho ví dụi minh họa.
  + 1. **Bài giải:**
* Định nghĩa:
  + Hồi quy Logistic là một mô hình thống kê trong máy học được sử dụng để phân loại nhị phân, dự đoán một đối tượng thuộc vào một trong hai nhóm (thường là có hoặc không). Hồi quy Logistic hoạt động dựa trên nguyên tắc của hàm sigmoid – một hàm phi tuyến tự chuyển đầu vào của nó thành xác suất thuộc về một trong hai lớp nhị phân.

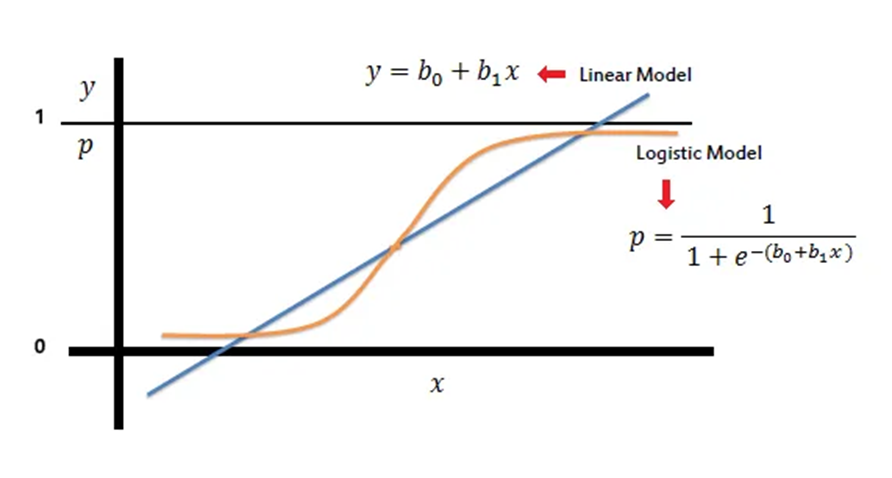


Figure 5. Mô hình hồi quy logistic

* + Hồi quy Logistic hoạt động dựa trên hàm Sigmoid, được biểu diễn như sau:
  + Hàm Sigmoid nhận đầu vào là một giá trị z bất kỳ, và trả về đầu ra là một giá trị xác suất nằm trong khoảng [0,1].

Ví dụ: Thực hiện hồi quy logistic trong Excel để xác định xem một tuyển thủ bóng rổ đại học có khả năng được lựa chọn vào NBA hay không. Bộ dữ liệu bao gồm các chỉ số hiệu suất cơ bản của các tuyển thủ từ mùa trước:

* Average Point.
* Rebounds.
* Assists.

Bởi vì hồi quy logistic phân loại nhị phân, mục tiêu là dự đoán khả năng có được chọn hay không (0 = không, 1 = có).

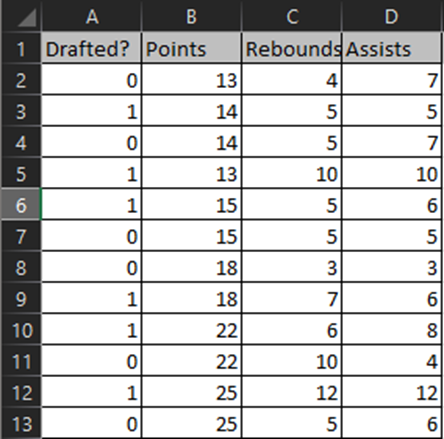
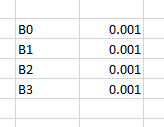
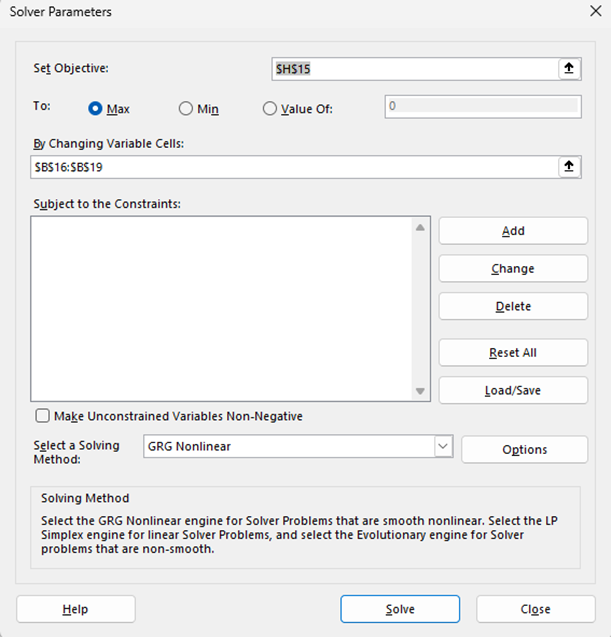


Figure 6. Khảo sát tuyển thủ bóng rổ nhập vào Excel

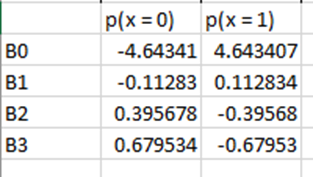
* + Xây dựng model tính ta có:
  + Gán giá trị tạm thời cho các β:



* + Sử dụng công cụ sẵn có của Excel, ta cần tìm cực đại của log likelihood bằng cách thay đổi giá trị của các β.



* + Sau đó ta được chỉ số các β:



* + Với tham số đã có, giờ đây model của chúng ta có thể dự đoán tỉ lệ phần trăm một tuyển thủ bóng rổ đó có được chọn tham gia giải NBA hay không.
  + Giả dụ ta có một tuyển thụ với chỉ số 15(Average Point), 4(Rebounds), 6(Assists). Áp dụng vào model ta được:
  + Như vậy tuyển thủ này sẽ có tỉ lệ là 66.28% được chọn vào giải NBA.
  + **Lưu ý**: Lý giải vì sao ta cần tìm giá cực dại của log likelihood: Đối với một bài toán Machine Learning tổng quát, việc giải quyết bài toán thường gồm 3 bước chính:
    - **Modeling**: Đi tìm mô hình có thể mô tả tốt nhất bài toán
    - **Learning**: Tối ưu tham số cho mô hình dựa vào dữ liệu có sẵn
    - **Inference**: Sử dụng mô hình đã tối ưu để dự đoán kết quả với đầu vào chưa biết
  + Ở bước Learning, có hai phương pháp được sử dụng phổ biến để tối ưu bộ tham số, đó là Maximum Likelihood Estimation(MLE) và Maximum A Posteriori Estimation.
  + MLE là phương pháp dự đoán tham số của một mô hình thống kê dựa trên những “quan sát” có sẵn, bằng cách tìm bộ tham số sao cho có thể tối đa hoá khả năng mà mô hình với bộ tham số đó sinh ra các “quan sát” có sẵn.

1. **Phần 2**
   1. **Câu a**
      1. **Yêu cầu**

* Sử dụng MS Excel, ngôn ngữ R và ngôn ngữ Python để thực hiện hồi quy tuyến tính đa biến với tệp dữ liệu Colleges and Universities.
  + 1. **Bài giải**
* Cơ sở lí thuyết:
  + Mô hình hồi quy tuyến tính đa biến: Là một công cụ phân tích trong thống kê và machine learning, được sử dụng để dự đoán giá trị của một biến phụ thuộc dựa trên các biến độc lập. Nó dựa trên giả định rằng mối quan hệ giữa các biến là tuyến tính.
  + Phương trình hồi quy tuyến tính đa biến có dạng:
    - Giả sử
      * Sai số có phân phối chuẩn
      * Các biến đôc lập có sai số bằng nhau
      * Các biến độc lập là độc lập
    - Ta có:
* S bé nhát nếu là nghiệm của hệ phương trình

, 0, …, 0

* + - Thực hiện đạo hàm riêng, ta có:

…

* + - Biến đổi, ta có:

…

* + - Với các ma trận:

* + - Ta có hệ phương trình được viết lại:
* Sử dụng MS Excel:
  + Với tính năng *Data Analysis* của MS Excel. Ta tích hợp bằng cách
    - Chọn *File -> Options -> Add-ins.* Ở muc *Manage* chọn *Go. Chọn Analysis ToolPak -> OK.*
  + Kết quả khi sử dụng *Regression* trong *Analysis Tools.*

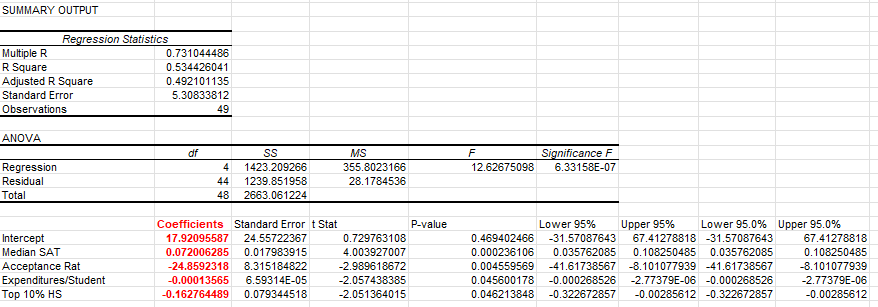
**

Figure Kết quả thực hiện MS Excel bằng Analysis Tools

* + Ngoài ra, ta có thể sử dụng các hàm *MMULT*, *MINVERSE*, *TRANSPOSE* để tính toán.

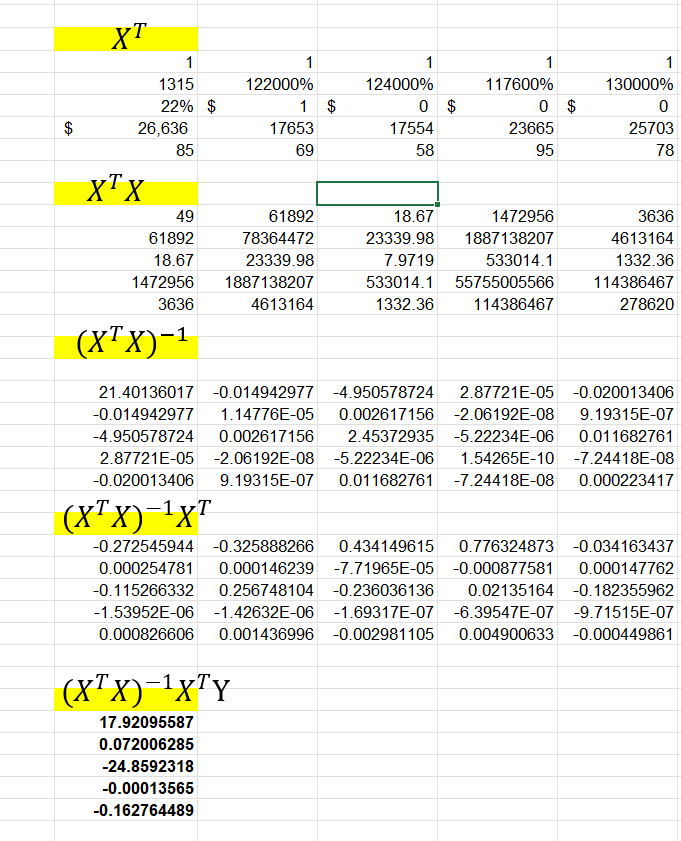


Figure Kết quả áp dụng công thức theo Slide

* Sử dụng ngôn ngữ R:

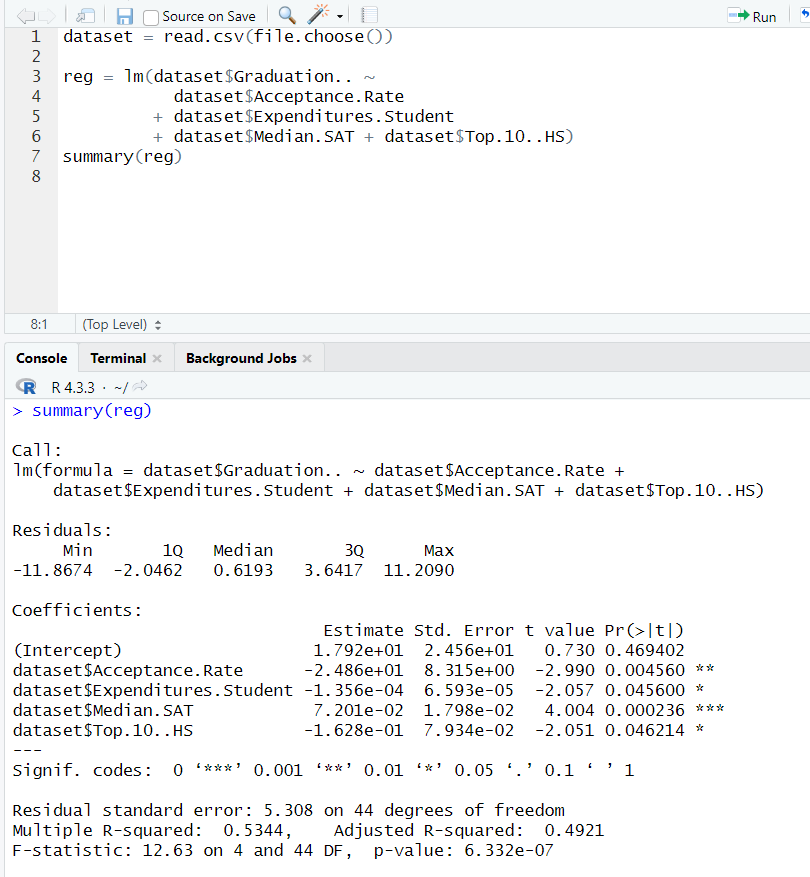


Figure Kết quả thực hiện trên R

* Sử dụng ngôn ngữ Python:

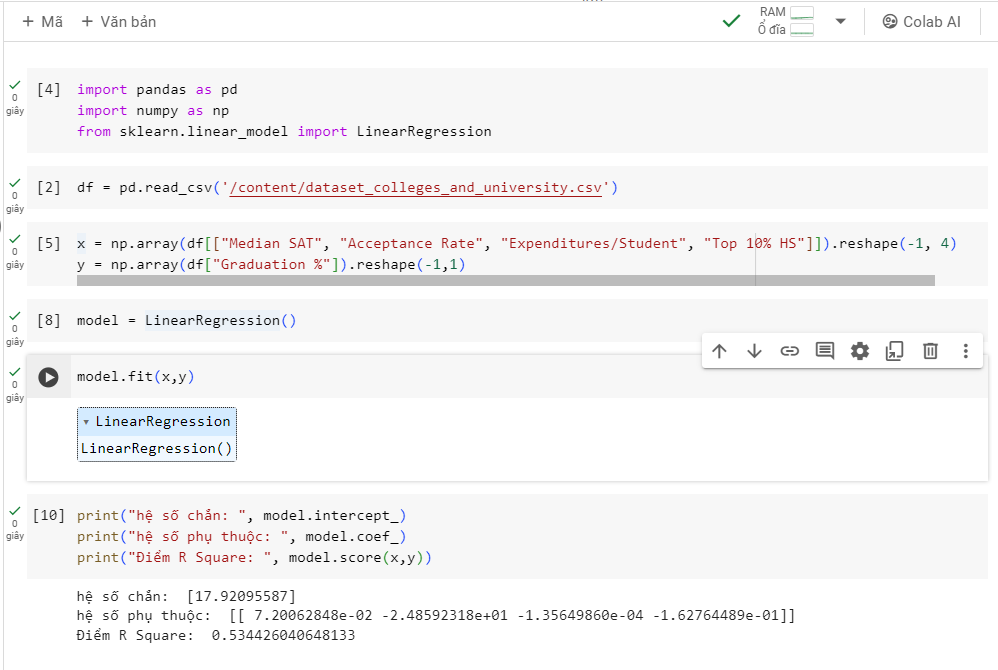


Figure Kết quả thực hiện trên Python

* Nhận xét:
  + Từ kết quả tính toán, ta có phương trình biểu diễn *Graduation %* theo các biến độc lập *Median SAT, Acceptance Rate, Expenditures/Student, Top 10% HS là:*
  + Multiple R = 0.731, có thể diễn giải là một mối quan hệ tương đối mạnh mẽ giữa biến phụ thuộc.
  + *Multiple R-squared* có giá trị là 0.5344 cho ta thấy phương trình đáp ứng được 53.44% hay các biến độc lập trong phương trình (*Median SAT, Acceptance Rate, Expenditures/Student, và Top 10% HS*) cùng đóng góp vào việc dự đoán *tỷ lệ tốt nghiệp* với tỷ lệ 53.44%.
  + Thực hiện kiểm định F-test:
  + Ta có *F-stat = 12.62675098 và .* Từ đó suy ra *F-stat > => Bác bỏ =>* Mô hình tổng thể có ý nghĩa.
  1. **Câu b**
     1. **Yêu cầu:**
* Sử dụng MS Excel, R, Python để thực hiện hồi quy bội phi tuyến trên 1 tệp dữ liệu tự chọn thực tế của Việt Nam.
  + 1. **Bài giải:**
* Sử dụng tệp dữ liệu chất lượng không khí ở vũng tàu:
  + vũng-tàu\_ngã tư giếng nước - tp.vũng tàu-air-quality.csv

**A screenshot of a table

Description automatically generated**

Figure Dataset chất lượng không khí ở Vũng Tàu

* + pm10 là tổng các hạt bụi lơ lửng có đường kính khí động học nhỏ hơn hoặc bằng 10 µm.
  + pm25 là tổng các hạt bụi lơ lửng có đường kính khí động học nhỏ hơn hoặc bằng 2,5 µm.
  + o3 là khí ozon
  + so2 là khí lưu huỳnh dioxit
  + Co là khí Carbon monoxide
* Như đã chứng minh từ 2.1. ta có hệ phương trình được viết lại:
* Trong tệp dữ liệu, đặt các biến không phụ thuộc mới:

A screenshot of a spreadsheet

Description automatically generated

Figure Thêm các cột tính toán mới

* Chọn pm10 làm biến phụ thuộc và các cột vừa tạo làm biến độc lập.
* Dùng MS Excel:
  + Kết quả sử dụng dụng cụ Regression trong excel:

A screenshot of a computer

Description automatically generated

Figure Kết quả sử dụng công cụ Regression của Excel

* + Hoặc dùng các hàm tính toán ma trận trong excel để tìm ma trận hệ số:

A white background with black and blue text

Description automatically generated with medium confidence

Figure Quá trình tìm ma trận hệ số trong Excel

* Dùng R:
  + Kết quả chạy code R:

A screenshot of a computer

Description automatically generated

Figure Sử dụng R thực hiện hồi quy bội phi tuyến tính

* Dùng python:
  + Kết quả chạy code python:

A screenshot of a computer

Description automatically generated

Figure Sử dụng python thực hiện hồi quy bội phi tuyến tính

* Với:
  + X1:
  + X2:
  + X3:
  + X4:
* Từ kết quả tính từ các công cụ trên ta nhận thấy:
  + R-squared = 0.714 -> 71.4% số lượng dữ liệu phù hợp với tập thể
  + P-value của intercept = 7.95e-0.8(<0.05)
  + P-value của = 0.00506(<0.05)
  + P-value của =0.01514 (<0.05)
  + P-value của = 0.15699 (>0.05)
  + P-value của = 1.15e-05 (<0.05)
* => Mô hình phù hợp với tập thể
* => Phương trình biễu diễn của pm10 theo các biến độc lập:

Pm10 = 17.3022 + 0.2978\* + (-3.3649)\* + 0.7023\*(

* 1. **Câu c**
     1. **Yêu cầu bài toán.**

Dùng MS Excel, ngôn ngữ R và ngôn ngữ Python thực hiện Hồi quy Logistic với dữ liệu thực tế tùy chọn về/của Việt Nam.

* + 1. **Bài giải:**
* **Giới thiệu dataset**
  + Dự đoán mưa ở Bạc Liêu thông qua tốc độ gió và tỉ lệ mây.
  + Link dataset:<https://www.kaggle.com/datasets/vanviethieuanh/vietnam-weather-data/data>

A table with numbers and letters

Description automatically generated

Figure Giới thiệu dataset

* + Chúng ta sẽ quan tâm đến 3 cột: humidi, rain
* **Bằng Excel:**
  + Bước 1: Khởi tạo giá trị các biến β0, β1. Ở đây ta khởi tạo giá trị ban đầu là 0.01.

**A grid of numbers with black text

Description automatically generated with medium confidence**

Figure Khởi tạo giá trị β0, β1

* + Bước 2: Tính toán giá trị Logit theo công thức:

A screenshot of a computer

Description automatically generated

Figure Tính giá trị Logit

* + Bước 3: Tính Probability Value.

A screenshot of a computer

Description automatically generated

Figure Tính giá trị Probability Value

* + Bước 4: Tính các giá trị Log likelihood và sum của các giá trị đó.
    - Tính giá trị Log likelihood

A screenshot of a computer

Description automatically generated

Figure Tính giá trị Log likelihood

* + - Tính giá trị sum của Log likelihood

A screenshot of a computer

Description automatically generated

Figure Tính sum của Log likehood

* + Bước 5: Sử dụng Solver Analysis Tool.
    - Bật Solver Analysis Tool. Trong Add-ins, chọn Excel Add-ins cho manage và tích chọn Solver Add-in và nhấn OK

A screenshot of a computer

Description automatically generated

Figure Bật Solver Analysis Tool

* + - Bật Solver Parameter và tính toán, ở ô Set objective ta chọn sum của các giá trị Log likelihood, ở ô by changing variable cells là các giá trị của biến β0, β1. Sau khi xong ta nhấn Solve

A screenshot of a computer

Description automatically generated

Figure Sử dụng Solver Parameters và tính toán

* + - Kết quả ta nhận được

A screenshot of a computer

Description automatically generated

Figure Kết quả tính toán với Excel

* **Bằng R**
  + Import dữ liệu

A screenshot of a computer

Description automatically generated

Figure Import dữ liệu

* + Sử dụng attach() để sử dụng biến sau khi import dữ liệu và dùng hàm glm() để phân tích hồi quy.

A screenshot of a computer

Description automatically generated

Figure Phân tích hồi quy trên RStudio

* **Bằng Python:**
  + Import các thư viện và file csv cần thiết

A screenshot of a computer

Description automatically generated

Figure Import các thư viện và dữ liệu cần thiết

* + Lấy ra biến phụ thuộc Y và biến độc lập X:

A screenshot of a computer

Description automatically generated

Figure Lấy ra các biến x, y

* + Dùng hàm LogisticRegression() trong thư viện sklearn() để đưa ra mô hình theo biến x và y

A screenshot of a computer program

Description automatically generated

Figure Thực hiện phân tích hồi quy Logistic

* + Kết quả nhận được

A screenshot of a computer error

Description automatically generated

Figure Kết quả nhận được với Python

* Bảng kết quả và nhận xét:
  + Từ kết quả ta có được từ phương trình hồi quy Logistic như sau:
  + Ta suy được kết quả như sau:
  + Ta đặt hệ số là odd
  + Đặt thì
  + Đặt thì
  + Tỉ số = 0.798
  + Lúc này ta có thể diễn dịch, cứ độ ẩm tăng thêm 1% thì khả năng mưa tăng lên 0.798 lần. Nếu tăng thêm 0.1% thì khả năng mưa tăng thêm 0.0798 lần.

# **CÁC TÀI NGUYÊN KHÁC**

|  |  |
| --- | --- |
| Tên tài nguyên | Đường dẫn |
| Mã nguồn | [Ấn vào đây](https://uithcm-my.sharepoint.com/:f:/g/personal/21522541_ms_uit_edu_vn/Enbj4oad8xxLqqmoexrRhb4BraxqaBmqVEU9PV-i4nBQLw?e=etEGKr) |

**DANH MỤC TÀI LIỆU THAM KHẢO**

|  |  |
| --- | --- |
| [1] | M. Tranmer, J. Murphy, M. Elliot and M. Pampaka, Multiple Linear Regression, 2nd ed., Cathie Marsh Institue, 2020. |
| [2] | Z. Bobbitt, "A Simple Guide to Understanding the F-Test of Overall Significance in Regression," Statology, 26 03 2019. [Online]. Available: https://www.statology.org/a-simple-guide-to-understanding-the-f-test-of-overall-significance-in-regression/. |
| [3] | W3School, "Machine Learning - Logistic Regression," 21 03 2000. [Online]. Available: https://www.w3schools.com/python/python\_ml\_logistic\_regression.asp. |