**VIET NAM NATIONAL UNIVERSITY HO CHI MINH CITY**

**UNIVERSITY OF INFORMATION TECHNOLOGY**

**INFORMATION SYSTEM FACULTY**



**LAB 3 REPORT**

**BUSINESS ANALYSIS**

Lecturer: [Nguyen Dinh Thuan](https://courses.uit.edu.vn/user/view.php?id=490&course=1)

Instructor: Nguyen Minh Nhut

Student performance: Team 15

Nguyen Cam Tu 20520837

Nguyen Thi Thu Thuy 20520797

Nguyen Thanh Son 20521847

**HO CHI MINH CITY,** **APRIL 2023**

# LỜI NHẬN XÉT CỦA GIẢNG VIÊN

………………………………………………………………………………………………….

………………………………………………………………………………………………….

………………………………………………………………………………………………….

………………………………………………………………………………………………….

………………………………………………………………………………………………….

………………………………………………………………………………………………….

………………………………………………………………………………………………….

………………………………………………………………………………………………….

………………………………………………………………………………………………….

………………………………………………………………………………………………….

………………………………………………………………………………………………….

………………………………………………………………………………………………….

………………………………………………………………………………………………….

………………………………………………………………………………………………….

………………………………………………………………………………………………….

………………………………………………………………………………………………….

………………………………………………………………………………………………….

………………………………………………………………………………………………….

………………………………………………………………………………………………….

………………………………………………………………………………………………….

………………………………………………………………………………………………….

………………………………………………………………………………………………….

………………………………………………………………………………………………….

………………………………………………………………………………………………….

………………………………………………………………………………………………….

………………………………………………………………………………………………….

………………………………………………………………………………………………….

………………………………………………………………………………………………….

………………………………………………………………………………………………….

………………………………………………………………………………………………….

# MỤC LỤC

[LỜI NHẬN XÉT CỦA GIẢNG VIÊN 2](#_Toc132846675)

[MỤC LỤC 3](#_Toc132846676)

[BẢNG PHÂN CÔNG CÔNG VIỆC 4](#_Toc132846677)

[YÊU CẦU BÀI LÀM 5](#_Toc132846678)

[Task 1 6](#_Toc132846679)

[1. Hồi quy tuyến tính đa biến 6](#_Toc132846680)

[2. Hồi quy phi tuyến đa biến 7](#_Toc132846681)

[3. Hồi quy Logistic 9](#_Toc132846682)

[Task 2 11](#_Toc132846683)

[1. Hồi quy tuyến tính đa biến 11](#_Toc132846684)

[1.1 Sử dụng MS Excel 11](#_Toc132846685)

[1.2 Sử dụng ngôn ngữ lập trình R 16](#_Toc132846686)

[1.3 Sử dụng ngôn ngữ lập trình Python 18](#_Toc132846687)

[2. Hồi quy phi tuyến đa biến 19](#_Toc132846688)

[2.1 Sử dụng MS Excel 19](#_Toc132846689)

[2.2 Sử dụng ngôn ngữ lập trình R 26](#_Toc132846690)

[2.3 Sử dụng ngôn ngữ lập trình Python 28](#_Toc132846691)

[3. Logistic Regression 34](#_Toc132846692)

[3.1 Sử dụng MS Excel 34](#_Toc132846693)

[3.2 Sử dụng ngôn ngữ lập trình R 39](#_Toc132846694)

[3.3 Sử dụng ngôn ngữ lập trình Python 40](#_Toc132846695)

[THAM KHẢO 42](#_Toc132846696)

# BẢNG PHÂN CÔNG CÔNG VIỆC

|  |  |  |  |
| --- | --- | --- | --- |
| **Thành viên**  **Công việc** | **Nguyen Cam Tu (Leader)** | **Nguyen Thi Thu Thuy** | **Nguyen Thanh Son** |
| Đọc và hiểu yêu cầu của bài báo cáo | ✓ | ✓ | ✓ |
| Xây dựng template bài báo cáo | ✓ | ✓ | ✓ |
| Tóm tắt và chỉnh sửa bài bảo cáo | ✓ | ✓ | ✓ |
| Làm bài tập Task 1.1 and Task 2.1 (Hồi quy tuyến tính đa biến và thực hiện trên Excel, lập trình R, lập trình Python) |  | ✓ |  |
| Làm bài tập Task 1.2 and Task 2.2 (Hồi quy phi tuyến đa biến và thực hiện trên Excel, lập trình R, lập trình Python) | ✓ |  |  |
| Làm bài tập Task 1.3 and Task 2.3 (Hồi quy logistic và thực hiện trên Excel, lập trình R, lập trình Python) |  |  | ✓ |
| Completion (%) | 100% | 100% | 100% |

# YÊU CẦU BÀI LÀM

|  |
| --- |
| **Lab 3**  **Task 1**  Explanation (What, How, and Why) and example of:   1. [Multivariable Linear Regression.](https://www.itl.nist.gov/div898/handbook/eda/section3/eda35a.htm) 2. Multivariable Nonlinear Regression 3. Logistic Regression   **Task 2**   1. Using MS Excel, R language, and Python language to perform **Multivariable Linear Regression** with data files: Colleges and Universities 2. Using MS Excel, R language, and Python language to perform **Multivariable Nonlinear Regression** with optional real data about/of Vietnam. 3. Using MS Excel, R language and Python language perform **Logistic Regression** with optional real data about/of Vietnam. |

# Task 1

Explanation (What, How, and Why) and example of:

## Hồi quy tuyến tính đa biến

1. ***Hồi quy tuyến tính là gì?***

Hồi quy tuyến tính là phép hồi quy xem xét mối quan hệ tuyến tính – dạng quan hệ đường thẳng giữa biến độc lập với biến phụ thuộc. Khi biểu diễn bằng đồ thị scatter trên mặt phẳng Oxy để thể hiện mối quan hệ độc lập – phụ thuộc, các điểm dữ liệu có xu hướng tạo thành một đường thẳng. Đường thẳng càng rõ ràng, mô hình hồi quy càng phù hợp, càng tốt; đường thẳng càng không rõ ràng, các điểm dữ liệu phân tán càng rải rác thì mô hình hồi quy tuyến tính càng không phù hợp.

Ảnh có chứa biểu đồ

Mô tả được tạo tự động

Trong nghiên cứu, chúng ta thường phải kiểm định các giả thuyết về mối quan hệ giữa hai hay nhiều biến, trong đó có một biến phụ thuộc và một hay nhiều biến độc lập. Nếu chỉ có một biến độc lập, mô hình được gọi là mô hình hồi quy đơn biến SLR (Simple Linear Regression). Trường hợp có từ hai biến độc lập trở lên, mô hình được gọi là hồi quy bội MLR (Multiple Linear Regression). Những nội dung tiếp theo ở tài liệu này chỉ đề cập đến hồi quy bội, hồi quy đơn biến tính chất tương tự với hồi quy bội.

* Phương trình hồi quy đơn biến: *Y = β0 + β1X +* ℇ
* Phương trình hồi quy bội: *Y = β0 + β1X1 + β2X2 + … + βnXn +* ℇ

Với Y là một biến phụ thuộc, là biến chịu tác động của biến khác

X là biến độc lập, là biến tác động lên biến khác

ꞵ0 = hằng số hồi quy, hay còn được gọi là hệ số chặn. Đây là chỉ số nói lên giá trị của Y sẽ là bao nhiêu nếu tất cả X cùng bằng 0. Nói cách khác, chỉ số này cho chúng ta biết giá trị của Y là bao nhiêu nếu không có các X. Khi biểu diễn trên đồ thị Oxy, β0 là điểm trên trục Oy mà đường hồi quy cắt qua.

ꞵn = hệ số hồi quy, hay còn được gọi là hệ số góc. Chỉ số này cho chúng ta biết về mức thay đổi của Y gây ra bởi X tương ứng. Nói cách khác, chỉ số này nói lên có bao nhiêu đơn vị Y sẽ thay đổi nếu X tăng hoặc giảm một đơn vị.

ℇ sai số. Chỉ số này càng lớn càng khiến cho khả năng dự đoán của hồi quy trở nên kém chính xác hơn hoặc sai lệch nhiều hơn so với thực tế. Sai số trong hồi quy tổng thể hay phần dư trong hồi quy mẫu đại diện cho hai giá trị, một là các biến độc lập ngoài mô hình, hai là các sai số ngẫu nhiên.

1. ***Hồi quy tuyến tính đa biến (Multiple Linear Regression) được sử dụng như thế nào?***

Để tìm đường phù hợp nhất cho mỗi biến độc lập, hồi quy đa tuyến tính tính toán ba điều:

* Các hệ số hồi quy dẫn đến lỗi mô hình tổng thể nhỏ nhất.
* Thống kê *t* của mô hình tổng thể.
* Giá trị p liên quan (khả năng thống kê t sẽ xảy ra tình cờ nếu giả thuyết null không có mối quan hệ giữa các biến độc lập và phụ thuộc là đúng).

 Hồi quy tuyến tính bội lập mô hình nhiều biến và tác động của chúng lên một kết quả:

* Lượng mưa, nhiệt độ và mức độ sử dụng phân bón đối với sản lượng cây trồng
* Chế độ ăn uống và tập thể dục đối với bệnh tim
* Tăng trưởng tiền lương và lạm phát trên lãi suất cho vay gia đình

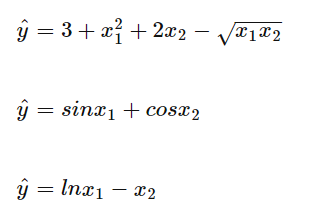
1. ***Tại sao chúng ta nên sử dụng hồi quy tuyến tính đa biến?***

* Phân tích đa biến: Hồi quy tuyến tính đa biến cho phép ta phân tích tương quan giữa nhiều biến độc lập và biến phụ thuộc một cách đồng thời. Điều này làm cho nó trở thành một phương pháp phân tích đa biến hiệu quả.
* Tính linh hoạt: Hồi quy tuyến tính đa biến có thể được sử dụng để phân tích nhiều biến độc lập và biến phụ thuộc khác nhau. Điều này cho phép ta thực hiện phân tích cho nhiều loại dữ liệu khác nhau, từ dữ liệu y học đến dữ liệu kinh tế.
* Độ chính xác: Hồi quy tuyến tính đa biến cung cấp các giá trị hệ số và giá trị p cho từng biến độc lập, giúp ta đánh giá mức độ ảnh hưởng của từng biến đến biến phụ thuộc. Điều này giúp ta xác định được mức độ ảnh hưởng của từng biến và đưa ra các quyết định phù hợp.

## Hồi quy phi tuyến đa biến

1. ***Hồi quy phi tuyến đa biến là gì?***

Thuật toán tìm ra phương trình giả thuyết của các điểm dữ liệu không phải dạng tuyến tính gọi là thuật toán Nonlinear Regression (Hồi quy phi tuyến tính).

Ví dụ các phương trình phi tuyến tính:

Mô hình hồi qui tuyến tính đa biến có dạng:

Trong phân tích hồi qui tuyến tính đa biến, ta cần biết mức độ ảnh hưởng của từng yếu tố lên biến kết cục y. Muốn biết mức độ ảnh hưởng cần lưu ý đến các trị số sau:

1. Hệ số tương quan R (coefficient of correlation): yếu tố nào có R càng lớn thì ảnh hưởng càng nhiều.
2. Bình phương của R (R square): yếu tố nào có R 2 càng lớn thì mối quan hệ giữa yếu tố đó và biến y càng chặt chẽ.
3. Hệ số hồi qui β (regression coefficient): yếu tố nào có β cao thì ảnh hưởng nhiều hơn, tuy nhiên các yếu tố có đơn vị khác nhau nên không thể so sánh mức ảnh hưởng giữa các yếu tố. Nếu muốn so sánh phải đổi các yếu tố có cùng đơn vị là độ lệch chuẩn, lúc đó ta có hệ số hồi qui chuẩn hóa:

(Với Sx là độ lệch chuẩn của x tương ứng và Sy là độ lệch chuẩn của y.)

1. Trị số p (p value): càng nhỏ mức ảnh hưởng càng mạnh.
2. ***Làm thế nào để sử dụng Hồi quy phi tuyến đa biến?***

Điều kiện để thực hiện hồi quy phi tuyến tính là các biến X và Y tuân theo phân phối bình thường. chúng ta có hai cách để có thể tính toán:

1. Chuyển đổi biến và đưa nó về hồi quy tuyến tính bội, sau đó sử dụng Công cụ Regression trong Data Analysis.

2. Vẽ biểu đồ phân tán và tìm các đường ngoại suy với Add TrendLine.

1. ***Tại sao chúng ta nên sử dụng Hồi quy phi tuyến đa biến?***

Hồi quy phi tuyến:

* Cung cấp tính linh hoạt cao hơn để phù hợp với các đường cong so với hồi quy tuyến tính.
* Nhưng Hồi quy phi tuyến tính sử dụng thuật toán lặp thay vì giải trực tiếp bằng các phương trình ma trận như Hồi quy tuyến tính.

Hơn nữa, các mô hình hồi quy phi tuyến có thể nhận hầu như vô số các hình thức:

* Không hạn chế về cách bạn có thể sử dụng các tham số trong phương trình hồi quy phi tuyến.
* Tính linh hoạt này cung cấp hồi quy phi tuyến với khả năng điều chỉnh đường cong linh hoạt nhất có thể.
* Nhưng không thể có một bài kiểm tra giả thuyết duy nhất cho tất cả các tham số.
* Giá trị giả thuyết H0 của mỗi tham số phụ thuộc vào hàm phi tuyến, vị trí của tham số trong đó.

1. ***Ví dụ***

Sử dụng Hồi qui phi tuyến tính đa biến để tìm một phương trình phù hợp cho tập dữ liệu. Bài toán là áp dụng một tập dữ liệu của Việt Nam, cụ thể là Tiki Books, để trả lời câu hỏi liệu tổng số lượng sách bán ra có bị ảnh hưởng bởi các yếu tố đo lường (Task2.2)

## Hồi quy Logistic

1. ***Hồi quy Logistic là gì?***

Hồi quy logistic là một loại hồi quy trong đó biến độc lập được sử dụng để dự đoán biến phụ thuộc. Khi biến phụ thuộc có hai loại, thì đó là một hồi quy logistic nhị phân. Khi biến phụ thuộc có nhiều hơn hai loại, thì đó là một hồi quy logistic đa thức. Khi danh mục biến phụ thuộc được xếp hạng, thì đó là một hồi quy logistic thứ tự (ordinal logistic regression - OLS). Để có được ước tính khả năng xảy ra tối đa, hãy biến đổi biến phụ thuộc trong hàm logit. Logit về cơ bản là một bản ghi tự nhiên của biến phụ thuộc và cho biết liệu sự kiện có xảy ra hay không. Hồi quy logistic thông thường không giả định mối quan hệ tuyến tính giữa biến phụ thuộc và biến độc lập. Nó không giả định tính đồng biến.

Trong hồi quy tuyến tính đơn, biến độc lập x và biến phụ thuộc y là biến số liên tục liên hệ qua phương trình: y = α + βx + ε

1. ***Làm thế nào để sử dụng Hồi quy Logistic?***

Điều kiện để sử dụng mô hình hồi quy Logistic là:

* Biến đầu ra (outcome) có giá trị là 1-0 hoặc có-không
* Biến đầu vào (input) là biến phân loại hoặc biến liên tục.

1. ***Tại sao chúng ta nên sử dụng Hồi quy Logistic?***

Mô hình hồi quy Logistic được ứng dụng rất rộng rãi như: phân tích dự báo, trong lĩnh vực học máy – Machine learning.

Các ví dụ về ứng dụng mô hình hồi quy Logistic trong các lĩnh vực đời sống:

● Trong lĩnh vực y tế: Điểm số mức độ nghiêm trọng của chấn thương và thương tật (TRISS).

● Trong lĩnh vực chính trị: Hồi quy logistic cũng có thể được sử dụng để dự đoán các cuộc bầu cử.

● Thử nghiệm sản phẩm: Hồi quy logistic có thể được sử dụng trong kỹ thuật để dự đoán sự thành công hay thất bại của một hệ thống đang được thử nghiệm hoặc một sản phẩm nguyên mẫu.

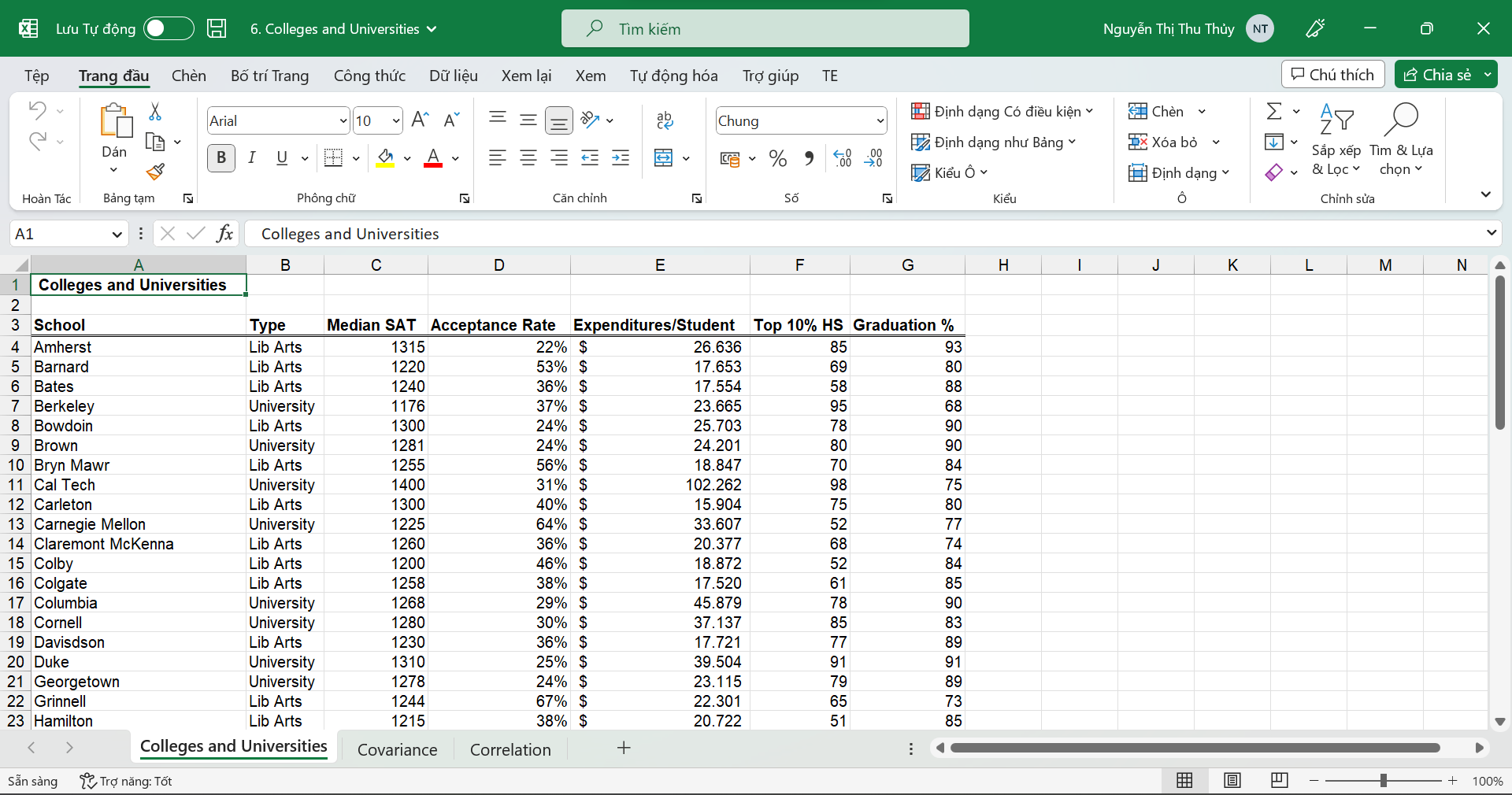
● Trong lĩnh vực tiếp thị: có thể được sử dụng để dự đoán khả năng yêu cầu của khách hàng chuyển thành bán hàng, khả năng bắt đầu, chấm dứt đăng ký hoặc thậm chí là sự quan tâm của khách hàng tiềm năng đối với một dòng sản phẩm mới.

# Task 2

## Hồi quy tuyến tính đa biến

### Sử dụng MS Excel

Sử dụng file dữ liệu “Colleges and Universities”, dự đoán tỷ lệ tốt nghiệp(Gradution%) dựa trên MedianSAT, Acceptance Rate, Expenditures/Student, Top 10% HS là có hợp lý hay không?



* Biểu đồ phân tán được dùng để so sánh ít nhất hai tập giá trị hoặc cặp dữ liệu, để thể hiện mối quan hệ giữa các tập giá trị.

Ảnh có chứa biểu đồ

Mô tả được tạo tự động

* Sử dụng chức năng Data Analysis trên thanh công cụ:

Ảnh có chứa bàn

Mô tả được tạo tự động

Graduation% = 17.92 + 0.072 SAT - 24.859 ACCEPTANCE

- 0.000136 EXPENDITURES

- 0.163 TOP10% HS

Với Graduation%: phần trăm tỉ lệ tốt nghiệp

Median SAT: điểm trung bình SAT

Acceptance Rate: tỉ lệ chấp thuận

Expenditures/Student: chi phí của mỗi sinh viên

Top 10% HS: 10% điểm cao nhất trong các sinh viên

Chọn 2 biến phụ thuộc x1, x2 và 1 biến độc lập y

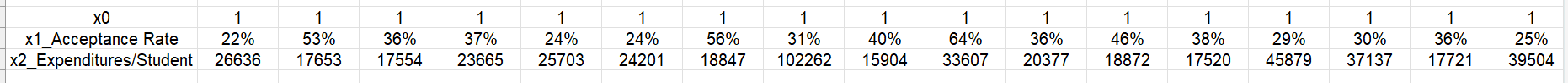
Ảnh có chứa bàn

Mô tả được tạo tự động

**Bước 1:** Tính ma trận nghịch đảo XT thông qua “Hoán đổi” trong Excel

Ảnh có chứa bàn

Mô tả được tạo tự động



**Bước 2:** Tính ma trận XT \* X bằng hàm MMULT

Trên thanh công cụ nhấn chọn “Công thức” 🡪 “Chèn hàm” 🡪 Chọn “MMULT”

Ảnh có chứa bàn

Mô tả được tạo tự động

Chọn vùng cần tính toán và nhấp chọn OK

Ta được kết quả như hình

Ảnh có chứa bàn

Mô tả được tạo tự động

**Bước 3:** Tính ma trận nghịch đảo (XT \* X) -1 bằng hàm MINVERSE

Trên thanh công cụ vào “Công thức”, nhấp chọn “ Chèn hàm” 🡪 MINVERSE

Ảnh có chứa văn bản

Mô tả được tạo tự động

Chọn vùng cần tính toán và nhấp chọn OK

Ảnh có chứa bàn

Mô tả được tạo tự động

Ta được kết quả:

Ảnh có chứa bàn

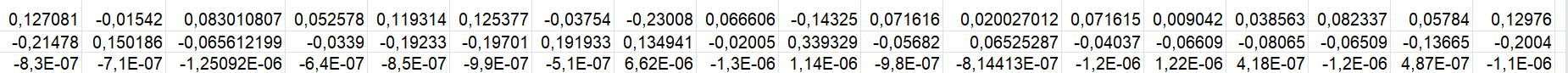
Mô tả được tạo tự động

**Bước 4:** Tính toán ma trận (XT \* X) -1 \* XT  bằng hàm MMULT

Ảnh có chứa bàn

Mô tả được tạo tự động

Ta được kết quả sau;



**Bước 5:** Tính ma trận (XT \* X) -1 \* XT  \* Y bằng hàm MMULT



Kết luận: kết quả tính thủ công giống kết quả từ chức năng Data Analysis

### Sử dụng ngôn ngữ lập trình R

**Bước 1:** Đọc file dữ liệu

Ảnh có chứa bàn

Mô tả được tạo tự động

**Bước 2:** Sử dụng hàm lm để tính hồi quy tuyến tính đa biến. Với reg1 = lm(Y~x1+x2+…)

Ảnh có chứa văn bản

Mô tả được tạo tự động

Ta thấy tại “Coefficients:”, giá trị Pr(>|t|) của Expenditures/Student > 0.05. Vì thế, chúng ta sẽ thực hiện việc lựa chọn lại mô hình loại đi Expenditures/Student xem kết quả có khả quan hơn hay không?

**Bước 3:** Tính toán lại

Ảnh có chứa văn bản

Mô tả được tạo tự động

Ta thấy độ tương quan R-squared, Adjusted R-squared của mô hình sau phù hợp hơn mô hình trước.

* Graduation = 94.926 – 30.658\*AcceptanceRate

**Bước 4:** Trực quan hóa bằng đồ thị

Ảnh có chứa văn bản

Mô tả được tạo tự động

Ảnh có chứa biểu đồ

Mô tả được tạo tự động

### Sử dụng ngôn ngữ lập trình Python

**Bước 1:** Sử dụng các thư viện và đọc file

Ảnh có chứa văn bản

Mô tả được tạo tự động

Ảnh có chứa bàn

Mô tả được tạo tự động

**Bước 2:** Tạo mảng x gồm nhiều biến và y

Ảnh có chứa Website

Mô tả được tạo tự động

**Bước 3:** Sử dụng hàm LinearRegression() và fit x và y



**Bước 4:** Đưa ra kết quả

Ảnh có chứa văn bản

Mô tả được tạo tự động

## Hồi quy phi tuyến đa biến

### Sử dụng MS Excel

Đầu tiên, chúng ta tiến hành trích xuất 29 dòng dữ liệu Tiki Books Dataset từ nguồn [Tiki Books Dataset | Kaggle](https://www.kaggle.com/datasets/biminhc/tiki-books-dataset?select=book_data.csv).

Sử dụng Hồi qui phi tuyến tính đa biến để tìm một phương trình phù hợp cho tập dữ liệu. Bài toán là áp dụng một tập dữ liệu của Việt Nam, cụ thể là Tiki Books, để trả lời câu hỏi liệu tổng số lượng sách bán ra có bị ảnh hưởng bởi các yếu tố đo lường. Trong trường hợp này, ta chọn các yếu tố đo lường là số lượt bình phẩm (n\_review) và trung bình tỉ lệ đánh giá quyển sách đó (avg\_rating).

*Đặt ra câu hỏi:* Liệu tổng số lượng sách bán ra có bị ảnh hưởng bởi số lượt bình phẩm và trung bình tỉ lệ đánh giá hay không?

Table

Description automatically generated with medium confidence Quantity = α + β1 \* n\_review + β2 \* avg\_rating + ε

Với:

book\_id: Mã sách

quantity: tổng số lượng sách được bán ra

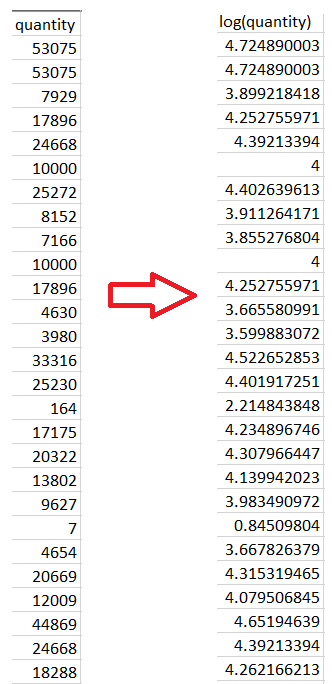
n\_review: số lượt bình phẩm

avg\_rating: trung bình đánh giá (được tính trên thang điểm 1-5)

**Bước 1:** Chúng ta chọn 2 biến phụ thuộc x1, x2 và 1 biến độc lập y.

Table

Description automatically generated with low confidence2 biến phụ thuộc: x1, x2

1 biến độc lập: y (quantity) sau đó đặt biến y = log(y)

**Bước 2:** Sử dụng công cụ Data Analysis của Excel

A picture containing timeline

Description automatically generated

Nhấn OK để có được bảng như sau:

Graphical user interface, application, table

Description automatically generated

Từ bảng kết quả, ta có những giá trị như sau:

+ R square = 0.802 => 89.04% dữ liệu phù hợp với mô hình.

+ Adjusted R Square = 0.786 => R square > 78.6% nghiên cứu được đánh giá tốt.

Bởi vì:

+ p-value of Intercept = 0.031 < 0.05

+ p-value of X1 = 0.00017 < 0.05

+ p-value of X2 = 3.097x10-8 < 0.05

* Mô hình dự đoán phù hợp với bài toán.

+ Log(quantity) = 0.8155 + 0.0001 \* n\_review + 0.6063 \* avg\_rating + ε

hay quantity = ε0.08155+0.0001\*n\_review+0.6063\*avg\_rating+ε

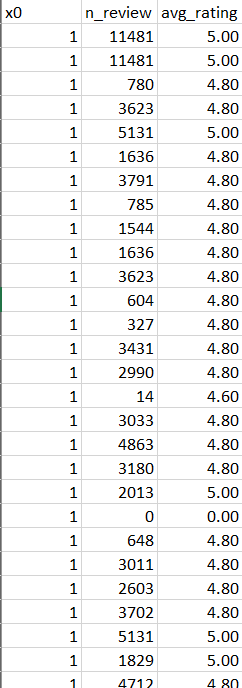
***Làm sao để tính Coefficient:***

Đầu tiên, chúng ta chọn ma trận bao gồm x1, x2 và y:

Table

Description automatically generated with medium confidence

Thêm cột x0:



**Bước 1:** Tính XT với phương thức Transpose trong Excel



**Bước 2:** Tính ma trận XT\*X bằng phương thức MMULT trong Excel



**Bước 3:** Tính ma trận (XT\*X)-1 bằng phương thức MINVERSE trong Excel



**Bước 4:** Tính ma trận (XT\*X)-1\*XT bằng phương thức MMULT trong Excel



**Bước 5:** Tính ma trận (XT\*X)-1XT\*Y bằng phương thức MMULT trong Excel

*Lưu ý:* Y ở đây là cột log(quantity) đã tính ở trên

Text

Description automatically generated

Table

Description automatically generatedGiá trị tại bảng Regression

So sánh giá trị vừa tính với giá trị tại bảng Regression, ta thấy được giá trị thu được giống nhau.

### Sử dụng ngôn ngữ lập trình R

**Bước 1:** Đọc dữ liệu

Table

Description automatically generated

**Bước 2:** Sử dụng chức năng lm() để thực hiện hồi quy phi tuyến đa biến

Attach(lab03) => Sử dụng câu lệnh lm() và đặt biến y = log(y)

Text

Description automatically generated

Sau khi tính toán xong, ta có được kết quả:

*1.878e+00 = 5.104933274*

*2.533e-04 = 2.885407871*

*1.396e+00 = 3.794721433*

Log(*quantity*) = 5.105 + 2.885\**n\_review* + 3.795\**avg\_rating* + e

*hay*

*Quantity* = e5.105 + 2.885\**n\_review* + 3.795\**avg\_rating*+e

### Sử dụng ngôn ngữ lập trình Python

**Bước 1:** Thêm thư viện cần thiết và đọc tập dữ liệu .csv

A picture containing graphical user interface

Description automatically generated

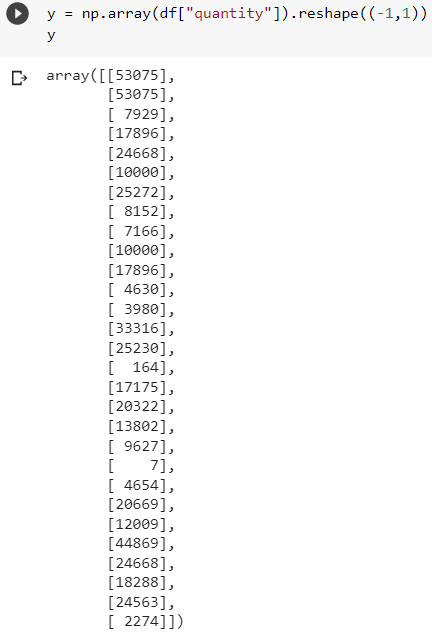
**Bước 2:** Ghi nhận 2 biến phụ thuộc là số lượt bình phẩm (n\_review) và trung bình bình chọn (avg\_rating); 1 biến độc lập là tổng số lượng sách bán ra (quantity).

2 biến phụ thuộc: x

Table

Description automatically generated

1 biến độc lập: y



**Bước 3:** Đặt biến y = log(y)

A picture containing table

Description automatically generated

Table

Description automatically generated

**Bước 4:** Sử dụng thuật toán Dự đoán tuyến tính (Linear Regression) để đạt được mô hình theo biến X và Y.

Graphical user interface, text, application

Description automatically generated

**Bước 5:** Lấy kết quả từ mô hình trên

Graphical user interface, text, application, email, Teams

Description automatically generated

**Bước 6:** Tạo bảng dự đoán

Table

Description automatically generated

## Logistic Regression

### Sử dụng MS Excel

Trích xuất 180 dòng dữ liệu về thời tiết của Hà Nội

Sử dụng Hồi qui logistic để tìm một phương trình phù hợp cho tập dữ liệu. Bài toán là áp dụng một tập dữ liệu của Việt Nam, cụ thể là VietNam weather data, để xem xét tác động của mây (%) đối với mưa ở Việt Nam.

**Bước 1:** Sử dụng công cụ Data Analysis của Excel

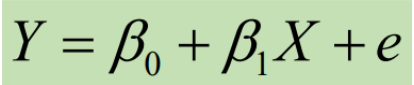
Graphical user interface, application, table, Excel

Description automatically generated

Graphical user interface, application, table, Excel

Description automatically generated

**Bước 2:** Sau khi có giá trị của Intercept và giá trị của biến X (cloud), chúng ta bắt đầu tính: linear score, probability, likelihood,...



Graphical user interface, application, table, Excel

Description automatically generated

**Bước 3:** Tính Probability:

A picture containing text

Description automatically generated

Graphical user interface, application, table, Excel

Description automatically generated

**Bước 4:** Tính Likelihood với hàm IF()

Graphical user interface, application, table, Excel

Description automatically generated

**Bước 5:** Log Likelihood = LOG(likelihood)

Graphical user interface, application, table, Excel

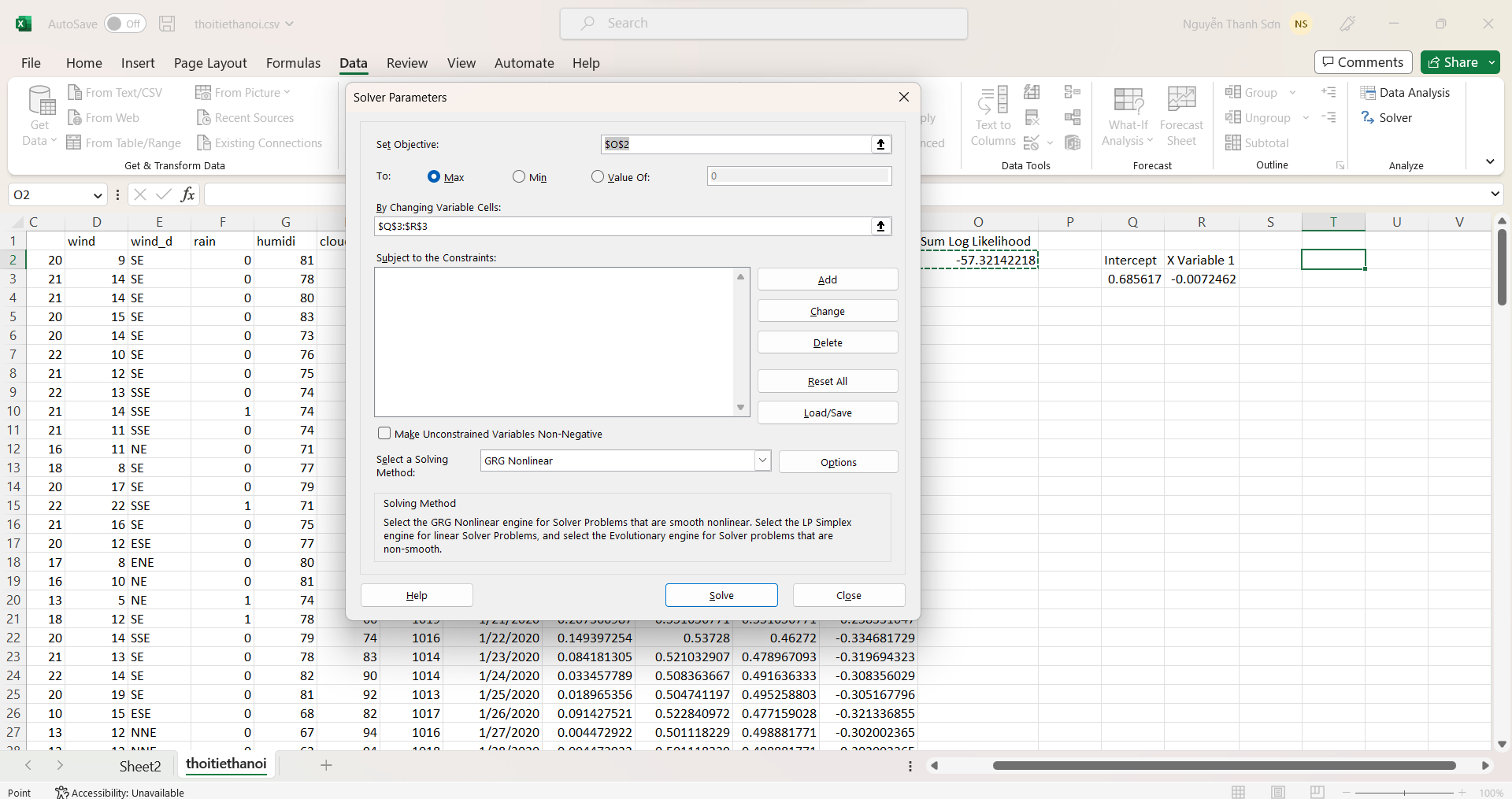
Description automatically generated

**Bước 6:** Tính Sum Log Likelihood

Graphical user interface, application, table, Excel

Description automatically generated

**Bước 7:** dùng Solver Parameters



**Bước 8:** ta được các hệ số đệ quy

Graphical user interface, application, table, Excel

Description automatically generated

### Sử dụng ngôn ngữ lập trình R

**Bước 1:** Nhập dữ liệu

Table

Description automatically generated

**Bước 2:** attach dữ liệu

Company name

Description automatically generated with medium confidence

**Bước 3:** Sử dụng hàm glm để thực hiện hồi quy logistic, với biến phụ thuộc là mưa và biến độc lập là mây.

Text

Description automatically generated

### Sử dụng ngôn ngữ lập trình Python

**Bước 1:** Cài đặt thư viện

A picture containing graphical user interface

Description automatically generated

**Bước 2:** Nhập dữ liệu

Table

Description automatically generated

**Bước 3:** đọc dữ liệu vào mảng numpy, chúng ta phải thay đổi dữ liệu biến độc lập (cloud) thành mảng 2 chiều để phù hợp với điều kiện đầu vào của thuật toán

A picture containing text

Description automatically generated

**Bước 4:** tạo model bằng hàm LogisticRegression, sau đó train bằng lệnh fit với x và y

Graphical user interface, text, application

Description automatically generated

**Bước 5:** In kết quả

Graphical user interface, application

Description automatically generated

# THAM KHẢO

[1]"Ước lượng hồi quy tuyến tính bằng OLS," Xulydinhluong, [Ước lượng hồi quy tuyến tính bằng OLS (xulydinhluong.com)](https://xulydinhluong.com/uoc-luong-hoi-quy-tuyen-tinh-bang-ols/#:~:text=Ph%C6%B0%C6%A1ng%20tr%C3%ACnh%20h%E1%BB%93i%20quy%20%C4%91%C6%A1n%20bi%E1%BA%BFn%3A%20Y%20%3D,%2B%20%CE%B22X2%20%2B%20%E2%80%A6%20%2B%20%CE%B2nXn%20%2B%20e) (accessed Apr. 20, 2023).

[2] <https://websitehcm.com/phan-tich-nonlinear-regression-trong-r/> (accessed Apr. 20, 2023)

[3] <https://bvag.com.vn/wp-content/uploads/2013/01/k2_attachments_PHAN-TICH-HOI-QUY-TUYEN-TINH-DA-BIEN.pdf> (accessed Apr. 20, 2023)

[4] “EXCEL SOLVER - HOW TO LOAD OR START SOLVER”, [Excel Solver - How to load or start Solver | solver](https://www.solver.com/excel-solver-how-load-or-start-solver) (accessed Apr. 20, 2023)

[5] Nguyen Minh Nhut, “LAB03\_THONGKEHOIQUY\_SLIDES-1.pdf”