

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**



**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA HỆ THỐNG THÔNG TIN**

****

**BÀI TẬP THỰC HÀNH**

**MÔN HỌC: PHÂN TÍCH DỮ LIỆU KINH DOANH**

**LAB 03**

*Giảng viên hướng dẫn:*

***PGS.TS Nguyễn Đình Thuân***

*Sinh viên thực hiện:*

|  |  |
| --- | --- |
| ***Nguyễn Hoàng Long*** | ***- 19521788*** |
| ***Phan Phạm Quỳnh Hoa***  ***Huỳnh Minh Thư*** | ***- 19521520***  ***- 19522304*** |

*TP HCM, Ngày 20 tháng 04 năm 2022*

LỜI CẢM ƠN

Em xin gửi lời cảm ơn chân thành và sự tri ân sâu sắc đối với các thầy cô của trường Trường Đại học Công nghệ thông tin – Đại học Quốc gia TP.HCM, đặc biệt là quý thầy cô khoa Hệ thống thông tin của trường đã giúp cho chúng em trang bị các kiến thức cơ bản, các kỹ năng thực tế và tạo điều kiện để chúng em có thể hoàn thành đồ án môn học của mình.

Đặc biệt chúng em xin chân thành cảm ơn thầy Nguyễn Đình Thuân đã nhiệt tình hướng dẫn hướng dẫn, quan tâm truyền đạt những kiến thức và kinh nghiệm, trực tiếp hướng dẫn tận tình, sửa chữa và đóng góp ý kiến quý báu cho chúng em trong suốt thời gian học tập để chúng em có thể hoàn thành tốt môn học này.

Trong thời gian thực hiện, nhóm tác giả đã vận dụng những kiến thức nền tảng đã tích lũy. Từ đó, nhóm tác giả vận dụng tối đa những gì đã thu thập được để hoàn thành bài tập thực hành một cách tốt nhất. Tuy nhiên, trong quá trình thực hiện, nhóm tác giả không tránh khỏi những thiếu sót. Chính vì vậy, nhóm tác giả rất mong nhận được những sự góp ý từ phía Thầy/Cô nhằm hoàn thiện những kiến thức mà nhóm tác giả đã học tập và là hành trang để nhóm tác giả thực hiện tiếp các đề tài khác trong tương lai.

Nhóm chúng em xin chân thành cảm ơn!

***Nhóm thực hiện***

**NHẬN XÉT CỦA GIẢNG VIÊN**

………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………

**MỤC LỤC**

[I. Thực hiện bài tập 5](#_Toc101562856)

[3a. Giải thích và ví dụ minh họa về Hồi qui tuyến tính đa biến 5](#_Toc101562857)

[3b. Giải thích và ví dụ minh họa về Hồi qui phi tuyến đa biến 5](#_Toc101562858)

[3c. Giải thích và ví dụ minh họa về Hồi qui logistic 6](#_Toc101562859)

[4d. Hồi quy tuyến tính với tập tin dữ liệu Colleges and Universities 7](#_Toc101562860)

[Phát biểu bài toán 7](#_Toc101562861)

[1. Sử dụng Excel 7](#_Toc101562862)

[2. Sử dụng R 14](#_Toc101562868)

[3. Sử dụng Python](#_Toc101562869)

[4e. Hồi quy phi tuyến đa biến với dữ liệu data housing của Việt Nam. 21](#_Toc101562870)

[Phát biểu bài toán 21](#_Toc101562871)

[1. Sử dụng Excel](#_Toc101562872)

[2. Sử dụng R](#_Toc101562873)

[3. Sử dụng Python](#_Toc101562874)

[4f. Hồi quy Logistic với dữ liệu Bank Customer Data in VietNam. 29](#_Toc101562875)

[Phát biểu bài toán 29](#_Toc101562876)

[4. Sử dụng Excel](#_Toc101562877)

[5. Sử dụng R](#_Toc101562878)

[6. Sử dụng Python](#_Toc101562879)

[II. Bảng phân công công việc](#_Toc101562880)

[III. Tài liệu tham khảo](#_Toc101562881)

1. **Thực hiện bài tập**

**3a. Giải thích và ví dụ minh họa về Hồi qui tuyến tính đa biến**

**Hồi quy tuyến tính đa biến** là một phần mở rộng của hồi quy tuyến tính đơn giản. Nó được sử dụng khi chúng ta muốn dự đoán giá trị của một biến dựa trên giá trị của hai hoặc nhiều biến khác. Biến chúng ta muốn dự đoán được gọi là biến phụ thuộc (hoặc đôi khi, biến kết quả, mục tiêu hoặc biến tiêu chí). Các biến chúng ta đang sử dụng để dự đoán giá trị của biến phụ thuộc được gọi là biến độc lập

Ví dụ:

Thu nhập, địa điểm sinh sống và số thành viên trong gia đình ảnh hưởng đến chi tiêu.

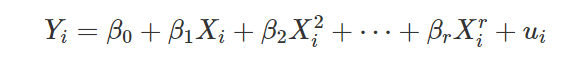
=> Biến độc lập:Thu nhập, địa điểm, số thành viên

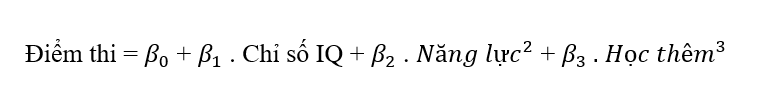
=> Biến phụ thuộc: Chi tiêu

**3b. Giải thích và ví dụ minh họa về Hồi qui phi tuyến đa biến**

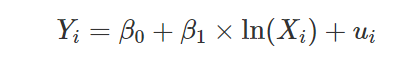
**Hồi quy phi tuyến** là một dạng phân tích hồi quy trong đó dữ liệu quan sát được mô hình hóa bằng một hàm là một sự kết hợp phi tuyến của các tham số mô hình và biến phụ thuộc vào nhiều biến độc lập. Các hàm phi tuyến như hàm đa thức, hàm logarithm, etc

1. Hàm đa thức



Ví dụ: 

2. Hàm logarithm



Ví dụ:

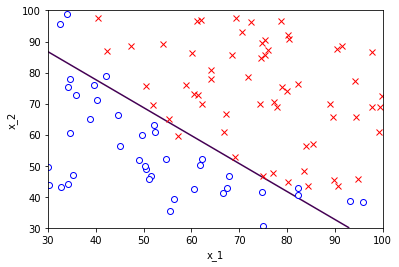
Ảnh có chứa văn bản

Mô tả được tạo tự động

**3c. Giải thích và ví dụ minh họa về Hồi qui logistic**

**Hồi quy logistic** là một mô hình hồi quy nhằm dự đoán giá trị đầu ra rời rạc (discrete target variable) y ứng với một véc-tơ đầu vào x. Việc này tương đương với chuyện phân loại các đầu vào x vào các nhóm y tương ứng

Ví dụ: Xem một bức ảnh có chứa một con mèo hay không. Thì ở đây ta coi đầu ra y=1 nếu bức ảnh có một con mèo và y=0 nếu bức ảnh không có con mèo nào. Đầu vào x ở đây sẽ là các pixel một bức ảnh đầu vào

[](https://res.cloudinary.com/dominhhai/image/upload/ml/logistic-regression_ex2_ret_1.png)

**4d. Hồi quy tuyến tính với tập tin dữ liệu** **Colleges and Universities**

**Phát biểu bài toán**

* Ý nghĩa nguồn dữ liệu: Một số thông tin về Colleges và Universities
* Phát biểu bài toán: Với độ tin cậy 95% có thể tìm ra mối liên hệ giữa Median SAT với Acceptance Rate, Expenditures, Top 10% HS, Graduation hay không
* Đặt giả thuyết
  + H0: Phương trình hồi qui không phù hợp.
  + H1: Phương trình hồi qui phù hợp
* Confident level = 95% tức α = 0.05

1. **Sử dụng Excel**

Tập dữ liệu Colleges and Universities

**Table

Description automatically generated**

* Giả sử phương trình hồi quy tuyến tính đa biến có dạng:

MedianSAT = b0 + b1AcceptanceRate + b2Expenditures + b3Top10%HS + b4Graduation

Hay y = b0 + b1x1 + b2x2 +b3x3 +b4x4

* Với các ma trận:

Ta có các ma trận tương ứng với tập dữ liệu

* Ma trận X bao gồm 5 cột, 49 hàng

Table

Description automatically generated

* Ma trận B bao gồm 1 cột, 4 hàng
* Ma trận Y bao gồm 1 cột, 49 hàng

Table

Description automatically generated with low confidence

Ta có hệ phương trình trước được viết lại:

XT.X.B = XT.Y

B=( XT.X)-1 XT.Y

* Chuyển vị của ma trận X, XT bao gồm 49 cột, 5 hàng

Table

Description automatically generated

* Trong excel sử dụng hàm MMULT(array1, array2) để nhân hai ma trận

Nhân hai ma trận XT, X thu được kết quả ma trận gồm 5 hàng, 5 cột

Table

Description automatically generated

* Trong excel sử dụng hàm MINVERSE(array) để nghịch đảo một ma trận

Kết quả của phép tính ( XT.X)-1 là một ma trận gồm 5 hàng, 5 cột

A screenshot of a computer

Description automatically generated with low confidence

* Kết quả của phép tính ( XT.X)-1 XT là một ma trận gồm 49 hàng, 5 cột

A screenshot of a computer

Description automatically generated with low confidence

* Kết quả của phép tính B=( XT.X)-1 XT.Y là một ma trận gồm 1 cột, 5 hàng

Table

Description automatically generated

* Vậy ta có được phương trình hồi quy:

**Y** = 887.78099 - 74.93338\***X1** + 0.0018198\***X2** + 0.5449401\***X3** +3.7087128\***X4**

Hay

**MedianSAT** = 887.78099 - 74.93338\***AcceptanceRate** + 0.0018198\***Expenditures** + 0.5449401\***Top10%HS** + 3.7087128\***Graduation**

* Phép kiểm định Fisher
* Giả thuyết:

H0: Phương trình hồi qui không phù hợp.

H1: Phương trình hồi qui phù hợp.

* Áp dụng các công thức sau:



* SSR: Tổng bình phương biến thiên độ lệch tiên lượng và giá trị trung bình
* SSE: Tổng bình phương biến thiên độ lệch tiên lượng và quan sát
* SST: Tổng bình phương biến thiên độ lệch quan sát và giá trị trung bình.
* MSR(mean square regression)

Text

Description automatically generated

* MSE (Mean Square Error) – trung bình bình phương sai số dự báo

Text

Description automatically generated

Logo

Description automatically generated

Bác bỏ H0 nếu Significance F < α

* Thực hiện tính toán trên Excel:

Table

Description automatically generated

Table

Description automatically generated

* Sử dụng hàm FDIST(F,hệ số độc lập, n-hệ số độc lập-1) để tính giá trị Significance F.

Graphical user interface

Description automatically generated with medium confidence

Kết quả:

Graphical user interface, application

Description automatically generated with medium confidence

* Ta thấy Significance F < α
* Vậy ta bác bỏ H0.
* Sử dụng công cụ Data Analysis trên Excel kiểm tra kết quả
* Data -> Data Analyst -> chọn Regression

Graphical user interface, application

Description automatically generated

* Hộp thoại Regression xuất hiện

Graphical user interface, application

Description automatically generated

Input Y Range: chọn vùng dữ liệu cột Median SAT.

Input Y Range: chọn vùng dữ liệu cột Acceptance Rate, Expenditures, Top 10% HS, Graduation.

Confidence Level: chọn độ tin cậy là 95%.

* Thu được kết quả như hình:

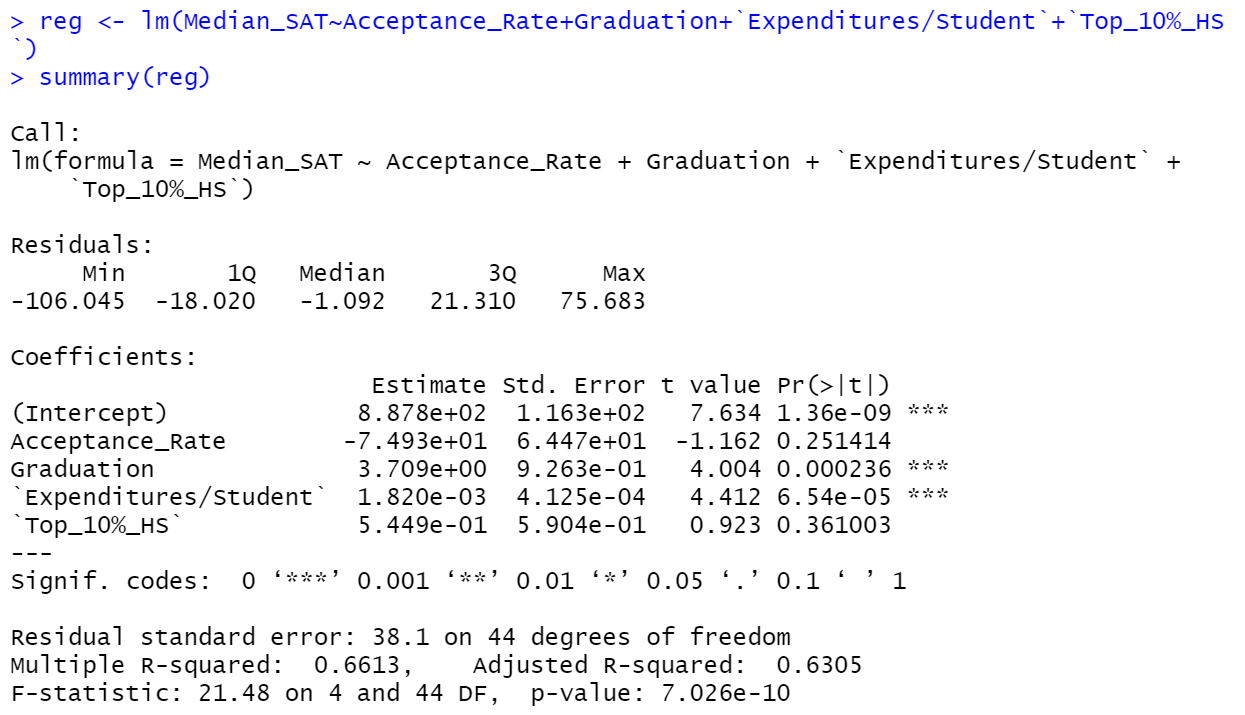
Table

Description automatically generated

5. **Sử dụng R**

Dùng ngôn ngữ R thực hiện các phép tính Hồi quy với tập tin dữ liệu Colleges and Universities

* Sử dụng hàm lm(y~x) để biểu diễn mô hình hồi quy tuyến tính linear model.



* Nhìn vào Coefficients ta thấy Pr(>|t|) cột Top 10% HS = 0.36 và cột Acceptance Rate P = 0.25 > 0.05. Nên ta sẽ chọn lại mô hình và loại 2 cột này.
* Thực hiện lại mô hình với 2 cột Graduation và cột Expenditures/Student.

Text

Description automatically generated

* Nhận sét về độ tương quan R-squared, Adjusted R-squared ta thấy mô hình sau là phù hợp:

**MedianSAT** = 817.3 + 0.002228\***Expenditures** + 4.550 \* **Graduation**

* + - **Kết luận:** Bác bỏ H0.
* Dùng hàm **plot(y~x)** để thực hiện mô hình hồi quy tuyến tính với giá trị Expenditures và Graduation

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

* Dùng hàm **abline(lm(y~x))** biểu diễn phương trình hồi qui MedianSAT và Expenditures/Student, Graduation

Chart, scatter chart

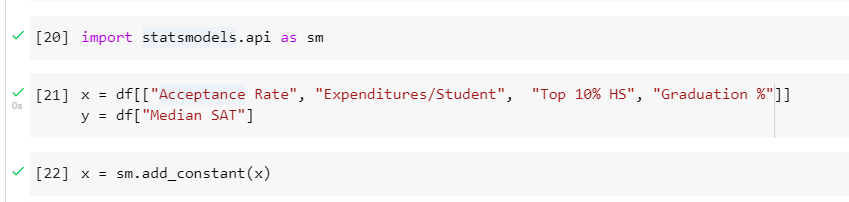
Description automatically generated

Chart, scatter chart

Description automatically generated

1. **Sử dụng Python**

**Bước 1:** Chuẩn bị dữ liệu



**Bước 2:** Fit model và xem kết quả

Ảnh có chứa văn bản

Mô tả được tạo tự động

Ta thấy được:

1. R-squared = 0.661 -> Phương trình đáp ứng được 66.1% hay Acceptance Rate, Expenditures, Top 10% HS, Graduation giải thích được 66.1% Median SAT
2. Prob (F-statistic) = 7.03e-10 < α = 0.05 -> bác bỏ H0

## **4e. Hồi quy phi tuyến đa biến với dữ liệu data housing của Việt Nam.**

### **Phát biểu bài toán**

* Nguồn dữ liệu: https://www.kaggle.com/datasets/ladcva/vietnam-housing-dataset-hanoi
* Ý nghĩa nguồn dữ liệu: Một số thông tin về giá nhà
* Phát biểu bài toán: Với độ tin cậy 95% có thể tìm ra mối liên hệ giữa price với num\_bed\_rooms, Squared\_meter\_area hay không
* Đặt giả thuyết
  + H0: Phương trình hồi qui không phù hợp.
  + H1: Phương trình hồi qui phù hợp
* Confident level = 95% tức α = 0.05

1. **Sử dụng Excel**

Tập dữ liệu data\_housing

**Table

Description automatically generated**

* Giả sử phương trình hồi phi tuyến tính đa biến có dạng:

price = b0 + b1 \*log(num\_bed\_rooms) + b2\* log(squared\_meter\_area)

Hay y = b0 + b1\*log(x1) + b2\*log(x2)

* Với các ma trận:

* Sử dụng hàm LN(number) để tính giá trị log(num\_bed\_rooms) và log(squared\_meter\_area)

Graphical user interface, table

Description automatically generated

Ta có các ma trận tương ứng với tập dữ liệu

* Ma trận F bao gồm 3 cột, 11313 hàng

Table

Description automatically generated

Ta có hệ phương trình trước được viết lại:

M=FT.F

Ma trận hệ số A=(FT.F)-1 .FT . Y

Chuyển vị của ma trận F, FT bao gồm 11313 cột, 3 hàng

Graphical user interface, text, application, table

Description automatically generated

* Trong excel sử dụng hàm MMULT(array1, array2) để nhân hai ma trận

Nhân hai ma trận FT, F thu được kết quả ma trận gồm 3 hàng, 3 cột Table

Description automatically generated

* Trong excel sử dụng hàm MINVERSE(array) để nghịch đảo một ma trận

Kết quả của phép tính ( FT.F)-1 là một ma trận gồm 3 hàng, 3 cột

Table

Description automatically generated

* Kết quả của phép tính ( FT.F)-1 FT là một ma trận gồm 11313 hàng, 3 cột

Table

Description automatically generated

* Kết quả của phép tính A=( FT.F)-1 FT.Y là một ma trận gồm 1 cột, 3 hàng

Table

Description automatically generated

* Vậy ta có được phương trình hồi quy phi tuyến:

**Y=** -300483 - 797629\***log(X1)** + 370679**\*log(X2)**

Hay

**Price=** - 300483 - 797629\***log(num\_bed\_rooms)** + 370679**\*log(squared\_meter\_area)**

* Phép kiểm định Fisher
* Giả thuyết:

H0: Phương trình hồi qui không phù hợp.

H1: Phương trình hồi qui phù hợp.

* Thực hiện tính toán trên Excel tương tự phần 4d.1.

Table

Description automatically generated

Table, Excel

Description automatically generated

* Ta thấy Significance F < α = 0.05
* Vậy ta bác bỏ H0.
* Phương trình hồi quy phù hợp.
* Sử dụng công cụ Data Analysis trên Excel kiểm tra kết quả
* Data -> Data Analyst -> chọn Regression

Graphical user interface, application

Description automatically generated

* Hộp thoại Regression xuất hiện

Graphical user interface, application

Description automatically generated

Input Y Range: chọn vùng dữ liệu cột price.

Input Y Range: chọn vùng dữ liệu cột log(num\_bed\_rooms), log(squared\_meter\_area).

Level: chọn độ tin cậy là 95%.

* Thu được kết quả như hình:

Application, table, Excel

Description automatically generated

1. **Sử dụng R**

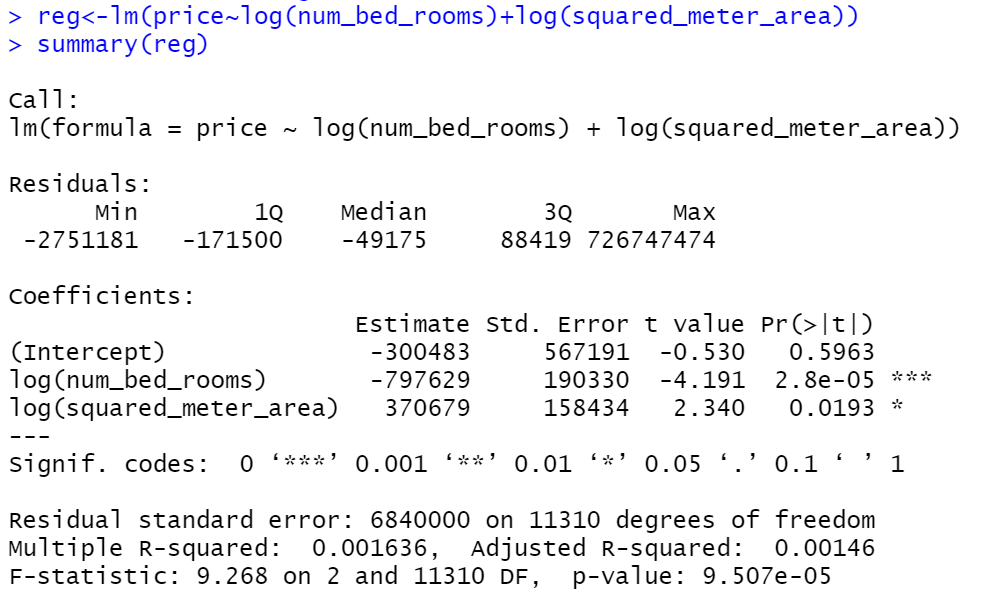
Dùng ngôn ngữ R thực hiện các phép tính Hồi quy với tập tin dữ liệu về Việt Nam.

* Import dataset

Table

Description automatically generated

* Sử dụng hàm **lm(y~log(x))** để biểu diễn mô hình hồi quy phi tuyến tính non-linear model.



* Nhìn vào Coefficients ta thấy Pr(>|t|) cột num\_bed\_rooms và cột squared\_meter\_area đều < 0.05.
* Nhận sét về độ tương quan R-squared, Adjusted R-squared ta thấy mô hình sau là phù hợp:

**Price** = -300483 – 797629 \* **num\_bed\_rooms** + 370679 \* **squared\_meter\_area**

* + **Kết luận:** Ta thấy p-value < α -> bác bỏ H0

1. **Sử dụng Python**

## **4f. Hồi quy Logistic với dữ liệu Bank Customer Data in VietNam.**

### **Phát biểu bài toán**

* Nguồn dữ liệu: https://www.kaggle.com/datasets/tomculihiddleston/bank-customer-data-in-vietnam
* Ý nghĩa nguồn dữ liệu: Thông tin của các khách hàng tại một ngân hàng Việt Nam
* Phát biểu bài toán: Kiểm tra mối liên hệ giữa output label **loan** với feature input **balance**

1. **Sử dụng Excel**

Tập dữ liệu Bank Customer Data in VietNam

Graphical user interface, application, table, Excel

Description automatically generated

Lấy cột dữ liệu *loan* tương ứng với biến Y, cột dữ liệu *balance* tương ứng với biến X.

Dạng tuyến tính của phương trình hồi quy logistic:

Y = b0 + b1X + e hay loan = b0 + b1\*balance + e

* Giả sử phương trình hồi quy logistic có dạng:

log() = b0 + b1X + ℇ

* Bước 1: tạo thêm cột biến chặn trước cột giá trị balance

Table

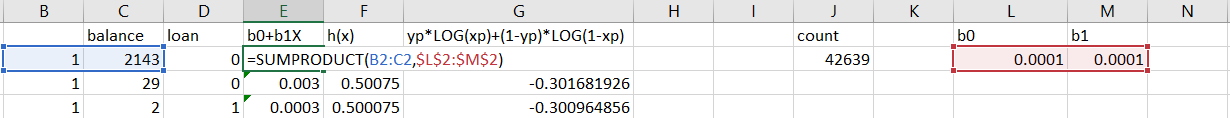
Description automatically generated with medium confidence

* Bước 2: gán cho b0, b1 hai giá trị bất kỳ, với bài này nhóm lấy giá trị 0.0001

Table

Description automatically generated

* Bước 3: Sử dụng hàm SUMPRODUCT(array1, array2) để tính giá trị cột b0+b1X



Ta được kết quả

Table

Description automatically generated with medium confidence

* Bước 4: áp dụng công thức

A picture containing text

Description automatically generated

Ta thu được kết quả cột h(x)

Text

Description automatically generated with low confidence

Table

Description automatically generated with medium confidence

* Bước 5: Tính log của hàm hợp lý

A picture containing diagram

Description automatically generated

Áp dụng công thức trên Excel

Timeline

Description automatically generated with low confidence

Ta được kết quả

A picture containing table

Description automatically generated

Sum: A picture containing text

Description automatically generated

* Bước 6: sử dụng công cụ Solver trên excel để tối ưu log(L) đạt giá trị lớn nhất

Chọn Data -> Solver -> xuất hiện hộp thoại Solver Parameters

Graphical user interface, application

Description automatically generated

Set Objecttive: giá trị cần tối ưu.

To: chọn Max, tối ưu giá trị đạt cực đại.

By Changing Variable Cells: chọn những giá trị cần thay đổi.

Bỏ chọn phần Make Unconstrained Variables Non-Negative

Graphical user interface, text, application

Description automatically generated

Nhấn chọn OK

* Ta thu được kết quả như hình

Table

Description automatically generated with medium confidence

* Từ kết quả trên ta có được phương trình hồi quy như sau:

Log() = -1.44827 - 0.001613 \* balance + ℇ

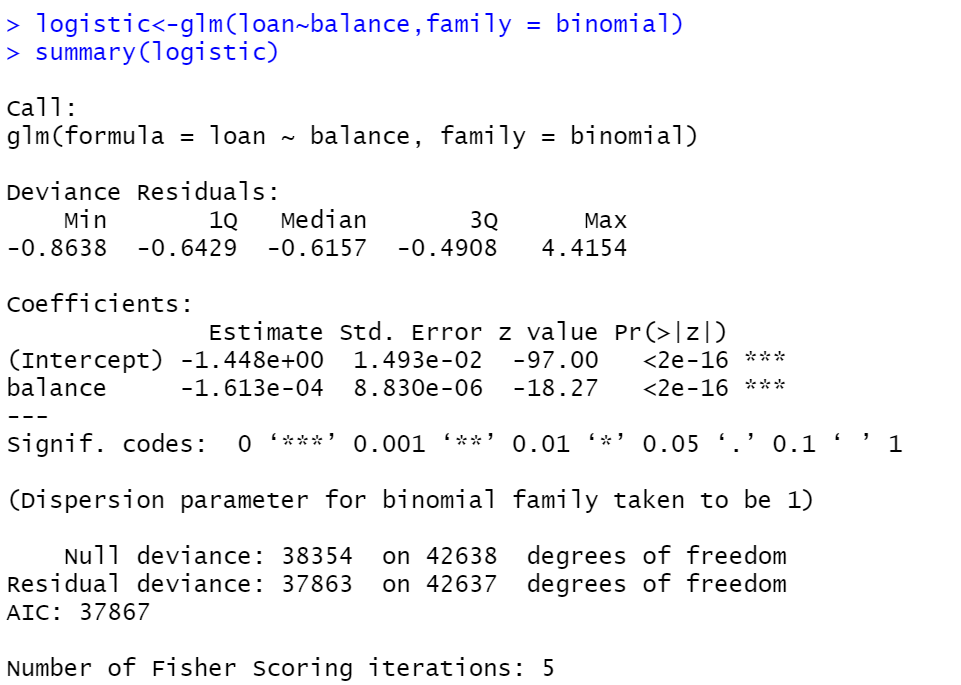
* Ta thu được kết quả sau:

() =

1. **Sử dụng R**

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

* Sử dụng hàm **glm(y~x)** để phân tích hồi quy logistic với tham số **family = binomial**



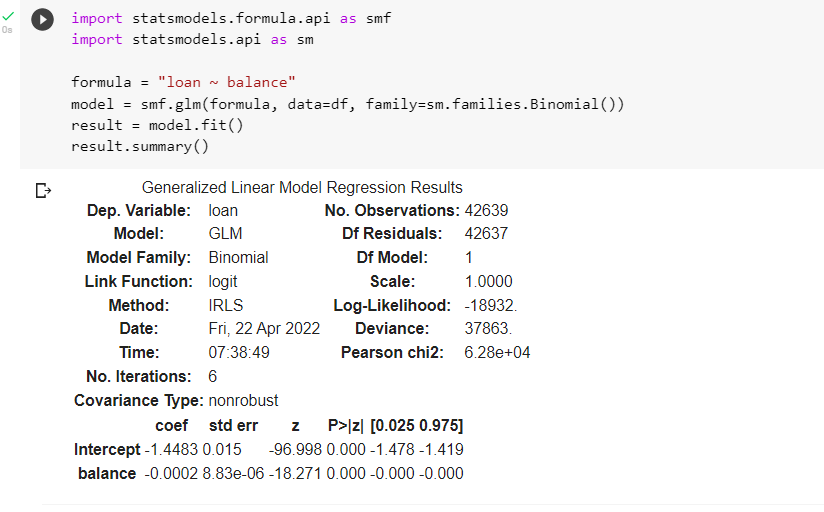
* Từ kết quả mô hình ta có được phương trình hồi quy như sau:

Log() = -1.448 -0.001613 \* balance + ℇ

* Ta thu được kết quả sau:

() =

1. **Sử dụng Python**



* Từ kết quả mô hình ta có được phương trình hồi quy như sau:

Log() = -1.448 -0.001613 \* balance + ℇ

* Ta thu được kết quả sau:

() =

1. **Bảng phân công công việc**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Nguyễn Hoàng Long | Phan Phạm Quỳnh Hoa | Huỳnh Minh Thư |
| Xây dựng Template báo cáo | X |  |  |
| Tìm hiểu và sử dụng Excel |  |  | X |
| Tìm hiểu và sử dụng R |  | X |  |
| Tìm hiểu và sử dụng Python | X |  |  |
| Làm bài tập 4d, 4e, 4f (Sử dụng Excel) |  |  | X |
| Làm bài tập 4d, 4e, 4f (Sử dụng R) |  | X |  |
| Làm bài tập 4d, 4e, 4f (Sử dụng Python) | X |  |  |

1. **Tài liệu tham khảo**

[1] <https://nguyenvantien0405.files.wordpress.com/2018/03/hd_excel_xstk_161223.pdf>

[2] <https://corporatefinanceinstitute.com/resources/excel/functions/>

[3] <https://www.statology.org/tukey-kramer-post-hoc-test-excel/>

[4] <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.f_oneway.html>

[5] <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2_contingency.html>