Định nghĩa

* Phân tích hồi qui logistic là phương pháp, kỹ thuật thống kê dùng để đánh giá mối liên hệ giữa biến độc lập (biến đó có thể là số hoặc phân loại) với một biến dạng nhị phân ( 0 và 1).
* Dạng tuyến tính của phương trình hồi logistic là, với biến độc lập Y và biến phụ thuộc X được biểu diễn qua phương trình sau đây:



* Trong phương trình trên Y là một biến độc lập với giá trị nhị phân là 0 (không tham gia bảo hiểm) và 1 (tham gia bảo hiểm)
* Nếu p là xác suất của một biến cố xảy ra (xác suất tham gia bảo hiểm), 1 – p là xác suất của một biến cố không xảy ra (xác suất không tham gia bảo hiểm) dạng như trên thì phương trình hồi qui logistic được biểu diễn qua phương trình sau đây:

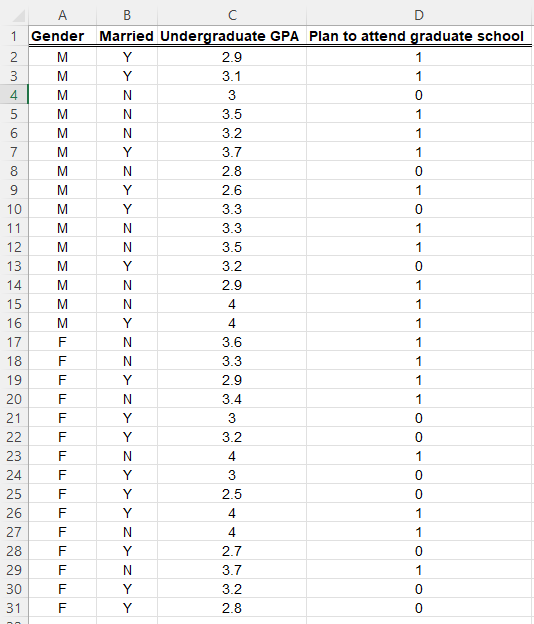


* Từ phương trình trên ta có thể tính được xác suất p theo giá trị X



Hồi qui Logistic đơn biến

* Nghiên cứu sự liên hệ giữa điểm GPA và tiên lượng có kế hoạch tham dự lễ tốt nghiệp của trường. Dữ liệu **Graduate School Survey.xlsx**



* Sử dụng công cụ Show ToolPak
  + Cài đặt
  + Chọn **Insert** > **Get Add-ins** > **Tìm XMLiner** > **XMLiner Analysis TookPak**  > **Add**

Graphical user interface, text, application, email

Description automatically generated

* Chọn công cụ **Show ToolPak** > **Logistic Regresstion**
* Input Y chọn: **Plan to attend graduate school**

Graphical user interface, application, table, Excel

Description automatically generated

* Input X chọn: **Undergraduate GPA**

Graphical user interface, application, table, Excel

Description automatically generated

Với Confidence Level là 95%

Chọn Output Range, OK

* BẢNG PHÂN TÍCH KẾT QUẢ NHƯ HÌNH MINH HỌA

Graphical user interface, application, table, Excel

Description automatically generated

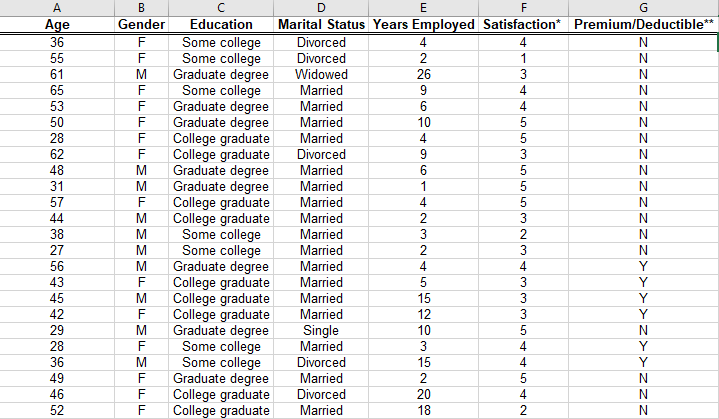
* Từ bảng trên ta có phương trình hồi qui logistic như sau:



* Ta có . Ta đặt hệ số p/(1-p) là odd
* Đặt Odd0 và undergraduateGPA = 0 thì 
* Đặt Odd1 và undergraduateGPA = 1 thì 
* Tỉ số 
* Giá trị cột Exp(B) ở bảng Variables in the Equation
* Lúc này ta có thể diễn dịch, cứ điểm undergraduateGPA lên 1 đơn vị thì khả năng đi dự tốt nghiệp tăng lên tỉ lệ có kế hoạch tham dự lễ tốt nghiệp tăng lên 36,359 lần, nếu tăng 0,1 điểm GPA thì tỉ lệ tham dự lễ tốt nghiệp tăng lên 3,6359 lần.

Hồi qui Logistic đa biến

* Tương tự như hồi qui logistic đơn biến, đa biến cũng tương tự như vậy nhưng biến covariates từ 2 giá trị trở lên
* Tập dữ liệu: **Insurance Survey**

****

*Tập dữ liệu Insurance Survey*

* Chuyển data về dạng binary: **Premium/Deductible\*\***

*Graphical user interface, application, table, Excel

Description automatically generated*

* Các bước tương tự như hồi qui Logistic đơn biến
* Input Y: **Premium/Deductible\*\***

*Graphical user interface, application, table, Excel

Description automatically generated*

* Input X: **Years Employed** và **Satisfaction\***

*Graphical user interface, application, table, Excel

Description automatically generated*

Với Confidence Level là 95%

Chọn Output Range, OK

* Ta được kết quả như hình minh họa:

*A picture containing calendar

Description automatically generated*

*Bảng kết quả phân tích hồi qui đa biến*

* Phương trình hồi quy logistic có dạng như sau:



* Tỉ số odd dựa vào bảng Variable in the Equation
  + Tỉ số odd của satisfaction là e-0,207 = 0,813
  + Tỉ số odd của Years Employed là e0,025 = 1,025
  + Vậy ta có thể kết luận rằng Year Employed càng lớn thì sẽ trả có xác suất trả bảo hiểm thấp (Y) sẽ cao với OddYear Employed  = 1,025

**Data**

* Tên data: **Bank Customer Data in VietNam**
* Dữ liệu liên quan đến các chiến dịch tiếp thị trực tiếp (gọi điện thoại) của một tổ chức ngân hàng Việt Nam. Mục tiêu phân loại là để dự đoán xem khách hàng có đăng ký một khoản tiền gửi có kỳ hạn hay không.
* Link data: <https://www.kaggle.com/datasets/tomculihiddleston/bank-customer-data-in-vietnam>
* Dữ liệu gồm 42640 dòng và 16 thuộc tính
* 1 số thuộc tính quan tâm:
  + **age** (numeric)
  + **loan**: có khoản vay cá nhân không? (categorical: 'no','yes','unknown')

**Graphical user interface, application, table, Excel

Description automatically generated**

**Phát biểu bài toán**

Kiểm tra sự phụ thuộc của việc xuất hiện các khoản vay có liên quan đến tuổi tác hay không.

H0: Các khoản vay không phụ thuộc vào tuổi tác.

H1: Các khoản vay có phụ thuộc vào tuổi tác.

**Trên Excel**

* Sử dụng công cụ Show ToolPak
  + Chọn công cụ **Show ToolPak** > **Logistic Regresstion**
  + Input Y chọn: **Plan to attend graduate school**

Graphical user interface, table, Excel

Description automatically generated

* + Input X chọn: **Plan to attend graduate school**

Graphical user interface, application, table

Description automatically generated

Với Confidence Level là 95%

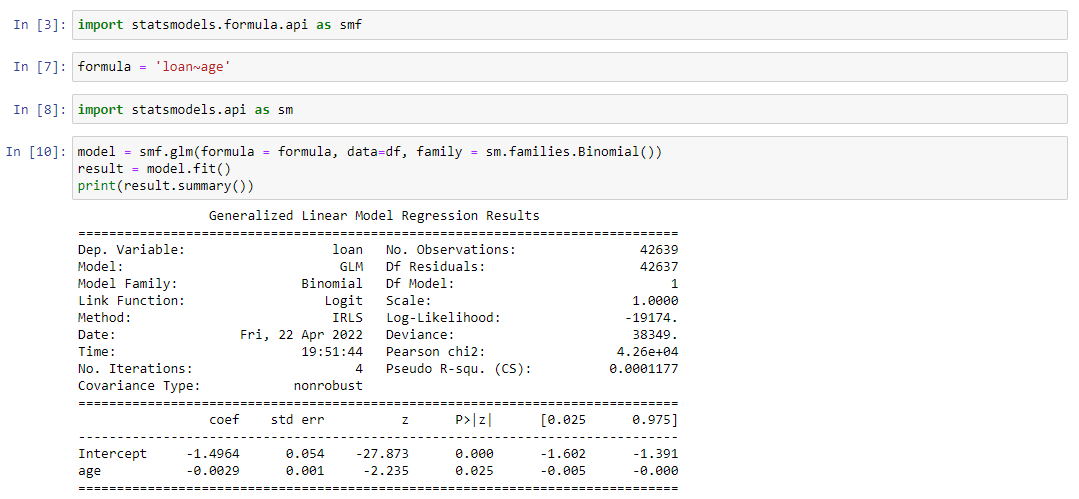
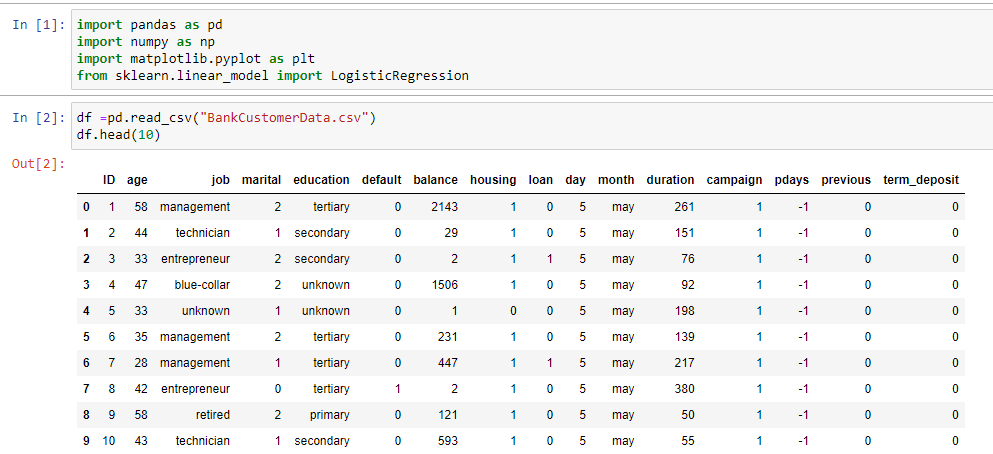
Chọn Output Range, OK

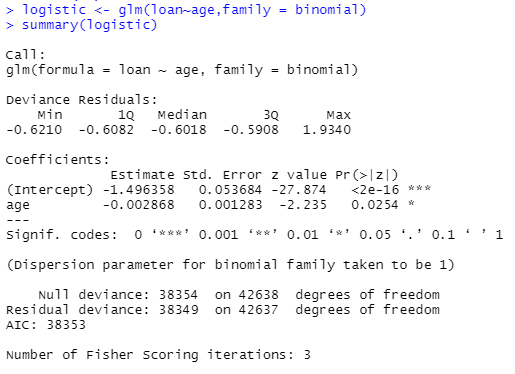
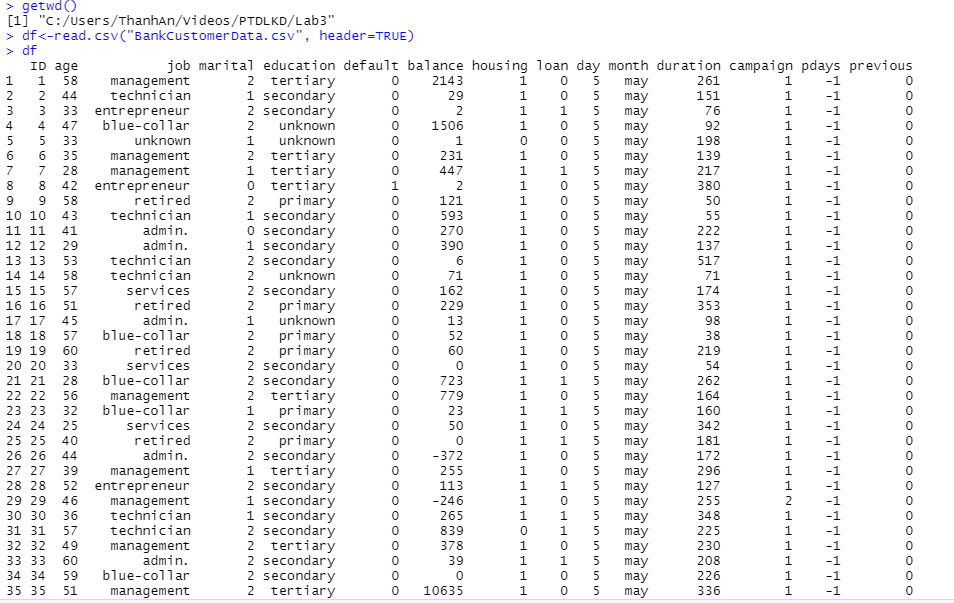
* Ta được kết quả như hình dưới:

Graphical user interface, application, table, Excel

Description automatically generated

**Trên Python**



**Trên R**

* Từ bảng trên ta có phương trình hồi qui logistic như sau:

log = -1.496358 – 0.002868 \* (age) +

* Ta có = -1.496358 – 0.002868 \* (age). Ta đặt hệ số là odd
* Đặt Odd0 và age = 0 thì Odd0 = -1.496358
* Đặt Odd1 và age = 1 thì Odd1 = -1.496358 – 0.002868
* Tỉ số 0,9971361088
* Lúc này ta có thể diễn dịch, khi age lên 1 đơn vị thì loan tăng lên tỉ lệ có kế hoạch tham dự lễ tốt nghiệp tăng lên 0,9971361088 lần