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

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**



**ĐỒ ÁN MÔN HỌC**

**MÔ HÌNH HỒI QUI TUYẾN TÍNH**

**MULTIVARIATE LINEAR REGRESSION MODELS**

**GVHD: LÝ QUỐC NGỌC**

TP HỒ CHÍ MINH, NGÀY 19 THÁNG 8 NĂM 2020

**MỤC LỤC**

1. **Thông tin nhóm** 
   1. Thông tin thành viên
   2. Phân công
2. **Hồi quy tuyến tính**

2.1 Mục đích

2.2 Mô hình

2.3 Ước lượng tham số cho mô hình

2.4 Least Square Estimate

1. **Đánh giá mô hình**
2. **Ước lượng giá trị mới**
3. **Dự đoán giá trị mới**
4. **Multivariate Linear Regression**
5. **Chương trình minh họa**
6. **Cài đặt**
7. **Tài liệu tham khảo**
8. **Thông tin nhóm**

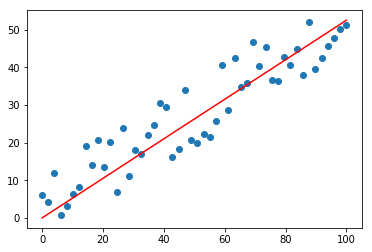
***1.1 Thông tin thành viên:***

Tên nhóm: **Bit64**

|  |  |  |
| --- | --- | --- |
| **MSSV** | **Họ và tên** | **Email** |
| 1712003 | Lê Nguyễn Đức Anh | lndanh23@gmail.com |
| 1712096 | Đặng Hồng Minh | minhdangit99@gmail.com |
| 1712491 | Lê Vũ Anh Huy | anhhuy.le.1999.tb@gmail.com |
| 1712532 | Nguyễn Anh Khoa | nguyenanhkhoa301199@gmail.com |

* 1. ***Phân công***

1. **Hồi quy tuyến tính**



***2.1 Mục đích***

Đưa ra được dự đoán khi có một giá trị mới được đưa vào mô hình

* Số lượng biến r = 1, mô hình sẽ có dạng một đường thẳng:
* Số lượng biến r = 2, mô hình sẽ có dạng một mặt phẳng:
* Số lượng biến r > 2, mô hình sẽ có dạng siêu phẳng
  1. ***Mô hình***

Dạng của mô hình:

Với n quan sát độc lập:

Trong đó:

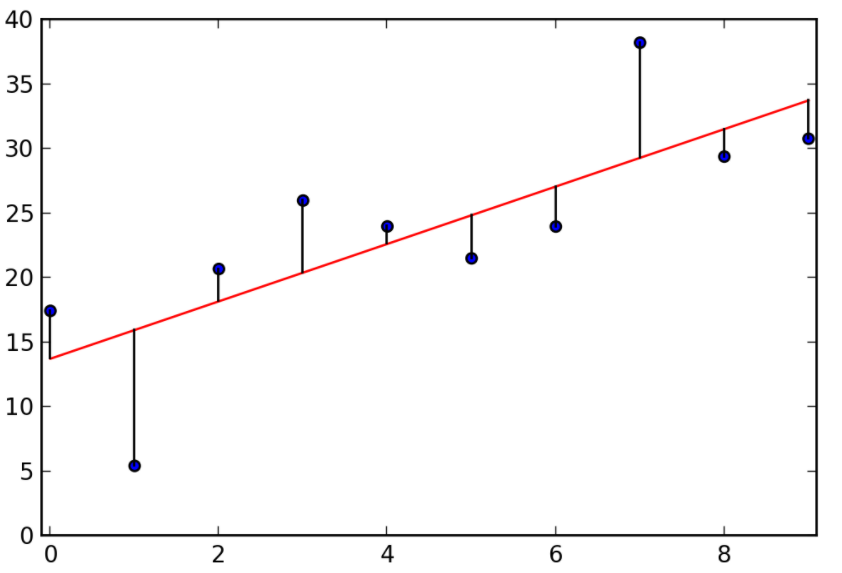
Nếu ta đặt :

Dạng ma trận:

Hay

Khi đó:

* 1. ***Ước lượng tham số cho mô hình***

******

Nguyên lý là cực tiểu hóa sai lệch giữa điểm dữ liệu và đường thẳng mô hình

*Phương pháp:*

* Giả sử ***b***là giá trị thử của
* Xét hiệu và là sai lệch giữa giá trị thực và giá trị tiên đoán của mô hình
* Nếu hiệu này bằng 0, điểm dữ liệu thuộc “đường thẳng” mô hình
* Thông thường hiệu này sẽ không bằng 0, vì các điểm dữ liệu thường dao động xung quanh “đường thẳng” do nhiễu (thể hiện qua tham số của mô hình)
* Do đó ta cần cực tiểu hóa sai lệch, dùng phương pháp **Least Square Estimation**
  1. ***Least Square Estimation***

Xét đạo hàm:

*Tính đạo hàm* :

* Với r = 1, ta có:

Khi đó:

* Với r > 1:

*Tính đạo hàm*

* Với r=1, n=3:

Do đó:

Mở rộng cho n, r:

Để tìm cực trị, xét đạo hàm bằng 0:

Hệ số ***b*** được gọi là ước lượng bình phương tối thiểu của tham số hồi quy , ***b*** được biểu diễn bởi để nhấn mạnh vai trò là ước lượng của

1. **Đánh giá mô hình**

Gọi là phần dư (residual) giữa giá trị thực và giá trị dự đoán của mô hình và đặt , ta có:

Hay

Trong đó: (“hat” matrix)

Một số tính chất của :

Chứng minh:

Chứng minh:

Chứng minh:

Trong thống kê sử dụng R2 (R-Square) để đánh giá mô hình hồi quy

*Trong đó:*

Và theo tính chất

Nghĩa là:

Do đó:

* là tổng bình phương phần dư:
* là tổng bình phương sai lệch giữa tiên đoán mô hình và giá trị trung bình của tập dữ liệu:
* Độ đo R2:

***Nhận xét:***

* Giá trị dự đoán của mô hình khớp hoàn toàn với các giá trị kết quả của dữ liệu mẫu Mô hình lý tưởng
* Các biến *zr* không có liên hệ với *yj* Mô hình không phù hợp

1. **Ước lượng giá trị mới**

Ta cần ước lượng giá trị *Y0* khi có một giá trị mới thuộc quần thể, nhưng các tham số từ quần thể ta không biết và cần suy ra từ tập mẫu. Khi đó cần thực hiện 2 giai đoạn:

* Ước lượng *Y0* từ mẫu (với lỗi đã biết từ tập mẫu)
* Thay đổi tăng giảm giá trị ước lượng một lượng do có thêm lỗi từ quần thể

**Ước lượng kì vọng**

Tạm thời không xét lỗi , ta có ***mean*** của *Y0* là giá trị của hàm hồi quy sau:

Khi đó, được ước lượng bằng phương pháp bình phương tối thiểu (Least Square Estimate) là , trong đó đã có được từ tập mẫu

Khi cho trước , ta xem như đây là một vector hằng, hay ma trận hằng. Do đó phương sai dự đoán:

Với có phân phối chuẩn (tham số chưa biết)

Người ta chứng minh được khoảng tin cậy cho là:

Trong đó là phân vị trên thứ 100 của phân phối student với bậc tự do

1. **Dự đoán giá trị mới**

Dự đoán giá trị mới thường ít chắc chắn hơn do sự đóng góp của nhiễu vào mô hình

Xét mô hình hồi quy tuyến tính

Tương tự, ta có từ tập mẫu. Tuy nhiên, ta không biết được . Giả thiết rằng , không phụ thuộc vào của tập mẫu, do đó cũng không phụ thuộc vào và

Ta có:

Lỗi dự đoán:

Lấy kì vọng 2 vế:

Vì và độc lập:

Tương tự, người ta chứng minh được khoảng dự đoán cho là:

Với là phân vị trên thứ của phân phối student với bậc tự do

**Vậy:**

* Khoảng tin cậy cho là:
* Khoảng tin cậy cho là:

1. **Multivariate Linear Regression**

Mô hình ban đầu:

Giả sử mỗi response tuân theo mô hình hồi quy riêng của nó, tức là:

Ta có thể thấy, với một tập dữ liệu gốc (thể hiện qua các biến z), ta mong muốn rằng ta có thể xây dựng được nhiều bộ tham số hơn-tương ứng nhiều mô hình hồi quy tuyến tính trên cùng một tập dữ liệu.

Dạng ma trận

Khi đó mô hình hồi quy tuyến tính đa biến là:

Ta tính được bằng phương pháp Least Square Estimate, tức:

Biểu diễn dạng ma trận tương ứng:

Hay:

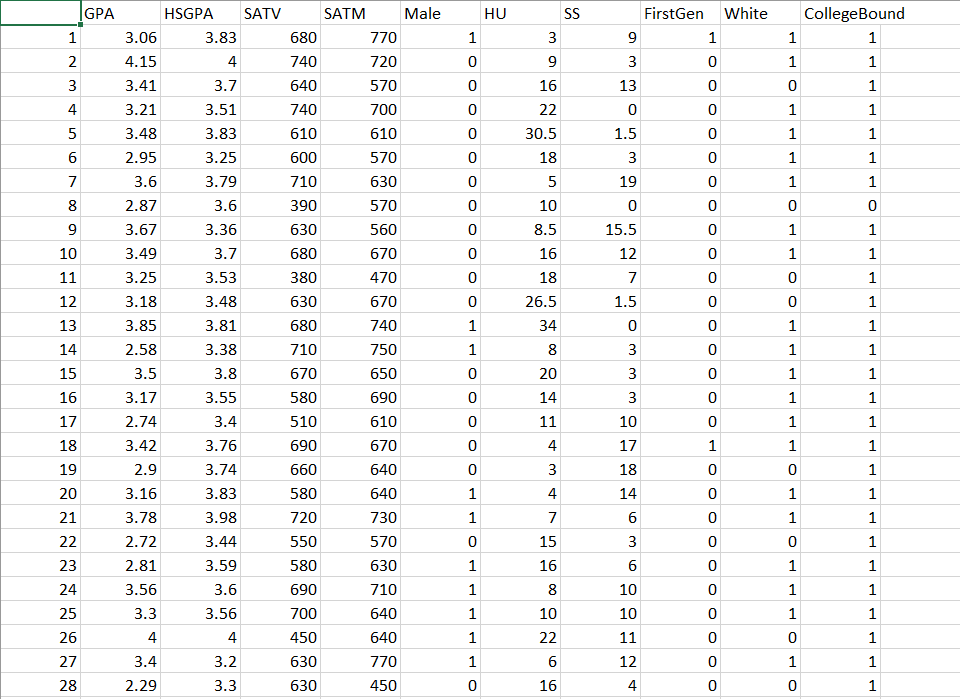
1. **Chương trình minh họa**

* Trong bài toán hồi quy bội, cần tìm kiếm một hàm có thể ánh xạ các điểm dữ liệu đầu vào thành các giá trị kết quả. Mỗi điểm dữ liệu là một vectơ đặc trưng bao gồm hai hoặc nhiều giá trị dữ liệu nắm bắt các tính năng khác nhau của đầu vào. Để biểu diễn tất cả dữ liệu đầu vào cùng với vectơ của các giá trị đầu ra, thiết lập ma trận đầu vào Z và vectơ đầu ra Y:

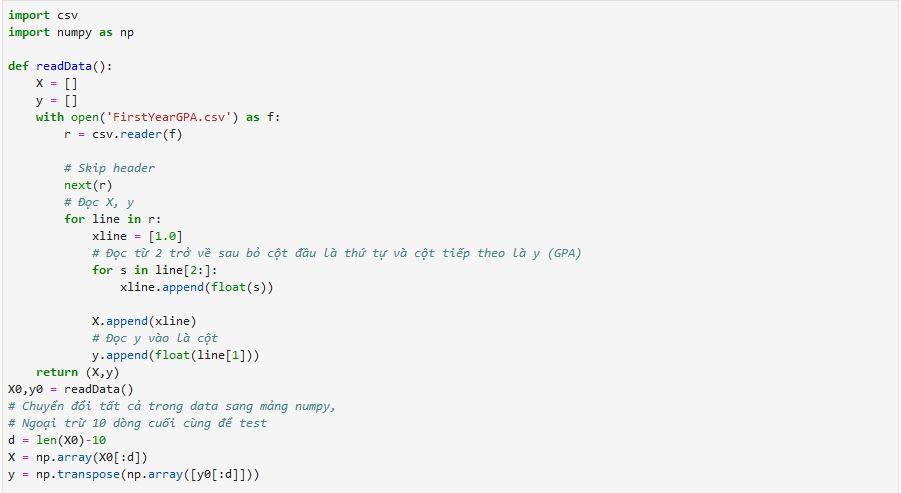
**Dữ liệu vào:**

* Tập dữ liệu với 219 quan sát trên 10 biến như sau:
* *GPA* Điểm trung bình đại học năm thứ nhất theo thang điểm 0,0 đến 4,0
* *HSGPA* Điểm trung bình trung học phổ thông theo thang điểm 0,0 đến 4.0
* *SATV* Điểm SAT đọc hiểu
* *SATM* Điểm SAT môn toán
* *Male* 1 = nam, 0 = nữ
* *HU* Số giờ tín dụng kiếm được trong các khóa học nhân văn ở trường trung học
* *SS* Số giờ tín dụng kiếm được trong các khóa học khoa học xã hội ở trường trung học
* *FirstGen* 1 = học sinh là người đầu tiên trong gia đình của cô hoặc anh ấy học đại học, 0 = nếu không
* *White* 1 = sinh viên da trắng, 0 = người khác
* *CollegeBound* 1 = học tại một trường trung học nơi 50% học sinh dự định vào đại học, 0 = nếu không
* Chi tiết:
* Dữ liệu trong FirstYearGPA chứa thông tin từ mẫu của 219 sinh viên năm thứ nhất tại một trường đại học trung tây có thể được sử dụng để xây dựng mô hình để dự đoán điểm trung bình năm đầu tiên của họ.
* Nguồn: Một mẫu từ một bộ dữ liệu lớn hơn được thu thập vào năm 1996 bởi một giáo sư tại trường đại học này.

**Bảng dữ liệu:**

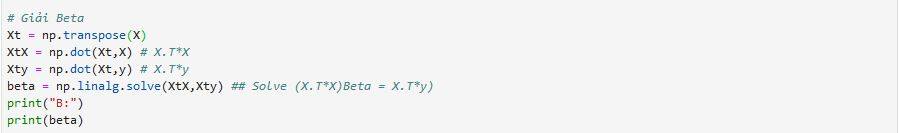


**Code đọc và xử lý dữ liệu đầu vào:**



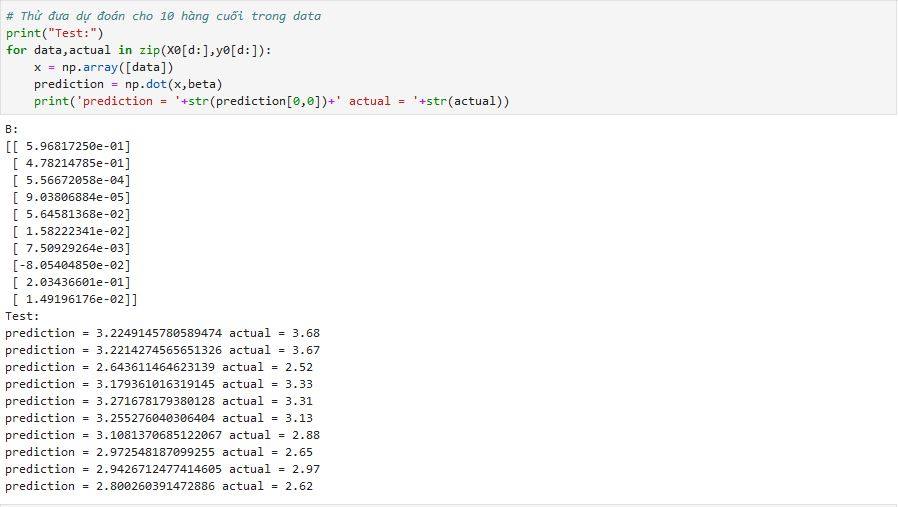
**Áp dụng công thức:**

* Công thức của đã được rút gọn như sau:
* Code tương ứng công thức:



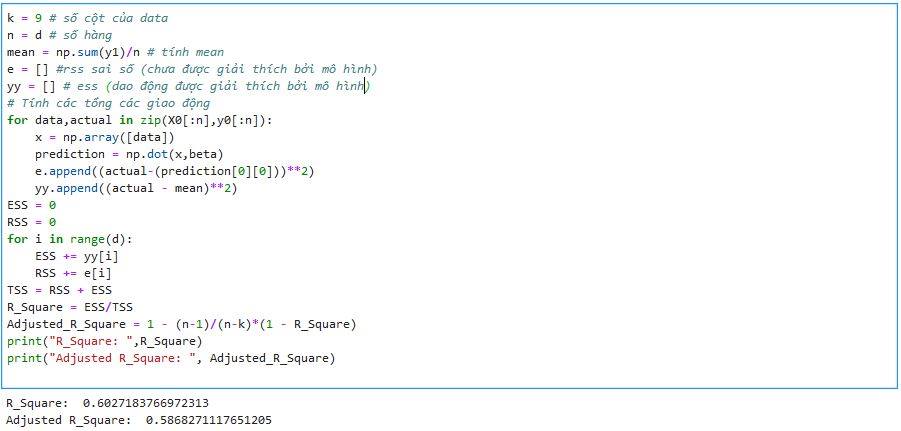
**Dự đoán kết quả:**

Sau khi đã tính toán xong, thử giả định dự đoán kết quả của 10 dòng cuối trong bộ dữ liệu bên trên



**Đo lường mức độ phù hợp của ước lượng theo OLS (Ordinary Least Squares)**

* Hệ số xác định
  + Gọi
    - TSS: Total sum of squares (Tổng tổng bình phương).
    - ESS: Explained sum of squares (Độ giao động có thể giải thích bằng mô hình hồi qui tuyến tính).
    - RSS: Residual sum of squares (Tổng bình phương phần dư mà mô hình không giải thích được).
* ***TSS = ESS + RSS***
  + - Với:
* **Hệ số xác định hiệu chỉnh**
  + - Đặt )
* Code tính và :



1. **Cài đặt**

* Chương trình minh họa xây dựng mô hình hồi quy tuyến tính trên tập dữ liệu mẫu được cho trước.
* Source code của chương trình được viết bằng ngôn ngữ Python 3.
  1. **Yêu cầu để chạy chương trình:** 
     + Python phiên bản 3.x
     + Thư viện được sử dụng: numpy.
     + Chương trình được chạy trên hệ điều hành window.
  2. **Tổ chức thư mục chạy chương trình:**
     + Tập tin FirstYearGPA.csv: tập dữ liệu được sử dụng trong chương trình này.
     + linear\_regression.py : tập tin chứa các hàm cài đặt để giải bài toán.
  3. **Hướng dẫn chạy chương trình:**
     + Chạy linear\_regression.py
     + Cú pháp lệnh chạy chương trình: command line

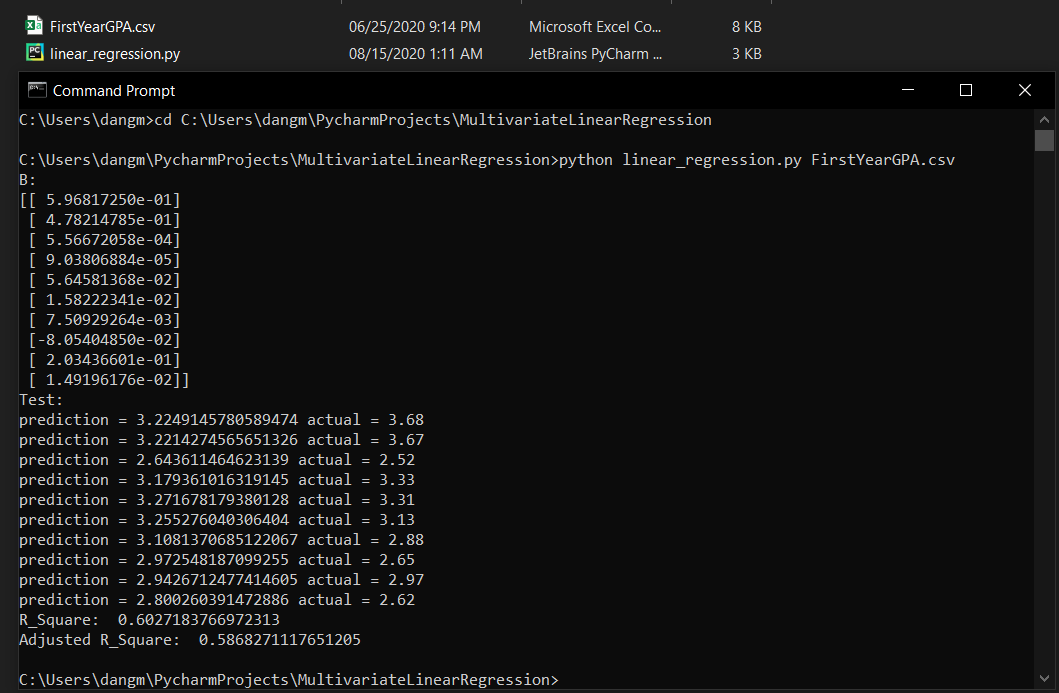
python linear\_regression.py <đường dẫn file dữ liệu>

vd:

* Dùng lệnh “**cd**” dẫn đến thư mục hiện hành

(**cd D:\Source\MultivariateLinearRegression**)

* **python linear\_regression.py FirstYearGPA.csv**
  1. **Màn hình kết quả:**

****

1. **Tài liệu tham khảo**
2. W. Richard A.Johnson, Applied Multivariate Statistical Analysis, US: Pearson Education, 2007.
3. Wolfgang Karl Härdle & Léopold Simar, Applied Multivariate Statistical Analysis,   
   Third Edition.
4. GS. Nguyễn Văn Tuấn, Phân tích dữ liệu với R, 2015.
5. Nguyễn Đình Huy, Giáo trình Xác suất và Thống kê, 2009.
6. Edouard Duchesnay, Tommy Löfstedt, Statistics and Machine Learning in Python, Release 0.1, 2017.
7. Wikipedia, "Ordinary Least Squares," [Online]. Available: <https://en.wikipedia.org/wiki/Ordinary_least_squares> [Accessed 15 July 2020].
8. Wikipedia, “Variance," [Online]. Available: https://en.wikipedia.org/wiki/Variance.   
   [Accessed 15 July 2020].
9. Wikipedia, "Maximum Likelihood Estimation," [Online]. Available: https://en.wikipedia.org/wiki/Maximum\_likelihood\_estimation. [Accessed 15 July 2020]
10. Wikipedia, "Gradient Descent," [Online]. Available: https://en.wikipedia.org/wiki/Gradient\_descent. [Accessed 15 July 2020]
11. Wikipedia, "Student t distribution," [Online]. Available: https://en.wikipedia.org/wiki/Student%27s\_t-distribution. [Accessed 15 July 2019]