

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	
1.1 Thông tin thành viên.....	
1.2 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	
3. Đánh giá mô hình	
4. Ước lượng giá trị mới	
5. Dự đoán giá trị mới.....	
6. Multivariate Linear Regression.....	
7. Chương trình minh họa.....	
8. Cài đặt.....	
9. Tài liệu tham khảo	

1. Thông tin nhóm

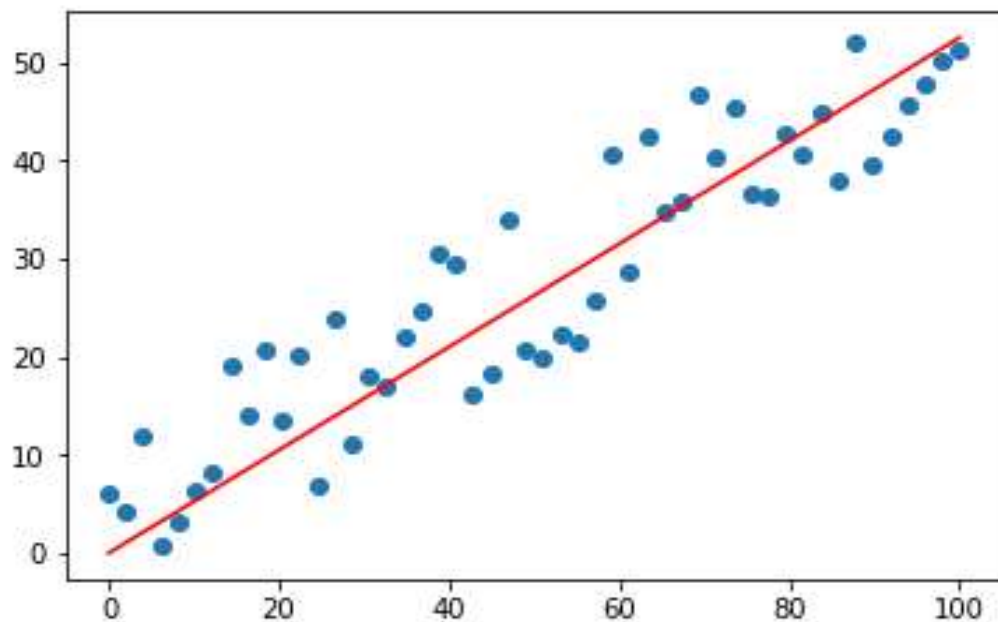
1.1 Thông tin thành viên:

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	anhhuyle.1999.tb@gmail.com
1712532	Nguyễn Anh Khoa	nguyenanhkhoa301199@gmail.com

1.2 Phân công

2. 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:

$$Y = \beta_0 + \beta_1 z_1$$

- Số lượng biến $r = 2$, mô hình sẽ có dạng một mặt phẳng:

$$Y = \beta_0 + \beta_1 z_1 + \beta_2 z_2$$

- Số lượng biến $r > 2$, mô hình sẽ có dạng siêu phẳng

2.2 Mô hình

Dạng của mô hình:

$$Y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_r z_r + \varepsilon$$

$$[Response] = [mean(depending\ on\ z_1, z_2, \dots, z_r)] + [error]$$

Với n quan sát độc lập:

$$Y_1 = \beta_0 + \beta_1 z_{11} + \beta_2 z_{12} + \cdots + \beta_r z_{1r} + \varepsilon_1$$

$$Y_2 = \beta_0 + \beta_1 z_{21} + \beta_2 z_{22} + \cdots + \beta_r z_{2r} + \varepsilon_2$$

...

$$Y_n = \beta_0 + \beta_1 z_{n1} + \beta_2 z_{n2} + \cdots + \beta_r z_{nr} + \varepsilon_n$$

Trong đó:

- $E(\varepsilon_j) = 0$
- $Var(\varepsilon_j) = \sigma^2$ (hằng số)
- $Cov(\varepsilon_j, \varepsilon_k) = 0, j \neq k$

Nếu ta đặt $z_{j0} = 1$:

$$Y = \beta_0 + \beta_1 z_{j1} + \beta_2 z_{j2} + \cdots + \beta_r z_{jr} = \beta_0 z_{j0} + \beta_1 z_{j1} + \beta_2 z_{j2} + \cdots + \beta_r z_{jr}$$

Dạng ma trận:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & z_{11} & z_{12} & \cdots & z_{1r} \\ 1 & z_{21} & z_{22} & \cdots & z_{2r} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & z_{n1} & z_{n2} & \cdots & z_{nr} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_r \end{bmatrix} + \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

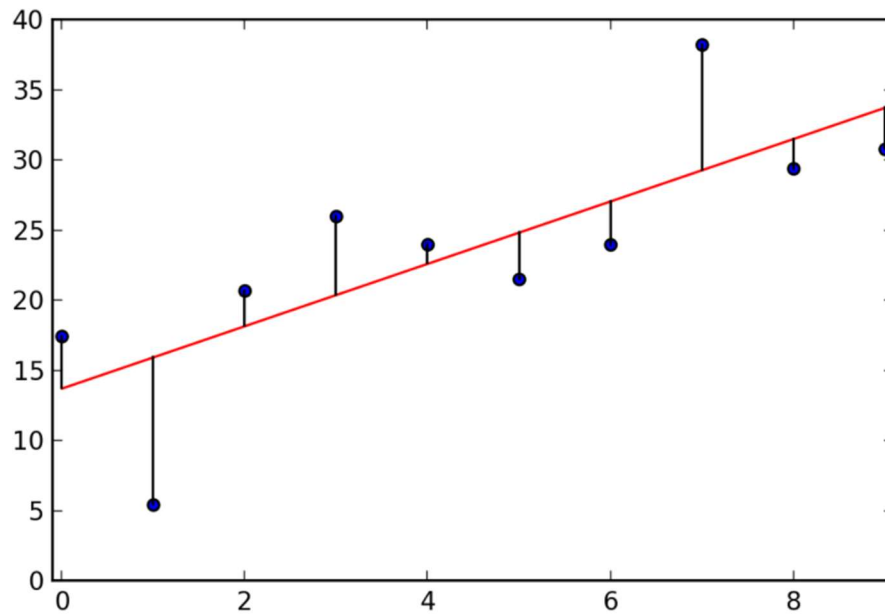
Hay

$$\begin{matrix} \mathbf{Y} \\ (n \times 1) \end{matrix} = \begin{matrix} \mathbf{Z} \\ (n \times (r+1)) \end{matrix} \begin{matrix} \boldsymbol{\beta} \\ ((r+1) \times 1) \end{matrix} + \begin{matrix} \boldsymbol{\varepsilon} \\ (n \times 1) \end{matrix}$$

\Rightarrow Khi đó:

- $E(\varepsilon) = 0$
- $Cov(\varepsilon) = E(\varepsilon \varepsilon') = \sigma I$

2.3 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 β
- Xét hiệu y_j và $b_0 + b_1 z_{j1} + \dots + b_r z_{jr}$ 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**

2.4 Least Square Estimation

$$\begin{aligned} S(b) &= \sum_{j=1}^n (y_j - b_0 - b_1 z_{j1} - \dots - b_r z_{jr})^2 \\ &= \sum_{j=1}^n (y_j - (b_0 z_{j0} + b_1 z_{j1} + \dots + b_r z_{jr}))^2 \quad (\text{với } z_{j0} = 1) \\ &= \sum_{j=1}^n (y_j^2 - 2y_j(b_0 z_{j0} + b_1 z_{j1} + \dots + b_r z_{jr}) + (b_0 z_{j0} + b_1 z_{j1} + \dots + b_r z_{jr})^2) \end{aligned}$$

$$\begin{aligned}
&= \sum_{j=1}^n y_j^2 - 2 \sum_{j=1}^n y_j (b_0 z_{j0} + b_1 z_{j1} + \dots + b_r z_{jr}) \\
&\quad + \sum_{j=1}^n (b_0 z_{j0} + b_1 z_{j1} + \dots + b_r z_{jr})^2 \\
&= \mathbf{y}^T \mathbf{y} - 2 \mathbf{b}^T \mathbf{Z}^T \mathbf{y} + (\mathbf{Zb})^T \mathbf{Zb} \\
&= (\mathbf{y}^T \mathbf{y} - \mathbf{b}^T \mathbf{Z}^T \mathbf{y}) - (\mathbf{b}^T \mathbf{Z}^T \mathbf{y} - \mathbf{b}^T \mathbf{Z}^T \mathbf{Zb}) \\
&= (\mathbf{y}^T \mathbf{y} - \mathbf{b}^T \mathbf{Z}^T \mathbf{y}) - (\mathbf{y}^T \mathbf{Zb} - \mathbf{b}^T \mathbf{Z}^T \mathbf{Zb}) \\
&= (\mathbf{y}^T - \mathbf{b}^T \mathbf{Z}^T) \mathbf{y} - (\mathbf{y}^T - \mathbf{b}^T \mathbf{Z}^T) \mathbf{Zb} \\
&= (\mathbf{y}^T - \mathbf{b}^T \mathbf{Z}^T) (\mathbf{y} - \mathbf{Zb}) \\
&= (\mathbf{y} - \mathbf{Zb})^T (\mathbf{y} - \mathbf{Zb})
\end{aligned}$$

Xét đạo hàm:

$$\begin{aligned}
\frac{\partial S(\mathbf{b})}{\partial \mathbf{b}} &= \frac{\partial (\mathbf{y}^T \mathbf{y} - 2 \mathbf{b}^T \mathbf{Z}^T \mathbf{y} + (\mathbf{Zb})^T \mathbf{Zb})}{\partial \mathbf{b}} \\
&= \frac{\partial (\mathbf{y}^T \mathbf{y})}{\partial \mathbf{b}} - \frac{\partial (2 \mathbf{b}^T \mathbf{Z}^T \mathbf{y})}{\partial \mathbf{b}} + \frac{\partial (\mathbf{b}^T \mathbf{Z}^T \mathbf{Zb})}{\partial \mathbf{b}} \\
&= -2 \frac{\partial (\mathbf{b}^T \mathbf{Z}^T \mathbf{y})}{\partial \mathbf{b}} + \frac{\partial (\mathbf{b}^T \mathbf{Z}^T \mathbf{Zb})}{\partial \mathbf{b}}
\end{aligned}$$

Tính đạo hàm $\frac{\partial (\mathbf{b}^T \mathbf{Z}^T \mathbf{Zb})}{\partial \mathbf{b}}$:

- Với $r = 1$, ta có:

$$\begin{aligned}
\mathbf{b}^T \mathbf{Z}^T \mathbf{Zb} &= [b_0 \quad b_1] \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} \\
&= [b_0 \quad b_1] \begin{bmatrix} a_{11}b_0 & a_{12}b_1 \\ a_{12}b_0 & a_{22}b_1 \end{bmatrix} \\
&= a_{11}b_0^2 + a_{22}b_1^2 + 2a_{12}b_0b_1
\end{aligned}$$

- $\frac{\partial (\mathbf{b}^T \mathbf{Z}^T \mathbf{Zb})}{\partial b_0} = \frac{\partial (a_{11}b_0^2 + a_{22}b_1^2 + 2a_{12}b_0b_1)}{\partial b_0} = 2a_{11}b_0 + 2a_{12}b_1$
- $\frac{\partial (\mathbf{b}^T \mathbf{Z}^T \mathbf{Zb})}{\partial b_1} = \frac{\partial (a_{11}b_0^2 + a_{22}b_1^2 + 2a_{12}b_0b_1)}{\partial b_1} = 2a_{12}b_0 + 2a_{22}b_1$

Khi đó:

$$\begin{aligned}\frac{\partial(\mathbf{b}^T \mathbf{Z}^T \mathbf{Z} \mathbf{b})}{\partial \mathbf{b}} &= \begin{bmatrix} \frac{\partial(\mathbf{b}^T \mathbf{Z}^T \mathbf{Z} \mathbf{b})}{\partial b_0} \\ \frac{\partial(\mathbf{b}^T \mathbf{Z}^T \mathbf{Z} \mathbf{b})}{\partial b_1} \end{bmatrix} = \begin{bmatrix} 2a_{11}b_0 + 2a_{12}b_1 \\ 2a_{22}b_1 + 2a_{12}b_0 \end{bmatrix} \\ &= 2 \begin{bmatrix} a_{11}b_0 + a_{12}b_1 \\ a_{12}b_0 + a_{22}b_1 \end{bmatrix} \\ &= 2 \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} \\ &= 2\mathbf{Z}^T \mathbf{Z} \mathbf{b}\end{aligned}$$

- Với $r > 1$:

$$\frac{\partial(\mathbf{b}^T \mathbf{Z}^T \mathbf{Z} \mathbf{b})}{\partial \mathbf{b}} = \begin{bmatrix} \frac{\partial(\mathbf{b}^T \mathbf{Z}^T \mathbf{Z} \mathbf{b})}{\partial b_0} \\ \frac{\partial(\mathbf{b}^T \mathbf{Z}^T \mathbf{Z} \mathbf{b})}{\partial b_1} \\ \vdots \\ \frac{\partial(\mathbf{b}^T \mathbf{Z}^T \mathbf{Z} \mathbf{b})}{\partial b_r} \end{bmatrix} = 2\mathbf{Z}^T \mathbf{Z} \mathbf{b}$$

Tính đạo hàm $\frac{\partial(\mathbf{b}^T \mathbf{Z}^T \mathbf{y})}{\partial \mathbf{b}}$

- Với $r=1, n=3$:

$$\begin{aligned}\mathbf{b}^T \mathbf{Z}^T \mathbf{y} &= \begin{bmatrix} b_0 & b_1 \end{bmatrix} \begin{bmatrix} z_{10} & z_{20} & z_{30} \\ z_{11} & z_{21} & z_{31} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \\ &= \begin{bmatrix} b_0 & b_1 \end{bmatrix} \begin{bmatrix} z_{10}y_1 + z_{20}y_2 + z_{30}y_3 \\ z_{11}y_1 + z_{21}y_2 + z_{31}y_3 \end{bmatrix}\end{aligned}$$

Do đó:

$$\frac{\partial(\mathbf{b}^T \mathbf{Z}^T \mathbf{y})}{\partial \mathbf{b}} = \begin{bmatrix} \frac{\partial(\mathbf{b}^T \mathbf{Z}^T \mathbf{y})}{\partial b_0} \\ \frac{\partial(\mathbf{b}^T \mathbf{Z}^T \mathbf{y})}{\partial b_1} \end{bmatrix} = \begin{bmatrix} z_{10}y_1 + z_{20}y_2 + z_{30}y_3 \\ z_{11}y_1 + z_{21}y_2 + z_{31}y_3 \end{bmatrix} = \mathbf{Z}^T \mathbf{y}$$

Mở rộng cho n, r :

$$\frac{\partial(\mathbf{b}^T \mathbf{Z}^T \mathbf{y})}{\partial \mathbf{b}} = \begin{bmatrix} \frac{\partial(\mathbf{b}^T \mathbf{Z}^T \mathbf{y})}{\partial b_0} \\ \frac{\partial(\mathbf{b}^T \mathbf{Z}^T \mathbf{y})}{\partial b_1} \\ \vdots \\ \frac{\partial(\mathbf{b}^T \mathbf{Z}^T \mathbf{y})}{\partial b_r} \end{bmatrix} = \mathbf{Z}^T \mathbf{y}$$

$$\Rightarrow \frac{\partial S(\mathbf{b})}{\partial \mathbf{b}} = -2 \frac{\partial(\mathbf{b}^T \mathbf{Z}^T \mathbf{y})}{\partial \mathbf{b}} + \frac{\partial(\mathbf{b}^T \mathbf{Z}^T \mathbf{Z} \mathbf{b})}{\partial \mathbf{b}} = -2\mathbf{Z}^T \mathbf{y} + 2\mathbf{Z}^T \mathbf{Z} \mathbf{b}$$

Để tìm cực trị, xét đạo hàm bằng 0:

$$\frac{\partial S(\mathbf{b})}{\partial \mathbf{b}} = 0$$

$$\Leftrightarrow -2\mathbf{Z}^T \mathbf{y} + 2\mathbf{Z}^T \mathbf{Z} \mathbf{b} = \mathbf{0}$$

$$\Leftrightarrow \mathbf{Z}^T \mathbf{Z} \mathbf{b} = \mathbf{Z}^T \mathbf{y}$$

$$\Leftrightarrow \mathbf{b} = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{y}$$

Hệ số \mathbf{b} được gọi là ước lượng bình phương tối thiểu của tham số hồi quy $\boldsymbol{\beta}$, \mathbf{b} được biểu diễn bởi $\hat{\boldsymbol{\beta}}$ để nhấn mạnh vai trò là ước lượng của $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{y}$$

3. Đánh giá mô hình

Gọi $\hat{\varepsilon}$ là phần dư (residual) giữa giá trị thực và giá trị dự đoán của mô hình và đặt $\hat{\mathbf{y}} = \mathbf{Z}\hat{\boldsymbol{\beta}}$, ta có:

$$\begin{aligned} \hat{\varepsilon}_j &= y_j - \hat{y}_j \\ &= y_j - \hat{\beta}_0 - \hat{\beta}_1 z_{j1} - \dots - \hat{\beta}_r z_{jr} \quad (j = 1, 2, \dots, n) \end{aligned}$$

Hay

$$\begin{aligned} \hat{\varepsilon} &= \mathbf{y} - \hat{\mathbf{y}} \\ &= \mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\beta}} \\ &= \mathbf{y} - \mathbf{Z}(\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{y} \\ &= \mathbf{y} - \mathbf{H}\mathbf{y} \end{aligned}$$

$$= [I - H]y$$

Trong đó: $H = Z(Z^T Z)^{-1} Z^T$ (“hat” matrix)

Một số tính chất của $[I - H]$:

- $[I - H]^T = [I - H]$

Chứng minh:

$$\begin{aligned} H^T &= \left[Z(Z^T Z)^{-1} Z^T \right]^T \\ &= (Z^T)^T ((Z^T Z)^{-1})^T Z^T \\ &= Z((Z^T Z)^T)^{-1} Z^T \\ &= Z(Z^T Z)^{-1} Z^T \\ &= H \end{aligned}$$

- $[I - H][I - H] = [I - H]$

Chứng minh:

$$\begin{aligned} [I - H][I - H] &= [I - Z(Z^T Z)^{-1} Z^T][I - Z(Z^T Z)^{-1} Z^T] \\ &= I - 2Z(Z^T Z)^T Z^T + Z(Z^T Z)^{-1} Z^T Z(Z^T Z)^{-1} Z^T \\ &= I - 2Z(Z^T Z)^T Z^T + Z[(Z^T Z)^{-1} Z^T Z](Z^T Z)^{-1} Z^T \\ &= I - 2Z(Z^T Z)^T Z^T + Z(Z^T Z)^{-1} Z^T \\ &= I - Z(Z^T Z)^T Z^T \\ &= I - H \end{aligned}$$

- $Z^T [I - H] = 0$

Chứng minh:

$$\begin{aligned} Z^T [I - H] &= Z^T [I - Z(Z^T Z)^{-1} Z^T] \\ &= Z^T - Z^T Z(Z^T Z)^{-1} Z^T \\ &= Z^T - [(Z^T Z)(Z^T Z)^{-1}] Z^T \\ &= Z^T - I Z^T \\ &= Z^T - Z^T \\ &= 0 \end{aligned}$$

Trong thống kê sử dụng R^2 (R-Square) để đánh giá mô hình hồi quy ($0 \leq R^2 \leq 1$)

$$R^2 = 1 - \frac{\text{Unexplained Variation (residuals)}}{\text{Total Variation}} = 1 - \frac{\text{Tổng bình phương phần dư}}{\text{Tổng phương sai}} = 1 - \frac{SS_{res}}{SS_{tot}}$$

Trong đó:

$$\begin{aligned}
SS_{tot} &= \sum_{j=1}^n (y_j - \bar{y})^2 \\
&= \sum_{j=1}^n ((\hat{y}_j - \bar{y}) + (y_j - \hat{y}_j))^2 \\
&= \sum_{j=1}^n ((\hat{y}_j - \bar{y}) + \hat{\varepsilon}_j)^2 \\
&= \sum_{j=1}^n ((\hat{y}_j - \bar{y})^2 + \hat{\varepsilon}_j^2 + 2(\hat{y}_j - \bar{y})\hat{\varepsilon}_j) \\
&= \sum_{j=1}^n (\hat{y}_j - \bar{y})^2 + \sum_{j=1}^n \hat{\varepsilon}_j^2 + 2 \sum_{j=1}^n (\hat{y}_j - \bar{y})\hat{\varepsilon}_j \\
&= \sum_{j=1}^n (\hat{y}_j - \bar{y})^2 + \sum_{j=1}^n \hat{\varepsilon}_j^2 + 2 \sum_{j=1}^n \hat{\varepsilon}_j (\hat{\beta}_0 + \hat{\beta}_1 z_{j1} + \hat{\beta}_2 z_{j2} + \dots + \hat{\beta}_r z_{jr} - \bar{y}) \\
&= \sum_{j=1}^n (\hat{y}_j - \bar{y})^2 + \sum_{j=1}^n \hat{\varepsilon}_j^2 + 2(\hat{\beta}_0 - \bar{y}) \sum_{j=1}^n \hat{\varepsilon}_j + \sum_{j=1}^n \hat{\varepsilon}_j (\hat{\beta}_1 z_{j1} + \hat{\beta}_2 z_{j2} + \dots + \hat{\beta}_r z_{jr})
\end{aligned}$$

Và theo tính chất $\mathbf{Z}^T \hat{\varepsilon} = \mathbf{Z}^T (\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{Z}^T [\mathbf{I} - \mathbf{H}] \mathbf{y} = 0$

Nghĩa là:

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ z_{11} & z_{21} & \dots & z_{n1} \\ z_{12} & z_{22} & \dots & z_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ z_{1r} & z_{2r} & \dots & z_{nr} \end{bmatrix} \begin{bmatrix} \hat{\varepsilon}_1 \\ \hat{\varepsilon}_2 \\ \vdots \\ \hat{\varepsilon}_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$((r+1) \times n) \times (n \times 1) = ((r+1) \times 1)$$

Do đó:

$$\begin{aligned}
\sum_{j=1}^n \hat{\varepsilon}_j &= 0 \text{ và } \sum_{j=1}^n \hat{\varepsilon}_j z_{jr} = 0 \\
\rightarrow SS_{tot} &= \sum_{j=1}^n (\hat{y}_j - \bar{y})^2 + \sum_{j=1}^n \hat{\varepsilon}_j^2 = SS_{regression} + SS_{res}
\end{aligned}$$

- $\sum_{j=1}^n \hat{\varepsilon}_j^2$ là tổng bình phương phần dư:

$$SS_{res} = \sum_{j=1}^n \hat{\varepsilon}_j^2$$

- $\sum_{j=1}^n (\hat{y}_j - \bar{y})^2$ là tổng bình phương sai lệch giữa tiên đoán mô hình và giá trị trung bình của tập dữ liệu:

$$SS_{regression} = \sum_{j=1}^n (\hat{y}_j - \bar{y})^2$$

- Độ đo R^2 :

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = \frac{(SS_{tot} - SS_{res})}{SS_{tot}} = \frac{SS_{regression}}{SS_{tot}} = \frac{\sum_{j=1}^n (\hat{y}_j - \bar{y})^2}{\sum_{j=1}^n (y_j - \bar{y})^2}$$

Nhận xét:

- $0 \leq R^2 \leq 1$
- $R^2 = 1 \Rightarrow \hat{y}_j = y_j \rightarrow$ Giá trị dự đoán của mô hình khớp hoàn toàn với các giá trị kết quả của dữ liệu mẫu \Rightarrow Mô hình lý tưởng
- $R^2 = 0 \Rightarrow \hat{y}_j = \bar{y} \Leftrightarrow \hat{\beta}_0 + \hat{\beta}_1 z_{j1} + \hat{\beta}_2 z_{j2} + \dots + \hat{\beta}_r z_{jr} = \bar{y} \Leftrightarrow \begin{cases} \hat{\beta}_0 = \bar{y} \\ \hat{\beta}_r = 0 \end{cases} \rightarrow$ Các biến z_r không có liên hệ với $y_j \Rightarrow$ Mô hình không phù hợp

4. Ước lượng giá trị mới

Ta cần ước lượng giá trị Y_0 khi có một giá trị mới $z_0 \in \mathbb{R}^{(p+1) \times 1}$ 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 Y_0 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 ε_0 , ta có **mean** của Y_0 là giá trị của hàm hồi quy sau:

$$E(Y_0|z_0) = \beta_0 + \beta_1 z_{01} + \dots + \beta_r z_{0r} = z_0^T \beta$$

Khi đó, $E(Y_0|z_0)$ được ước lượng bằng phương pháp bình phương tối thiểu (Least Square Estimate) là $z_0^T \hat{\beta}$, trong đó $\hat{\beta}$ đã có được từ tập mẫu

Khi cho trước z_0^T , ta xem như đây là một vector hằng, hay ma trận hằng. Do đó phương sai dự đoán:

$$Var(z_0^T \hat{\beta}) = (z_0^T) Var(\hat{\beta}) (z_0^T)^T = \sigma^2 z_0^T (\mathbf{Z}^T \mathbf{Z})^{-1} z_0$$

Với ε có phân phối chuẩn $N(\mu, \sigma^2)$ (tham số σ^2 chưa biết)

Người ta chứng minh được khoảng tin cậy $100(1 - \alpha)\%$ cho $E(Y_0|z_0)$ là:

$$z_0^T \hat{\beta} \pm \left[t_{n-r-1} \left(\frac{\alpha}{2} \right) \right] \sqrt{(z_0^T (\mathbf{Z}^T \mathbf{Z})^{-1} z_0) s^2}$$

Trong đó $t_{n-r-1} \left(\frac{\alpha}{2} \right)$ là phân vị trên thứ $100 \left(\frac{\alpha}{2} \right)$ của phân phối student với $n - r - 1$ bậc tự do

5. 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

$$Y_0 = z_0^T \hat{\beta} + \varepsilon_0$$

Tương tự, ta có $\hat{\beta}$ từ tập mẫu. Tuy nhiên, ta không biết được ε_0 . Giả thiết rằng $\varepsilon_0 \sim N(0, \sigma^2)$, không phụ thuộc vào ε của tập mẫu, do đó cũng không phụ thuộc vào $\hat{\beta}$ và s^2
Ta có:

$$E(z_0^T \hat{\beta}) = z_0^T \beta \text{Var}(z_0^T \hat{\beta}) = z_0^T (\mathbf{Z}^T \mathbf{Z})^{-1} z_0 \sigma^2$$

Lỗi dự đoán:

$$Y_0 - z_0^T \hat{\beta} = z_0^T \beta + \varepsilon_0 - z_0^T \hat{\beta} = \varepsilon_0 + z_0^T (\beta - \hat{\beta})$$

Lấy kì vọng 2 vế:

$$\begin{aligned} E(Y_0 - z_0^T \hat{\beta}) &= E(\varepsilon_0 + z_0^T (\beta - \hat{\beta})) \\ &= E(\varepsilon_0) + E(z_0^T (\beta - \hat{\beta})) \\ &= 0 + z_0^T (E(\beta) - E(\hat{\beta})) \\ &= z_0^T (E(\beta) - E(\beta)) \\ &= 0 \end{aligned}$$

Vì ε_0 và $\hat{\beta}$ độc lập:

$$\begin{aligned} \text{Var}(Y_0 - z_0^T \hat{\beta}) &= \text{Var}(\varepsilon_0 + z_0^T (\beta - \hat{\beta})) \\ &= \text{Var}(\varepsilon_0) + \text{Var}(z_0^T \beta - z_0^T \hat{\beta}) \\ &= \sigma^2 + z_0^T (\mathbf{Z}^T \mathbf{Z})^{-1} z_0 \sigma^2 \\ &= \sigma^2 (1 + z_0^T (\mathbf{Z}^T \mathbf{Z})^{-1} z_0) \end{aligned}$$

Tương tự, người ta chứng minh được khoảng dự đoán $100(1 - \alpha)\%$ cho Y_0 là:

$$z_0^T \hat{\beta} \pm \left[t_{n-r-1} \left(\frac{\alpha}{2} \right) \right] \sqrt{s^2 (1 + z_0^T (\mathbf{Z}^T \mathbf{Z})^{-1} z_0)}$$

Với $t_{n-r-1} \left(\frac{\alpha}{2} \right)$ là phân vị trên thứ $100 \left(\frac{\alpha}{2} \right)$ của phân phối student với $n - r - 1$ bậc tự do

Vậy:

- Khoảng tin cậy $100(1 - \alpha)\%$ cho $E(Y_0|z_0)$ là:

$$z_0^T \hat{\beta} \pm \left[t_{n-r-1} \left(\frac{\alpha}{2} \right) \right] \sqrt{(z_0^T (\mathbf{Z}^T \mathbf{Z})^{-1} z_0) s^2}$$

- Khoảng tin cậy $100(1 - \alpha)\%$ cho Y_0 là:

$$z_0^T \hat{\beta} \pm \left[t_{n-r-1} \left(\frac{\alpha}{2} \right) \right] \sqrt{s^2 (1 + z_0^T (\mathbf{Z}^T \mathbf{Z})^{-1} z_0)}$$

6. Multivariate Linear Regression

Mô hình ban đầu:

$$Y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \dots + \beta_r z_r + \varepsilon$$

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

$$Y_1 = \beta_{01} + \beta_{11} z_1 + \dots + \beta_{r1} z_r + \varepsilon_1$$

$$Y_2 = \beta_{02} + \beta_{12} z_1 + \dots + \beta_{r2} z_r + \varepsilon_2$$

\vdots

$$Y_m = \beta_{0m} + \beta_{1m} z_1 + \dots + \beta_{rm} z_r + \varepsilon_m$$

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 β hơn-tương ứng nhiều mô hình hồi quy tuyến tính trên cùng một tập dữ liệu.

Dạng ma trận

$$\begin{matrix} \mathbf{Y} \\ (n \times m) \end{matrix} = \begin{bmatrix} Y_{11} & Y_{12} & \dots & Y_{1m} \\ Y_{21} & Y_{22} & \dots & Y_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n1} & Y_{n2} & \dots & Y_{nm} \end{bmatrix} = \begin{bmatrix} Y_{(1)} & \vdots & Y_{(2)} & \vdots & \dots & \vdots & Y_{(m)} \end{bmatrix}$$

$$\begin{matrix} \beta \\ ((r+1) \times m) \end{matrix} = \begin{bmatrix} \beta_{01} & \beta_{02} & \dots & \beta_{0m} \\ \beta_{11} & \beta_{12} & \dots & \beta_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{r1} & \beta_{r2} & \dots & \beta_{rm} \end{bmatrix} = [\beta_{(1)} \quad \vdots \quad \beta_{(2)} \quad \vdots \quad \dots \quad \vdots \quad \beta_{(m)}]$$

$$\begin{matrix} \varepsilon \\ (n \times m) \end{matrix} = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \dots & \varepsilon_{1m} \\ \varepsilon_{21} & \varepsilon_{22} & \dots & \varepsilon_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \dots & \varepsilon_{nm} \end{bmatrix} = [\varepsilon_{(1)} \quad \vdots \quad \varepsilon_{(2)} \quad \vdots \quad \dots \quad \vdots \quad \varepsilon_{(m)}] = \begin{bmatrix} \varepsilon'_1 \\ \vdots \\ \varepsilon'_2 \\ \vdots \\ \vdots \\ \vdots \\ \varepsilon'_n \end{bmatrix}$$

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

$$\begin{matrix} Y \\ (n \times m) \end{matrix} = \begin{matrix} Z \\ (n \times (r+1)) \end{matrix} \begin{matrix} \beta \\ ((r+1) \times m) \end{matrix} + \begin{matrix} \varepsilon \\ (n \times m) \end{matrix}$$

Ta tính được $\hat{\beta}_i$ bằng phương pháp Least Square Estimate, tức:

$$\hat{\beta}_i = (Z^T Z)^{-1} Z^T Y_i$$

Biểu diễn dạng ma trận tương ứng:

$$\hat{\beta} = [\hat{\beta}_{(1)} \quad \vdots \quad \hat{\beta}_{(2)} \quad \vdots \quad \dots \quad \vdots \quad \hat{\beta}_{(m)}] = (Z^T Z)^{-1} Z^T [Y_{(1)} \quad \vdots \quad Y_{(2)} \quad \vdots \quad \dots \quad \vdots \quad Y_{(m)}] = (Z^T Z)^{-1} Z^T Y$$

Hay:

$$\hat{\beta} = (Z'Z)^{-1} Z'Y$$

7. 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 (z_1, z_2, \dots, z_n) 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 :

$$Z = \begin{bmatrix} 1 & z_{11} & z_{12} & \dots & z_{1m} \\ 1 & z_{21} & z_{22} & \dots & z_{2m} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & z_{n1} & z_{n2} & \dots & z_{nm} \end{bmatrix} \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}$$

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
- *HS GPA* 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 kiểm được trong các khóa học nhân văn ở trường trung học
- *SS* Số giờ tín dụng kiểm 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 $\geq 50\%$ 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:

	GPA	HSGPA	SATV	SATM	Male	HU	SS	FirstGen	White	CollegeBound
1	3.06	3.83	680	770	1	3	9	1	1	1
2	4.15	4	740	720	0	9	3	0	1	1
3	3.41	3.7	640	570	0	16	13	0	0	1
4	3.21	3.51	740	700	0	22	0	0	1	1
5	3.48	3.83	610	610	0	30.5	1.5	0	1	1
6	2.95	3.25	600	570	0	18	3	0	1	1
7	3.6	3.79	710	630	0	5	19	0	1	1
8	2.87	3.6	390	570	0	10	0	0	0	0
9	3.67	3.36	630	560	0	8.5	15.5	0	1	1
10	3.49	3.7	680	670	0	16	12	0	1	1
11	3.25	3.53	380	470	0	18	7	0	0	1
12	3.18	3.48	630	670	0	26.5	1.5	0	0	1
13	3.85	3.81	680	740	1	34	0	0	1	1
14	2.58	3.38	710	750	1	8	3	0	1	1
15	3.5	3.8	670	650	0	20	3	0	1	1
16	3.17	3.55	580	690	0	14	3	0	1	1
17	2.74	3.4	510	610	0	11	10	0	1	1
18	3.42	3.76	690	670	0	4	17	1	1	1
19	2.9	3.74	660	640	0	3	18	0	0	1
20	3.16	3.83	580	640	1	4	14	0	1	1
21	3.78	3.98	720	730	1	7	6	0	1	1
22	2.72	3.44	550	570	0	15	3	0	0	1
23	2.81	3.59	580	630	1	16	6	0	1	1
24	3.56	3.6	690	710	1	8	10	0	1	1
25	3.3	3.56	700	640	1	10	10	0	1	1
26	4	4	450	640	1	22	11	0	0	1
27	3.4	3.2	630	770	1	6	12	0	1	1
28	2.29	3.3	630	450	0	16	4	0	0	1

Code đọc và xử lý dữ liệu đầu vào:

```
import csv
import numpy as np

def readData():
    X = []
    y = []
    with open('FirstYearGPA.csv') as f:
        r = csv.reader(f)

        # Skip header
        next(r)
        # Đọc X, y
        for line in r:
            xline = [1.0]
            # Đọc từ 2 trở về sau bỏ cột đầu là thứ tự và cột tiếp theo là y (GPA)
            for s in line[2:]:
                xline.append(float(s))

            X.append(xline)
            # Đọc y vào là cột
            y.append(float(line[1]))

    return (X,y)
X0,y0 = readData()
# Chuyển đổi tất cả trong data sang mảng numpy,
# Ngươi trừ 10 dòng cuối cùng để test
d = len(X0)-10
X = np.array(X0[:d])
y = np.transpose(np.array([y0[:d]]))
```


Áp dụng công thức:

- Công thức của β đã được rút gọn như sau:

$$\hat{\beta} = (Z'Z)^{-1}(Z'Y)$$

- Code tương ứng công thức:

```
# Giải Beta
Xt = np.transpose(X)
XtX = np.dot(Xt,X) # X.T*X
Xty = np.dot(Xt,y) # X.T*y
beta = np.linalg.solve(XtX,Xty) ## Solve (X.T*X)Beta = X.T*y
print("B:")
print(beta)
```

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

```
# Thử đưa dự đoán cho 10 hàng cuối trong data
print("Test:")
for data,actual in zip(X0[d:],y0[d:]):
    x = np.array([data])
    prediction = np.dot(x,beta)
    print('prediction = '+str(prediction[0,0])+' actual = '+str(actual))
```

```
B:
[[ 5.96817250e-01]
 [ 4.78214785e-01]
 [ 5.56672058e-04]
 [ 9.03806884e-05]
 [ 5.64581368e-02]
 [ 1.58222341e-02]
 [ 7.50929264e-03]
 [-8.05404850e-02]
 [ 2.03436601e-01]
 [ 1.49196176e-02]]
Test:
prediction = 3.2249145780589474 actual = 3.68
prediction = 3.2214274565651326 actual = 3.67
prediction = 2.643611464623139 actual = 2.52
prediction = 3.179361016319145 actual = 3.33
prediction = 3.271678179380128 actual = 3.31
prediction = 3.255276040306404 actual = 3.13
prediction = 3.1081370685122067 actual = 2.88
prediction = 2.972548187099255 actual = 2.65
prediction = 2.9426712477414605 actual = 2.97
prediction = 2.800260391472886 actual = 2.62
```

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

- Hệ số xác định

$$Y_i = \hat{Y}_i + e_i$$

$$\Leftrightarrow Y_i - \bar{Y} = \hat{Y}_i - \bar{Y} + e_i$$

$$\Rightarrow \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n e_i^2$$

Tổng dao động của Y so với giá trị trung bình *Dao động được giải thích bởi mô hình* *Dao động chưa được giải thích bởi mô hình – sai số*

- 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:
- $ESS = \sum (\hat{Y}_i - \bar{Y})^2$
- $RSS = \sum_{i=1}^n e_i^2$
- Đặt $R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$ $0 \leq R^2 \leq 1$

- **Hệ số xác định hiệu chỉnh**

- Đặt $\bar{R}^2 = 1 - \frac{\frac{RSS}{n-k}}{\frac{TSS}{n-1}} = 1 - \frac{n-1}{n-k} (1 - R^2)$

- Code tính R^2 (R_Square) và \bar{R}^2 ($Adjusted_R_Square$):

```
k = 9 # số cột của data
n = d # số hàng
mean = np.sum(y1)/n # tính mean
e = [] # rss sai số (chưa được giải thích bởi mô hình)
yy = [] # ess (dao động được giải thích bởi mô hình)
# Tính các tổng các giao động
for data,actual in zip(X0[:n],y0[:n]):
    x = np.array([data])
    prediction = np.dot(x,beta)
    e.append((actual-(prediction[0][0]))**2)
    yy.append((actual - mean)**2)
ESS = 0
RSS = 0
for i in range(d):
    ESS += yy[i]
    RSS += e[i]
TSS = RSS + ESS
R_Square = ESS/TSS
Adjusted_R_Square = 1 - (n-1)/(n-k)*(1 - R_Square)
print("R_Square: ",R_Square)
print("Adjusted R_Square: ", Adjusted_R_Square)
```

R_Square: 0.6027183766972313

Adjusted R_Square: 0.5868271117651205

8. 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.

9.1 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.

9.2 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.

9.3 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:

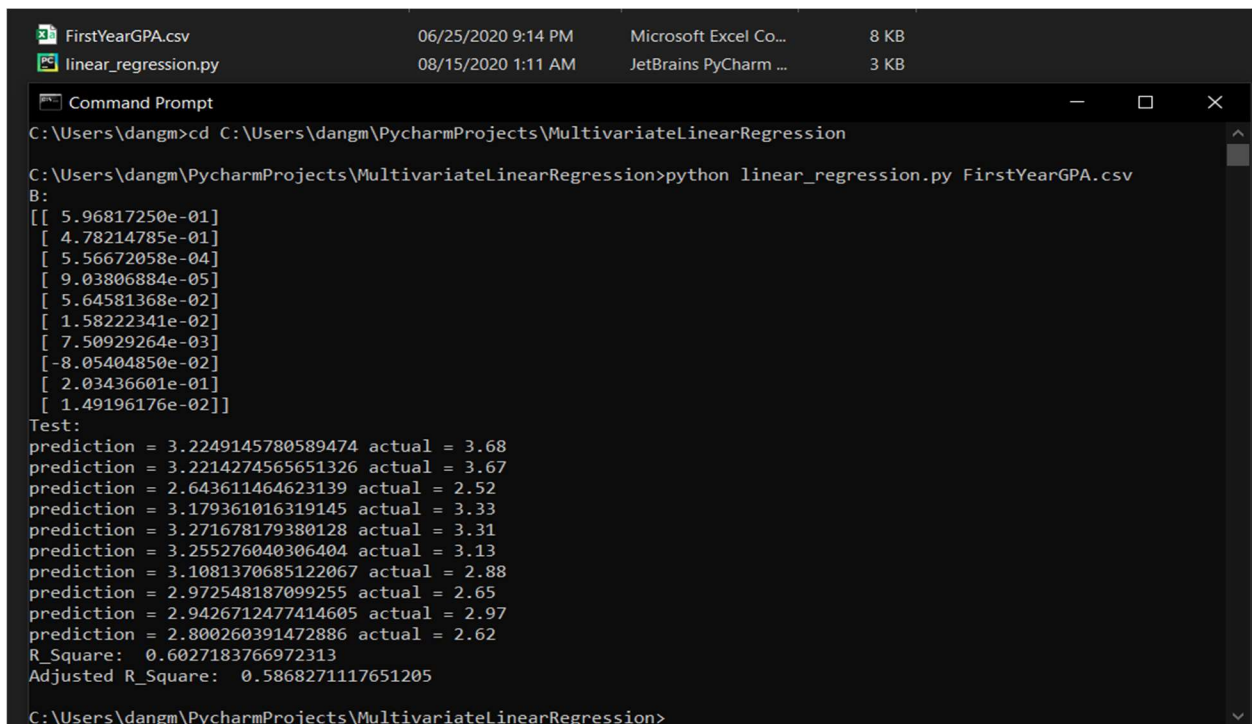
⇒ Dùng lệnh “cd” dẫn đến thư mục hiện hành

(cd D:\Source\MultivariateLinearRegression)

⇒ **python linear_regression.py FirstYearGPA.csv**

⇒

9.4 Màn hình kết quả:



```
FirstYearGPA.csv      06/25/2020 9:14 PM  Microsoft Excel Co...  8 KB
linear_regression.py  08/15/2020 1:11 AM  JetBrains PyCharm ...  3 KB

Command Prompt
C:\Users\dangm>cd C:\Users\dangm\PycharmProjects\MultivariateLinearRegression

C:\Users\dangm\PycharmProjects\MultivariateLinearRegression>python linear_regression.py FirstYearGPA.csv
B:
[[ 5.96817250e-01]
 [ 4.78214785e-01]
 [ 5.56672058e-04]
 [ 9.03806884e-05]
 [ 5.64581368e-02]
 [ 1.58222341e-02]
 [ 7.50929264e-03]
 [-8.05404850e-02]
 [ 2.03436601e-01]
 [ 1.49196176e-02]]
Test:
prediction = 3.2249145780589474 actual = 3.68
prediction = 3.2214274565651326 actual = 3.67
prediction = 2.643611464623139 actual = 2.52
prediction = 3.179361016319145 actual = 3.33
prediction = 3.271678179380128 actual = 3.31
prediction = 3.255276040306404 actual = 3.13
prediction = 3.1081370685122067 actual = 2.88
prediction = 2.972548187099255 actual = 2.65
prediction = 2.9426712477414605 actual = 2.97
prediction = 2.800260391472886 actual = 2.62
R_Square: 0.6027183766972313
Adjusted R_Square: 0.5868271117651205
C:\Users\dangm\PycharmProjects\MultivariateLinearRegression>
```

9. Tài liệu tham khảo

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