Unity
Thống nhất
Excellence
Vượt trội
Leadership
Tiên phong

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT

PHÂN TÍCH DỮ LIỆU VỚI R/PYTHON

GV: ThS Nguyễn Quang Phúc

www.uel.edu.vn

Phân tích dữ liệu với R/Python:

Phân tích dữ liệu với Python (phần 2): MỘT SỐ MÔ HÌNH CƠ BẢN TRONG PHÂN TÍCH DỰ BÁO

ThS. Nguyễn Quang Phúc
phucnq@uel.edu.vn

NỘI DUNG

1. Các dạng dữ liệu:
   - Chuỗi thời gian
   - Dữ liệu chéo
   - Dữ liệu bảng
2. Mô hình h ồi quy:
   - Tuyến tính
   - Logistic

1. Các dạng dữ liệu

Chuỗi thời gian

Chuỗi thời gian là một bảng dữ liệu với nhiều cột khác nhau, nhưng *bắt buộc phải có cột dữ liệu thể hiện thời gian* và *các biến cố thay đổi theo cột thời gian đó*.

	Date	High	Low	Open	Close	Avg	Volume
0	2016-01-04	47.5	45.2	45.6	47.5	46.81	4809120.0
1	2016-01-05	47.7	46.8	47.0	47.5	47.29	2480100.0
2	2016-01-06	47.9	46.7	47.4	47.5	47.16	2001950.0
3	2016-01-07	48.2	46.5	46.8	48.0	47.44	2852010.0
4	2016-01-08	48.0	47.0	47.5	48.0	47.86	1641950.0
	...	...	...	...	...	...	...
1246	2020-12-25	106.4	105.0	105.7	105.9	105.83	370300.0
1247	2020-12-28	106.6	105.4	105.9	105.9	105.96	711710.0
1248	2020-12-29	106.6	105.8	106.1	106.5	106.25	612360.0
1249	2020-12-30	109.6	106.7	106.7	108.5	108.74	1528950.0
1250	2020-12-31	109.3	105.5	108.5	108.2	108.60	656040.0

1. Các dạng dữ liệu

Chuỗi thời gian

Chuỗi thời gian là một bảng dữ liệu với nhiều cột khác nhau, nhưng *bắt buộc phải có cột dữ liệu thể hiện thời gian* và *các biến cố thay đổi theo cột thời gian đó*.

Year	CPI	Lai_suat	GTSX_CN
2007M1	111.0	6.5	49,212.0
2007M2	113.4	6.5	35,392.0
2007M3	113.1	6.5	45,154.0
2007M4	113.7	6.5	47,344.6
2007M5	114.5	6.5	47,953.4

1. Các dạng dữ liệu

Dữ liệu chéo

Dữ liệu chéo là một bảng dữ liệu với nhiều cột khác nhau, nhưng *bắt buộc phải có cột dữ liệu thể hiện chủ thể nghiên cứu khác nhau* và *các biến khác cần xem xét*. Các chủ thể có thể được hiểu là "không gian" cần phân tích, đánh giá.

Tỉnh	GTSX_CN	GTSX_TM
Hà Nội	2345	1244
HCM	2436	1242
Đà Nẵng	3454	1222
Hải Phòng	2333	1111

1. Các dạng dữ liệu

Dữ liệu bảng

Dữ liệu bảng là sự kết hợp giữa cấu trúc dữ liệu chuỗi thời gian và dữ liệu chéo.

Năm	Tỉnh	GTSX_TM	GTSX_CN
2008	Ha Noi	1244	4577
2009	Ha Noi	1242	4575
2010	Ha Noi	1222	4555
2011	Ha Noi	1111	4444
2008	HCM	2244	5577
2009	HCM	2242	5575
2010	HCM	1422	4755
2011	HCM	1151	4484

2. Mô hình hồi quy

Regression

Regression (hồi quy): là một trong những kỹ thuật thống kê và học máy cơ bản. Hồi quy giúp tìm ra ***mối quan hệ giữa các biến*** → mối quan hệ này được sử dụng để ***dự đoán*** (tiên lượng) các giá trị trong tương lai.

Ví dụ: phân tích tìm hiểu mức lương các nhân viên của một số công ty phụ thuộc vào các yếu tố nào, chẳng hạn như kinh nghiệm, trình độ học vấn, thành phố họ làm việc, …

Tương tự, chúng ta có thể thiết lập một sự phụ thuộc toán học của giá nhà vào diện tích, số phòng ngủ, thời gian xây dựng, khoảng cách đến trung tâm thành phố, …

2. Mô hình hồi quy

Linear Regression

Linear Regression (hồi quy tuyến tính): là một trong những kỹ thuật hồi quy được sử dụng rộng rãi. Đây là phương pháp hồi quy đơn giản nhất được ứng dụng trong phân tích dự báo.

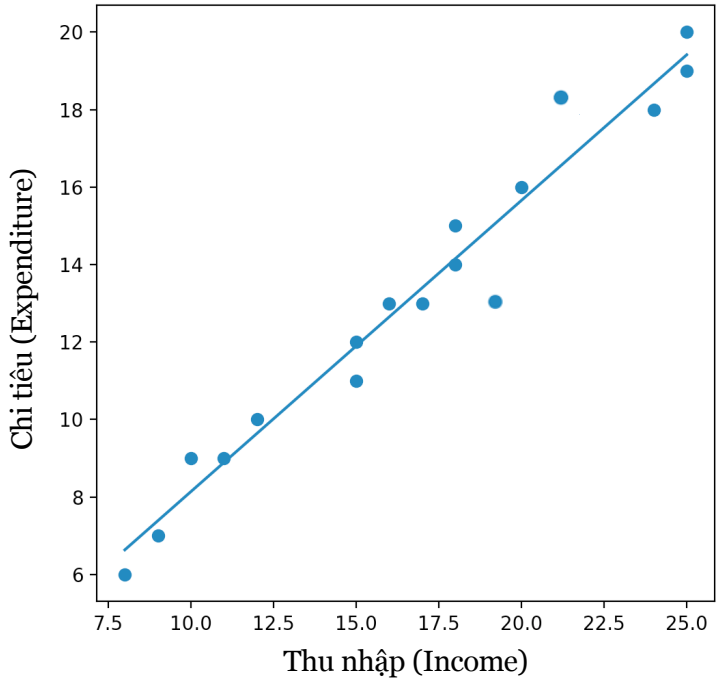
$$Y = \alpha + \beta X + \varepsilon$$

α : intercept

β : gradient / slope

ε : sai số ngẫu nhiên (những giao động về Y trong mỗi giá trị X)

Income	Expenditure
8	6
9	7
10	9
11	9
12	10
15	12
15	11
16	13
17	13
18	15
18	14
20	16
24	18
25	20
25	19



2. Mô hình hồi quy

Linear Regression

Giả định:

- Mối liên quan giữa X và Y là ***tuyến tính về tham số***.
- X không có sai số ngẫu nhiên.
- Giá trị của Y là độc lập với nhau.
- Sai số ngẫu nhiên (ε): có phân bố chuẩn, trung bình 0, phương sai bất biến.

$$\varepsilon \sim N(0, \sigma^2)$$

2. Mô hình hồi quy

Linear Regression

Mô hình hồi quy tổng thể (PRF):

$$Y = \alpha + \beta X + \varepsilon$$

$$E(Y|X_i) = \alpha + \beta X_i$$

Chúng ta không biết α và β nhưng có thể dùng dữ liệu thực nghiệm để ước tính 2 tham số đó.

2. Mô hình hồi quy

Linear Regression

Mô hình hồi quy mẫu (SRF):

$$\hat{Y}_i = a + bX_i$$

- $\hat{Y}_i$ là ước lượng của $E(Y_i|X_i)$.
- a, b là ước lượng của α và β.

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

2. Mô hình hồi quy

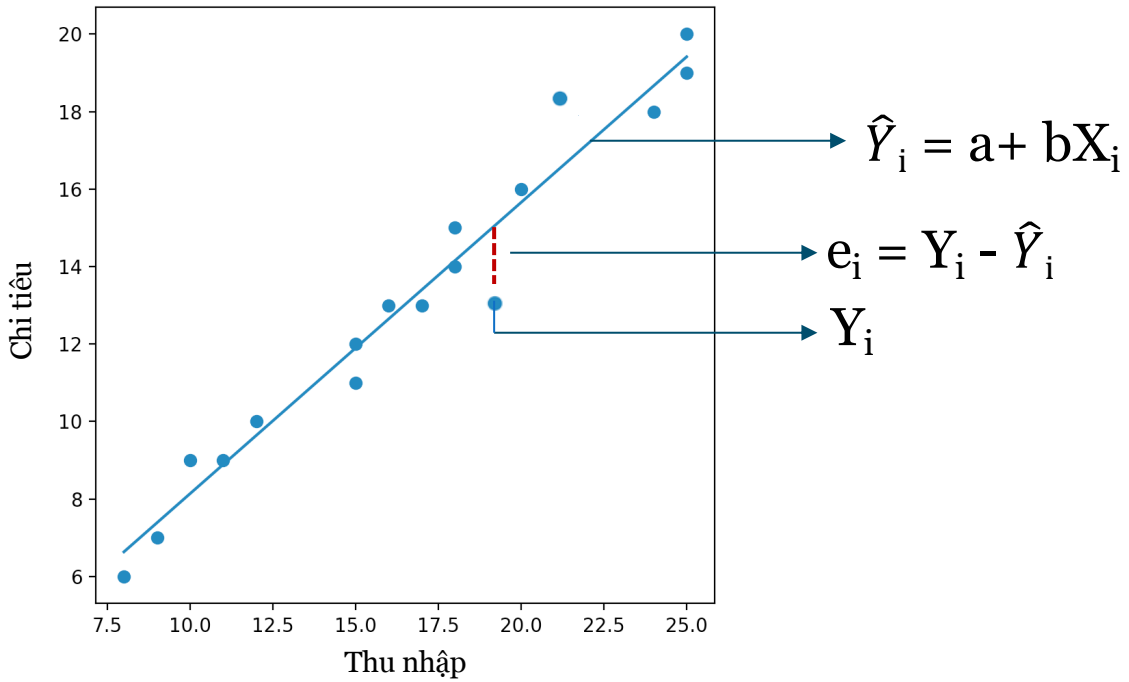
Linear Regression

Mô hình hồi quy mẫu (SRF):

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

Tìm a, b sao cho $\sum e_i^2 \rightarrow \min$.

Ordinary Least Square (OLS)

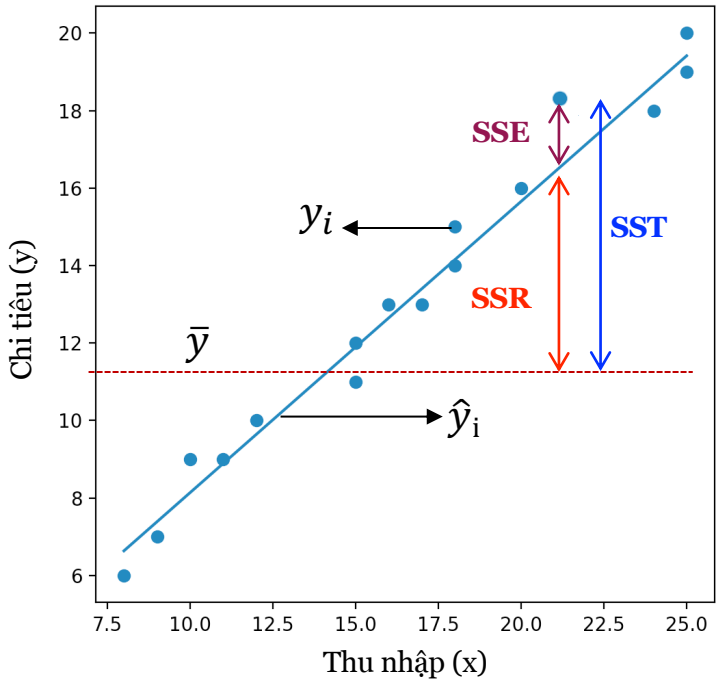


2. Mô hình hồi quy

Linear Regression

Mô hình hồi quy mẫu (SRF):

do sai lech, do loi thong qua mo hinh hoi quy, ki vong SSE nho lai, toi da hoa SSR, toi thieu hoa SSE



SST (Total sum of squares): $\sum_{i=1}^{n}(y_i - \bar{y})^2$

→ Thể hiện sự thay đổi của y_i so với $\bar{y}$

SSR (Regression sum of squares): $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

→ Thể hiện sự thay đổi của $\hat{y}_i$ so với $\bar{y}$

SSE (Error (residual) sum of squares): $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

→ Thể hiện sự thay đổi của y_i so với $\hat{y}_i$

R^2 (coefficient of determination): $R^2 = SSR/SST \ (\in [0, 1])$ => mo hinh hoi quy có the giai thich duoc

r (coefficient of correlation): $r = \sqrt{R^2} \ (\in [-1, 1])$

he so tuong quan, tuong quan giua cac bien vs nhau, cang tien gan ve 1, su tuong quan cang manh

2. Mô hình hồi quy

Linear Regression

Vd: với dữ liệu quan sát thu nhập (triệu đồng) và chi tiêu (triệu đồng) của các hộ gia đình hãy ước tính mối liên quan giữa thu nhập và chi tiêu.

	Income	Expenditure
0	8	6
1	9	7
2	10	9
3	11	9
4	12	10
5	15	12
6	15	11
7	16	13
8	17	13
9	18	15
10	18	14
11	20	16
12	24	18
13	25	20
14	25	19

$$\hat{Y}_i = a + bX_i$$

Tuy du lieu cua minh ma nen xem xet co he so a hay khong

$$\widehat{Expenditure} = a + b^*\text{Income}$$

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.983
Model:                            OLS   Adj. R-squared:                  0.982
Method:                 Least Squares   F-statistic:                     747.5
Date:                Mon, 10 May 2021   Prob (F-statistic):           7.13e-13
Time:                        10:30:00   Log-Likelihood:                -11.998
No. Observations:                  15   AIC:                             28.00
Df Residuals:                      13   BIC:                             29.41
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.6207      0.470      1.321      0.209      -0.394       1.636
x1             0.7518      0.027     27.341      0.000       0.692       0.811
==============================================================================
Omnibus:                        1.567   Durbin-Watson:                   2.716
Prob(Omnibus):                  0.457   Jarque-Bera (JB):                0.855
Skew:                           0.045   Prob(JB):                        0.652
Kurtosis:                       1.834   Cond. No.                         53.9
==============================================================================
```

mo hinh giai thich duoc 98% chi tieu cua thu nhap

neu de bai 5% thi gia tri nay co y nghia: 0,....713

Khi thu nhap tang len 1 trieu, thi chi tieu tang len 751 nghin

2. Mô hình hồi quy

Linear Regression

Vd: dự báo giá căn hộ theo diện tích và số lượng phòng ngủ.

$\widehat{Price} = a + b*Area$

$\widehat{Price} = a + b*Bedrooms$?

$\widehat{Price} = a + b*Area + c*Bedrooms$

tinh da cong tuyen, su phu thuoc giua cac bien doc lap,

hai bien doc lap co su tuong quan thi co so khong con dung nua, nên loai mo hinh do
=> lam ro co so lua chon mo hinh bien doc lap.

	Price	Area	Bedrooms
0	4.70000	74	2.50000
1	5.30000	80	3.00000
2	6.60000	100	3.00000
3	2.60000	50	1.00000
4	5.50000	80	2.50000
5	2.30000	48	1.50000
6	4.50000	74	2.00000
7	7.50000	113	3.00000
8	5.60000	80	2.50000
9	2.40000	48	1.50000
10	4.10000	69	2.50000
11	6.20000	95	3.00000
12	8.30000	135	3.00000
13	3.20000	56	2.00000
14	2.80000	50	1.50000
15	6.30000	95	2.00000
16	5.40000	80	2.00000
17	2.20000	48	1.00000
18	7.70000	113	3.00000

2. Mô hình hồi quy

Linear Regression

➢ Mô hình hồi quy tuyến tính đa biến

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

Diễn giải mô hình: → *cho biết lượng thay đổi của biến phụ thuộc khi biến giải thích thứ j thay đổi một đơn vị, trong điều kiện các biến giải thích khác và sai số không đổi.*

***Lưu ý:** với dữ liệu thực tế thì các biến giải thích $x_1, x_2, \ldots, x_k$ có thể tương quan (chịu tác động) lẫn nhau → **hiện tượng đa cộng tuyến**.

Cách phát hiện đa cộng tuyến và phương pháp xử lý?

2. Mô hình hồi quy

Linear Regression

➢ Lựa chọn mô hình

Dữ liệu nghiên cứu thường có nhiều biến → số mô hình có thể rất nhiều, với ***k*** biến thì số mô hình tối thiểu là 2^k-1

k = 2, số mô hình tối thiểu là 3; k = 3, số mô hình tối thiểu là 7; …

k = 10, số mô hình tối thiểu là 1023

→ Chọn mô hình sao cho có **ít biến giải thích** nhưng có thể **"giải thích" tối đa** dữ liệu.

***Tiêu chuẩn**: R^2, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC)

***Phương pháp**: Stepwise, Backward, Forward Regression

2. Mô hình hồi quy

Linear Regression

BT: dự đoán lượng khí thải CO_2 của một chiếc xe hơi dựa trên trọng lượng, dung tích xi lanh của động cơ.

$\widehat{CO_2}$ = a + b*Weight

$\widehat{CO_2}$ = a + b*Volume

?

$\widehat{CO_2}$ = a + b*Weight + c*Volume

	Car	Model	Volume	Weight	CO2
0	Toyoty	Aygo	1000	790	99
1	Mitsubishi	Space Star	1200	1160	95
2	Skoda	Citigo	1000	929	95
3	Fiat	500	900	865	90
4	Mini	Cooper	1500	1140	105
5	VW	Up!	1000	929	105
6	Skoda	Fabia	1400	1109	90
7	Mercedes	A-Class	1500	1365	92
8	Ford	Fiesta	1500	1112	98
9	Audi	A1	1600	1150	99
10	Hyundai	I20	1100	980	99
11	Suzuki	Swift	1300	990	101
12	Ford	Fiesta	1000	1112	99
13	Honda	Civic	1600	1252	94
14	Hundai	I30	1600	1326	97
15	Opel	Astra	1600	1330	97
16	BMW	1	1600	1365	99
17	Mazda	3	2200	1280	104
18	Skoda	Rapid	1600	1119	104
19	Ford	Focus	2000	1328	105
20	Ford	Mondeo	1600	1584	94
21	Opel	Insignia	2000	1428	99
22	Mercedes	C-Class	2100	1365	99
23	Skoda	Octavia	1600	1415	99
24	Volvo	S60	2000	1415	99
25	Mercedes	CLA	1500	1465	102
26	Audi	A4	2000	1490	104
27	Audi	A6	2000	1725	114
28	Volvo	V70	1600	1523	109
29	BMW	5	2000	1705	114
30	Mercedes	E-Class	2100	1605	115

2. Mô hình hồi quy

Logistic Regression

Logistic Regression (hồi quy logistic): là một kỹ thuật thống kê xem xét mối liên hệ giữa *biến độc lập (biến liên tục hoặc nhị phân)* và *biến phụ thuộc (biến nhị phân)*.

Linear Regression

- Giá nhà
- Giá cổ phiếu
- Doanh số
- Hàng tồn kho
- Sức mua của KH
- ...

Logistic Regression

- Phân loại KH
- Phân loại SP
- Thư spam?
- ...

Khach hang co nhu cau mua san pham co hay khong?

Bien phu thuoc: bien phan loai

Voi do tuoi do co hay khong co kha nang mua may tinh

2. Mô hình hồi quy

Logistic Regression

Logistic Regression (hồi quy logistic): là một kỹ thuật thống kê xem xét mối liên hệ giữa *biến độc lập (biến liên tục hoặc nhị phân)* và *biến phụ thuộc (biến nhị phân)*.

$$y = \alpha + \beta x + \varepsilon$$

y: biến phụ thuộc với 2 trạng thái (0/1; true/false; yes/no)

→ Mô hình hồi quy logistic được phát biểu như sau:

logit(p) $\log(\frac{p}{1-p}) = \alpha + \beta x + \varepsilon$ => Odds ratio = exp(β)

p là xác suất biến cố xảy ra và 1-p là xác suất biến cố không xảy ra

→ Xác suất tiên lượng theo trị số của x:

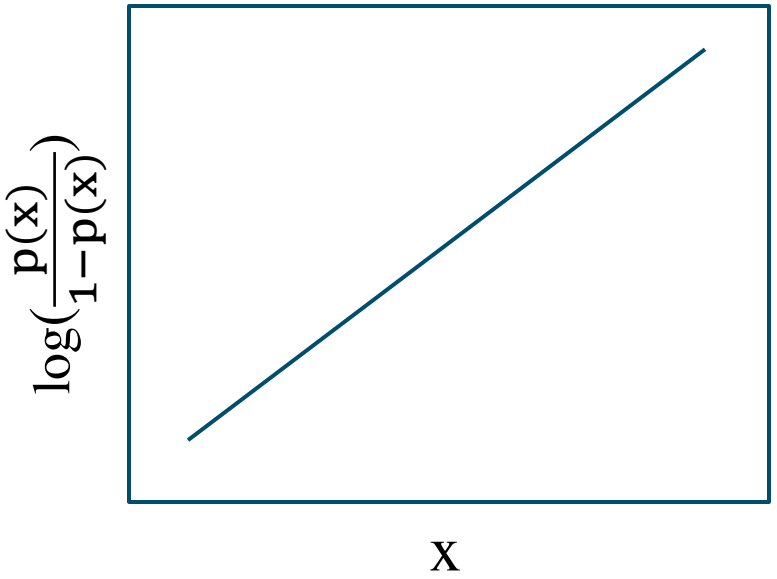
can cu odds ratio de giai thich mo hinh

Odds $\frac{p}{1-p} = e^{\alpha + \beta x}$ $p = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}$

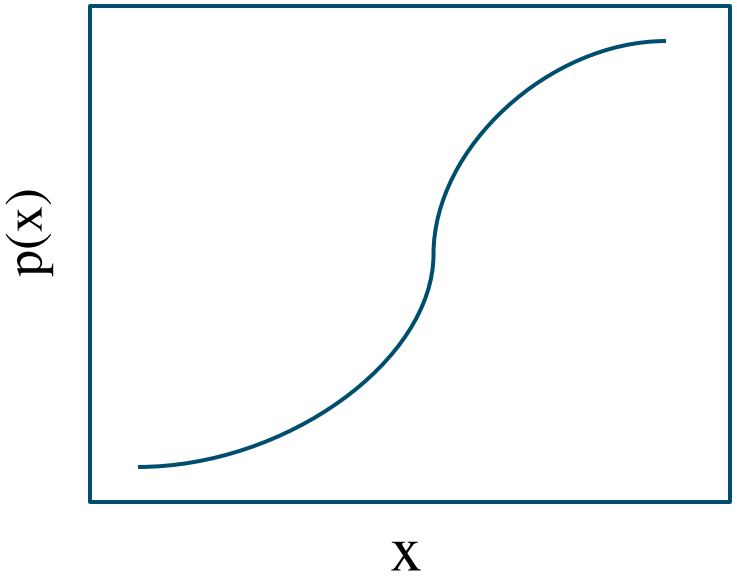
2. Mô hình hồi quy

Logistic Regression

$$\log\left(\frac{p}{1-p}\right)$$



$$p = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}$$



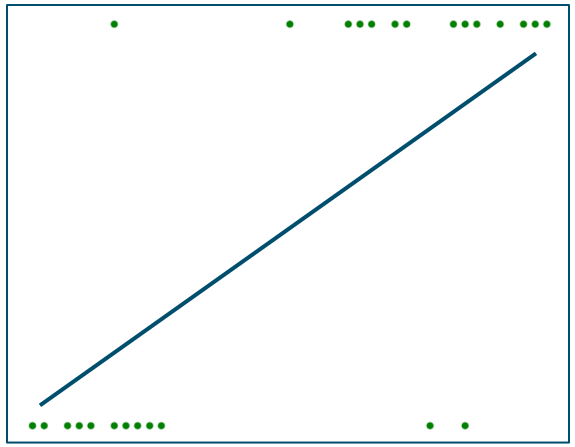
2. Mô hình hồi quy

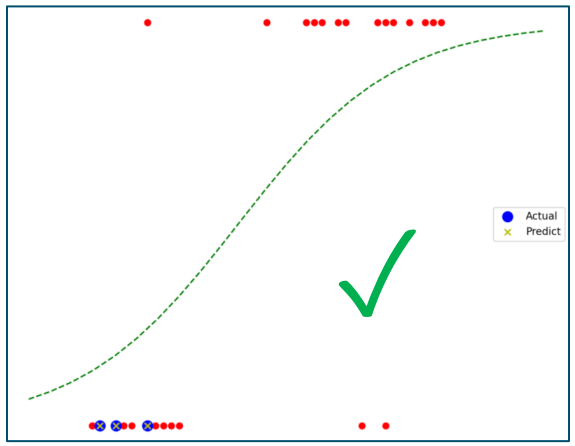
Logistic Regression

Vd: dự đoán khả năng mua bảo hiểm của khách hàng dựa theo độ tuổi:

logit(p) = a + b*age

Với p là xác suất mua bảo hiểm





logit(p) = -4.0389 + 0.1042*age

=> Odds ratio = exp(0.1042) = **1.11**

?

	age	bought_insurance
0	22	0
1	25	0
2	47	1
3	52	0
4	46	1
5	56	1
6	55	0
7	60	1
8	62	1
9	61	1
10	18	0
11	28	0
12	27	0
13	29	0
14	49	1
15	55	1
16	25	1
17	58	1
18	19	0
19	18	0
20	21	0
21	26	0
22	40	1
23	45	1
24	50	1
25	54	1
26	23	0

🏠: SV tìm hiểu thêm về mô hình hồi quy Logistic với biến độc lập là *biến nhị phân*, biến độc lập là *biến thứ bậc*.

2. Mô hình hồi quy

Logistic Regression

Vd: dự đoán khả năng mua bảo hiểm của khách hàng dựa theo độ tuổi:

$$\text{logit}(p) = a + b*\text{age}$$

Với p là xác suất mua bảo hiểm

Logit Regression Results

Dep. Variable:	bought_insurance	**No. Observations:**	27
Model:	Logit	**Df Residuals:**	25
Method:	MLE	**Df Model:**	1
Date:	Fri, 30 Jul 2021	**Pseudo R-squ.:**	0.4543
Time:	19:37:50	**Log-Likelihood:**	-10.203
converged:	True	**LL-Null:**	-18.696
Covariance Type:	nonrobust	**LLR p-value:**	3.764e-05

	coef	std err	z	P>|z|	[0.025	0.975]
const	-5.2729	1.814	-2.907	0.004	-8.828	-1.718
age	0.1357	0.044	3.118	0.002	0.050	0.221

	age	bought_insurance
0	22	0
1	25	0
2	47	1
3	52	0
4	46	1
5	56	1
6	55	0
7	60	1
8	62	1
9	61	1
10	18	0
11	28	0
12	27	0
13	29	0
14	49	1
15	55	1
16	25	1
17	58	1
18	19	0
19	18	0
20	21	0
21	26	0
22	40	1
23	45	1
24	50	1
25	54	1
26	23	0


Q
A

THANK YOU

028 37244555 www.uel.edu.vn

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT

Số 669, đường Quốc lộ 1, khu phố 3, phường Linh Xuân,
quận Thủ Đức, Thành phố Hồ Chí Minh.

UNITY - EXCELLENCE - LEADERSHIP