

LAB03. PHÂN TÍCH HỒI QUY

IS403 – PHÂN TÍCH DỮ LIỆU KINH DOANH



TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA HỆ THỐNG THÔNG TIN

Người trình bày: Nguyễn Minh Nhựt

Số điện thoại: 0939013911 - 0981734105

TÓM TẮT NỘI DUNG LAB3

- Sinh viên nắm được hồi quy **tuyến tính đa biến**, hồi quy **phi tuyến đa biến**.
- Quá trình tự **Logistics**
- Hướng dẫn quy trình thực hiện **đồ án môn học**
- **Đánh giá mô hình hồi quy**
- Thực hành trên các công cụ: **Python, R, Excel**



CHỦ ĐỀ 1

HỒI QUY TUYẾN TÍNH, PHI TUYẾN ĐƠN BIỀN, ĐA BIỀN

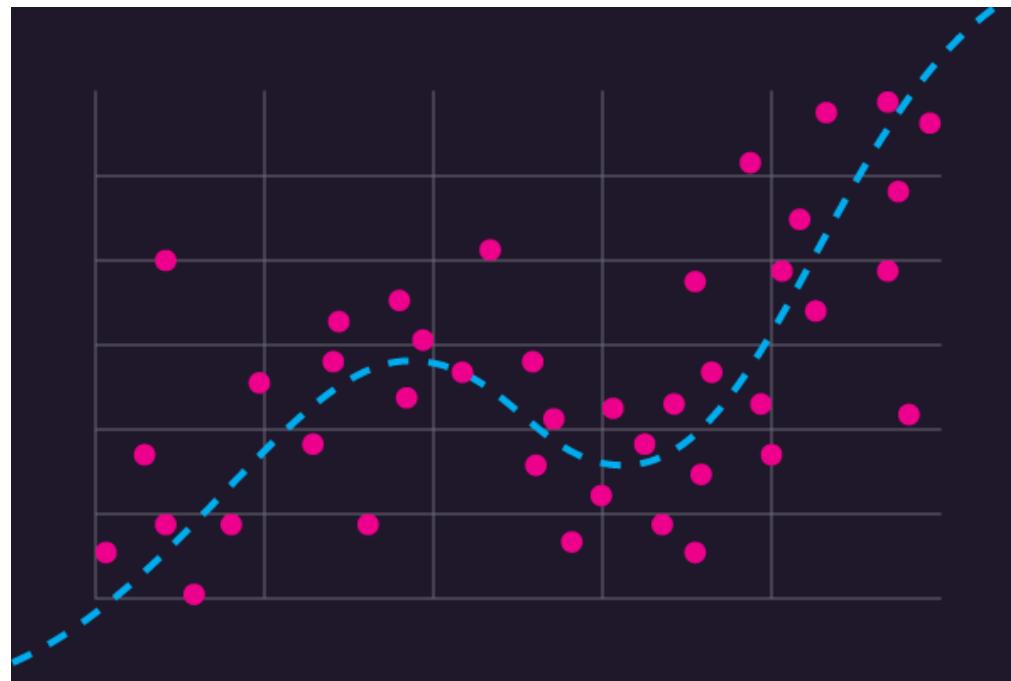
(LINEAR REGRESSION, NON-LINEAR REGRESSION)

- *Linear Regression, Non-Linear*
- *Multivariables*
- *Thực hành trên R, Python và Excel*

1 - DẪN NHẬP VỀ HỒI QUY

REGRESSION

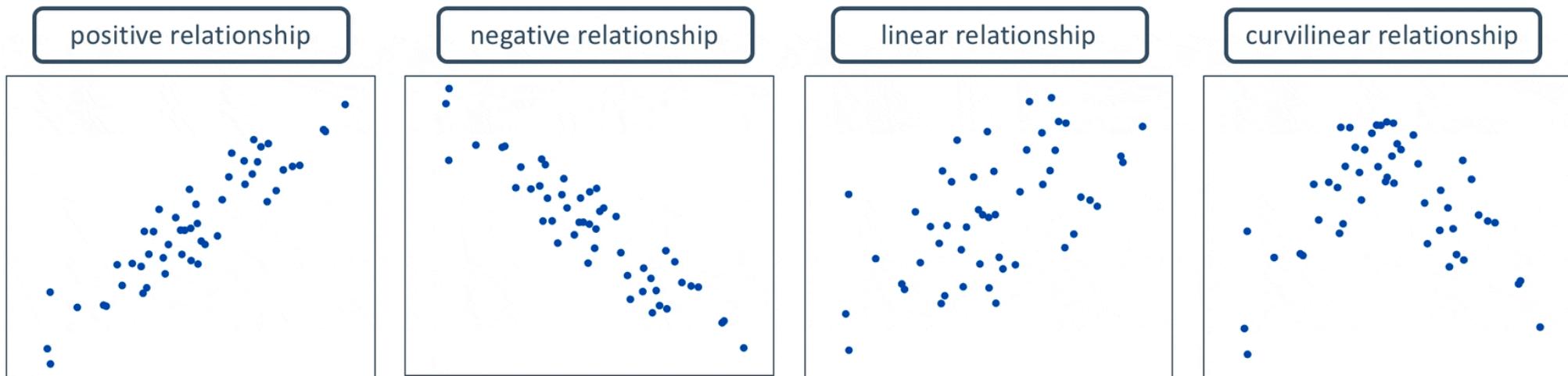
**Đưa ra mối
quan hệ giữa
biến phụ
thuộc Y và
một dãy biến
độc lập X**



**Để đưa ra quyết
định dự đoán
dựa vào hàm
phụ thuộc**

1- DẪN NHẬP VỀ HỒI QUY

CATEGORY REGRESSION

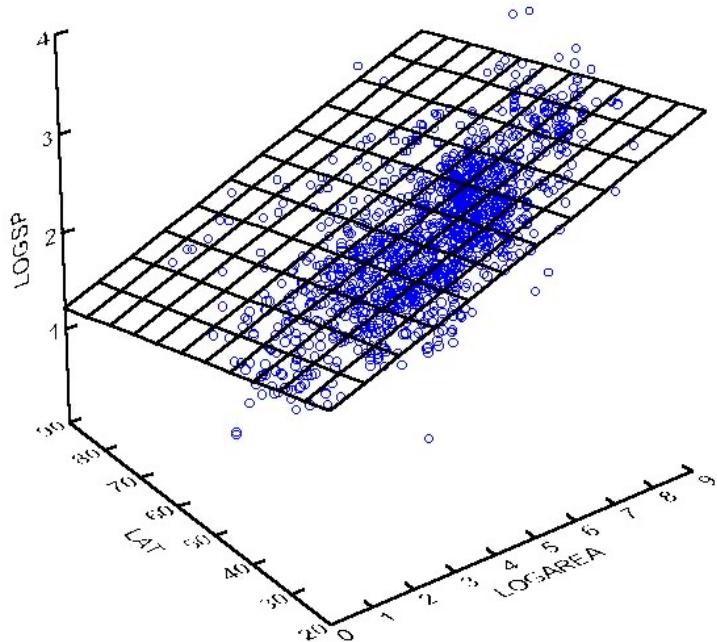


$$Y = \beta_0 + \beta_1 X + e$$

SLR (Simple Linear Regression)

1 - DẪN NHẬP VỀ HỒI QUY

CATEGORY REGRESSION



$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + e$$

MLR (Multiple Linear Regression)

2- HỒI QUY TUYẾN TÍNH ĐƠN BIỀN

EXAMPLE SLR

Time (Minutes)	60	120	200	90	10	20	30	50	80
Score	7	8	9	8,5	4,5	5	6,5	7	8

$$\text{Score} = 5,436 + 0,022 * \text{Time}$$

Cách tính như thế nào?

2- HỒI QUY TUYẾN TÍNH ĐƠN BIÊN

EXAMPLE MLR

Giá nhà = 0.1227 + 0.0352*[Diện tích] + 0.8436*[Số phòng ngủ] - 0.0148*[Độ tuổi ngôi nhà]

Cách tính như thế nào?

Diện tích (m ²)	Số phòng ngủ	Độ tuổi của nhà	Giá bán (tỷ đồng)
120	3	5	3.8
80	2	2	2.5
150	4	10	5.2
100	3	7	3.0
200	5	15	7.0

3- THỰC HÀNH TRÊN NGÔN NGỮ R

DỮ LIỆU HOME MARKET VALUE

	A	B	C	D	E	F	G
1	Home Market Value						
2							
3	House Age	Square Feet	Market Value				
4	33	1.812	\$90.000,00				
5	32	1.914	\$104.400,00				
6	32	1.842	\$93.300,00				
7	33	1.812	\$91.000,00				
8	32	1.836	\$101.900,00				
9	33	2.028	\$108.500,00				
10	32	1.732	\$87.600,00				
11	33	1.850	\$96.000,00				
12	32	1.791	\$89.200,00				
13	33	1.666	\$88.400,00				
14	32	1.852	\$100.800,00				
15	32	1.620	\$96.700,00				
16	32	1.692	\$87.500,00				
17	32	2.372	\$114.000,00				
18	32	2.372	\$113.200,00				
19	33	1.666	\$87.500,00				
20	32	2.123	\$116.100,00				
21	32	1.620	\$94.700,00				
22	32	1.731	\$86.400,00				
23	32	1.666	\$87.100,00				
24	28	1.520	\$83.400,00				

3- THỰC HÀNH TRÊN NGÔN NGỮ R

THỰC HÀNH HỒI QUY TUYẾN TÍNH BẰNG R

```
> reg1 = lm(Market.Value~Square.Feet+House.Age)
> summary(reg1)

Call:
lm(formula = Market.Value ~ Square.Feet + House.Age)

Residuals:
    Min      1Q  Median      3Q     Max 
 -9164   -4220   -2175   2487  30968 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 47331.382 13884.347   3.409  0.00153 ** 
Square.Feet    40.911     6.697   6.109 3.65e-07 ***  
House.Age   -825.161    607.313  -1.359  0.18205    
---
Signif. codes:  0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 ' ' 1

Residual standard error: 7212 on 39 degrees of freedom
Multiple R-squared:  0.5558,    Adjusted R-squared:  0.533 
F-statistic: 24.4 on 2 and 39 DF,  p-value: 1.344e-07
```

- Thực hiện hồi quy tuyến tính trong R

`reg = lm(Y~x1 + x2 + ...)`

- Công thức: $Y \sim x_1 + x_2$ (Cho trường hợp đa biến)

- Giải thích bảng kết Home Market Value

- Tham khảo file **Excel** mẫu

→ Nhìn vào Coefficients ta thấy $Pr(>|t|)$ của House.Age > 0.05. Nên ta có thể thực hiện việc chọn lại mô hình loại đi House.Age xem kết quả có khả quan hơn hay không?

3- THỰC HÀNH TRÊN NGÔN NGỮ R

THỰC HÀNH HỒI QUY TUYẾN TÍNH BẰNG R

```
> req2 = lm(Market.Value~Square.Feet)
> summary(req2)

Call:
lm(formula = Market.Value ~ Square.Feet)

Residuals:
    Min      1Q  Median      3Q     Max 
 -8067   -4327   -1923    3097   32634 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 32673.220   8831.951   3.699  0.00065 ***
Square.Feet    35.036     5.167   6.780  3.8e-08 ***  
---
Signif. codes:  0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 '.' 1 

Residual standard error: 7288 on 40 degrees of freedom
Multiple R-squared:  0.5347,    Adjusted R-squared:  0.5231 
F-statistic: 45.97 on 1 and 40 DF,  p-value: 3.798e-08
```

Nhận xét về độ tương quan R-squared, Adjusted R-squared ta thấy mô hình sau là phù hợp. Nên Ta có

$$\text{Market.Value} = 32673.220 + 35.036 * \text{Square.Feet}$$

- Biểu diễn phương trình hồi qui Market.Value và Square.Feet

```
> abline(lm(y~x))
```

4- THỰC HÀNH TRÊN NGÔN NGỮ PYTHON

THỰC HÀNH HỒI QUY TUYẾN TÍNH BẰNG PYTHON

```
✓ [31] # TRƯỜNG HỢP ĐA BIẾN  
0s   x, y = \  
     np.array(df[["Square Feet", "House Age"]]).reshape((-1, 2)), \  
     np.array(df[["Market Value"]]).reshape((-1, 1))
```

```
✓ [32] ⏎ model = LinearRegression()  
0s   model.fit(x,y)  
    ↓  
    ▾ LinearRegression  
    LinearRegression()
```

```
✓ [33] ⏎ print("Hệ số chấn:", model.intercept_)  
0s   print("Hệ số phụ thuộc:", model.coef_)  
   print("Điểm R Square:", model.score(x, y))
```

```
Hệ số chấn: [47331.38153562]  
Hệ số phụ thuộc: [[40911.06844844 -825.16122035]]  
Điểm R Square: 0.5557624615954795
```

- Tạo mảng X gồm nhiều biến và Y là biến market value

```
x, y = np.array(df[["Square Feet", "House Age"]]).reshape((-1, 2)), \  
np.array(df[["Market Value"]]).reshape((-1, 1))
```

- Sử dụng hàm LinearRegression và fit x và y

```
model = LinearRegression()  
  
model.fit(x,y)
```

- Đưa ra kết quả

```
print("Hệ số chấn:", model.intercept_)  
print("Hệ số phụ thuộc:", model.coef_)  
print("Điểm R Square:", model.score(x, y))
```

5- HỒI QUY PHI TUYẾN ĐƠN BIỀN

EXAMPLE SNLR

Mô hình hồi quy phi tuyến khi ít nhất một tham số là hàm số phi tuyến. Qua một số ví dụ như sau

$$\log(Y) = \beta_0 + \beta_1 X + e$$

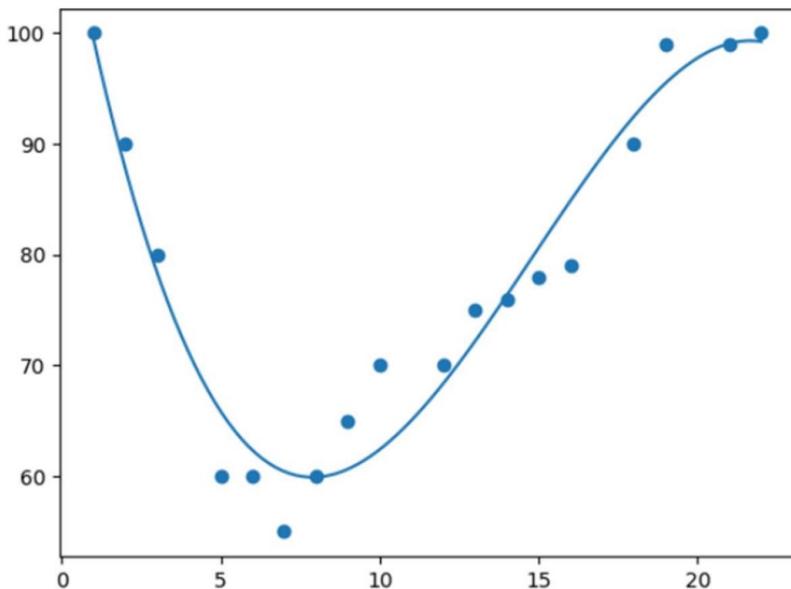
$$Y = \exp(\beta_0 + \beta_1 X + e)$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2^2 + e$$

Xác định hàm phi tuyến nào là hợp lý?

5- HỒI QUY PHI TUYẾN ĐƠN BIỀN

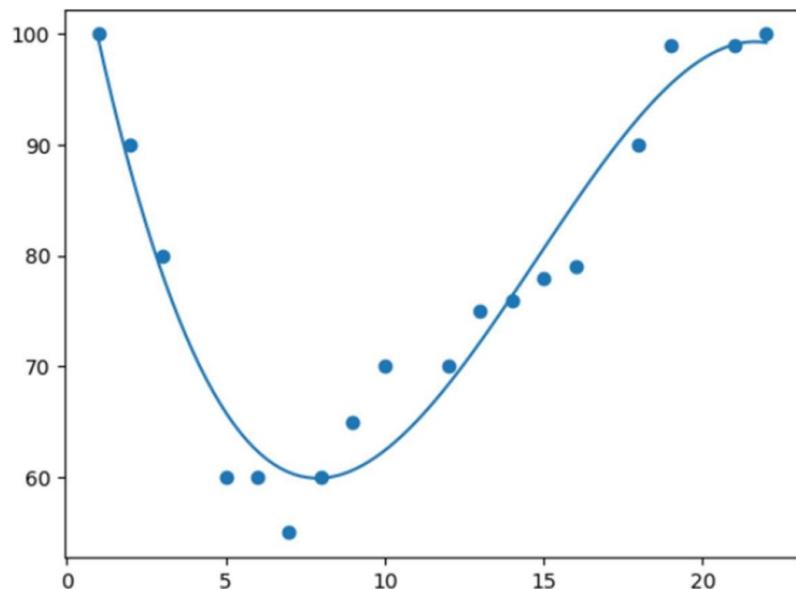
MỘT SỐ HÀM PHI TUYẾN TRONG R



Tên hàm	Formula	Giải thích
poly	<code>poly(x, degree=z, raw=TRUE)</code>	Hàm đa thức
logarithms	<code>log(x)</code>	Hàm ln()
sin	<code>sin(x)</code>	Hàm sin
cos	<code>cos(x)</code>	Hàm cos
	<code>log(x, base=y)</code>	<code>logy(x)</code>

5- HỒI QUY PHI TUYẾN ĐƠN BIỀN

MỘT SỐ HÀM PHI TUYẾN TRONG PYTHON



- Để sử dụng các hàm toán học trong python dùng thư viện math được cài đặt sẵn:

```
import math
```

- Ví dụ một số hàm toán học

```
math.cos(5)
```

```
math.log(5)
```

Tham khảo:

https://www.w3schools.com/python/module_math.asp



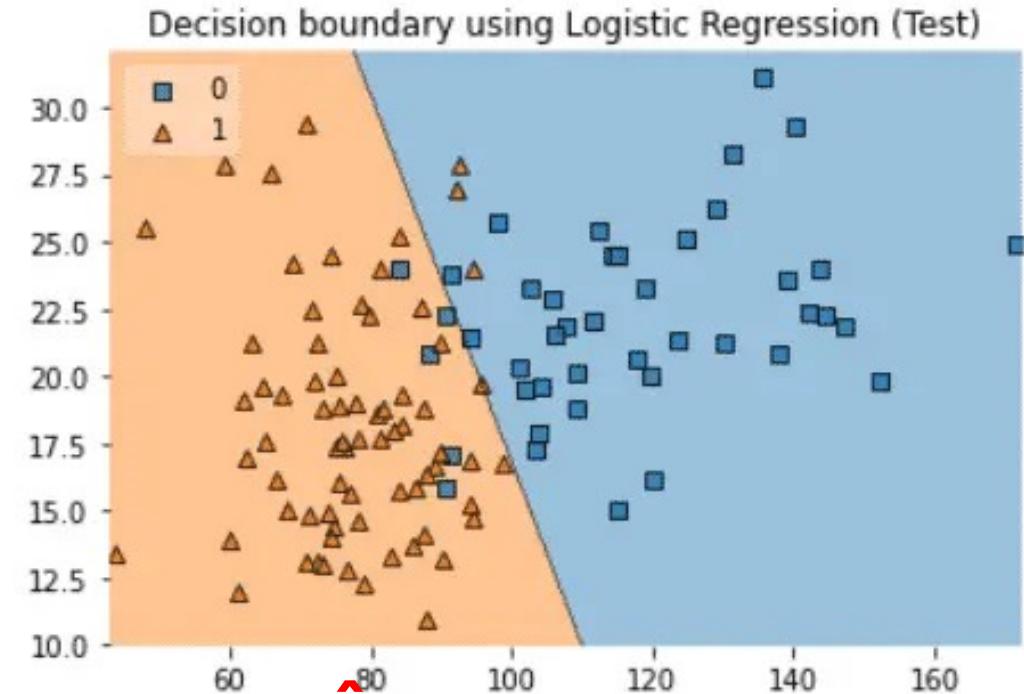
CHỦ ĐỀ 2

HỒI QUY LOGISTICS (LOGISTICS REGRESSION)

- *Hồi quy Logistics*
- *Thực hành trên R, Python và Excel*

1- DẪN NHẬP VỀ HỒI QUY LOGISTICS

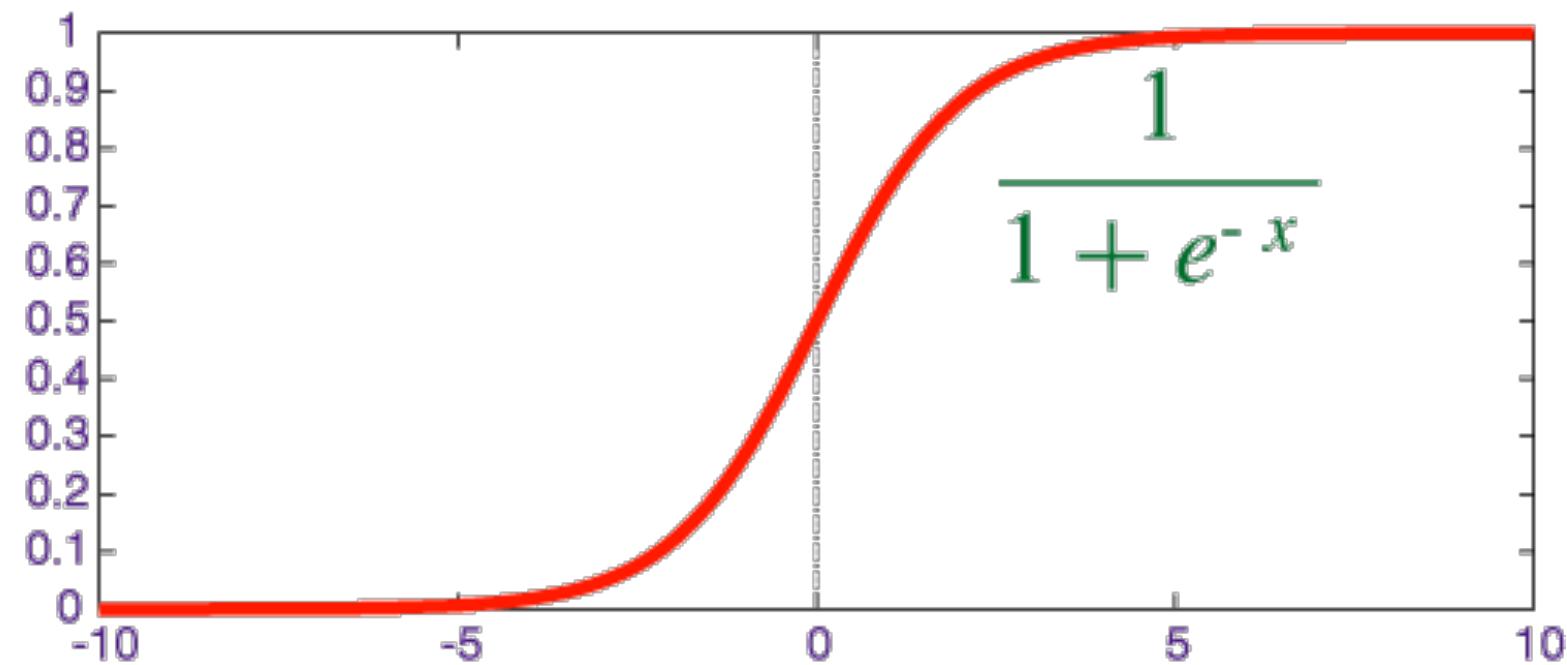
LOGISTICS



BIÉN PHÂN LOẠI NHỊ PHÂN

1- DẪN NHẬP VỀ HỒI QUY LOGISTICS

SIGMOID FUNCTION



1 - DẪN NHẬP VỀ HỒI QUY LOGISTICS

THINKING

- Gọi $P(X = 1) = \tilde{y}$
- Gọi $P(X = 0) = 1 - \tilde{y}$
- Như những phần trước

$$Y = \beta_0 + \beta_1 X + e$$

Để cho Y quy về xác suất 0 đến 1

Gọi P là *xác suất xảy ra sự kiện A* và 1-P là *biến cõ đối của sự kiện A*.

- Chỉ số ODDs = $P/(1-P)$
 - Nếu **ODDs > 1** xác suất biến cõ A xảy ra **khả năng cao** hơn biến cõ đối của nó.
 - Nếu **ODDs < 1** xác suất biến cõ A xảy ra **khả năng thấp** hơn biến cõ đối của nó.
 - Nếu **ODDs = 1** xác suất biến cõ A xảy ra **khả năng bằng** biến cõ đối của nó.
- Từ chỉ số ODDs ta chuyển Y trong **phương trình hồi quy tuyến tính** → **log(ODDs)**. Phương trình hồi quy logistic có dạng khác:

$$\log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X + e$$

2- THỰC HÀNH HỒI QUY LOGISTICS TRÊN R

THỰC HÀNH HỒI QUY LOGISTICS BẰNG R

```
> logistic <- glm(Plan.to.attend.graduate.school~Undergraduate.GPA,family = binomial)
> summary(logistic)

Call:
glm(formula = Plan.to.attend.graduate.school ~ Undergraduate.GPA,
     family = binomial)

Deviance Residuals:
    Min      1Q   Median      3Q      Max 
-1.5976 -0.8444  0.2483  0.7797  1.8741 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -10.909     4.585  -2.379  0.0174 *  
Undergraduate.GPA  3.593     1.463   2.456  0.0140 *  
...
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 39.429  on 29  degrees of freedom
Residual deviance: 29.494  on 28  degrees of freedom
AIC: 33.494

Number of Fisher Scoring iterations: 5
```

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

$$\log\left(\frac{P}{1-P}\right) = -10.909 + 3.593 * \text{UndergraduateGPA} + \varepsilon$$

=> Ta suy ra được kết quả như sau:

$$\left(\frac{p}{1-p}\right) = e^{-10.909+3.593*(undergraduateGPA)}$$

- Tập dữ liệu **Graduate School Survey**.
- Trong ngôn ngữ R dùng hàm **glm()** để phân tích hồi quy logistic. Với tham số **family = binomial**.

2- THỰC HÀNH HỒI QUY LOGISTICS TRÊN R

THỰC HÀNH HỒI QUY LOGISTICS BẰNG R

- Với ví dụ 3.11 làm ở phía trên hãy thử với trường hợp Odd0 và Odd1 tức UndergraduateGPA = 0 và 1 rồi lập tỉ lệ giữa hai phần này để xem xét khi tăng 1 điểm tỉ lệ tham dự lễ tốt nghiệp là bao nhiêu lần?

- Ta có $\left(\frac{p}{1-p}\right) = e^{-10,909+3,593*(undergraduateGPA)}$. Ta đặt hệ số p/(1-p) là odd
- Đặt Odd₀ và undergraduateGPA = 0 thì Odd₀ = $e^{-10,909}$
- Đặt Odd₁ và undergraduateGPA = 1 thì Odd₁ = $e^{-10,909+3,593}$
- Tí số $\frac{Odd_1}{Odd_0} = \frac{e^{-10,909+3,593}}{e^{-10,909}} \approx 36,359$
- 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

$$\log\left(\frac{P}{1-P}\right) = -10.909 + 3.593 * UndergraduateGPA + \varepsilon$$

- Tập dữ liệu Graduate School Survey.**
- Trong ngôn ngữ R dùng hàm **glm()** để phân tích hồi quy logistic. Với tham số family = binomial.**

2- THỰC HÀNH HỒI QUY LOGISTICS TRÊN R

THỰC HÀNH HỒI QUY LOGISTICS BẰNG PYTHON

3.12 Thực hiện mô hình hồi quy logistic trên Python

- Tập dữ liệu **Graduate School Survey**.
- Để thực hiện hồi quy logistic trên **Python** dùng thông qua hàm **LogisticRegression()**.
- Cũng giống như **LinearRegression()** cách import cũng tương tự.
from sklearn.linear_model import LogisticRegression
- Chuẩn bị dữ liệu Y là giá trị nhị phân, X là giá trị số hoặc phân loại.
- Thực hiện các bước như **hồi quy tuyến tính đơn biến, đa biến**.
- **Cách gọi tham khảo ví dụ bên dưới:**

Python

```
model = LogisticRegression(solver='liblinear', C=10.0, random_state=0)
model.fit(x, y)
```

- Về tìm các giá trị kiểm định làm tương tự như **LinearRegression**.
- So sánh kết quả Odd1 và Odd0 cũng làm tương tự như trong ngôn ngữ R.

1 - DẪN NHẬP VỀ HỒI QUY LOGISTICS

LOGISTICS DATA

Age	Income	Loan Amount	Credit Score	Approved
32	40000	20000	650	0
28	60000	30000	720	1
35	80000	40000	750	1
26	35000	15000	600	0
29	55000	25000	680	0
33	70000	35000	710	1
27	32000	12000	580	0
31	45000	18000	620	0
30	50000	22000	670	1



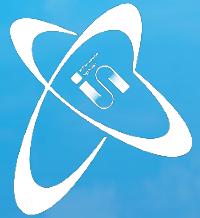
CHỦ ĐỀ 3 (KHÔNG NẰM TRONG LAB) HƯỚNG DẪN CÁC BƯỚC THỰC HIỆN ĐỒ ÁN

- *Project Description*
- *Tiêu chuẩn Project*

TÀI LIỆU THAM KHẢO.

1. <https://www.investopedia.com/>.
2. Roxy Peck and el, “Introduction to Statistics and Data Analysis”, 6th Edition
3. Phil Simon, “The Hundred-Page Machine Learning Book”

THANK YOU



TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA HỆ THỐNG THÔNG TIN

