


Regression

Trịnh Thành

thanh.trinh@phenikaa-uni.edu.vn

Phenikaa

- 
- ▶ Introduction to Regression
 - ▶ Simple Linear Regression
 - ▶ Multivariate Linear Regressions
 - ▶ Logistic Regressions
 - ▶ Nonlinear Regressions

- ▶ Introduction to Regression
- ▶ Simple Linear Regression
- ▶ Multivariate Linear Regressions
- ▶ Logistic Regressions
- ▶ Nonlinear Regressions

Regression - Hồi quy

► Vấn đề!

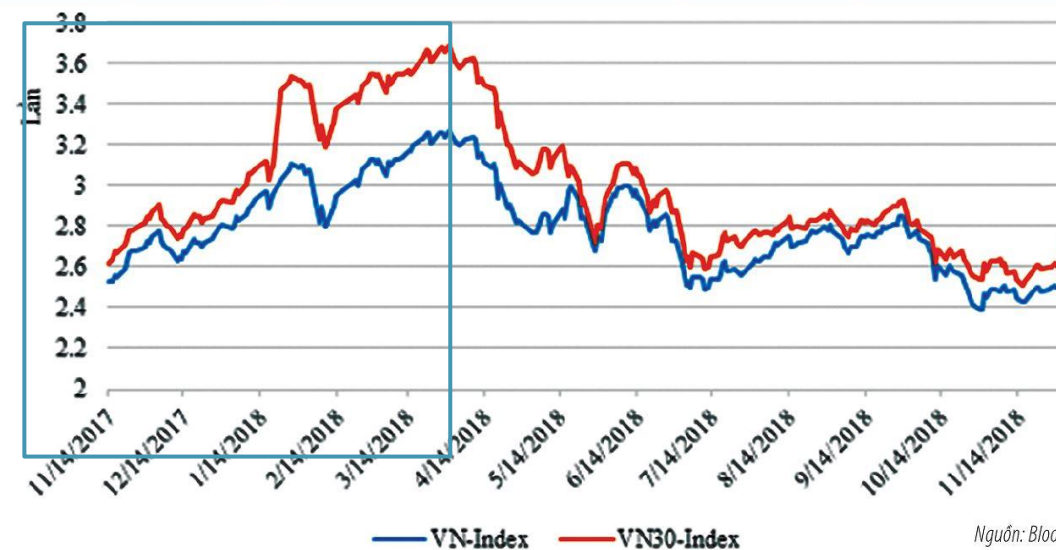


Top 10 cổ phiếu có giá cao nhất trên thị trường tại đỉnh 1,170

Mã CK	Giá (21/03/2018)	KLGD	LNST 2017 (tỷ đồng)	EPS 2017	Mã CK	Giá (12/03/2007)	KLGD	LNST 2017 (tỷ đồng)	EPS 2017
VCS	238,500	46,618	1,121.8	14,022	BVS	93,124	3,786	121.9	1,689
SAB	226,000	106,190	4,840.2	7,548	TLC	84,106	72,595	-	-
VNM	209,000	907,550	10,295.7	7,094	NKD	83,190	15,722	-	-
VJC	203,000	916,790	4,742.0	8,765	HRC	77,665	5,877	8.5	283
PNJ	183,000	197,960	724.9	6,705	HSC	76,021	96	-	-
VCF	181,000	280	372.5	14,015	SJS	75,565	12,197	131.3	1,325
CTD	171,800	369,730	1,652.7	21,100	BMI	66,472	9,599	164.0	1,795
WCS	165,000	100	61,177.0	24,471	PVD	60,197	23,376	26.2	68
ROS	155,800	1,694,630	848.5	1,794	GMD	59,613	43,482	1,239.5	1,748
BHN	134,000	5,570	751.4	3,242	FPT	58,711	8,403	2,931.5	5,522

Nguồn: VietstockFinance

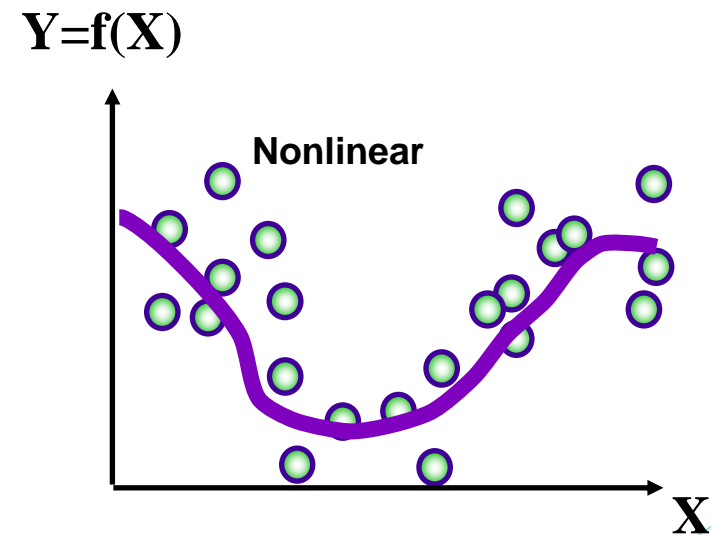
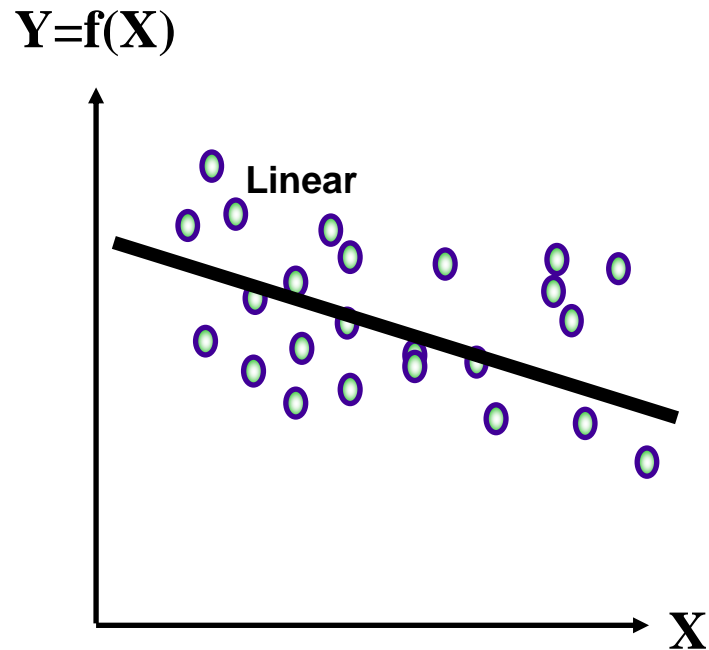
BIẾN ĐỘNG P/B CỦA VN-INDEX VÀ VN30-INDEX TỪ CUỐI NĂM 2017 ĐẾN NAY



Nguồn: Bloomberg

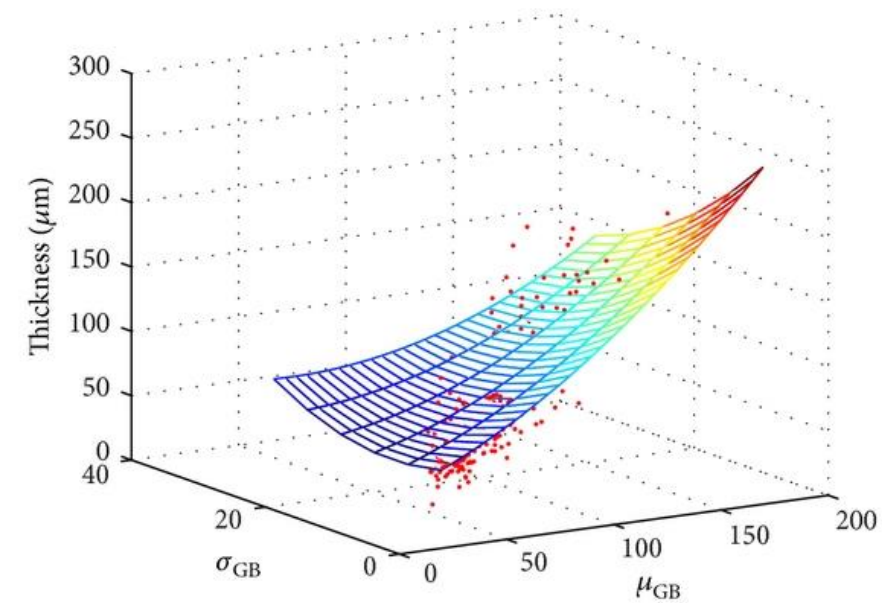
Hồi quy đơn biến

- ▶ Cho một sự phân bố một tập hợp các điểm, tìm một đường thẳng hoặc một đường cong $y = f(X)$, mà có thể phù hợp nhất với các điểm đấy.
- ▶ Như vậy nếu cho trước X , thì có thể dùng hàm $f(x)$ để dự đoán Y



Hồi quy đa biến

- ▶ Cho trước:
 - ▶ Một lượng lớn dữ liệu data D: N (mẫu - sample) * M (thuộc tính - biến)
 - ▶ Một thuộc tính Y có thể được dự đoán từ M-1 thuộc tính.
 - ▶ Sẽ có hàm $Y = f(x_1, x_2, \dots, x_{m-1})$
- ▶ Yêu cầu:
 - ▶ Xây dựng một hàm từ dữ liệu training data và sử dụng hàm này để dự đoán giá trị của các sample mới.
- ▶ Được là mô hình hồi quy - regression model



Kỹ thuật hồi quy

- ▶ Có 3 kỹ thuật thông thường:
- ▶ Hồi quy tuyến tính - Linear regression: dự đoán một giá trị của Y (mang giá trị liên tục), Linear regression là một hàm tuyến tính của một hay nhiều biến độc lập.
- ▶ Hồi quy phi tuyến - Nonlinear regression: dự đoán một giá trị của Y (mang giá trị liên tục), Nonlinear regression là một hàm phi tuyến tính của một hay nhiều biến độc lập.
- ▶ Hồi quy logistic - Logistic regression dự đoán xác suất của Y (giá trị nhị phân hoặc theo thứ tự), xác suất sẽ thu được từ một hàm của một hay nhiều biến độc lập.

Các mô hình hồi quy

- ▶ Mô hình hồi quy - regression model :

$$Y = \mu + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

- ▶ Sẽ không thích hợp cho biến Y là dạng 0 -1;
- ▶ Mô hình hồi quy logistic - logistic regression model.

$$P = P(Y = 1 | X_1, X_2, \cdots, X_n)$$

$$\ln \frac{P}{1-P} = \mu + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$


$$P = \frac{e^{(\mu + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}{1 + e^{(\mu + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}$$

Xác định mô hình phù hợp

- ▶ Nhìn vào biến đích (Y)
 - ▶ Nếu giá trị biến đích là liên tục (continuous), thì linear regression sẽ được chọn để tính toán. Nếu linear regression không thoả mãn kết quả, có thể chọn nonlinear regression để tính.
 - ▶ Nếu biến đích Y mang kiểu giá trị 0-1, thì logistic regression nên được chọn

Các vấn đề

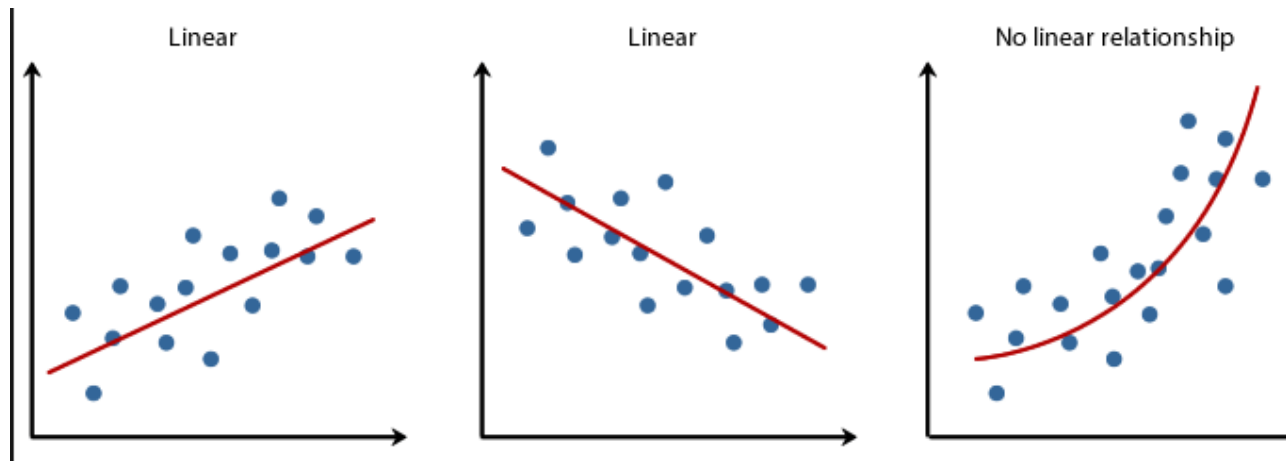
- ▶ Xác định cấu trúc mô hình
- ▶ Học từ mô hình
- ▶ Training dữ liệu
- ▶ Tính chất không tuyến tính

- 
- ▶ Introduction to Regression
 - ▶ **Simple Linear Regression**
 - ▶ Multivariate Linear Regressions
 - ▶ Logistic Regressions
 - ▶ Nonlinear Regressions

Dạng Tuyến Tính - linear regression

▶ Hồi quy tuyến tính:

- ▶ Là phương pháp học máy có giám sát đơn giản, được sử dụng để dự đoán (predict) giá trị đầu ra (liên tục, dạng số).
- ▶ Là phương pháp dựa trên thống kê để thiết lập mối quan hệ giữa một biến phụ thuộc và một nhóm tập hợp các biến độc lập.



Hồi quy tuyến tính đơn giản và tương quan - Simple linear regression and Correlation

- ▶ Hồi quy tuyến tính đơn giản và tương quan:
- ▶ Định lượng được mối quan hệ giữa hai biến liên tục
- ▶ Dự đoán giá trị của một biến từ việc **HIỂU** về giá trị của một biến khác.
- ▶ Trong Simple Linear Regression, dễ dàng tạo ra một phương trình để tìm giá trị của một biến phụ thuộc (Y) từ một biến độc lập (X).

$$Y = a X + b;$$

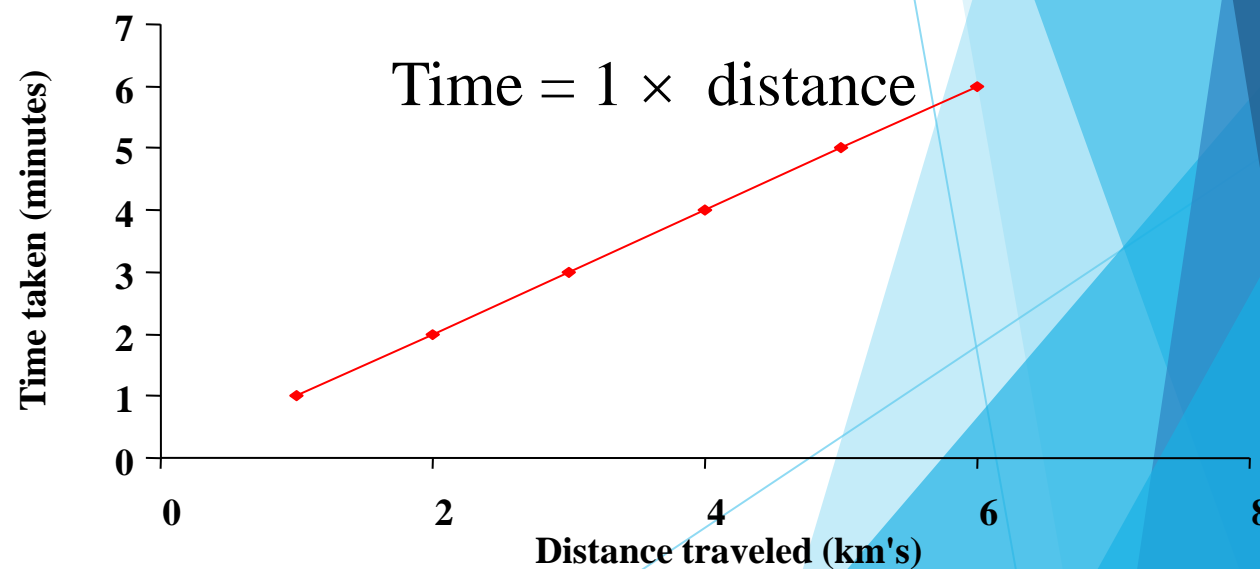
Regression model

- ▶ Ví dụ: Lái xe đi làm với vận tốc trung bình 60km/giờ.
- ▶ Như vậy. Có một mô hình toán học diễn tả được mối quan hệ giữa hai biến: quãng đường (distance) và thời gian (time).

$$\text{Time} = 1 * \text{distance}; \quad (1)$$

- ▶ Hay là:

$$Y = a * X$$



Regression model

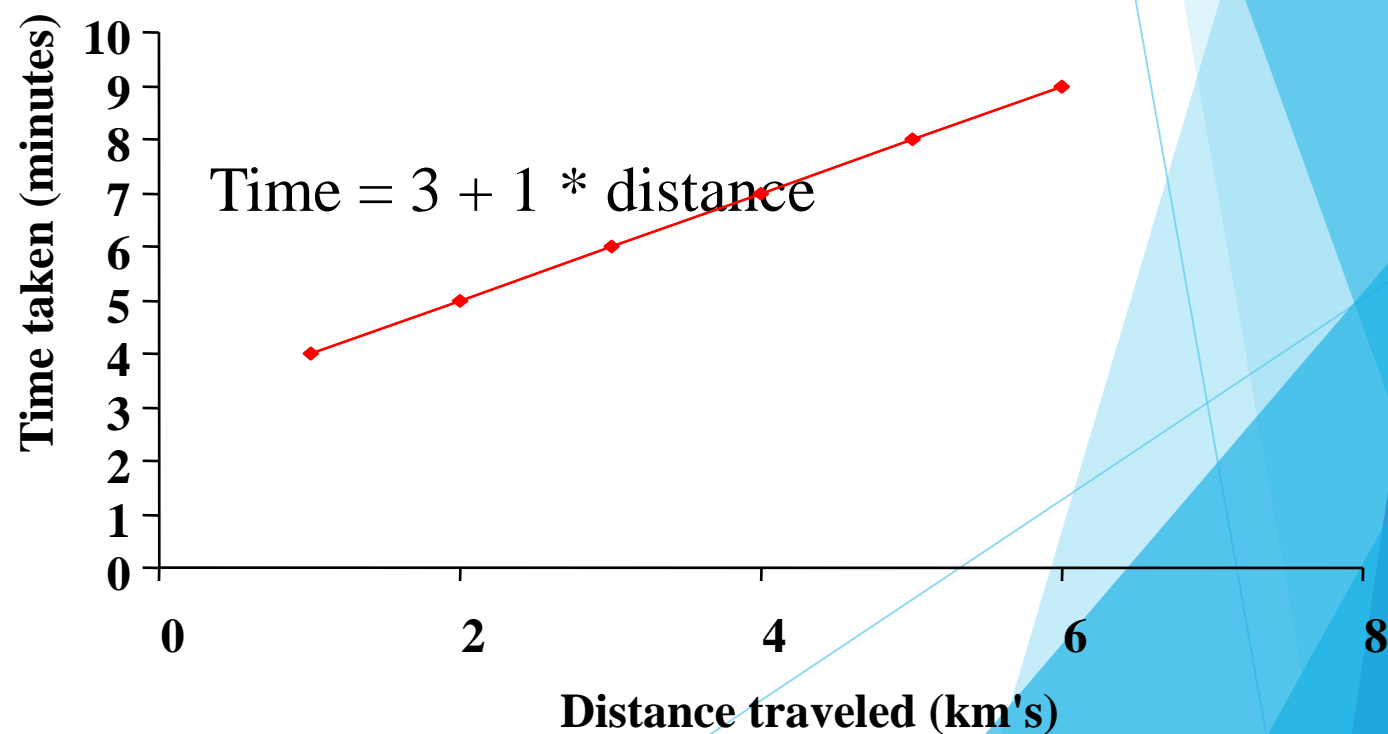
- ▶ Tuy nhiên: Nếu mất thêm 3 phút mỗi ngày từ nhà tới nơi để xe. Sau đấy mới lái tới công ty.
- ▶ Vậy mô hình (1) sẽ thành

$$\text{Time} = 3 + 1 * \text{distance} \quad (2)$$

$$\Leftrightarrow Y = b + aX$$

$$Y = 3 + 1 * X$$

- ▶ Simple linear regression.



Regression model

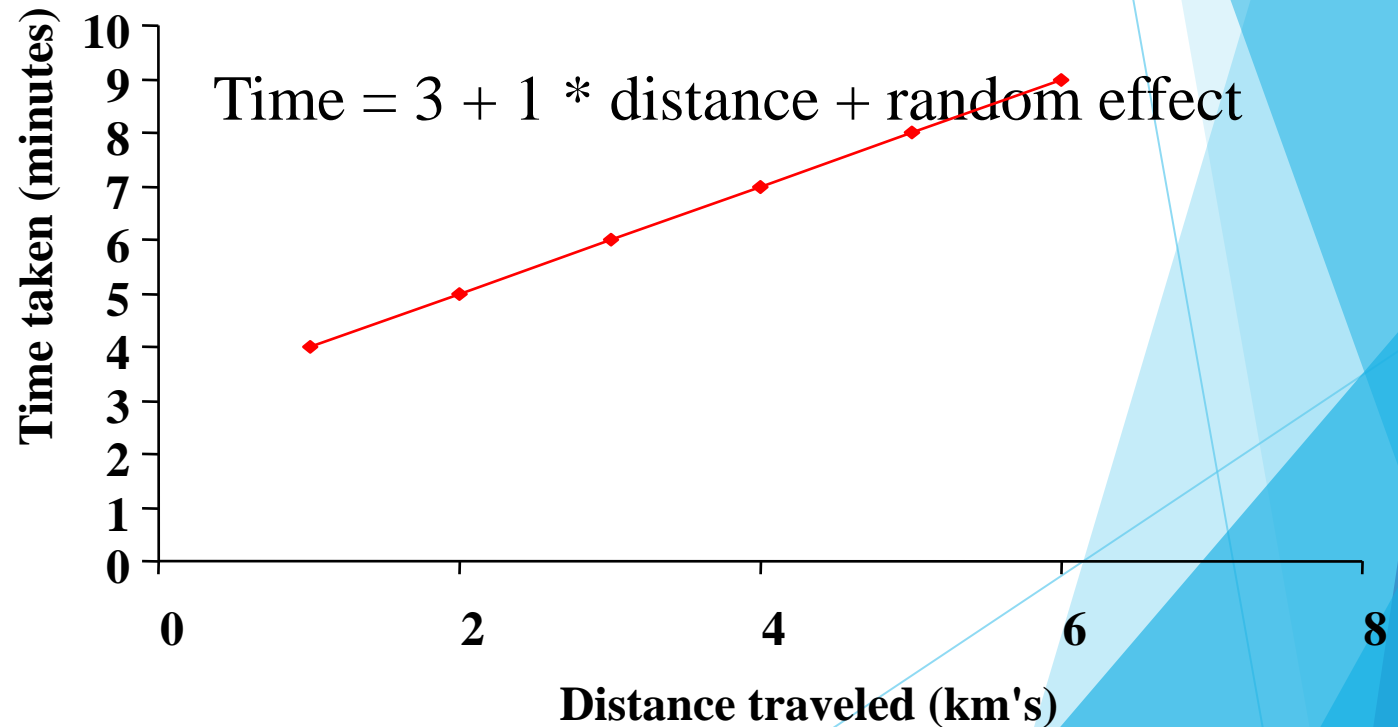
- ▶ Tuy nhiên: Quãng đường đi làm cho mỗi phút là không chính xác bởi vì giao thông, làm đường,... Lúc này mô hình (2) trở thành:

Time = 3 + 1 * distance + random effect (3)

⇔ $Y = b + a * X + e$;

với $b = 3$; $a = 1$;

e : hệ số lỗi.



Hồi quy tuyến tính hai biến

- ▶ Hồi quy tuyến tính hai biến - Bi-variate linear regression model.

- ▶ Mô hình hồi quy hai biến có dạng như sau:

$$y = \beta_0 + \beta_1 x + e$$

- ▶ y = là biến phụ thuộc
 - ▶ x = là biến độc lập the independent variable
 - ▶ β_0 = Giá trị chặn Y (intercept)
 - ▶ β_1 = Góc dốc của đường thẳng
 - ▶ e = Hệ số lỗi ngẫu nhiên (random error term)

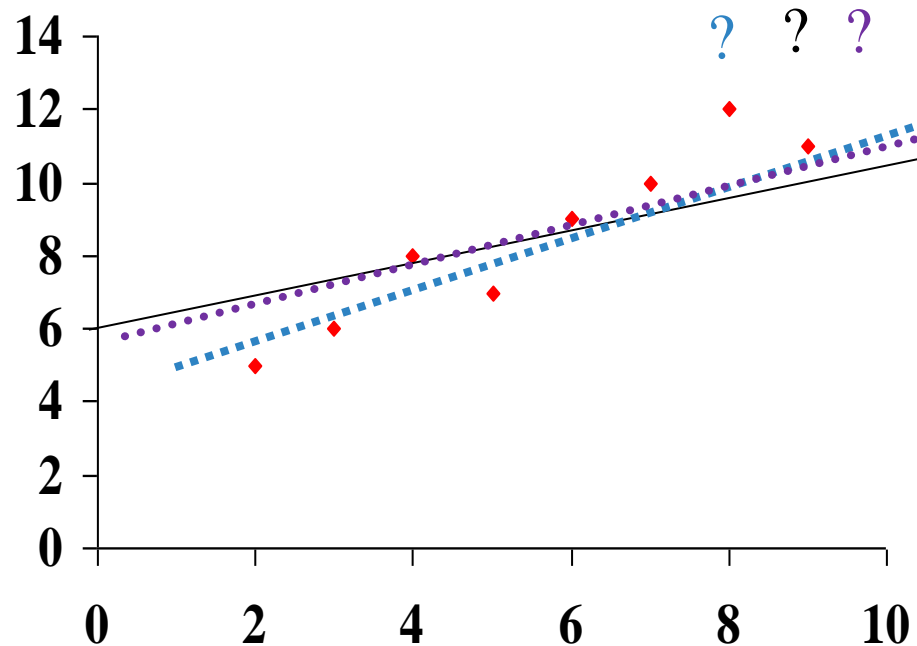
Chọn mô hình tốt nhất

► Với dữ liệu như hình vẽ.

- Xác định cách tính các parameters của phương trình

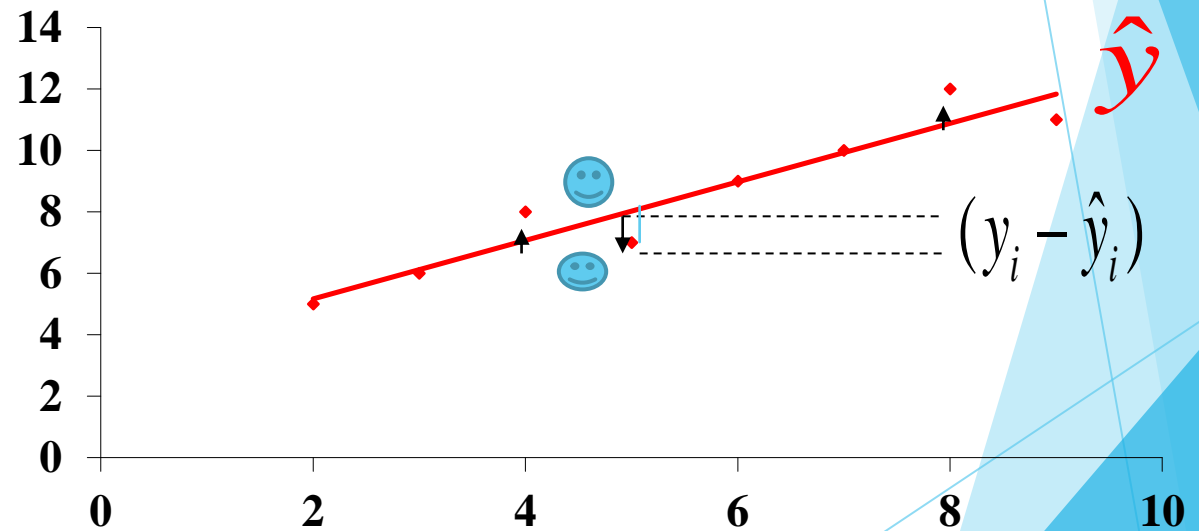
$$y = \beta_0 + \beta_1 x + e$$

- Chúng ta cần tìm đường phù hợp với dữ liệu nhất. (the line of best fit)



The line of best fit - đường phù hợp nhất

- ▶ Rõ ràng, đường thẳng rất hiếm phù hợp (chạy qua) toàn bộ dữ liệu một cách chính xác, do đó luôn tồn tại lỗi liên quan đến đường đấy
- ▶ Như vậy, chúng ta có thể xem:
 - ▶ The line of best fit là đường phù hợp nhất là đường giảm tối đa sự mở rộng của các lỗi



Error - lỗi

- ▶ $y_i - \hat{y}$ được gọi là lỗi (error) hay là phần dư thừa (residual)

$$e_i = (y_i - \hat{y})$$

- ▶ Đường line of best fit sẽ xác định được khi tổng lỗi bình phương (SSE) được giảm tối đa:

$$SSE = \sum_{i=1}^n (y_i - \hat{y})^2$$

- ▶ Trong đấy SSE (Sum of the Squared Errors): Tổng lỗi bình phương

Ước lượng các tham số

- ▶ Góc dốc của đường thẳng :

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_x}$$

- ▶ Trong đấy:

$$SS_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

$$SS_x = \sum_{i=1}^n (x_i - \bar{x})^2$$

Ước lượng các tham số

- ▶ Chặn Y (intercept):

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- ▶ Trong đây:

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$$

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

Ví dụ:

- ▶ Cho sự phân bố giữa X (kilos) và Y (cost) như sau:

X (kg)	Y(cost)
17	132
21	150
35	160
39	162
50	149
65	170

- ▶ Yêu cầu tìm mô hình

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

- ▶ Góc dốc của đường thẳng :

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_x}$$

- ▶ Trong đấy:

$$SS_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

$$SS_x = \sum_{i=1}^n (x_i - \bar{x})^2$$

- ▶ Chặn Y (intercept):

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

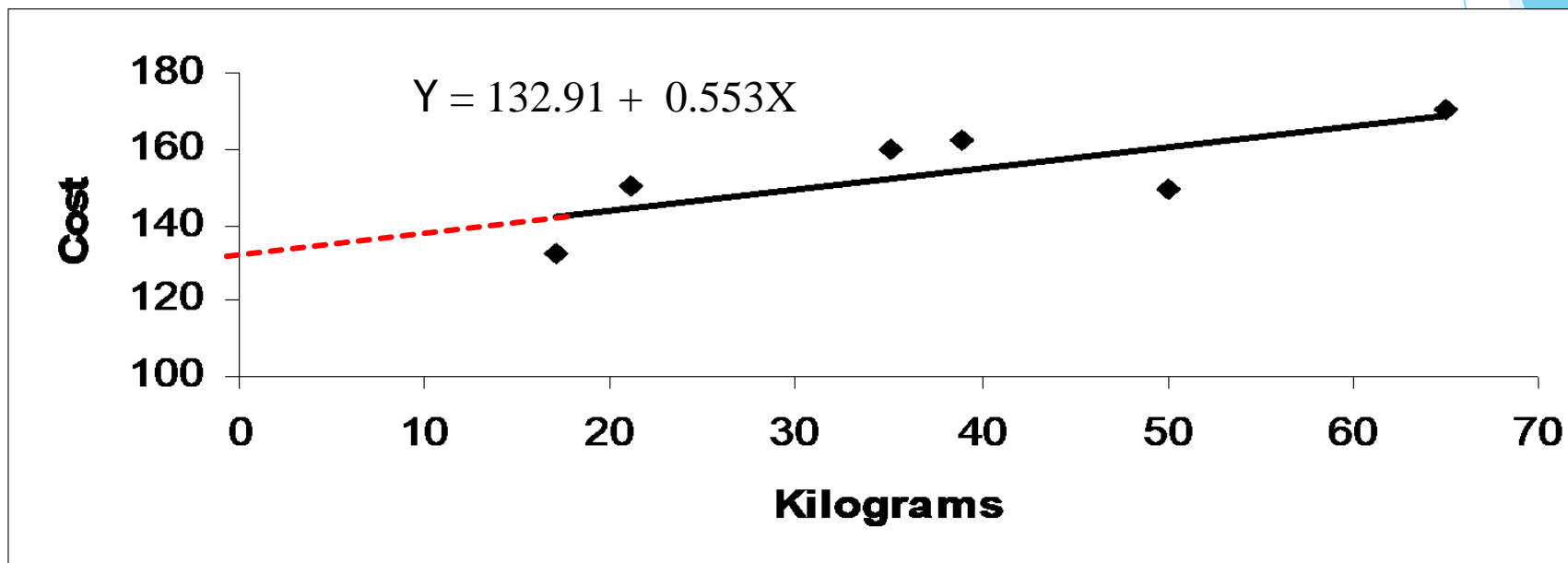
- ▶ Trong đấy:

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} \quad \bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

Ví dụ:

- ▶ X trung bình: $\bar{x} = 37.83$
- ▶ Y trung bình $\bar{y} = 153.83$
- ▶ Tổng lỗi x và y $SS_{xy} = 891.83$
- ▶ Tổng lỗi bình phương X: $SS_x = 1612.83$
- ▶ Vậy góc dốc: $\hat{\beta}_1 = \frac{SS_{xy}}{SS_x} = \frac{891.83}{1612.83} = 0.533$
- ▶ Ta có $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 153.83 - 0.553 \times 37.83 = 132.91$
- ▶ Phương trình (mô hình): $Y = 132.91 + 0.553X$


- ▶ Dự đoán được góc dốc là 0.553. Nghĩa là mỗi thay đổi của X trong phạm vi 1kg thì Y sẽ thay đổi là 0.553 \$.



- ▶ Và Nếu $X = 0$; thì $Y = ?$

Ghi nhớ

- ▶ Phương pháp xác định β_0 và β_1 của linear regression là phương pháp **least square estimation (LSE)**
- ▶ Các phân tích đi cùng
 - ▶ Vẽ Y và X trên đồ thị để kiểm tra tính chất tuyến tính
 - ▶ Khoảng tin cậy của β_0 và β_1
 - ▶ Kiểm chứng thống kê
 - ▶ Phân tích tương quan

- 
- ▶ Introduction to Regression
 - ▶ Simple Linear Regression
 - ▶ **Multivariate Linear Regressions**
 - ▶ Logistic Regressions
 - ▶ Nonlinear Regressions

Hồi quy nhiều lần - Multiple Regression

- ▶ Vấn đề:
 - ▶ Chúng ta có nhiều biến đầu vào với các kiểu khác nhau?
 - ▶ Ví dụ:
 - ▶ Age, gender
 - ▶ Education, income

- ▶ Mọi biến đầu vào đều có ảnh hưởng lên kết quả
 - ▶ Ước lượng Kết quả sẽ thay đổi bao nhiêu nếu biến đầu vào này tăng lên một đơn vị và những biến đầu vào khác giữ nguyên.
- ▶ Độ tuyến tính cùng nhau tối thiểu (minimal co-linearity)
 - ▶ Các biến đầu vào không quá phụ thuộc lẫn nhau. Nếu có 2 biến phụ thuộc hoàn toàn nhau. Có thể dùng một trong hai biến.
 - ▶ Kiểm tra các giả thuyết trước.

Hồi quy tuyến tính đa biến- Multivariate Linear Regression

- ▶ Hồi quy tuyến tính đa biến: Phân tích giữa biến phụ thuộc Y và một tập các biến độc lập $X = (x_1, x_2, \dots, x_k)$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon$$

- ▶ Cho một dataset X và Y, chúng ta cần tính các hệ số: $\beta_0, \beta_1, \beta_2, \dots, \beta_k$
- ▶ Sau khi tính được các hệ số: $\beta_0, \beta_1, \beta_2, \dots, \beta_k$
- ▶ Với một sample mới của X, có thể tính được biến Y tương ứng.

Các giả định: Assumptions for Multivariate Linear Regression

- ▶ Cho trước: $(k+1)$ biến $\{(Y, X_1, \dots, X_k)\}$
- ▶ 1- Giả định về mẫu:
 - ▶ Trung bình μ_Y của tập mẫu con của các giá trị Y với $X_1=x_1, \dots, X_k=x_k$, được tính như sau:

$$\mu_Y(x_1, \dots, x_k) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

- ▶ Có nghĩa là: Bất kỳ dòng (sample $(X_1=x_1, \dots, X_k=x_k)$) liên quan tới một tập các giá trị của biến Y , và giá trị từ mô hình hồi quy sẽ là giá trị trung bình μ_Y của các biến Y .

Các giả định:

▶ Giả định 2:

- ▶ Độ lệch chuẩn của các giá trị Y cho bất kỳ sample $(X_1=x_1, \dots, X_k=x_k)$ là giống nhau.

▶ Giả định 3:

- ▶ Một mẫu con của các giá trị Y đều có một phân bố Gaussian
- ▶ Như vậy $\{(Y, X_1, \dots, X_k)\}$ là một phân bố Gaussian với $k+1$ biến

▶ Giả định 4:

- ▶ Mọi mẫu dữ liệu được lấy bằng simple random sampling

▶ Giả định 5:

- ▶ Mọi bộ giá trị sample $y_i, x_{i1}, \dots, x_{ik}$ với $i = 1..,n$ được xác nhận không có lỗi.

Dữ liệu mẫu - Sample data

Y	X_1	X_2		X_k
y_1	$x_{1,1}$	$x_{1,2}$...	$x_{1,k}$
y_2	$x_{2,1}$	$x_{2,2}$...	$x_{2,k}$
.
.
.
y_i	$x_{i,1}$	$x_{i,2}$...	$x_{i,k}$
.
.
.
y_n	$x_{n,1}$	$x_{n,2}$...	$x_{n,k}$

Least Square Estimation Method

- ▶ Cho một dữ liệu và công thức hồi quy:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

- ▶ Chúng ta phải xác định công thức sau:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_k x_k$$

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

- ▶ Cần phải tính các giá trị $\hat{y}_1, \dots, \hat{y}_i, \dots, \hat{y}_n$
- ▶ So sánh các giá trị này với giá trị tương ứng thực tế $y_1, \dots, y_i, \dots, y_n$
- ▶ Như vậy, có thể tính được e : error (lỗi):

$$\hat{e}_i = y_i - \hat{y}_i = y_i - [\hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \dots + \hat{\beta}_k x_{i,k}]$$

Least Square Estimation Method

- ▶ Phương pháp LSE được chọn để ước tính $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$, như vậy chúng ta cần giảm thiểu tổng lỗi bình phương SSE (sum of square error)

$$SSE = \sum_{i=1}^n \hat{e}_i^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - \dots - \hat{\beta}_k x_{i,k})^2$$

Làm thế nào để tính $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$

LSE: biểu diễn dưới dạng matrix

- $\{y_1, \dots, y_i, \dots, y_n\}$ và $\{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k\}$ có thể biểu diễn dưới dạng vector (n x 1) và (k x 1); như vậy X là ma trận (n x (k+1))

Y	X_1	X_2		X_k
y_1	$x_{1,1}$	$x_{1,2}$...	$x_{1,k}$
y_2	$x_{2,1}$	$x_{2,2}$...	$x_{2,k}$
.
.
y_i	$x_{i,1}$	$x_{i,2}$...	$x_{i,k}$
.
.
.
y_n	$x_{n,1}$	$x_{n,2}$...	$x_{n,k}$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdot & \cdot & \cdot & x_{1,k} \\ 1 & x_{2,1} & x_{2,2} & \cdot & \cdot & \cdot & x_{2,k} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & x_{i,1} & x_{i,2} & \cdot & \cdot & \cdot & x_{i,k} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & x_{n,1} & x_{n,2} & \cdot & \cdot & \cdot & x_{n,k} \end{bmatrix}$$

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$$

LSE: biểu diễn dưới dạng matrix

Y	X_1	X_2		X_k
y_1	$x_{1,1}$	$x_{1,2}$...	$x_{1,k}$
y_2	$x_{2,1}$	$x_{2,2}$...	$x_{2,k}$
.
.
.
y_i	$x_{i,1}$	$x_{i,2}$...	$x_{i,k}$
.
.
.
y_n	$x_{n,1}$	$x_{n,2}$...	$x_{n,k}$

- Công thức hồi quy: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_k x_k$
có thể biểu diễn dưới dạng Matrix như sau:

$$\hat{y} = X \hat{\beta}$$

- Biểu diễn các lỗi error $\hat{e}_1, \hat{e}_2, \dots, \hat{e}_n$ như một vector (n x 1)

$$\hat{e} = \begin{bmatrix} \hat{e}_1 \\ \hat{e}_2 \\ . \\ . \\ . \\ \hat{e}_n \end{bmatrix}$$

- Ta được:

$$\hat{e} = y - \hat{y} = y - X \hat{\beta}$$

LSE: biểu diễn dưới dạng matrix

- ▶ Vậy: Với

$$y = X \hat{\beta}$$

- ▶ Trong dãy $Y = \{ y_1, \dots, y_i, \dots, y_n \}$, ước lượng $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$ như sau:

$$X^T X \hat{\beta} = X^T y$$

- ▶ $\Rightarrow \hat{\beta} = (X^T X)^{-1} X^T y$

- ▶ Trong dãy X^T là matrix chuyển vị

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

- ▶ $A = \text{transpose}(X) ; X^T$
- ▶ $B = A * X$
- ▶ $C = \text{invert}(B) ; \text{nghich dao}; B^{-1}$
- ▶ $D = C * A$
- ▶ $\text{Beta} = D * y$

Ví dụ

► Cho :

$$y = \begin{bmatrix} 6 \\ 9 \\ 12 \\ 5 \\ 13 \\ 2 \end{bmatrix} \quad X = \begin{bmatrix} 1 & 3 & 9 & 16 \\ 1 & 6 & 13 & 13 \\ 1 & 4 & 3 & 17 \\ 1 & 8 & 2 & 10 \\ 1 & 3 & 4 & 9 \\ 1 & 2 & 4 & 7 \end{bmatrix} \quad \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix}$$

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 3 & 6 & 4 & 8 & 3 & 2 \\ 9 & 13 & 3 & 2 & 4 & 4 \\ 16 & 13 & 17 & 10 & 9 & 7 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 6 & 26 & 35 & 72 \\ 26 & 138 & 153 & 315 \\ 35 & 153 & 295 & 448 \\ 72 & 315 & 448 & 944 \end{bmatrix}$$


$$X^T y = \begin{bmatrix} 47 \\ 203 \\ 277 \\ 598 \end{bmatrix}$$

$$\hat{\beta} = (X^T X)^{-1} X^T y = \begin{bmatrix} 2.59578 & -0.15375 & -0.01962 & -0.13737 \\ -0.15375 & 0.03965 & -0.00014 & -0.00144 \\ -0.01962 & -0.00014 & 0.01234 & -0.00431 \\ -0.13737 & -0.00144 & -0.00431 & 0.01406 \end{bmatrix} \begin{bmatrix} 47 \\ 203 \\ 277 \\ 598 \end{bmatrix}$$

$$= \begin{bmatrix} 3.20975 \\ -0.07573 \\ -0.11162 \\ 0.46691 \end{bmatrix}$$

$$\hat{\beta}_0 = 3.20975 \quad \hat{\beta}_1 = -0.07573 \quad \hat{\beta}_2 = -0.11162 \quad \hat{\beta}_3 = 0.46691$$

$$\hat{y} = 3.20975 - 0.07573x_1 - 0.11162x_2 + 0.46691x_3$$

- 
- ▶ Introduction to Regression
 - ▶ Simple Linear Regression
 - ▶ Multivariate Linear Regressions
 - ▶ **Logistic Regressions**
 - ▶ Nonlinear Regressions

Logistic regression

- ▶ Vấn đề:
- ▶ Mô hình mối quan hệ giữa một tập biến X
 - ▶ Dạng 2 giá trị ; ví dụ: yes/no
 - ▶ Dạng nhóm; ví dụ: social class
 - ▶ Số học; ví dụ tuổi
- ▶ Và
 - ▶ Biến đích Y là ở dạng nhị phân
 - ▶ Y : Respond or Not Respond, Risk or Not Risk, Claim or No claim
- ▶ Thường dùng trong phân lớp nhị phân. Có thể mở rộng cho multiclass (softmax regression)

Logistic regression

- ▶ Ví dụ: Tuổi (Age) và dấu hiệu của bệnh tim phình mạch vành (CD)

Age	CD
22	0
23	0
24	0
27	0
28	0
30	0
30	0
32	0
33	0
35	1
38	0

Age	CD
40	0
41	1
46	0
47	0
48	0
49	1
49	0
50	1
51	0
51	1
52	0

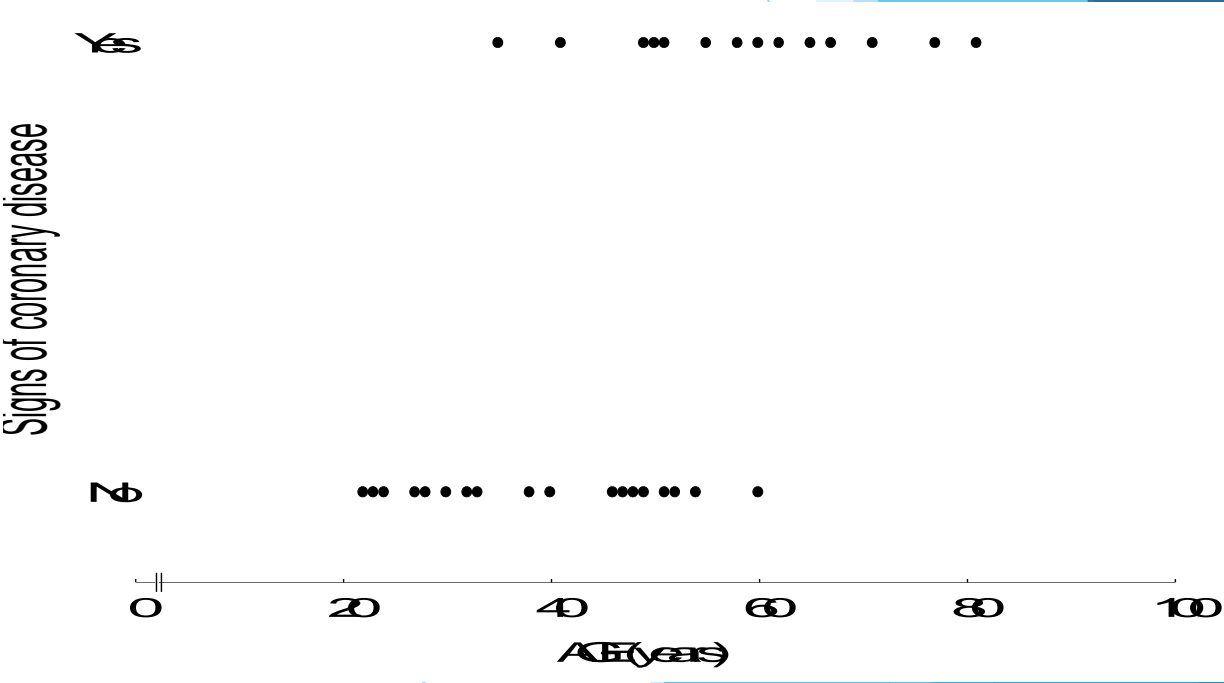
Age	CD
54	0
55	1
58	1
60	1
60	0
62	1
65	1
67	1
71	1
77	1
81	1

Chúng ta sẽ phân tích dữ liệu này như thế nào

Phân tích dữ liệu -

- ▶ So sánh độ tuổi trung bình mắc bệnh và không mắc bệnh
 - ▶ Non-diseased: 38.6 years
 - ▶ Diseased: 58.7 years
- ▶ Hình vẽ mô tả số người bị bệnh và không bị bệnh trải dài theo độ tuổi

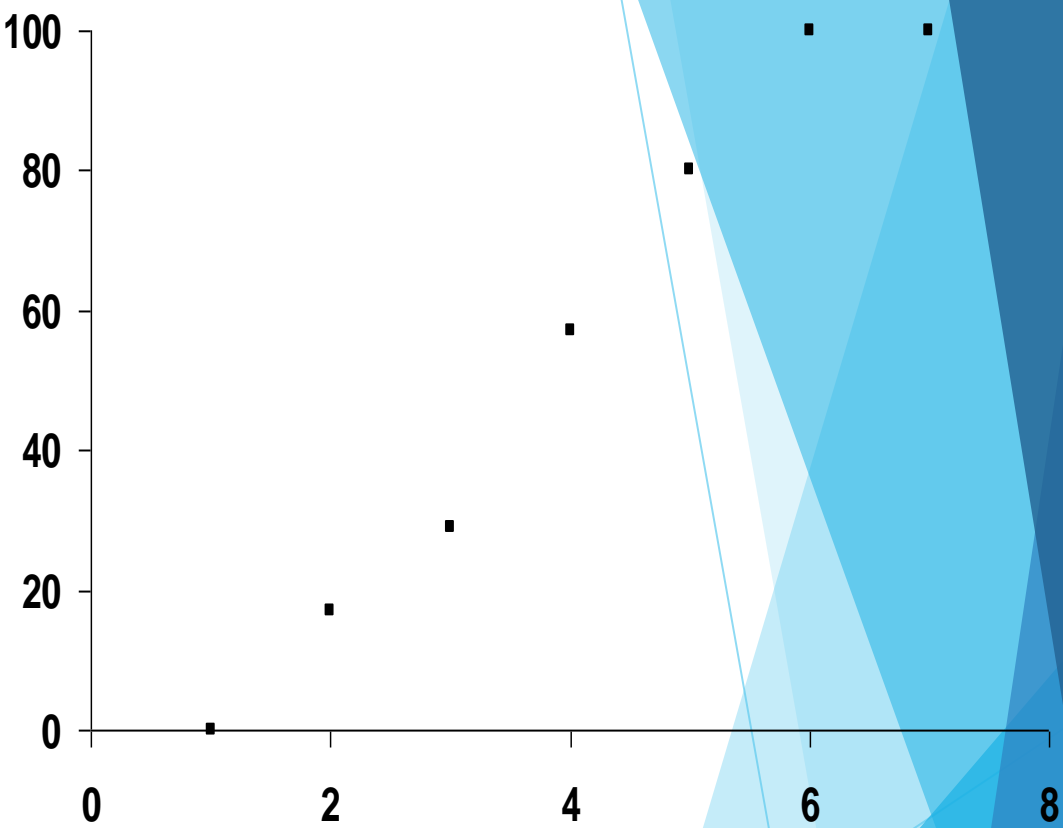
Age	CD	Age	CD	Age	CD
22	0	40	0	54	0
23	0	41	1	55	1
24	0	46	0	58	1
27	0	47	0	60	1
28	0	48	0	60	0
30	0	49	1	62	1
30	0	49	0	65	1
32	0	50	1	67	1
33	0	51	0	71	1
35	1	51	1	77	1
38	0	52	0	81	1



Phân tích dữ liệu

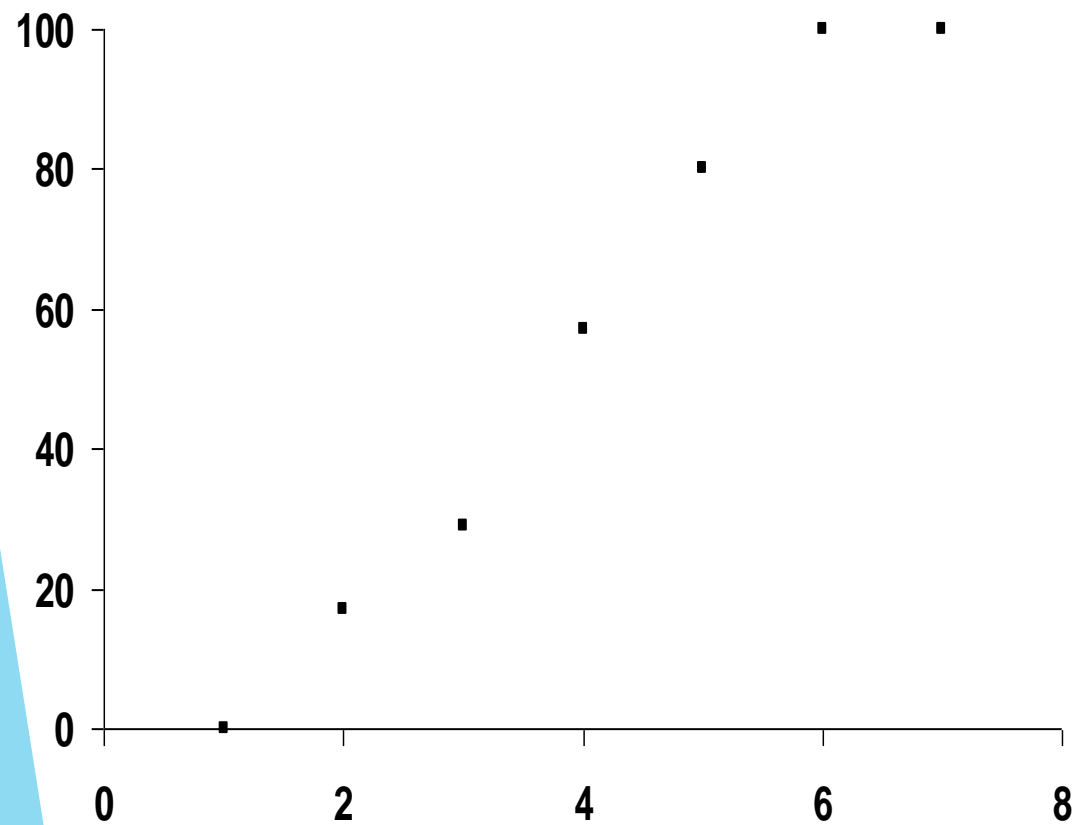
► Nhóm các độ tuổi, 10 năm một với nhau

Age group	# in group	Diseased	
		#	%
20 - 29	5	0	0
30 - 39	6	1	17
40 - 49	7	2	29
50 - 59	7	4	57
60 - 69	5	4	80
70 - 79	2	2	100
80 - 89	1	1	100

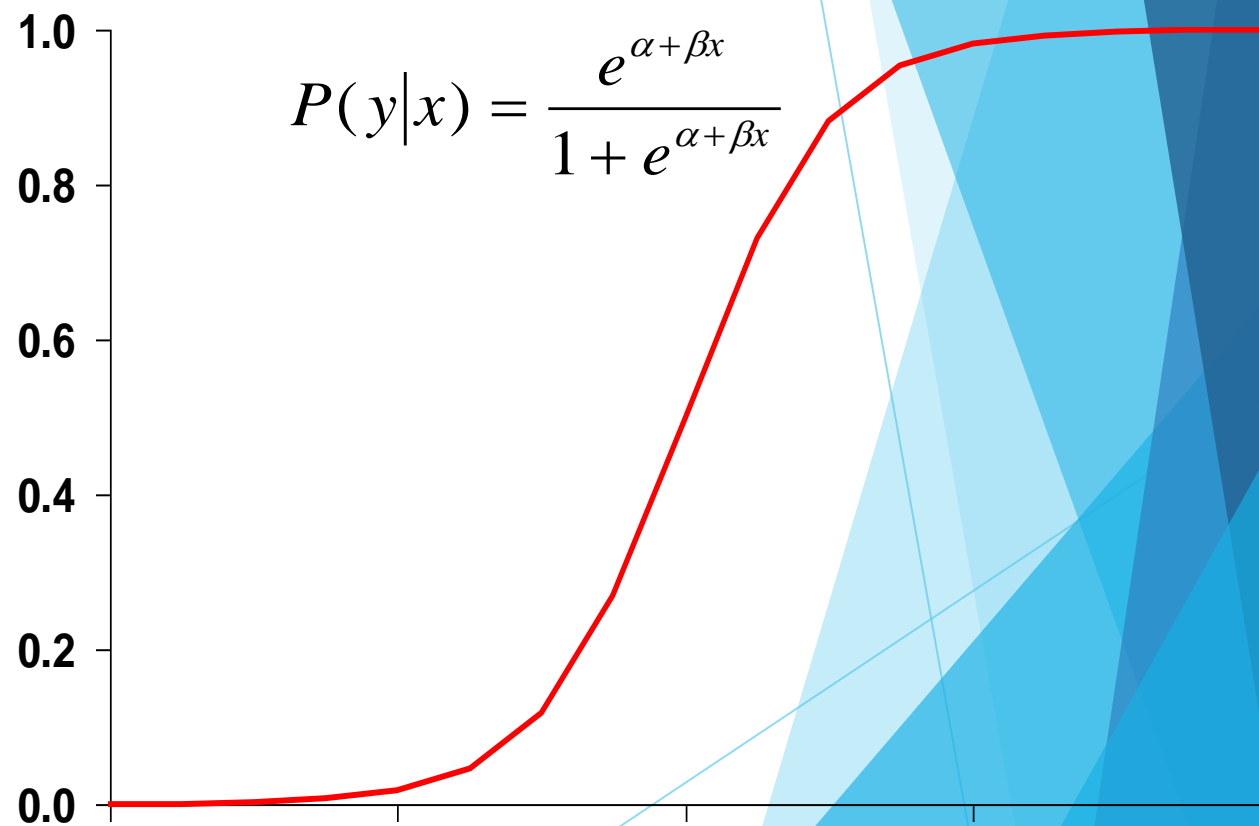


Age	CD	Age	CD	Age	CD
22	0	40	0	54	0
23	0	41	1	55	1
24	0	46	0	58	1
27	0	47	0	60	1
28	0	48	0	60	0
30	0	49	1	62	1
30	0	49	0	65	1
32	0	50	1	67	1
33	0	51	0	71	1
35	1	51	1	77	1
38	0	52	0	81	1

Hàm logistic



Xác
suất
bị
bệnh



Biến đổi - Logistic Transformation

- Công thức của hàm log

$$P(y|x) = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}} \quad (1)$$

- Biến đổi ta có

$$\ln \left[\frac{P(y|x)}{1 - P(y|x)} \right] = \alpha + \beta x \quad (2)$$

Biến đổi - Logistic Transformation

- ▶ Biến đổi từ mô hình phi tuyến sang mô hình hồi quy tuyến tính
- ▶ Logit từ $(-\infty ; +\infty)$
- ▶ Xác suất (P) ở khoảng giá trị 0 và 1;
- ▶ Biến đổi về tỷ lệ xác suất xảy ra;

$$\ln \left[\frac{P(y|x)}{1 - P(y|x)} \right] = \alpha + \beta x \longrightarrow \ln \left(\frac{P}{1-P} \right) = \alpha + \beta x \longrightarrow \frac{P}{1-P} = e^{\alpha + \beta x}$$

Xác định hệ số - β

- ▶ Y: yes - mắc bệnh (disease)
- ▶ Y: no: - Không mắc bệnh
- ▶ $x = 1$; yes: Phơi nhiễm (exposure)
- ▶ $x = 0$; no: Không phơi nhiễm

(disease) y	exposure (x)	
	yes	no
yes	$P(y x = 1)$	$P(y x = 0)$
no	$1 - P(y x = 1)$	$1 - P(y x = 0)$

- ▶ P là xác suất bị bệnh khi phơi nhiễm;

- ▶ $\frac{P}{1-P} = e^{\alpha + \beta x}$ Ta có nguy cơ (odds) là $Odds_{d|e} = e^{\alpha + \beta}$ (3)

- ▶ P là xác suất bị bệnh khi không phơi nhiễm;

- ▶ $\frac{P}{1-P} = e^{\alpha + \beta x}$; Ta có nguy cơ (odds) là $Odds_{d|\bar{e}} = e^{\alpha}$ (4)

Xác định hệ số - β

- ▶ Ta có tỷ lệ nguy cơ bị bệnh (odds ratio - **OR**)

- ▶ $OR = (3) / (4)$

$$OR = \frac{e^{\alpha+\beta}}{e^{\alpha}} = e^{\beta} \quad (5)$$

$$\ln(OR) = \beta$$

- ▶ β : tăng lên trong hàm logarit của tỷ lệ OR cho mỗi đơn vị khi tăng lên trong x

- ▶ Khoảng tin cậy Confidence Interval (CI): Theo phân phối chuẩn

$$95\% \text{ CI} = e^{(\beta \pm 1.96SE_{\beta})}$$

- ▶ SE là sai số chuẩn

Ví dụ:

- ▶ Rủi ro của bệnh tim phình mạch vành theo độ tuổi Age(<55) và (55+)

- ▶ $\ln \left(\frac{P}{1-P} \right) = \alpha + \beta \times Age \quad (6)$

disease	55+ (1)	< 55 (0)
YES (1)	21	22
NO (0)	6	51

- ▶ Tỷ lệ bệnh nhóm 1: $Odd1 = 21/6$; $(21/27) / (6/27)$

- ▶ Tỷ lệ bệnh nhóm 2: $Odd2 = 22/51$;

- ▶ Vậy ta có tỷ lệ nguy cơ nhóm 1 và nhóm 2;

- ▶ $OR = 8.1$;

- ▶ (5) Ta có $\ln(OR) = \beta \rightarrow \ln(8.1) = \beta \rightarrow \beta = 2.094$;

- ▶ Với Age =55+; hay x=1;

- ▶ (6) trên $\rightarrow \ln(21/6) = \alpha + \beta * 1 \rightarrow \alpha = -0.841$;

(min; max); max <1

- ▶ $SE = \sqrt{1/21 + 1/6 + 1/22 + 1/51} = 0.529$

- ▶ Xác định khoảng tin cậy :

$$95\% CI = e^{(\beta \pm 1.96 SE_{\beta})}$$

(2.9; 22.9) ; (0.9;22.9)

= 2.9; 22.9

Multiple Logistic Regression

- ▶ Nhiều biến với các kiểu dữ liệu khác nhau;

$$\ln \left(\frac{P}{1-P} \right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots \beta_i x_i$$

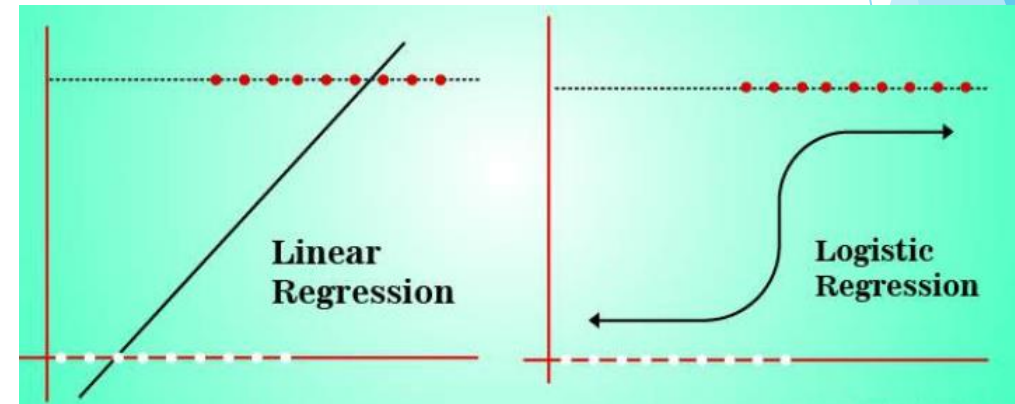
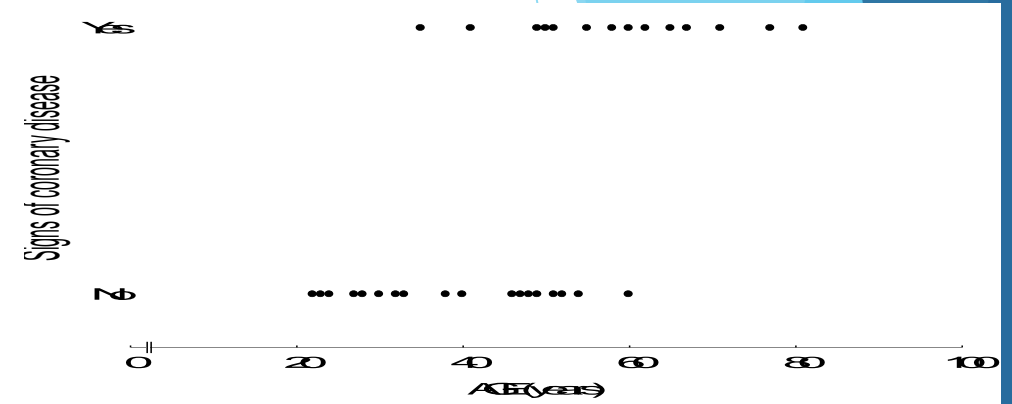
- ▶ Linear Regression

$$f(x) = w^T x$$

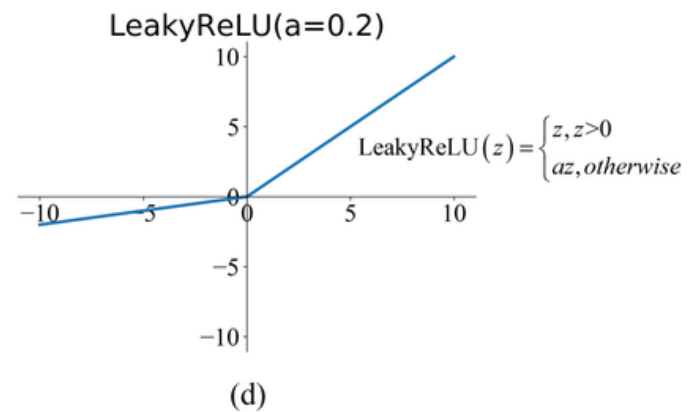
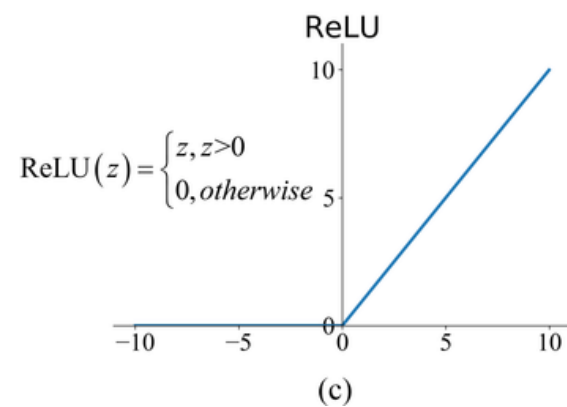
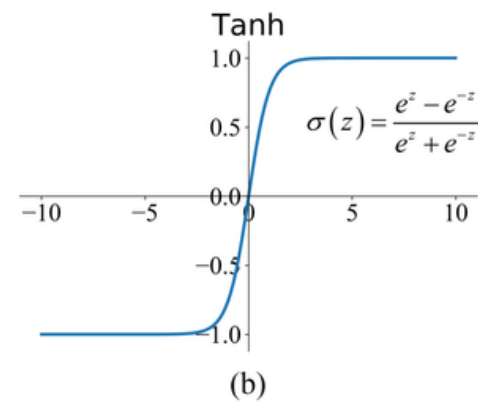
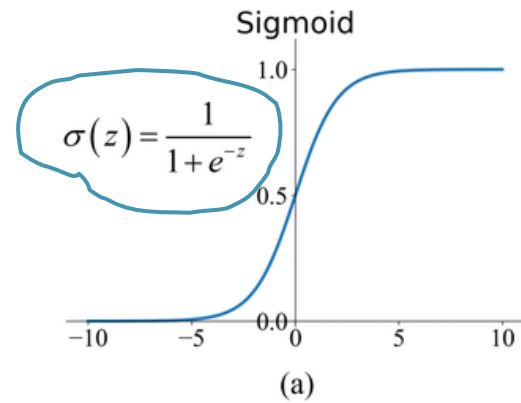
- ▶ Logistic regression

$$f(x) = \theta(w^T x)$$

- ▶ Trong đây θ là hàm activation function



Một số activation function



- ▶ Hàm sigmoid và hàm tanh được sử dụng nhiều nhất vì tính chất chặn của nó

Logistic regression

- ▶ Xét bài toán phân lớp nhị phân (0 ; 1)
- ▶ Giả sử rằng xác suất để một điểm dữ liệu \mathbf{x} rơi vào
- ▶ Class 1 là : $f(\mathbf{w}^T \mathbf{x})$
- ▶ Class 0 là : $1 - f(\mathbf{w}^T \mathbf{x})$
- ▶ Dựa vào dữ liệu training (đã biết \mathbf{y} và \mathbf{x}), ta có thể viết như sau

$$P(y_i=1 | x_i; \mathbf{w}) = f(\mathbf{w}^T \mathbf{x}_i) \quad (1)$$

$$P(y_i=0 | x_i; \mathbf{w}) = 1 - f(\mathbf{w}^T \mathbf{x}_i) \quad (2)$$

- ▶ $P(y_i=1 | x_i; \mathbf{w})$; Xác suất xảy ra $Y_i = 1$ khi đã biết \mathbf{w} và \mathbf{x}_i
- ▶ Như vậy cần tìm \mathbf{W} sao cho $P(y_i=1 | x_i; \mathbf{w}) \rightarrow 1$ và $P(y_j=0 | x_j; \mathbf{w}) \rightarrow 0$

▶ Đặt $z_i = f(w^T x_i)$;

$$P(y_i=1 | x_i; w) = f(w^T x_i) \quad (1)$$

$$P(y_i=0 | x_i; w) = 1 - f(w^T x_i) \quad (2)$$

▶ (1) và (2) \rightarrow $P(y_i=1 | x_i; w) = z_i \quad (3)$
 $P(y_i=0 | x_i; w) = 1 - z_i \quad (4)$

▶ Công thức xác suất Bernoulli;

$$f(k; p) = p^k (1 - p)^{1-k} \quad \text{for } k \in \{0, 1\}$$

▶ (3) (4) viết gộp lại:

$$P(y_i | x_i; w) = z_i^{y_i} (1 - z_i)^{1-y_i}$$

▶ Vậy với training set: $X = \{x_1, \dots, x_N\}$ và $Y = \{y_1, \dots, y_N\}$

▶ Ta chuyển về bài toán $\mathbf{w} = \arg \max_{\mathbf{w}} P(\mathbf{y} | \mathbf{X}; \mathbf{w})$

$$p(x_i | c) = p(i | c)^{x_i} (1 - p(i | c))^{1-x_i}$$

$$\mathbf{w} = \arg \max_{\mathbf{w}} P(\mathbf{y}|\mathbf{X}; \mathbf{w})$$

- ▶ Ta có bài toán maximum likelihood estimation;
- ▶ Hàm số phía sau gọi là likelihood function
- ▶ Để giải quyết bài toán này:
 - ▶ Giả sử các điểm dữ liệu được sinh ra một cách ngẫu nhiên độc lập với nhau (independent)

$$\begin{aligned} P(\mathbf{y}|\mathbf{X}; \mathbf{w}) &= \prod_{i=1}^N P(y_i|\mathbf{x}_i; \mathbf{w}) \quad \text{---} P(y_i|\mathbf{x}_i; \mathbf{w}) = z_i^{y_i}(1 - z_i)^{1-y_i} \\ &= \prod_{i=1}^N z_i^{y_i}(1 - z_i)^{1-y_i} \end{aligned}$$

$$P(\mathbf{y}|\mathbf{X}; \mathbf{w}) = \prod_{i=1}^N z_i^{y_i} (1 - z_i)^{1-y_i}$$

- Ta dùng hàm logarit hai bên; thêm vào dấu “-”; gọi đây là hàm mất mát ($J(\mathbf{w})$ - loss function);

$$J(\mathbf{w}) = -\log P(\mathbf{y}|\mathbf{X}; \mathbf{w}) = -\sum_{i=1}^N (y_i \log z_i + (1 - y_i) \log(1 - z_i))$$

- Vậy chuyển bài toán từ maximum likelihood thành minimum loss likelihood ;
- Và hàm này còn được gọi là negative log likelihood.

$$\mathbf{w} = \arg \max_{\mathbf{w}} P(\mathbf{y}|\mathbf{X}; \mathbf{w})$$

- ▶ Optimize loss function: Sử dụng phương pháp Stochastic gradient descent (SGD).

- ▶ Xem xét: Loss function với điểm dữ liệu (x_i, y_i) là

$$J(\mathbf{w}) = -\log P(\mathbf{y}|\mathbf{X}; \mathbf{w}) = -\sum_{i=1}^N (y_i \log z_i + (1 - y_i) \log(1 - z_i))$$

$$J(\mathbf{w}; \mathbf{x}_i, y_i) = -(y_i \log z_i + (1 - y_i) \log(1 - z_i))$$

- ▶ Đạo hàm hai bên theo \mathbf{w}

$$\frac{\partial J(\mathbf{w}; \mathbf{x}_i, y_i)}{\partial \mathbf{w}} = -\left(\frac{y_i}{z_i} - \frac{1 - y_i}{1 - z_i}\right) \frac{\partial z_i}{\partial \mathbf{w}} = \frac{z_i - y_i}{z_i(1 - z_i)} \frac{\partial z_i}{\partial \mathbf{w}} \quad (5)$$

- ▶ Với $z=f(\mathbf{w}^\top \mathbf{x})$; đặt $s=\mathbf{w}^\top \mathbf{x}$,

$$\frac{\partial z_i}{\partial \mathbf{w}} = \frac{\partial z_i}{\partial s} \frac{\partial s}{\partial \mathbf{w}} = \frac{\partial z_i}{\partial s} \mathbf{x} \quad (6)$$

▶ Hàm sigmoid; $f(s) = \frac{1}{1 + e^{-s}}$

▶ Ta đã có $z=f(w^T x)$; đặt $s=w^T x$; ta có hàm sigmoid

$$z(s) = \frac{1}{1 + e^{-s}}$$

▶ Với x_i ta có s_i ; vậy ta có;

$$z_i = \frac{1}{1 + e^{-s_i}}$$

$$\frac{\partial J(\mathbf{w}; \mathbf{x}_i, y_i)}{\partial \mathbf{w}} = - \left(\frac{y_i}{z_i} - \frac{1 - y_i}{1 - z_i} \right) \frac{\partial z_i}{\partial \mathbf{w}} = \frac{z_i - y_i}{z_i(1 - z_i)} \frac{\partial z_i}{\partial \mathbf{w}}$$

▶ Đạo hàm 2 vế theo s_i ; ta có $\frac{\partial z_i}{\partial s} = z_i(1 - z_i)$ (7)

$$\frac{\partial z_i}{\partial \mathbf{w}} = \frac{\partial z_i}{\partial s} \frac{\partial s}{\partial \mathbf{w}} = \frac{\partial z_i}{\partial s} \mathbf{x}$$

▶ Kết hợp phương trình (5) (6) (7); ta được

$$\frac{\partial J(\mathbf{w}; \mathbf{x}_i, y_i)}{\partial \mathbf{w}} = (z_i - y_i) \mathbf{x}_i$$

▶ Theo Stochastic Gradient Descent: Ta có công thức cập nhật cho logistic regression:

$$\mathbf{w} = \mathbf{w} + \eta(y_i - z_i) \mathbf{x}_i$$

- ▶ Introduction to Regression
- ▶ Simple Linear Regression
- ▶ Multivariate Linear Regressions
- ▶ Logistic Regressions
- ▶ **Nonlinear Regressions**

Nonlinear Regressions

- ▶ Một mô hình hồi quy tuyến tính là tuyến tính trong các tham số. Nghĩa là, **chỉ có một** tham số trong mỗi số hạng của mô hình và mỗi tham số là một hằng số nhân trên (các) biến độc lập của số hạng đó.
- ▶ Một mô hình phi tuyến là phi tuyến tính trong các tham số.

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

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

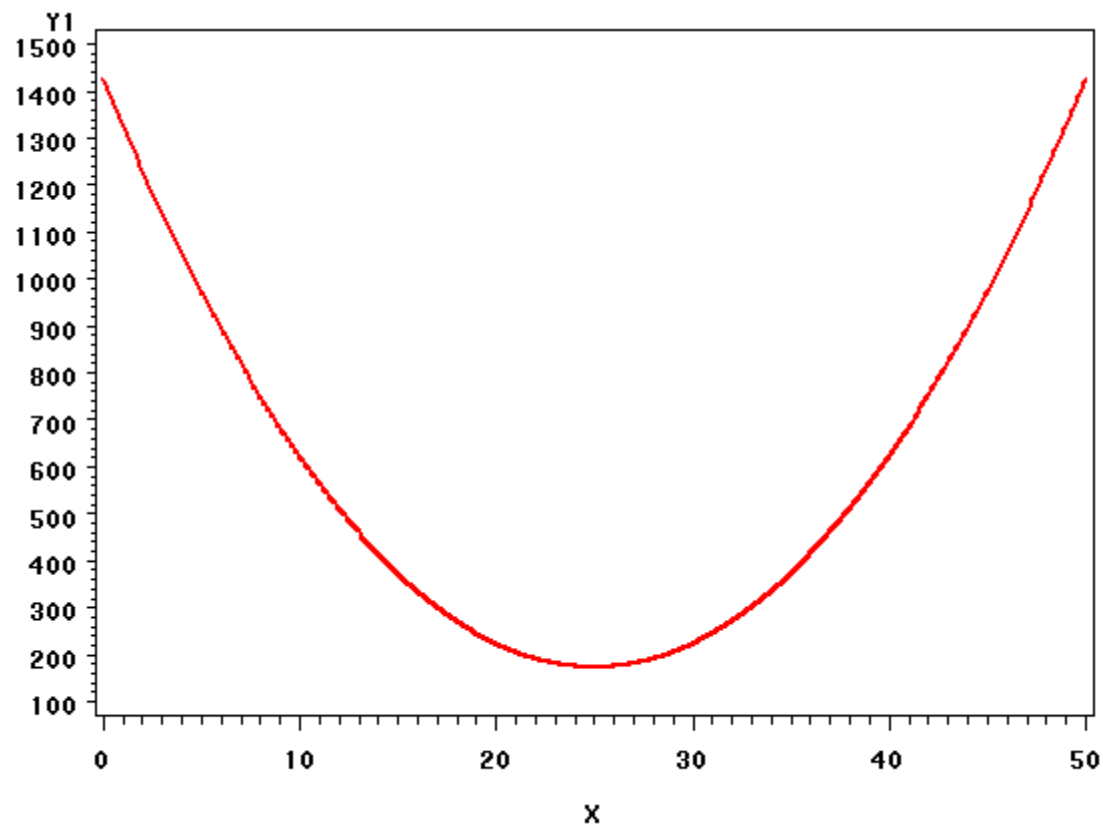
$$Y = \beta_0 + \beta_1 \ln(X) + \varepsilon$$

$$Y = e^{\beta_0 + \beta_1 X} + \varepsilon$$

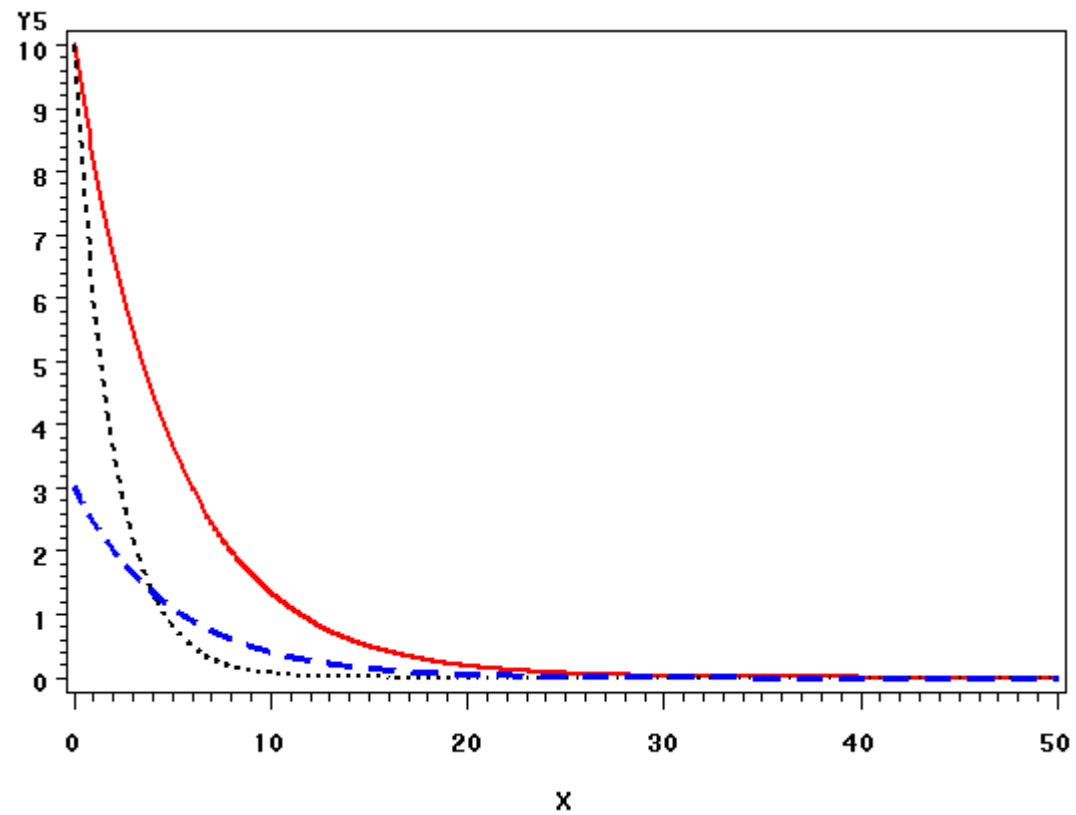
$$Y = \beta_1 e^{-\beta_2 X} + \varepsilon$$

$$Y = \beta_1 + (\beta_2 - \beta_1) e^{-\beta_3 X} + \varepsilon$$

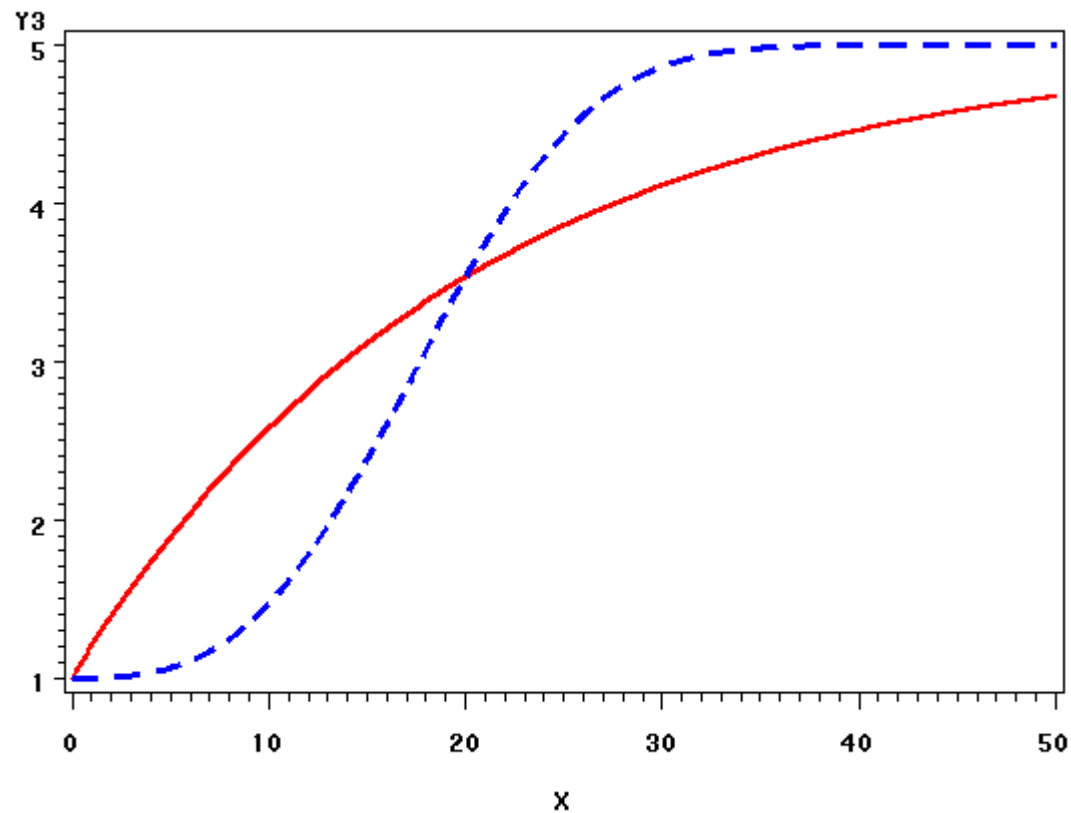
$$Y = \beta_1 X^{\beta_2} + \varepsilon$$



$$Y = \beta_1 (X - \beta_2)^2 + \beta_3 + \varepsilon$$



$$Y = \beta_1 X^{\beta_2} + \varepsilon$$



$$Y = \beta_1 + (\beta_2 - \beta_1)e^{-(\beta_3 X)^{\beta_4}} + \varepsilon$$

$$\mu_Y(x_1, \dots, x_k) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

- ▶ -1.04676230e-01 6.74894547e-02 2.34752467e-02 3.05815562e+00
- ▶ -2.36185956e+01 2.82984361e+00 -2.89871559e-04 -1.75361452e+00
- ▶ 3.40344058e-01 -1.40339781e-02 -7.98792622e-01 4.04104267e-03
- ▶ -5.69708381e-01]]
- ▶ [46.6031794]

$$MAE = \frac{1}{M} \sum_{i=1}^M |y_i - y_i^{predicted}|$$

- ▶ GRAPH algorithm

$$RMSE = \sqrt{\frac{1}{M} \sum_{i=1}^M (y_i - y_i^{predicted})^2}$$