

22.01.2019

# Statistical Methods in AI (CSE/ECE 471)

## Lecture-7:

- A short detour back to NB
- Linear Regression
- Logistic Regression

Ravi Kiran

Center for Visual Information Technology (CVIT), IIIT Hyderabad



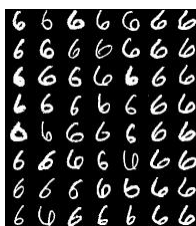
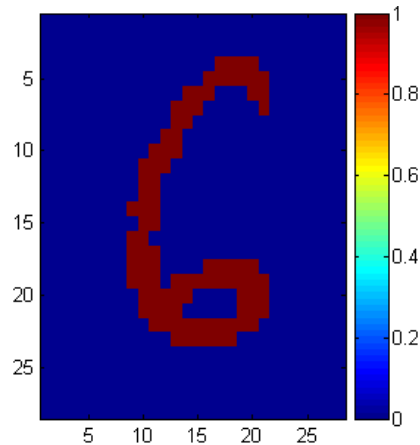
# Announcements

- A2 is due Feb 2, 11.59 pm
- SMAI Mid-1 will be on Feb 7 (Thursday)
  - Syllabus: Lec 1 – Lec 8 (this week's Friday lecture)

# A short detour back to Naïve Bayes

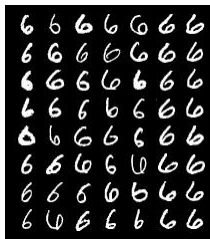
$$\arg \max_Y P(Y|X_1, \dots, X_n)$$

$$P(Y|X_1, \dots, X_n) = \frac{\overset{\text{Likelihood}}{P(X_1, \dots, X_n|Y)} \overset{\text{Prior}}{P(Y)}}{\underset{\text{Normalization Constant}}{P(X_1, \dots, X_n)}}$$



$$P(X_1, \dots, X_n|Y) = \prod_{i=1}^n P(X_i|Y)$$

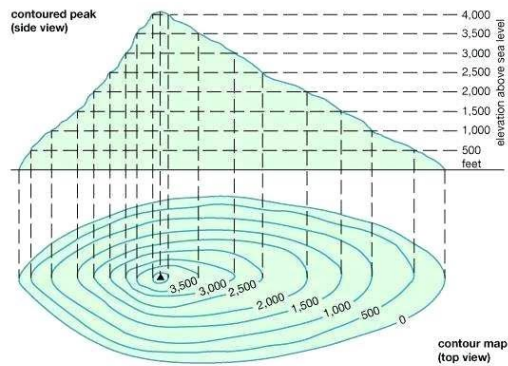
# Generative v/s Discriminative Models



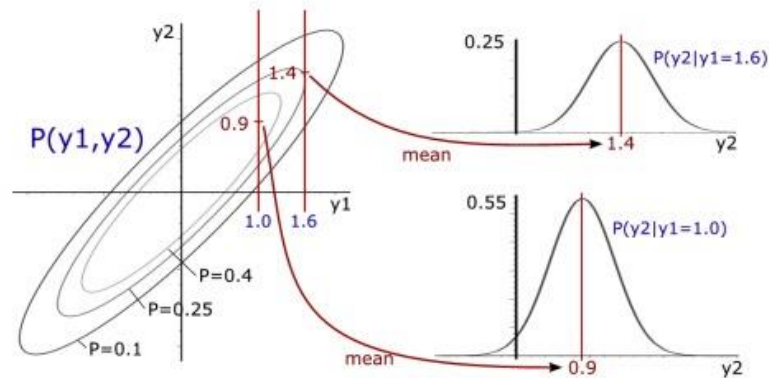
$$\arg \max_Y P(Y|X_1, \dots, X_n)$$

$$P(Y|X_1, \dots, X_n) = \frac{P(X_1, \dots, X_n|Y)P(Y)}{P(X_1, \dots, X_n)}$$

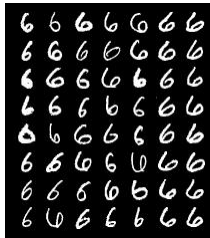
$$P(X_1, \dots, X_n|Y) = \prod_{i=1}^n P(X_i|Y)$$



© 2011 Encyclopedia Britannica, Inc.



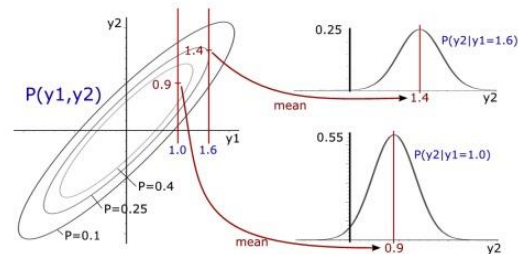
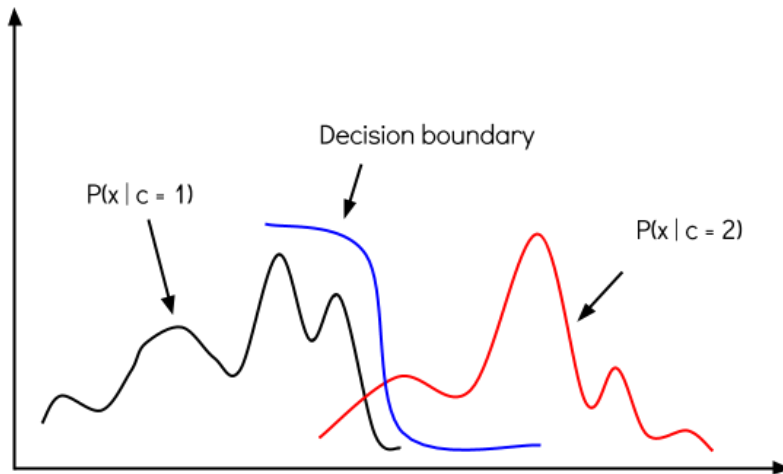
# Generative v/s Discriminative Models

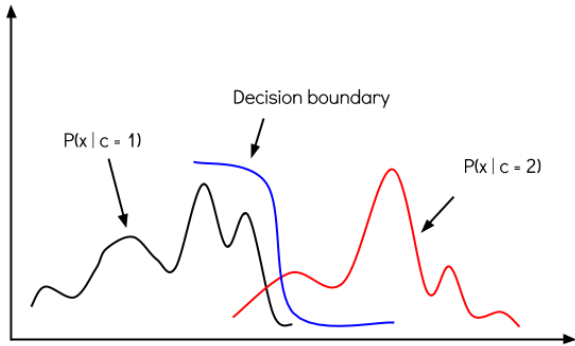


$$\arg \max_Y P(Y|X_1, \dots, X_n)$$

$$P(Y|X_1, \dots, X_n) = \frac{\overset{\text{Likelihood}}{P(X_1, \dots, X_n|Y)} \overset{\text{Prior}}{P(Y)}}{\underset{\text{Normalization Constant}}{P(X_1, \dots, X_n)}}$$

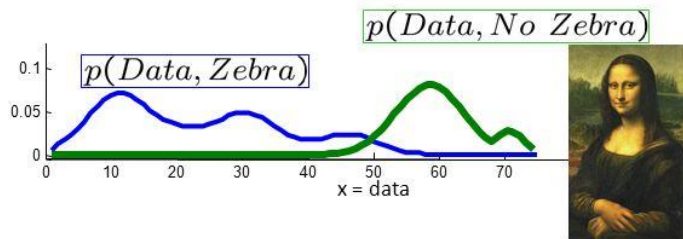
$$P(X_1, \dots, X_n|Y) = \prod_{i=1}^n P(X_i|Y)$$

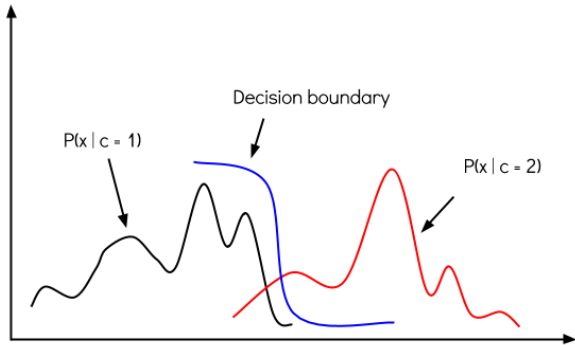




# Discriminative vs. generative

- Generative model  
(The artist)

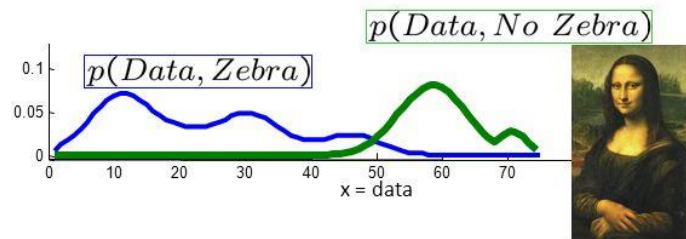




# Discriminative vs. generative

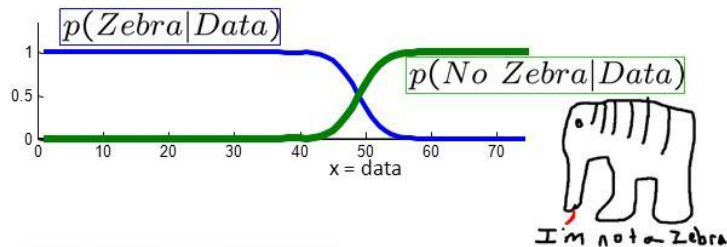
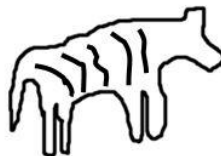
- Generative model

(The artist)

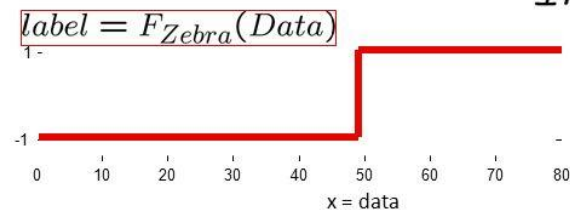


- Discriminative model

(The lousy painter)



- Classification function



# Supervised Learning

```
graph TD; A[Supervised Learning] --> B[Classification]; A --> C[Regression]; A --> D[Reinforcement Learning]; style C stroke-dasharray: 5 5
```

Classification

Regression

Reinforcement  
Learning



# Linear Regression Model

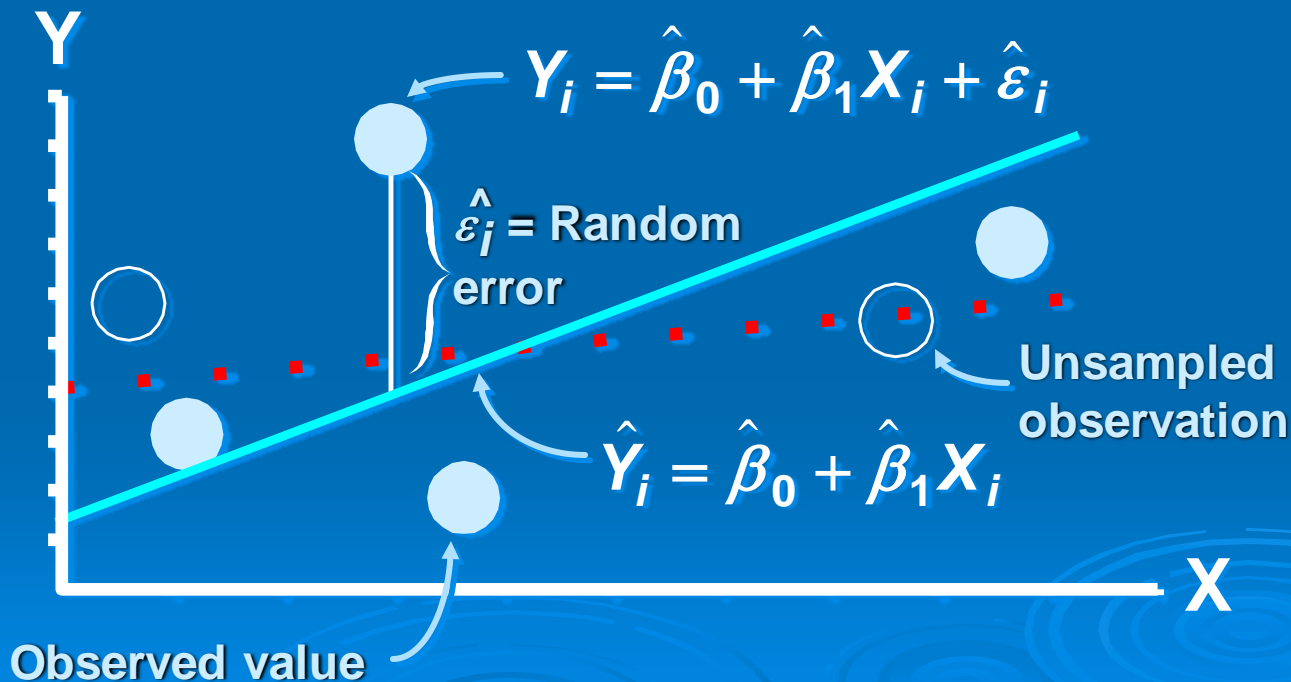
## ➤ 1. Relationship Between Variables Is a Linear Function

The diagram illustrates the Linear Regression Model equation  $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ . It features five labels with arrows pointing to the corresponding parts of the equation:

- Population Y-Intercept** points to  $\beta_0$ .
- Population Slope** points to  $\beta_1$ .
- Random Error** points to  $\varepsilon_i$ .
- Dependent (Response) Variable (e.g. Salary)** points to  $Y_i$ .
- Independent (Explanatory) Variable (e.g. Yrs of experience)** points to  $X_i$ .

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

# Sample Linear Regression Model



# Least Squares

- 1. 'Best Fit' Means Difference Between Actual Y Values & Predicted Y Values is a Minimum. *But* Positive Differences Off-Set Negative ones. **So square errors!**

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

# Least Squares

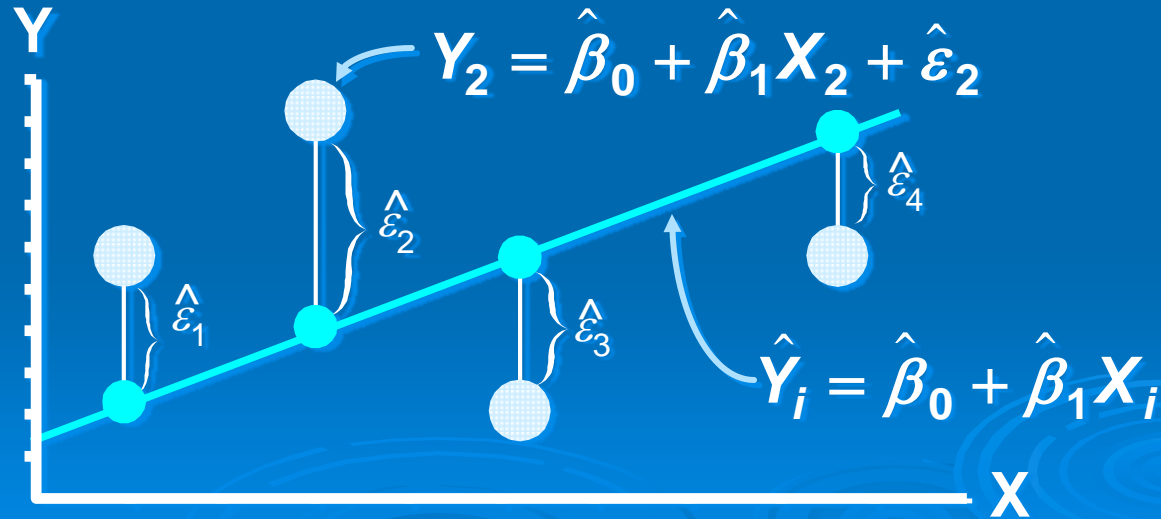
- 1. 'Best Fit' Means Difference Between Actual Y Values & Predicted Y Values Are a Minimum. *But* Positive Differences Off-Set Negative. So square errors!

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n \hat{\mathcal{E}}_i^2$$

- 2. LS Minimizes the Sum of the Squared Differences (errors) (SSE)

# Least Squares Graphically

LS minimizes  $\sum_{i=1}^n \hat{\varepsilon}_i^2 = \hat{\varepsilon}_1^2 + \hat{\varepsilon}_2^2 + \hat{\varepsilon}_3^2 + \hat{\varepsilon}_4^2$



# Coefficient Equations

## ➤ Prediction equation

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

## ➤ Sample slope

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

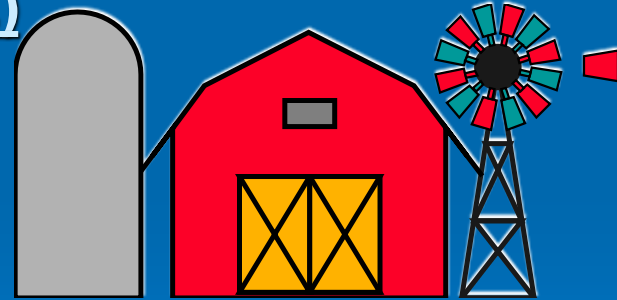
## ➤ Sample Y - intercept

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

# Parameter Estimation Thinking Challenge

- You're a Vet epidemiologist for the county cooperative. You gather the following data:

- | <u>Food (lb.)</u> | <u>Milk yield (lb.)</u> |
|-------------------|-------------------------|
| 4                 | 3.0                     |
| 6                 | 5.5                     |
| 10                | 6.5                     |
| 12                | 9.0                     |



© 1984-1994 T/Maker Co.

- What is the **relationship** between cows' food intake and milk yield?

# Coefficient Interpretation Solution\*

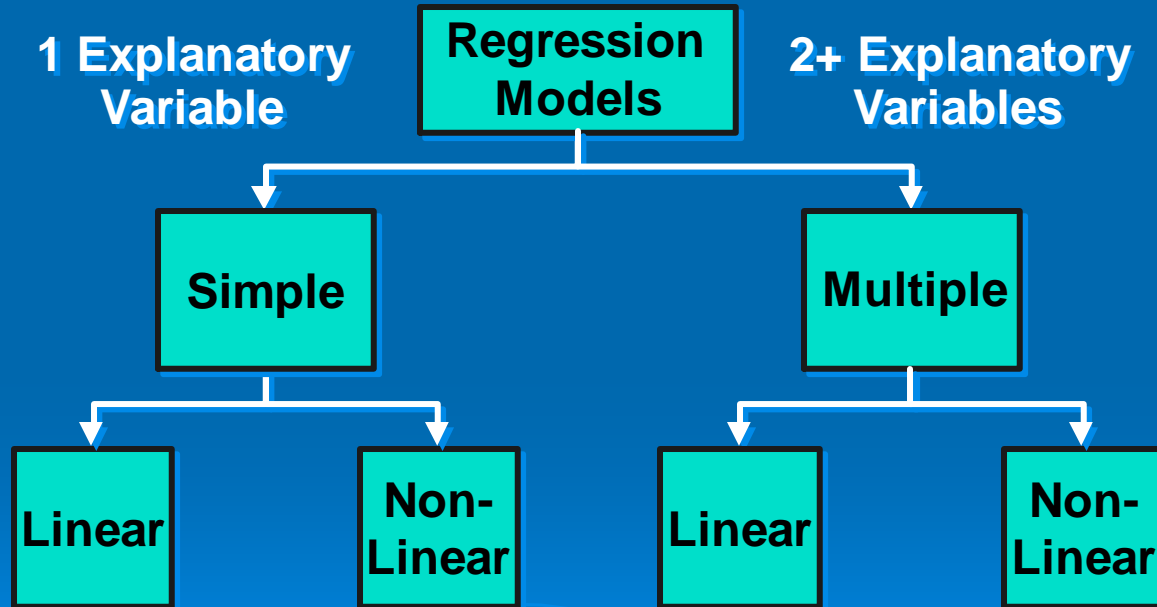
- 1. Slope ( $\hat{\beta}_1$ )
  - Milk Yield ( $Y$ ) Is Expected to Increase by .65 lb. for Each 1 lb. Increase in Food intake ( $X$ )



# Coefficient Interpretation Solution\*

- 1. Slope ( $\hat{\beta}_1$ )
  - **Milk** Yield ( $Y$ ) Is Expected to Increase by .65 lb. for Each 1 lb. Increase in **Food intake** ( $X$ )
  
- 2. Y-Intercept ( $\hat{\beta}_0$ )
  - Average Milk yield ( $Y$ ) Is Expected to Be 0.8 lb. When Food intake ( $X$ ) Is 0

# Types of Regression Models

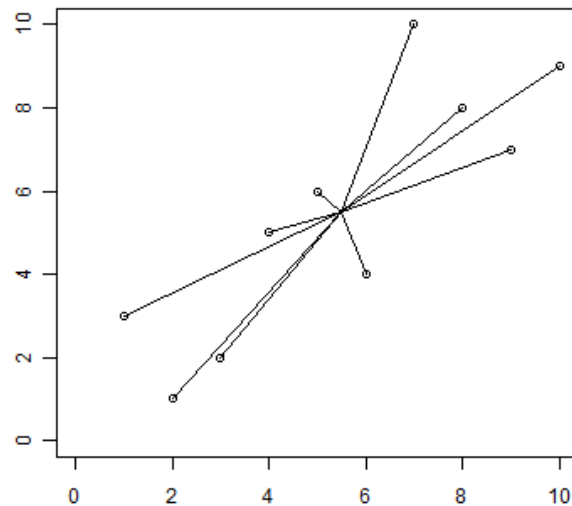
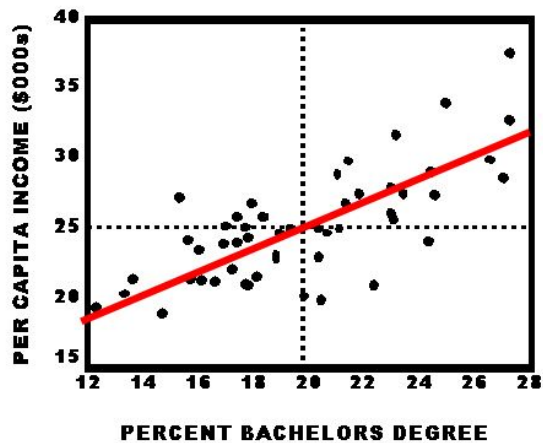


# Interpretation of coefficients

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

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

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$



# Regression – Error measures

$$MAE = \frac{1}{n} \sum |y - \hat{y}|$$

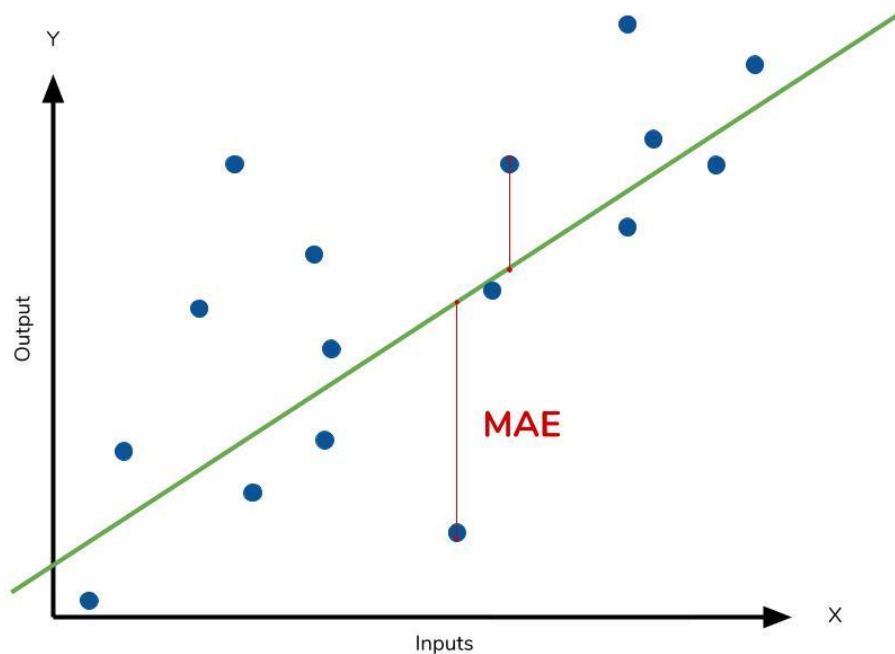
Divide by the total number of data points

Actual output value

Predicted output value

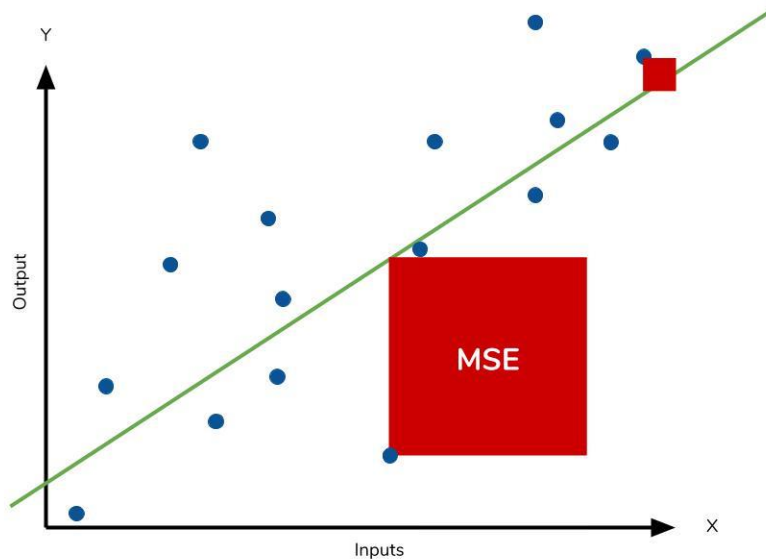
Sum of

The absolute value of the residual

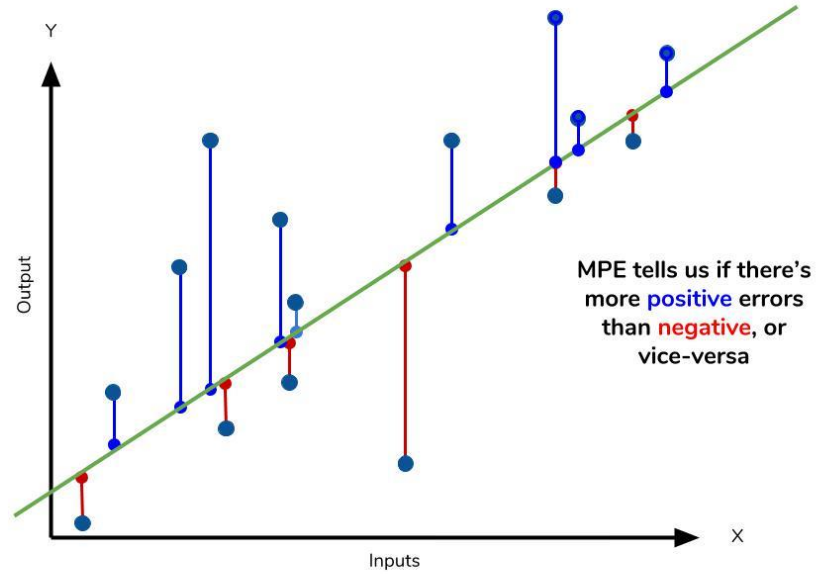


# Regression – Error measures

$$MSE = \frac{1}{n} \sum \left( \underbrace{y - \hat{y}}_{\substack{\text{The square of the difference} \\ \text{between actual and} \\ \text{predicted}}} \right)^2$$

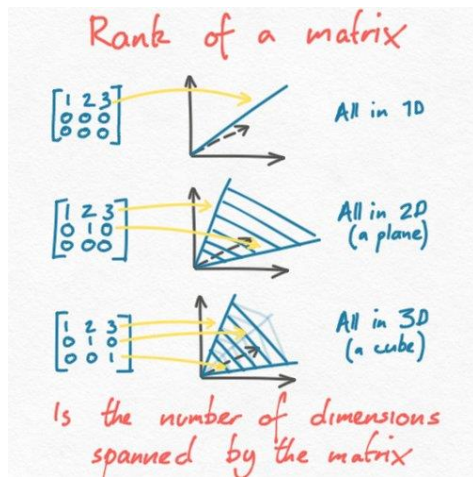


$$MPE = \frac{100\%}{n} \sum \left( \frac{y - \hat{y}}{y} \right)$$



# Linear Algebra in 1 slide

- Matrices – Square, Rectangular
- Matrix ops – Transpose, Inverse
- Rank of a matrix



# Linear Regression – Matrix Form



# Linear Regression – Matrix Form

Consider the model

$$Y = X\beta + \epsilon$$

where  $Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$   $X = \begin{pmatrix} 1 & X_{11} & X_{12} & \dots & X_{1p} \\ 1 & X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{np} \end{pmatrix}$   $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$   $\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$

Based on this model we get the following expansion for the first subject:

$$Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \dots + \beta_p X_{1p} + \epsilon_1$$

Then using matrix calculus we find that the least squares estimate for  $\beta$  is given by

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

Hence, the least squares regression line is  $\hat{Y} = X\hat{\beta}$ .

# Linear Regression – Matrix Form - Issues

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

Hence, the least squares regression line is  $\hat{Y} = X\hat{\beta}$ .

- N samples, p-dimensional (what if  $p > N$  ?)
- Complexity of matrix inversion (what if N very large ?)
- Collinearity

- Linear Regression → Linear in coefficients and NOT variables

- A second-order model (quadratic model):

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

- $\beta_1$ : Linear effect parameter.
- $\beta_2$ : Quadratic effect parameter.

*k*th order polynomial model in one variable

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \epsilon$$

## A quadratic polynomial regression function

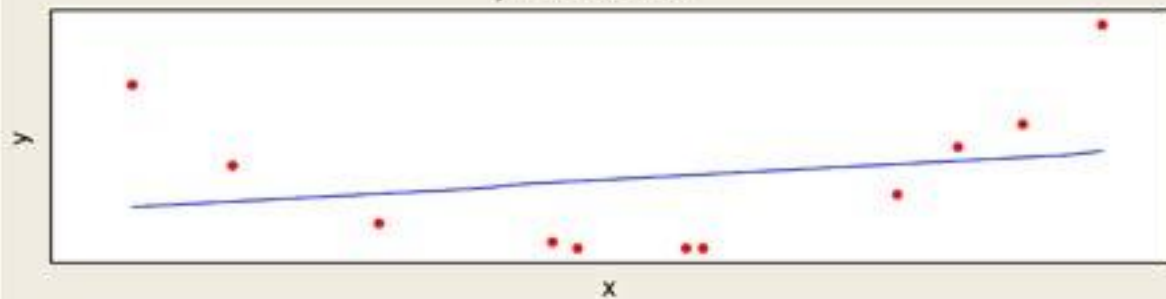
$$Y_i = \beta_0 + \beta_1 X_i + \beta_{11} X_i^2 + \varepsilon_i$$

where:

- $Y_i$  = amount of immunoglobulin in blood (mg)
- $X_i$  = maximal oxygen uptake (ml/kg)
- typical assumptions about error terms (“INE”)

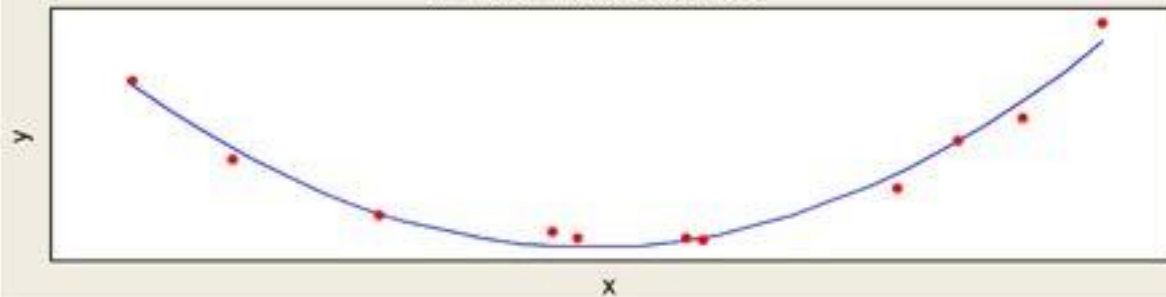
### Fitted Line Plot for Linear Model

$$y = 1.79 + 0.3869 x$$



### Fitted Line Plot for Quadratic Model

$$y = 113.8 - 11.63 x + 0.2967 x^2$$

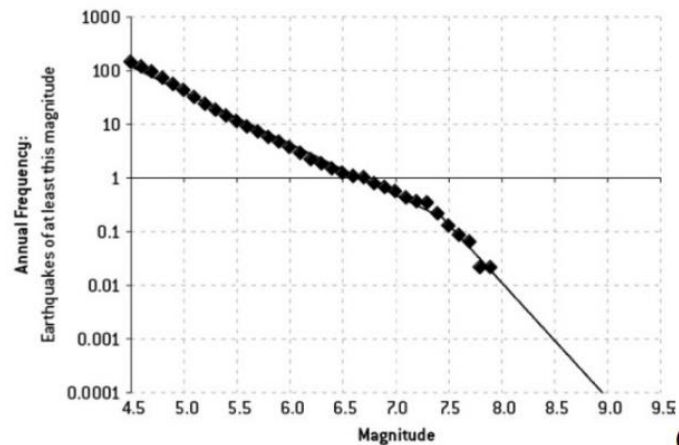


- Linear Regression → Linear in coefficients and NOT variables

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^m \\ 1 & x_2 & x_2^2 & \dots & x_2^m \\ 1 & x_3 & x_3^2 & \dots & x_3^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^m \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

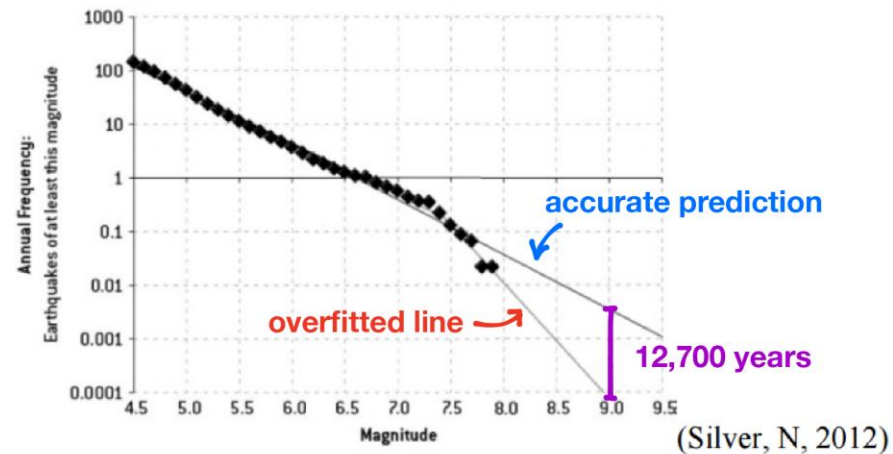


FIGURE 5-7C: TŌHOKU, JAPAN EARTHQUAKE FREQUENCIES  
CHARACTERISTIC FIT



(Silver, N, 2012)

FIGURE 5-7C: TŌHOKU, JAPAN EARTHQUAKE FREQUENCIES  
CHARACTERISTIC FIT



(Silver, N, 2012)

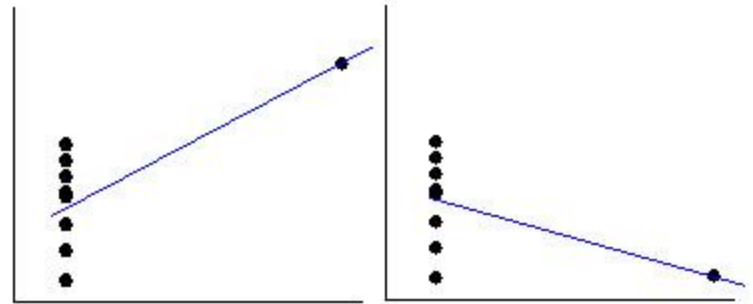
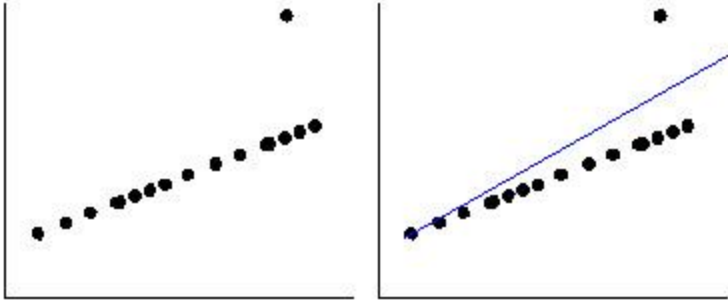


# Careful: X may not be **causing** y !

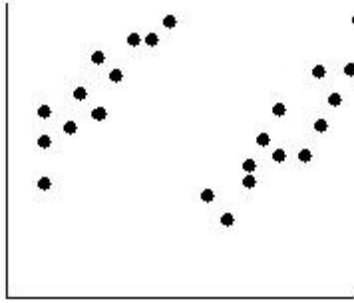
## CORRELATION DOES NOT MEAN CAUSATION



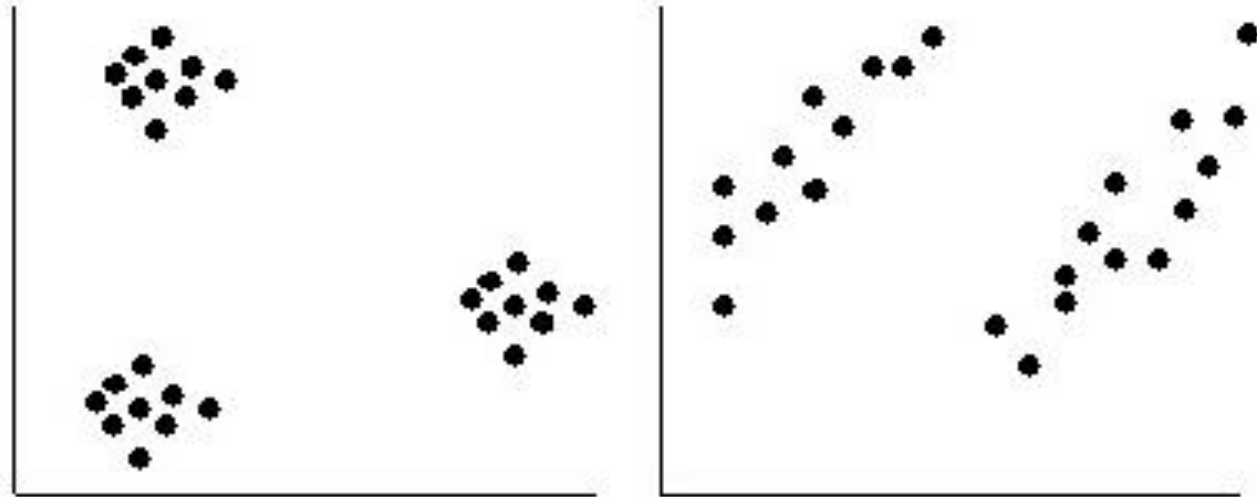
# Linear Regression – Outliers



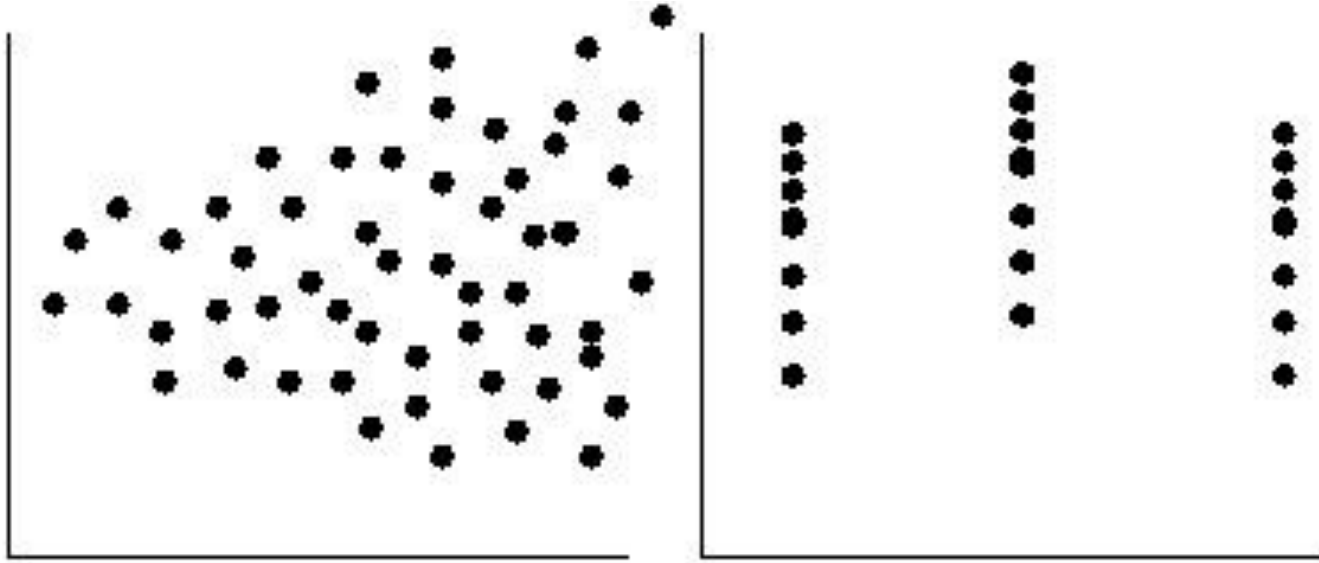
# Linear Regression is problematic in many other cases



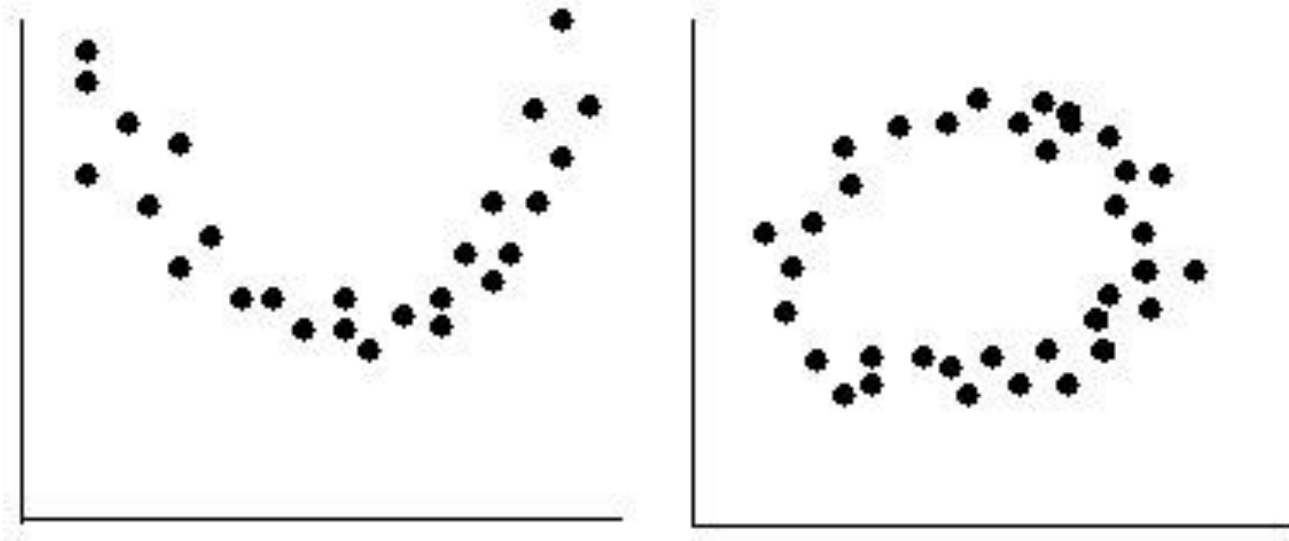
# Linear Regression is problematic in many other cases



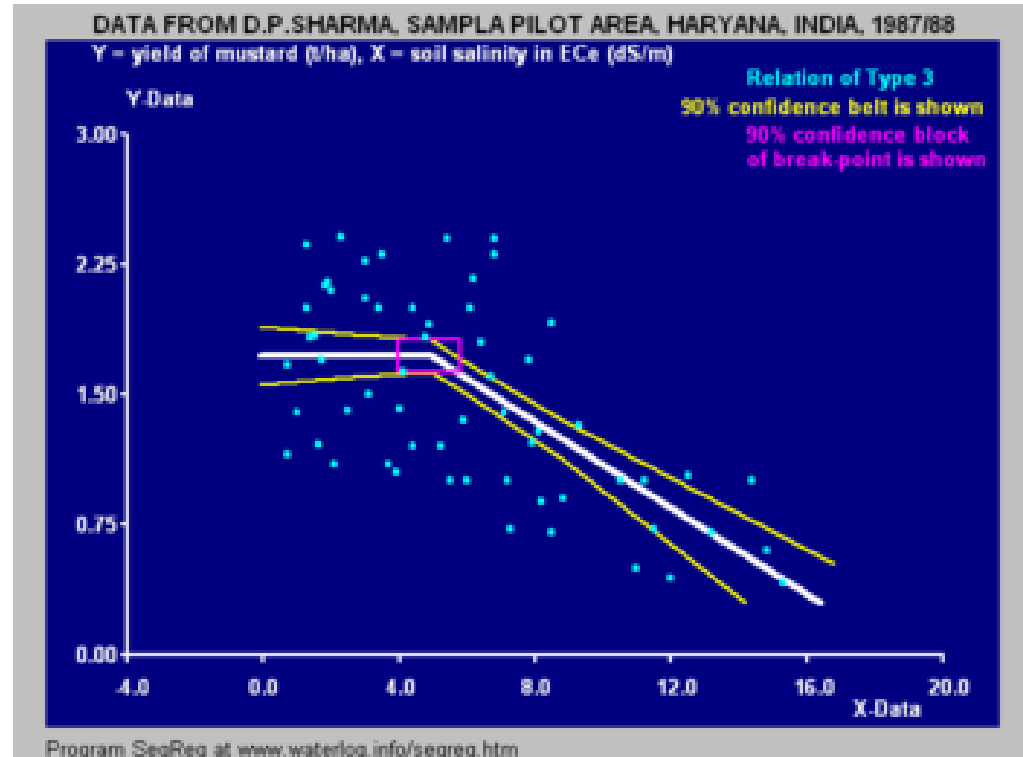
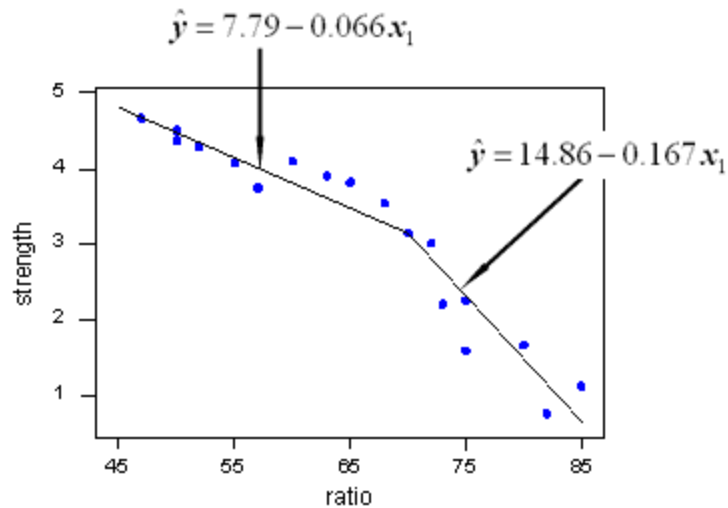
# Linear Regression is problematic in many other cases



# Linear Regression is problematic in many other cases

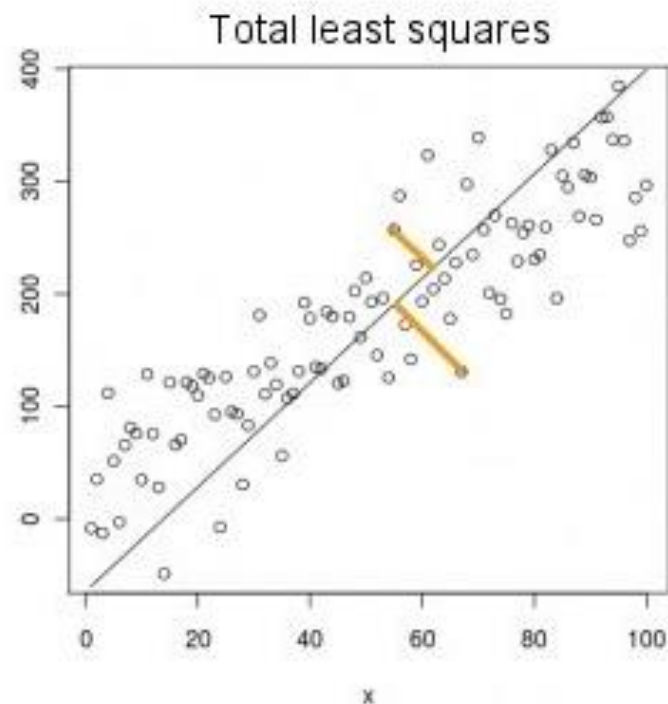
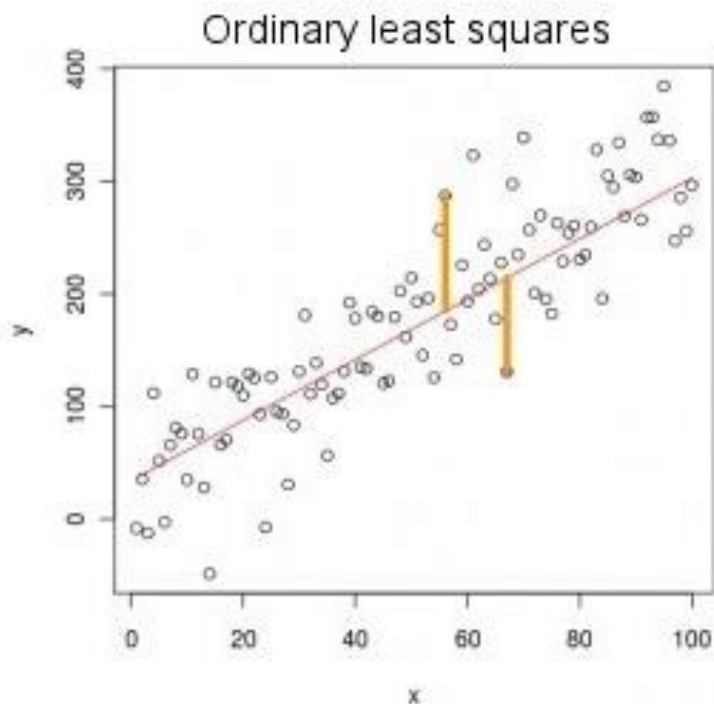


# Piecewise Linear Regression



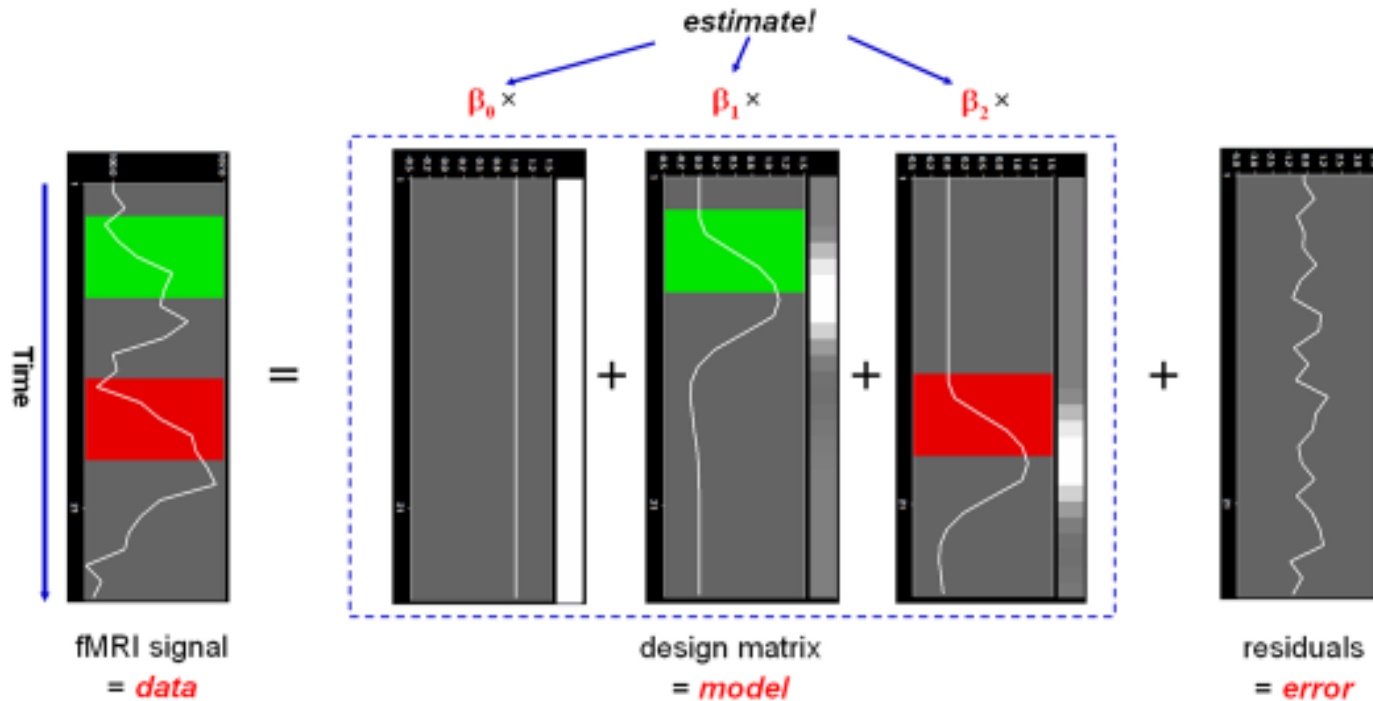
# Total Least Squares

(“Errors-in-variables” model)





# General Linear Models



# Generalized Linear Models

- For bounded or discrete data
  - Positive quantities (e.g. prices, populations)
  - Varying over a large scale (log-normal, Poisson distribution)
  - Categorical data (Bernoulli / Binomial / Multinomial)
  - Ordinal data (e.g. ratings – Ordered logit)

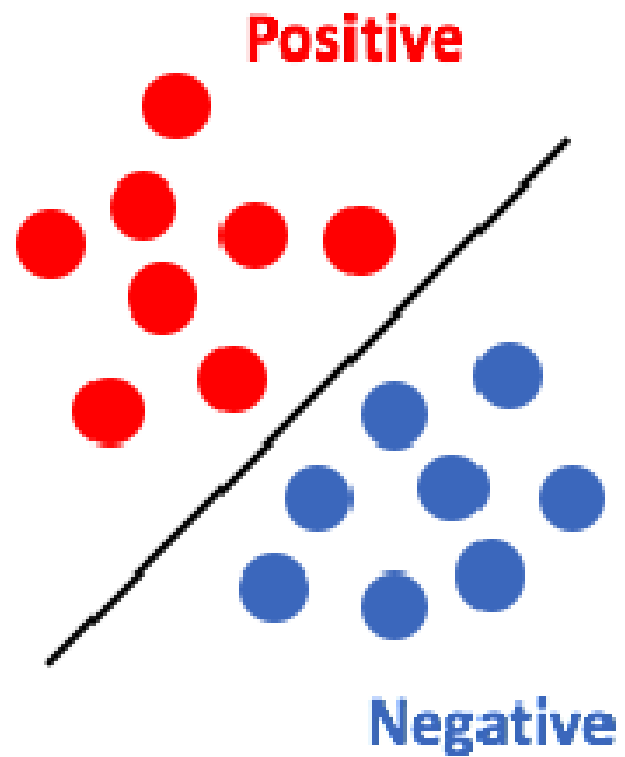
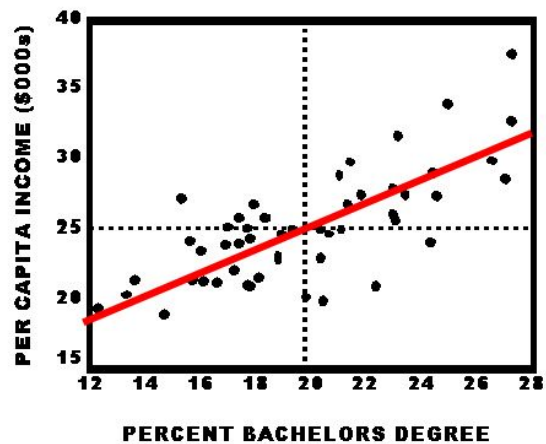
# Supervised Learning

```
graph TD; A[Supervised Learning] --> B[Classification]; A --> C[Regression]; A --> D[Reinforcement Learning];
```

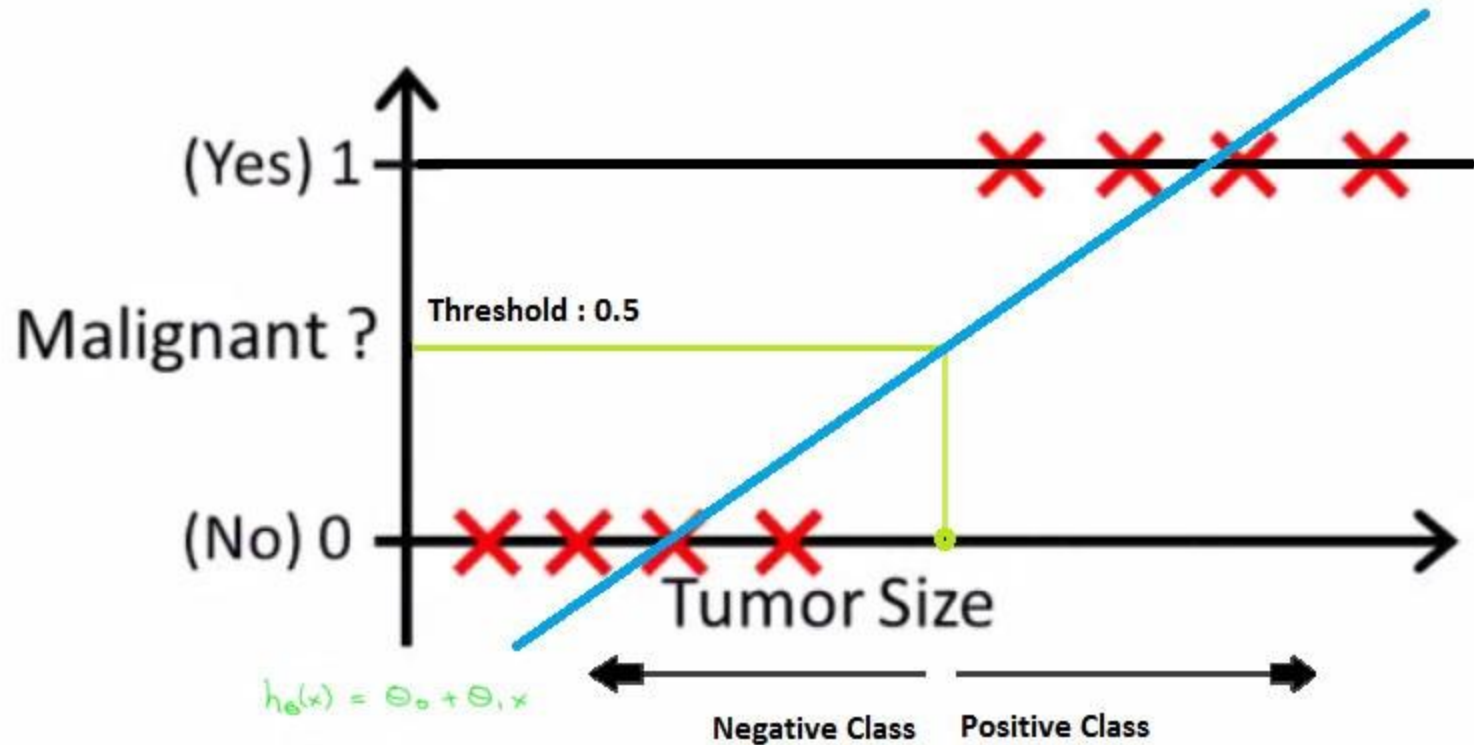
Classification

Regression

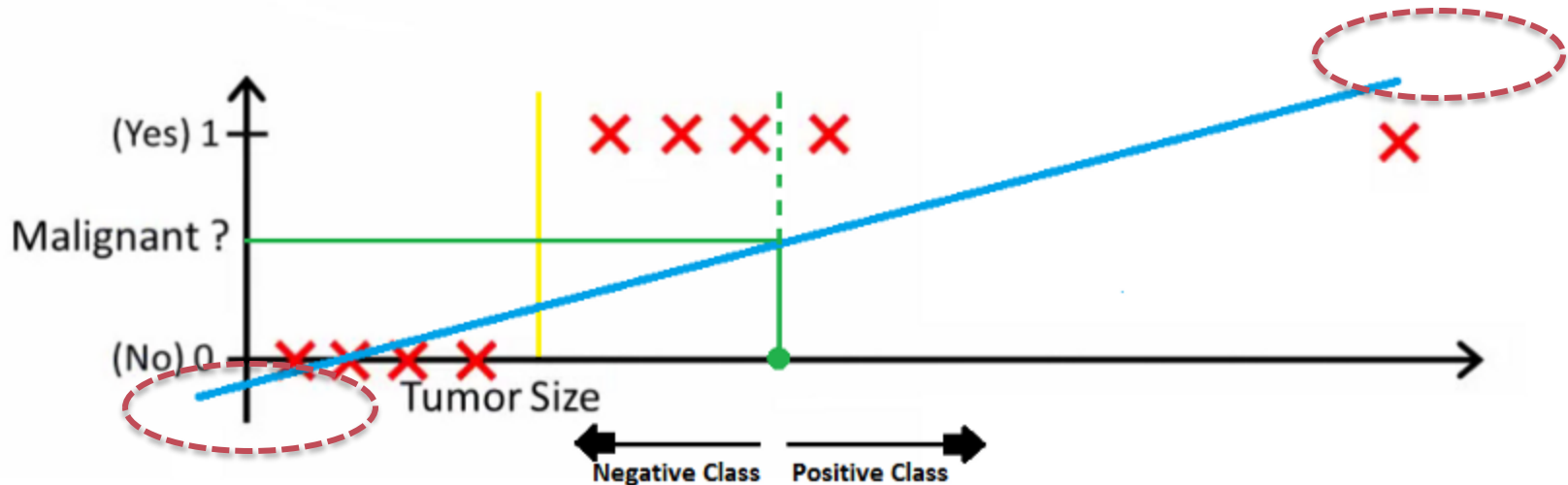
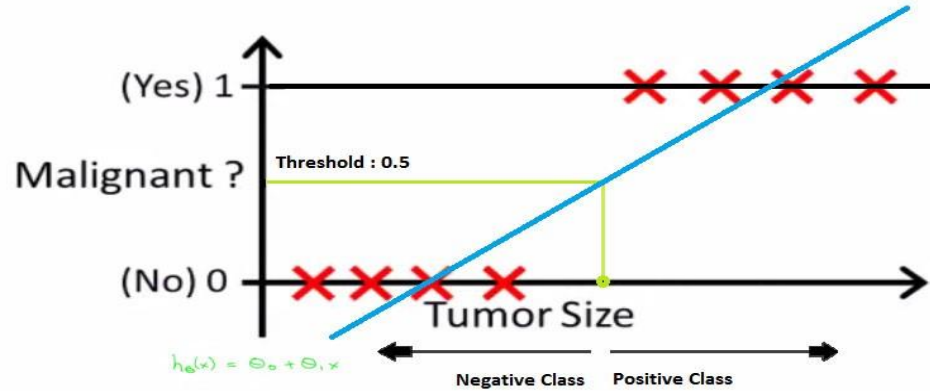
Reinforcement  
Learning



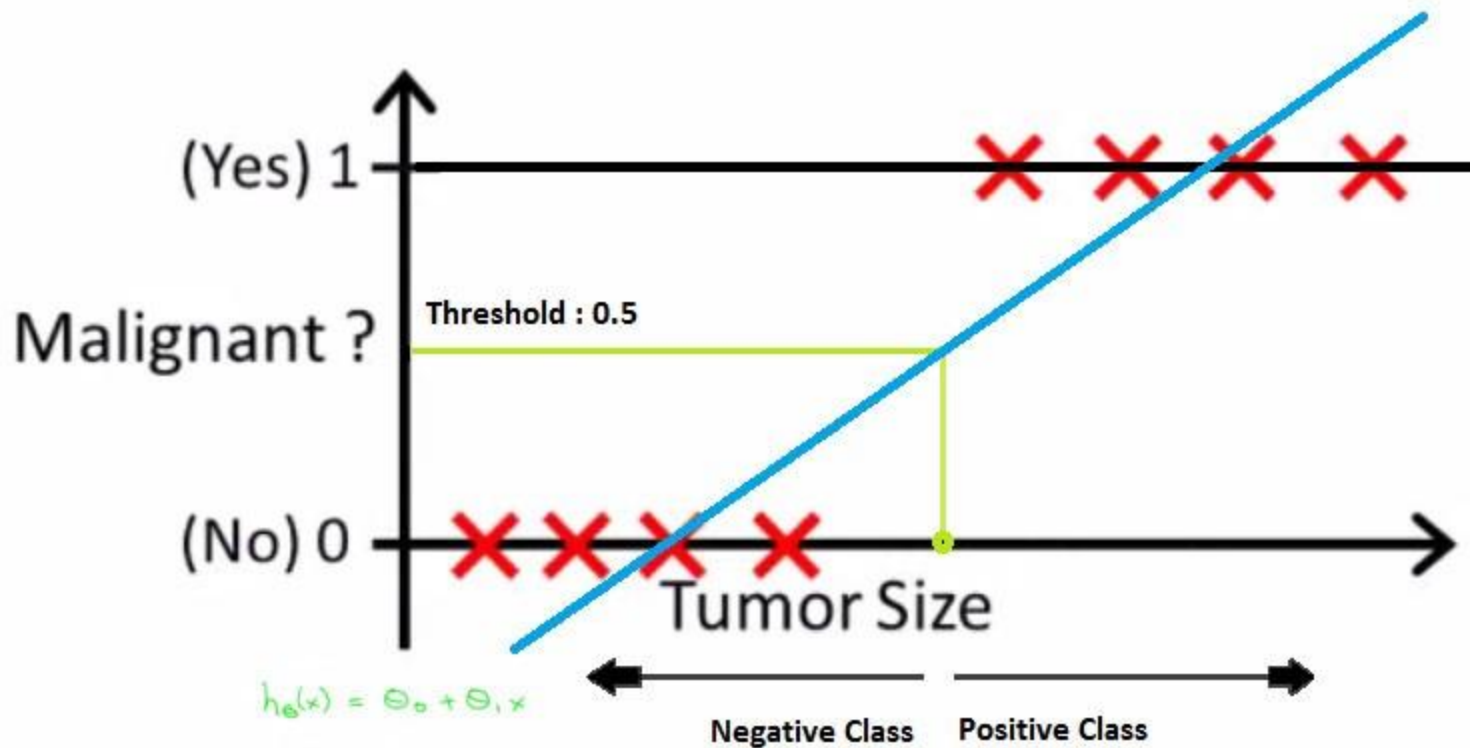
# Re-using linear regression ?



# Re-using linear regression ?



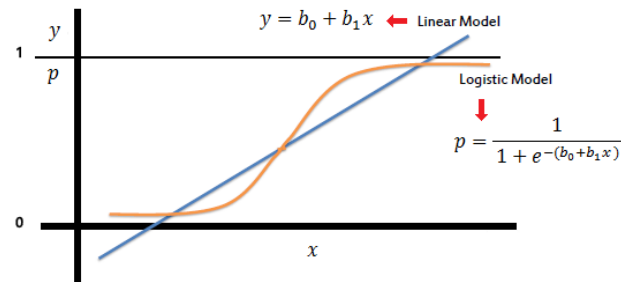
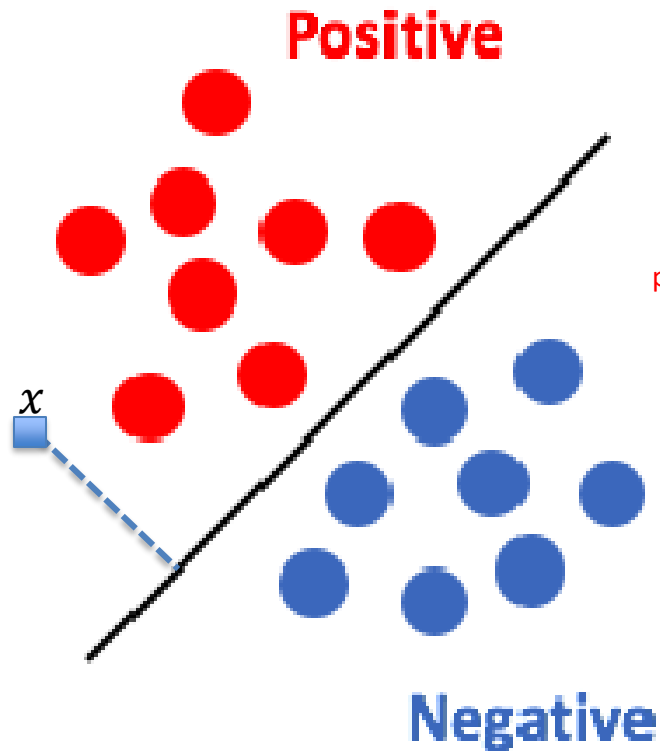
# What we really want – a step function !



- We want a step-function like behavior
  - But with nicer mathematical properties (e.g. like linear regression)!
- Probabilistic classification is also nice (Naïve Bayes)
- Combine all of these ?



# Logistic Regression - Intuition



$p(X) = p(Y=1|x) = \text{probability that } x \text{ belongs to positive class}$

$$\begin{aligned} \Rightarrow p(X) &= \frac{e^{(\beta_0 + \beta_1 x)}}{e^{(\beta_0 + \beta_1 x)} + 1} \\ \Rightarrow p(e^{(\beta_0 + \beta_1 x)} + 1) &= e^{(\beta_0 + \beta_1 x)} \\ \Rightarrow p \cdot e^{(\beta_0 + \beta_1 x)} + p &= e^{(\beta_0 + \beta_1 x)} \\ \Rightarrow p &= e^{(\beta_0 + \beta_1 x)} - p \cdot e^{(\beta_0 + \beta_1 x)} \\ \Rightarrow p &= e^{(\beta_0 + \beta_1 x)}(1 - p) \\ \Rightarrow \frac{p}{1 - p} &= e^{(\beta_0 + \beta_1 x)} \\ \Rightarrow \ln\left(\frac{p}{1 - p}\right) &= \beta_0 + \beta_1 x \end{aligned}$$

Distance of  $x$  from boundary

# Maximum Likelihood

- The likelihood function is the simultaneous density of the observation, as a function of the model parameters.

$$L(\Theta) = \Pr(Data|\Theta)$$

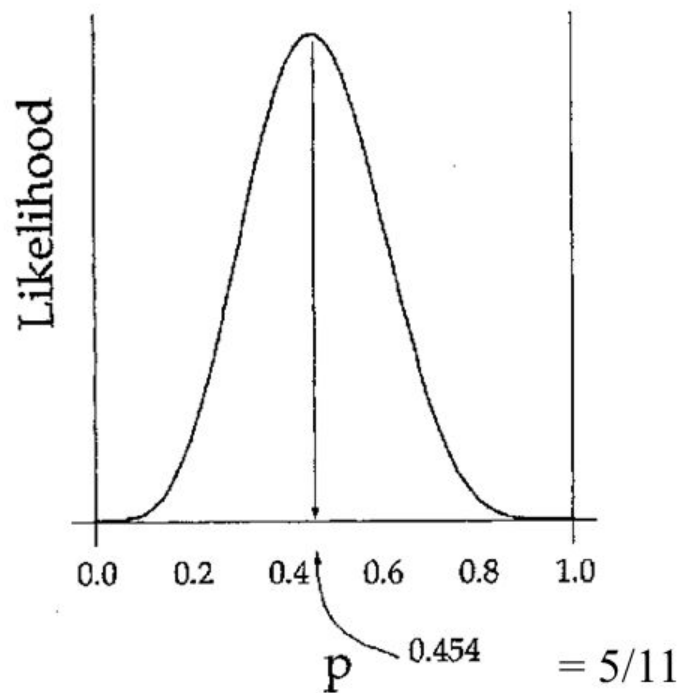
- If the observations are independent, we can decompose the term into

$$\Pr(Data | \Theta) = \prod_{i=1}^n \Pr(X_i | \Theta)$$

## An example

- Consider the estimation of heads probability of a coin tossed  $n$  times
- Heads probability  $p$
- Data = HHTTHTHHTTT
- $L(p) = \Pr(D|p) = pp(1-p)(1-p)p(1-p)pp(1-p)(1-p)(1-p) = p^5(1-p)^6$

$$L(p) = p^5(1-p)^6$$



# Maximum Likelihood

$$L(p) = p^5(1-p)^6$$

Take the derivative of  $L$  with respect to  $p$ :

$$\frac{dL}{dp} = 5 p^4 (1-p)^6 - 6 p^5 (1-p)^5$$

Equate it to zero and solve:

$$\hat{p} = 5/11$$

# Log Likelihood

$$L(p) = p^5(1-p)^6$$

- For computational reasons, we maximise the logarithm

$$\ln L = 5 \ln p + 6 \ln(1-p)$$

with derivative

$$\frac{d(\ln L)}{dp} = \frac{5}{p} - \frac{6}{(1-p)} = 0$$

$$\hat{p} = 5/11$$

Maximum Likelihood Estimator will maximize  $p(x_1)p(x_2)p(x_3)\dots$

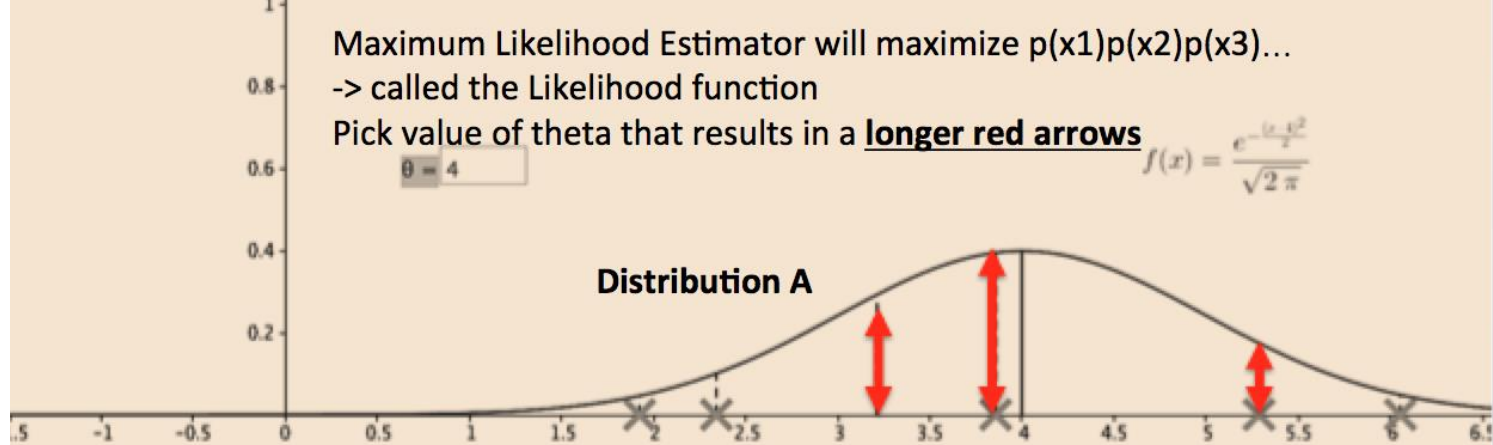
-> called the Likelihood function

Pick value of theta that results in a **longer red arrows**

$$f(x) = \frac{e^{-\frac{(x-\theta)^2}{2}}}{\sqrt{2\pi}}$$

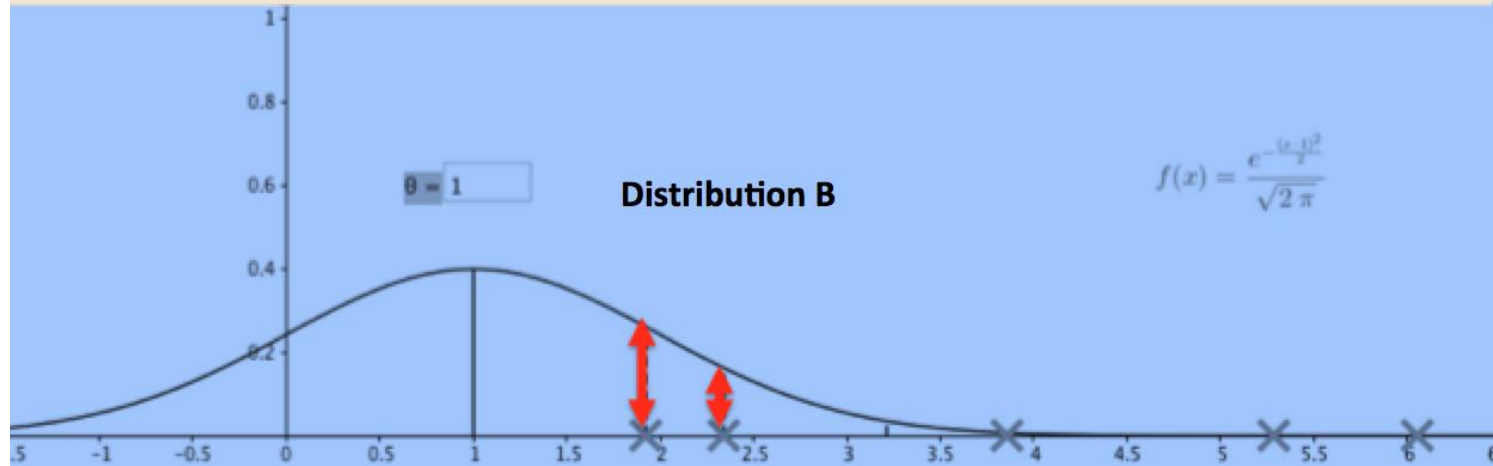
$\theta = 4$

Distribution A



$\theta = 1$

Distribution B



# Logistic Regression - Learning parameters

$$h_{\theta}(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}};$$

$$p(y \mid x; \theta) = (h_{\theta}(x))^y (1 - h_{\theta}(x))^{1-y}$$

$$\begin{aligned} L(\theta) &= p(\vec{y} \mid X; \theta) \\ &= \prod_{i=1}^m p(y^{(i)} \mid x^{(i)}; \theta) \\ &= \prod_{i=1}^m (h_{\theta}(x^{(i)}))^{y^{(i)}} (1 - h_{\theta}(x^{(i)}))^{1-y^{(i)}} \end{aligned}$$

$$\begin{aligned} \ell(\theta) &= \log L(\theta) \\ &= \sum_{i=1}^m y^{(i)} \log h(x^{(i)}) + (1 - y^{(i)}) \log(1 - h(x^{(i)})) \end{aligned}$$



# References and Reading

- Linear Regression
  - [https://en.wikipedia.org/wiki/Linear\\_regression](https://en.wikipedia.org/wiki/Linear_regression)
  - <http://www.stat.purdue.edu/~boli/stat512/lectures/topic3.pdf> (up to page 7)
- Logistic Regression
  - [https://en.wikipedia.org/wiki/Logistic\\_regression](https://en.wikipedia.org/wiki/Logistic_regression) (up to Section 6)