Παλινδρόμηση

Μοντέλα

- Το μαθηματικό μοντέλο είναι μια μαθηματική έκφραση κάποιου φαινομένου
- Συχνά περιγράφουν σχέσεις μεταξύ μεταβλητών
- Τύποι
  - Ντετερμινιστικά μοντέλα
  - Πιθανοτικά μοντέλα

Ντετερμινιστικά

- Υποθέτουμε ακριβείς σχέσεις
- Κατάλληλο όταν το σφάλμα πρόβλεψης είναι αμελητέο
- Παράδειγμα: η δύναμη είναι ακριβώς η μάζα επί την επιτάχυνση

  $F = m \cdot a$

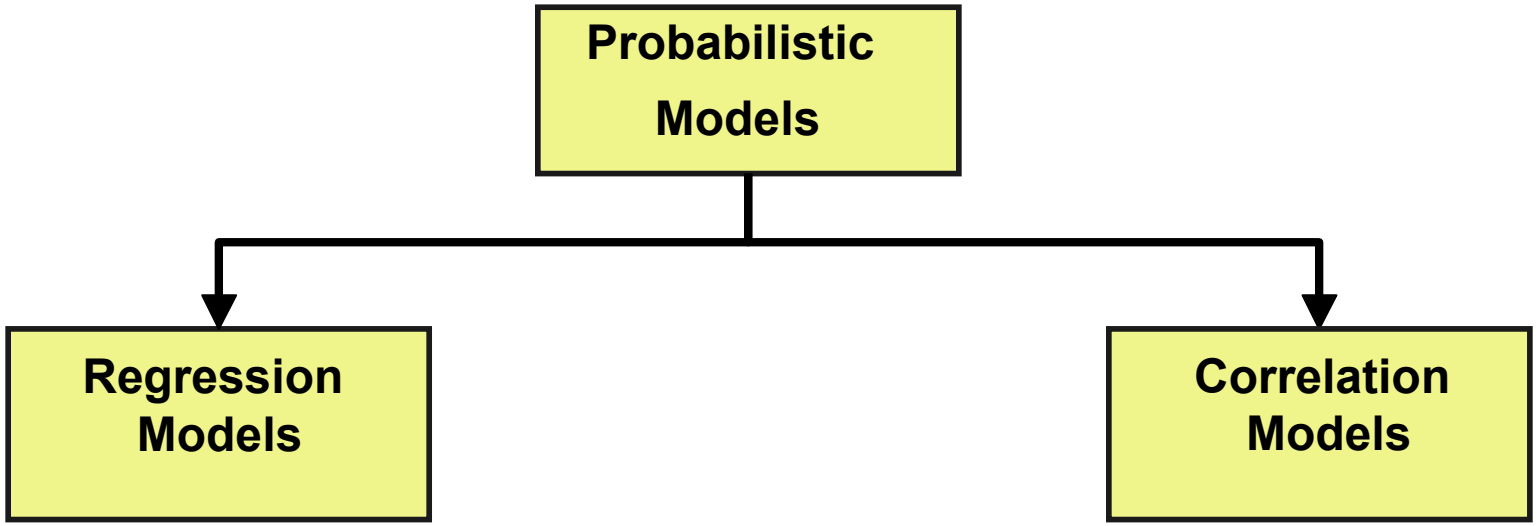
Πιθανοτικά

- Υποθέστε δύο συστατικά
  - Ντετερμινιστικό
  - Τυχαίο σφάλμα
- Παράδειγμα: ο όγκος πωλήσεων (y) είναι 10πλάσιος των διαφημιστικών δαπανών (x) + τυχαίο σφάλμα

$$y = 10x + \varepsilon$$

- Το τυχαίο σφάλμα μπορεί να οφείλεται σε άλλους παράγοντες εκτός από τη διαφήμιση

Τύποι


Probabilistic Models
Regression Models
Correlation Models

Παλινδρόμηση

- Απαντήσεις «Ποια είναι η σχέση μεταξύ των μεταβλητών;»
- Εξίσωση που χρησιμοποιείται
  - Μία αριθμητική εξαρτημένη (απόκριση) μεταβλητή
  - Τι είναι να προβλεφθεί
- Μία ή περισσότερες αριθμητικές ή κατηγορικές ανεξάρτητες (επεξηγηματικές) μεταβλητές
- Χρησιμοποιείται κυρίως για πρόβλεψη και εκτίμηση

Βήματα

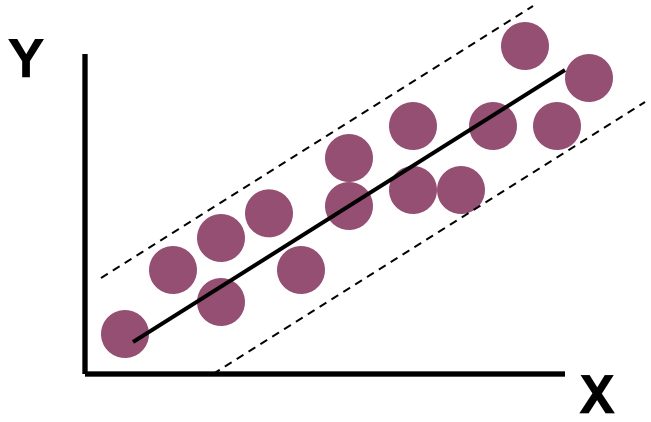
- Υποθέτουμε τη ντετερμινιστική συνιστώσα
- Εκτίμηση άγνωστων παραμέτρων του μοντέλου
- Καθορίζουμε την κατανομή πιθανότητας του τυχαίου σφάλματος
- Εκτίμηση τυπικής απόκλισης σφάλματος
- Αξιολογούμε το μοντέλο
- Χρησιμοποιούμε μοντέλο για πρόβλεψη και εκτίμηση

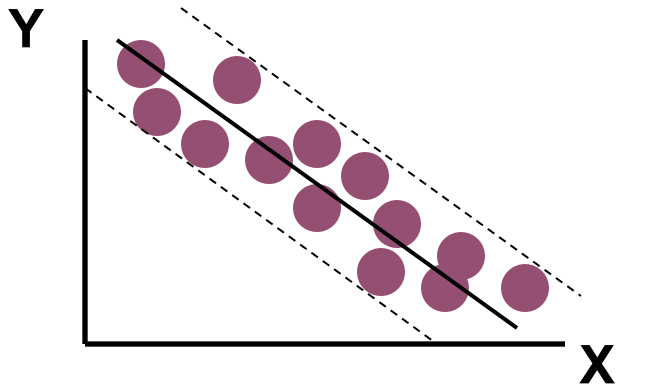
Ορισμός

- Ορίζουμε τις μεταβλητές
  - Εννοιολογική (π.χ. Διαφήμιση, τιμή)
  - Εμπειρική (π.χ. τιμή καταλόγου, κανονική τιμή)
  - Μέτρηση (π.χ. $, Μονάδες)
- Υποθέστε τη φύση της σχέσης
  - Αναμενόμενα αποτελέσματα (π.χ. πρόσημα συντελεστών)
  - Λειτουργική μορφή (γραμμική ή μη γραμμική)
  - Αλληλεπιδράσεις

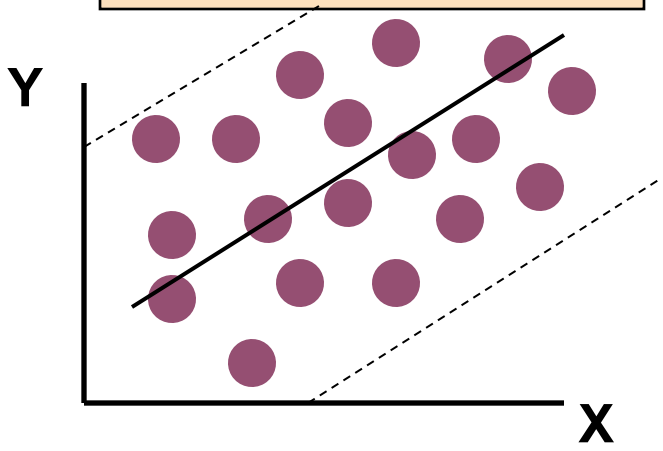
Συσχετίσεις

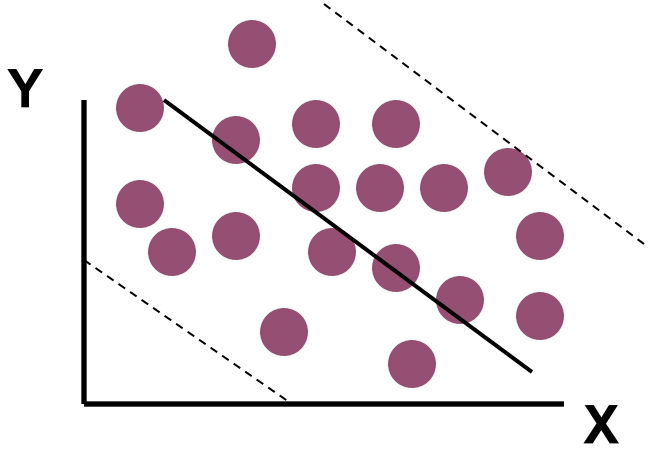
Strong relationships



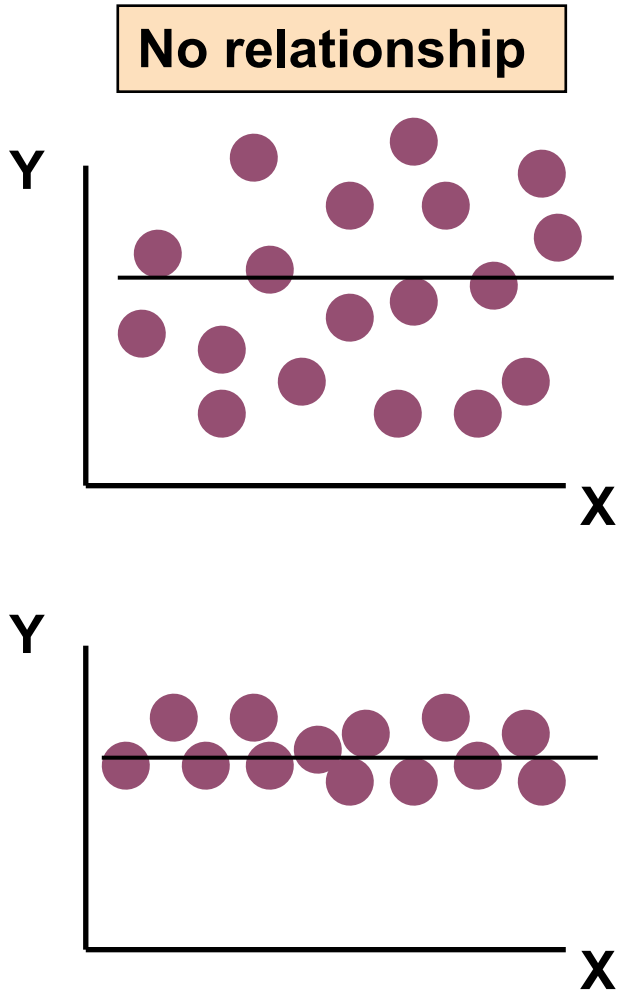


Weak relationships

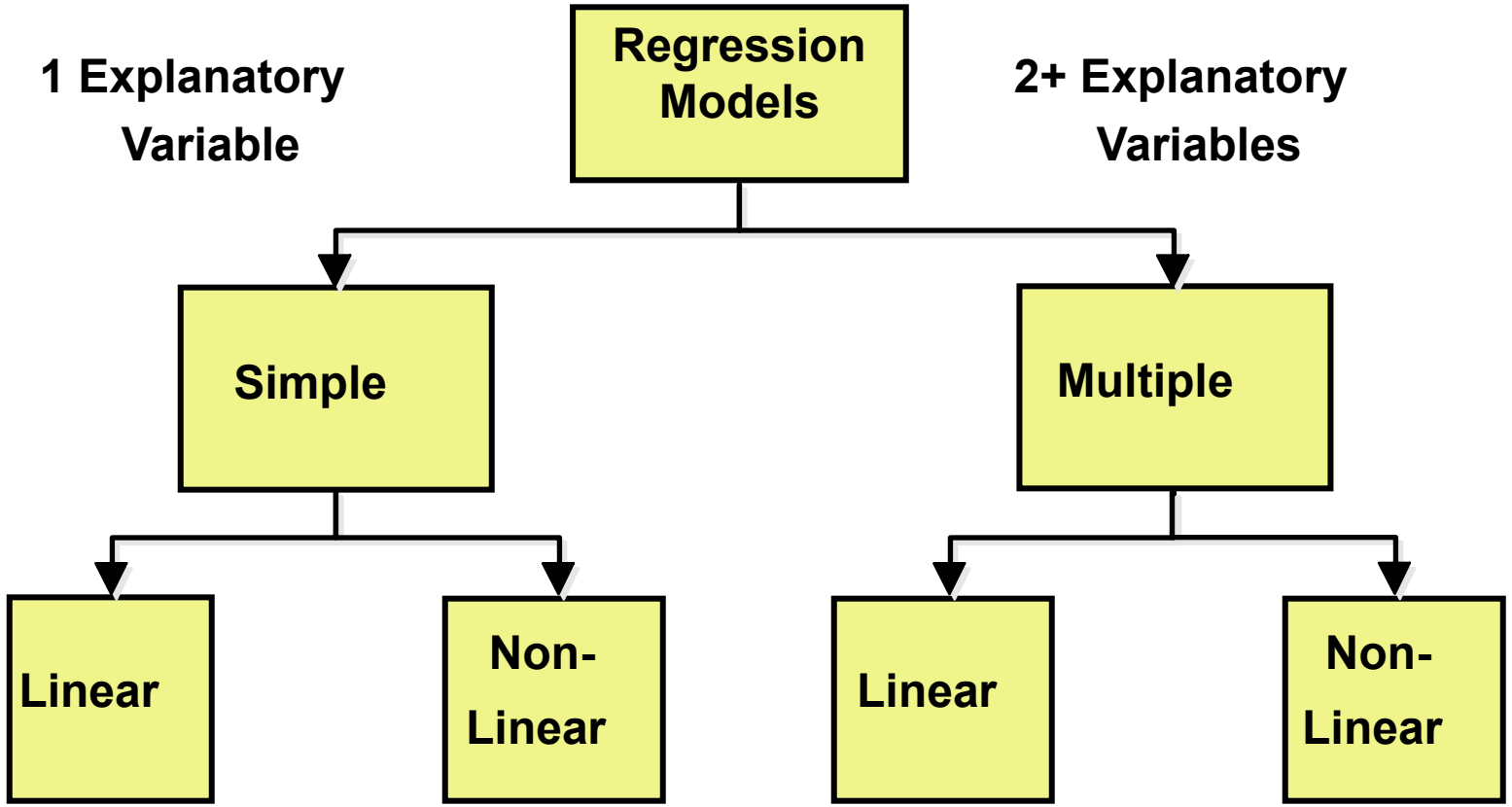




Συσχετίσεις

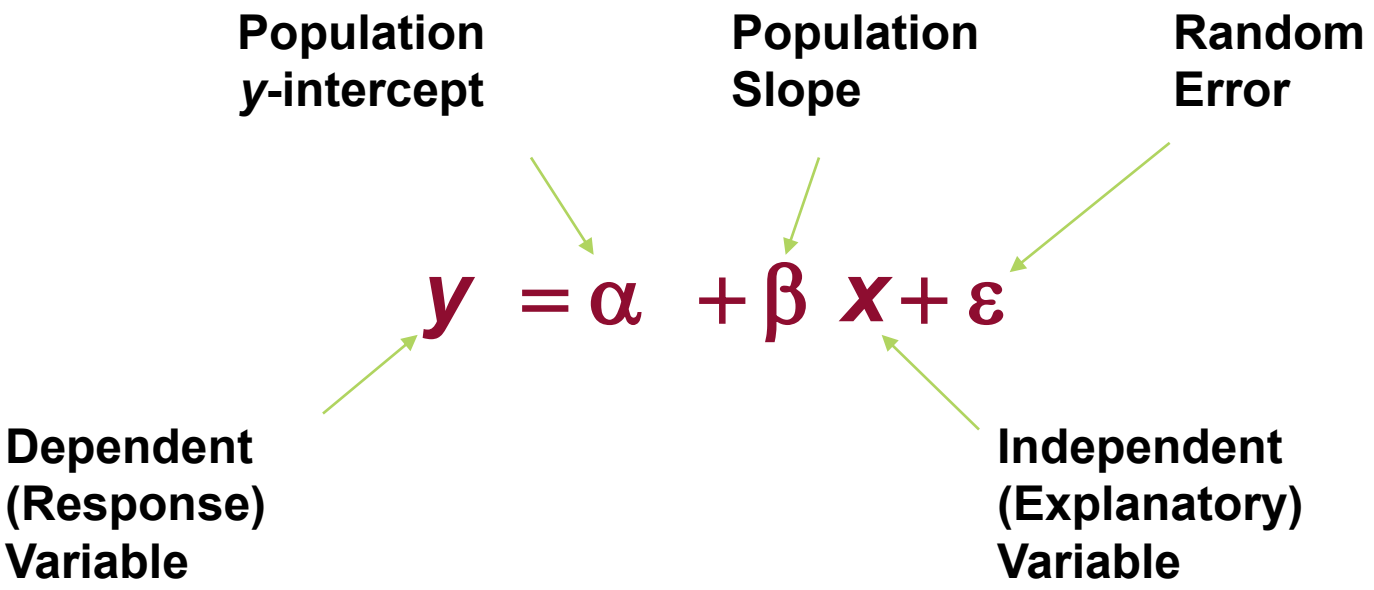


Τύποι


Regression Models
1 Explanatory Variable
2+ Explanatory Variables
Simple
Multiple
Linear
Non-Linear
Linear
Non-Linear

Μοντέλο

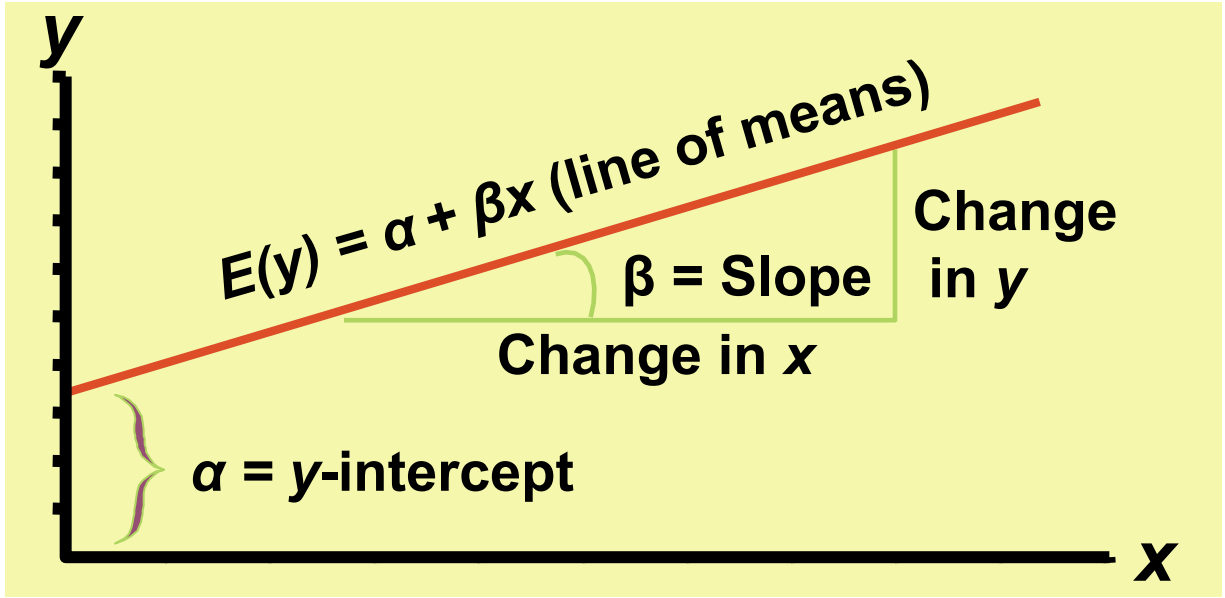
- Η σχέση μεταξύ των μεταβλητών είναι μια γραμμική συνάρτηση



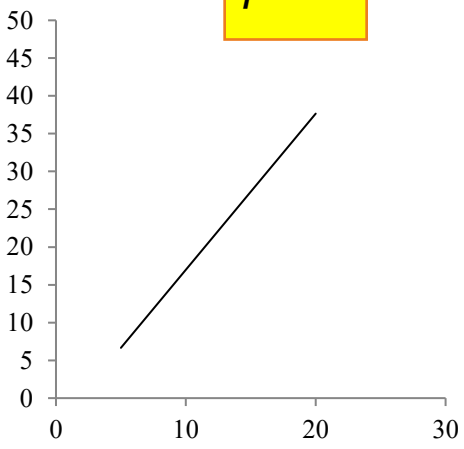
$$f(X) = a + \beta \cdot X$$

Μια ευθεία γραμμή

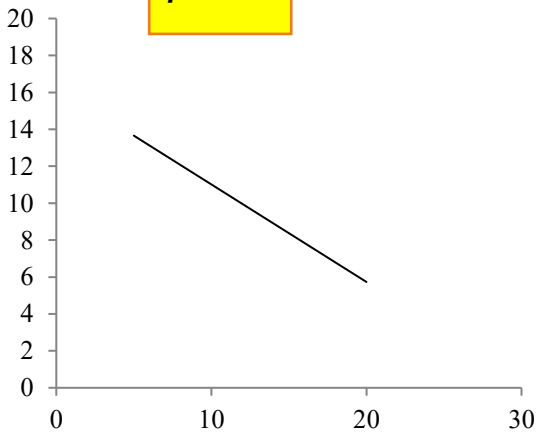
Μοντέλο



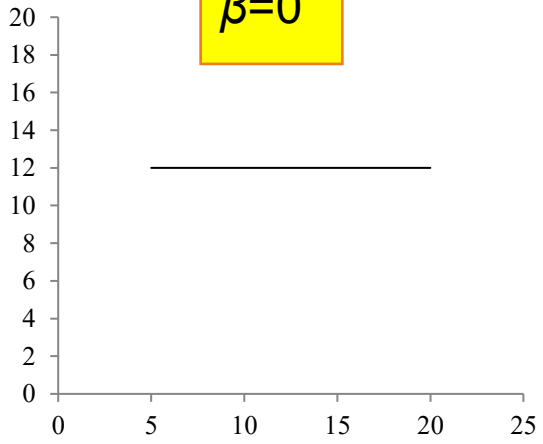
β>0



β<0

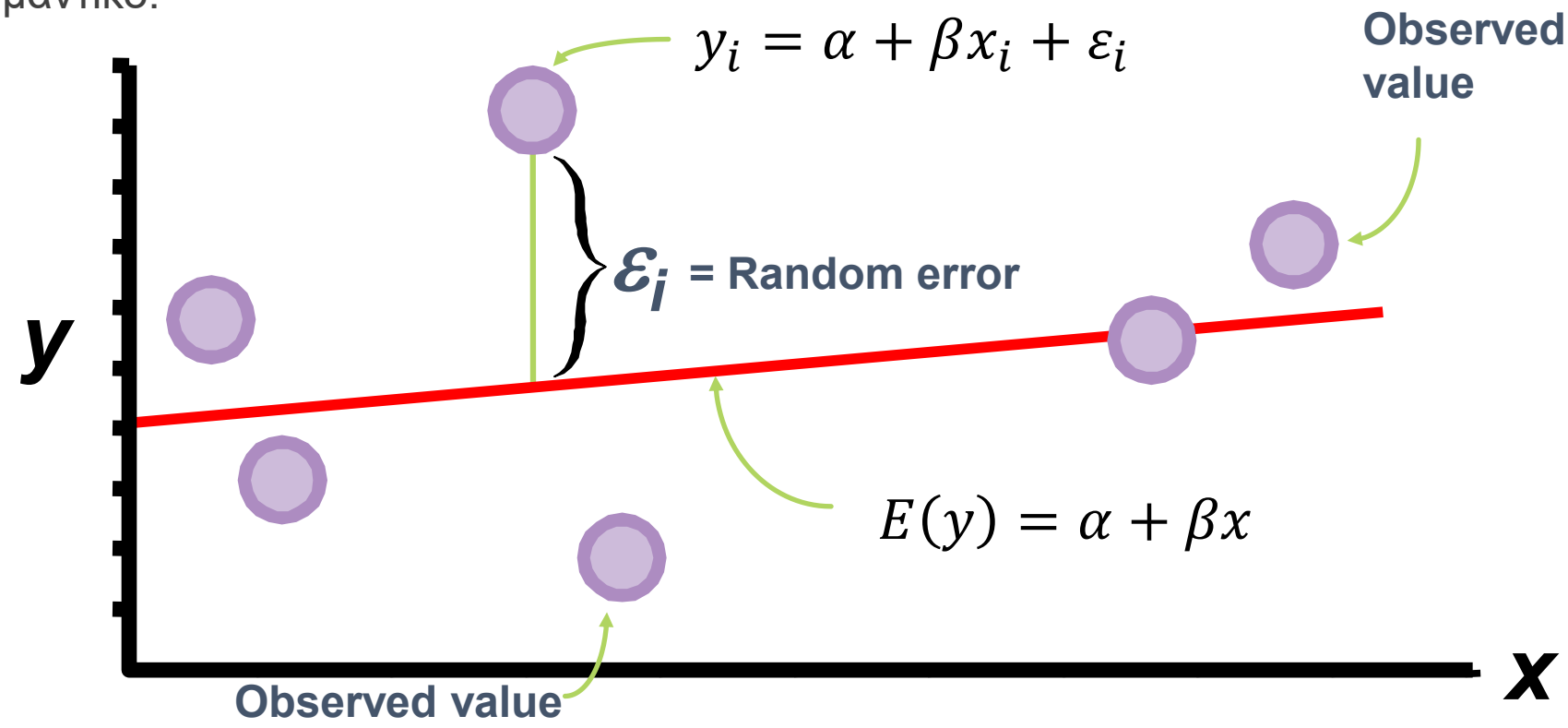


β=0



Μοντέλο

- α είναι το σημείο τομής (με τον άξονα Y). Συνήθως δεν μας ενδιαφέρει πολύ.
- β είναι η κλίση (η εφαπτομένη της γωνίας που δημιουργεί η ευθεία με τον άξονα X) – λέγεται και ρυθμός μεταβολής του Y όταν μεταβάλλεται το X
- Αν η κλίση είναι 0, τότε η X δεν μπορεί να μας βοηθήσει στην εκτίμηση της Y. Το μοντέλο δεν είναι σημαντικό.



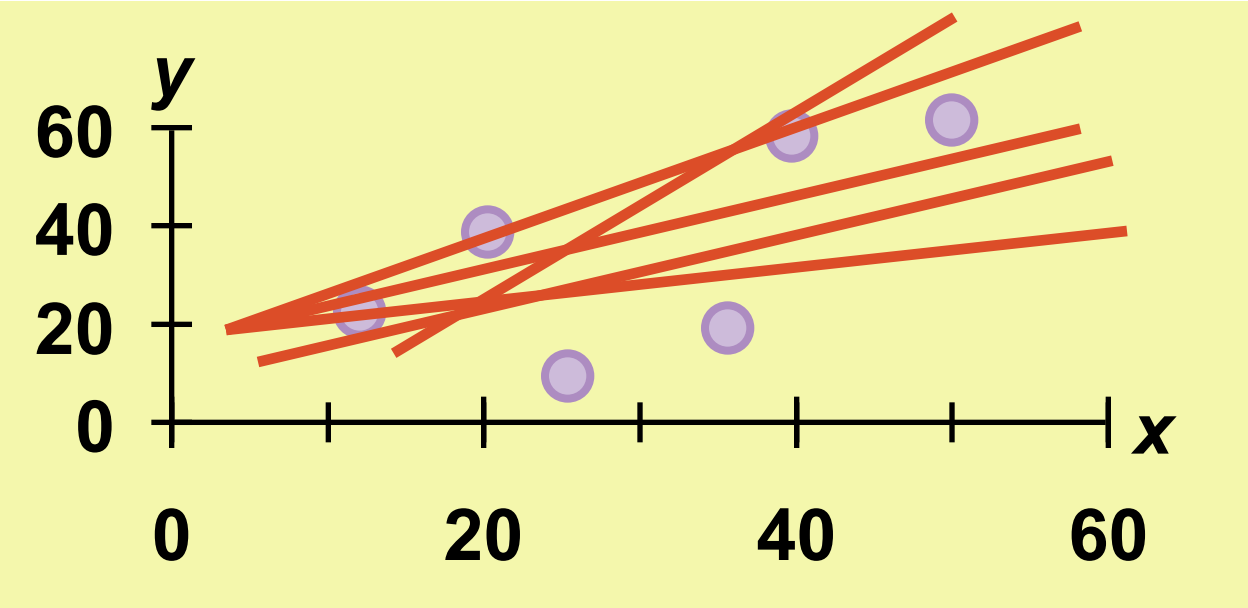
Μοντέλο

X (ανεξάρτητη)	**Y (εξαρτημένη)**
X_1	Y_1
X_2	Y_2
…	…
X_n	Y_n

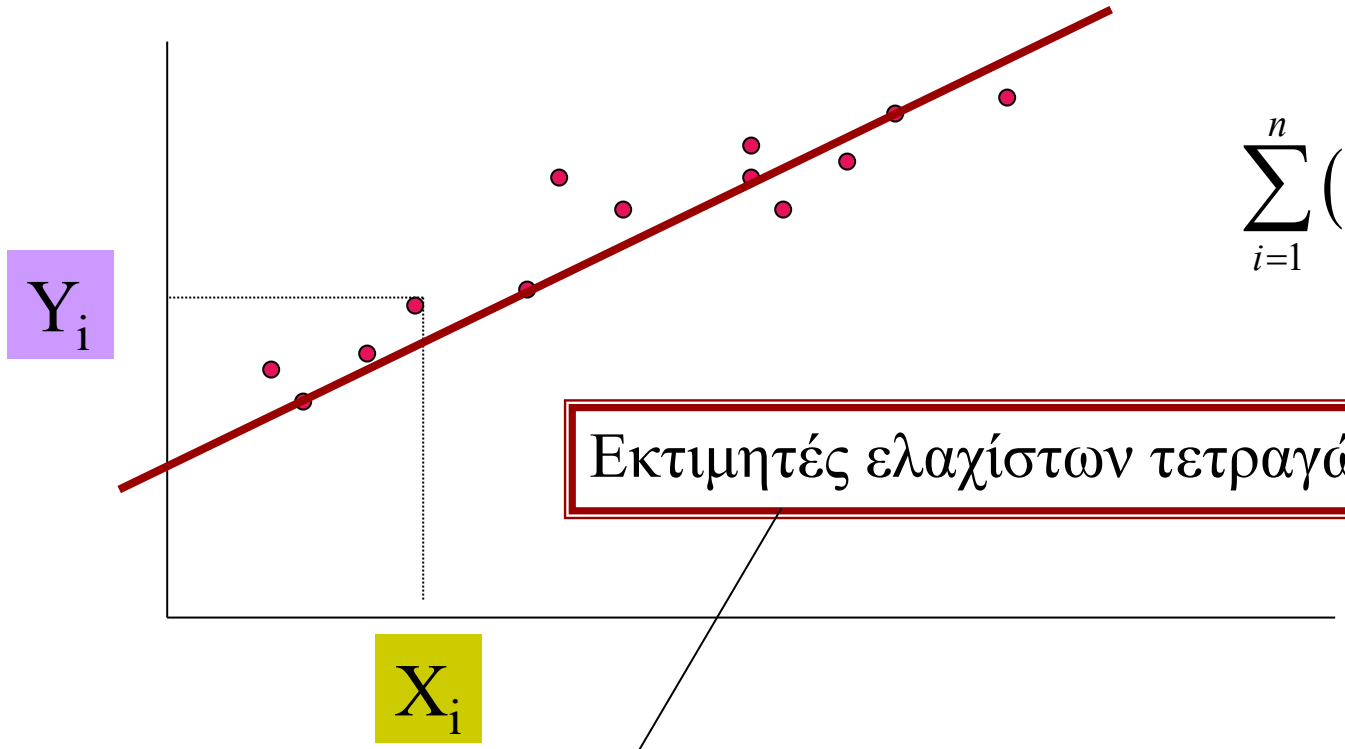
Ένα δείγμα μεγέθους n

Μοντέλο

- Πώς θα σχεδιάζατε μια γραμμή μέσα από τα σημεία;
- Πώς προσδιορίζετε ποια γραμμή «ταιριάζει καλύτερα»;



Μοντέλο



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

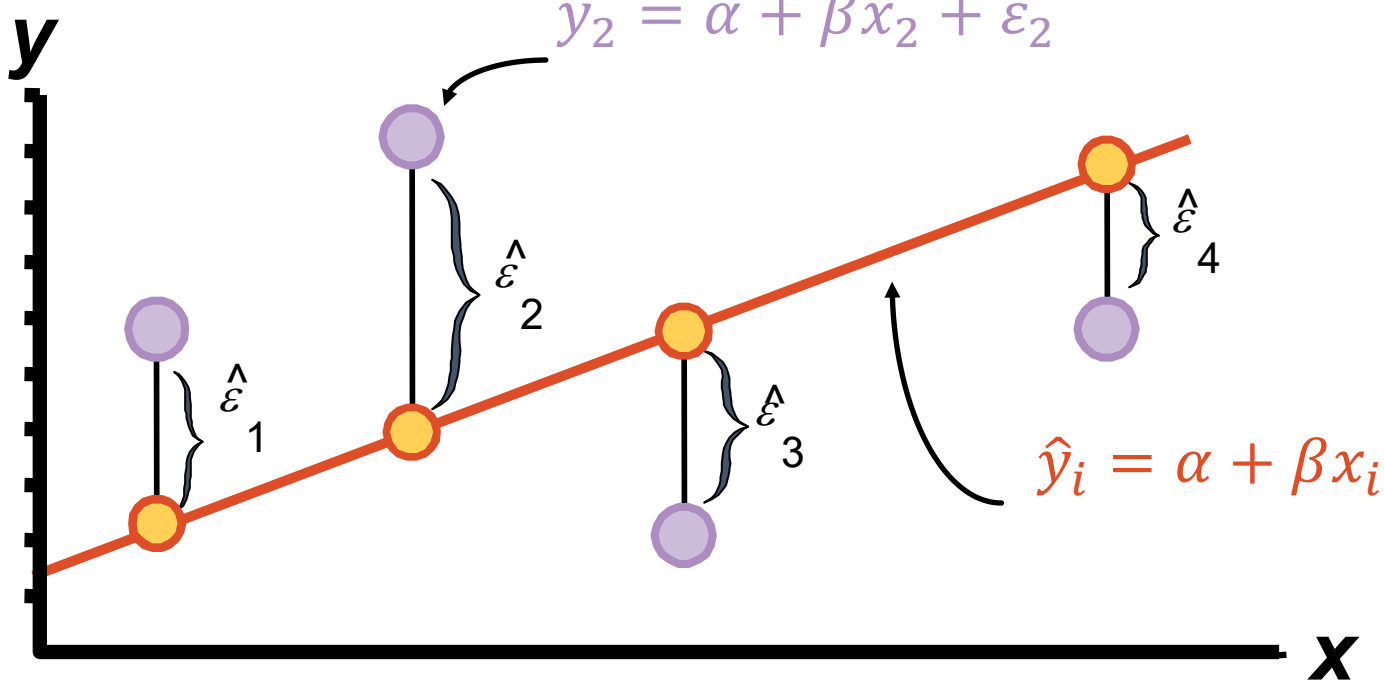
Εκτιμητές ελαχίστων τετραγώνων

$$\hat{Y}=a+b\cdot X \longleftarrow$$

Εκτιμάμε τα α και β
(με a και b αντίστοιχα)

Μοντέλο

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



Η γραμμή που προσαρμόζεται με τα ελάχιστα τεράγωνα είναι εκείνη που κάνει το άθροισμα των τετραγώνων <u>όλων αυτών των αποκλίσεων</u> όσο το δυνατό μικρότερο.

Μοντέλο

$$f\,(a,\,b) = a + b\,x,$$

$$R^2 \equiv \sum [y_i - f\,(x_i,\,a_1,\,a_2,\,\ldots,\,a_n)]^2$$

$$R^2\,(a,\,b) \equiv \sum_{i=1}^{n} [y_i - (a + b\,x_i)]^2 \qquad \frac{\partial(R^2)}{\partial a_i} = 0$$

$$\frac{\partial(R^2)}{\partial a} = -2 \sum_{i=1}^{n} [y_i - (a + b\,x_i)] = 0$$

$$\frac{\partial(R^2)}{\partial b} = -2 \sum_{i=1}^{n} [y_i - (a + b\,x_i)]\,x_i = 0.$$

Μοντέλο

$$\frac{\partial(R^2)}{\partial a} = -2\sum_{i=1}^{n}[y_i - (a + b\,x_i)] = 0$$

$$\frac{\partial(R^2)}{\partial b} = -2\sum_{i=1}^{n}[y_i - (a + b\,x_i)]\,x_i = 0.$$

$$n\,a + b\sum_{i=1}^{n} x_i$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2$$

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i\,y_i \end{bmatrix},$$

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}^{-1}\begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i\,y_i \end{bmatrix}.$$

Μοντέλο

$$\mathsf{A} \equiv \begin{bmatrix} a & b \\ c & d \end{bmatrix},$$

$$\mathsf{A}^{-1} \equiv \frac{1}{|\mathsf{A}|} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

$$= \frac{1}{a\,d - b\,c} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i\, y_i \end{bmatrix}.$$

$$\begin{bmatrix} a \\ b \end{bmatrix} = \frac{1}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \begin{bmatrix} \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i\, y_i \\ n \sum_{i=1}^{n} x_i\, y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i \end{bmatrix},$$

Μοντέλο

$$\begin{bmatrix} a \\ b \end{bmatrix} = \frac{1}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \begin{bmatrix} \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i \\ n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i \end{bmatrix},$$

$$\begin{aligned} a &= \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \\ &= \frac{\bar{y}\left(\sum_{i=1}^{n} x_i^2\right) - \bar{x}\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \\ b &= \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \\ &= \frac{(\sum_{i=1}^{n} x_i y_i) - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \end{aligned}$$

Παράδειγμα

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
x_1	y_1	x_1^2	y_1^2	$x_1 y_1$
x_2	y_2	x_2^2	y_2^2	$x_2 y_2$
:	:	:	:	:
x_n	y_n	x_n^2	y_n^2	$x_n y_n$
Σx_i	Σy_i	Σx_i^2	Σy_i^2	$\Sigma x_i y_i$

Παράδειγμα

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	1	1	1	1
2	1	4	1	2
3	2	9	4	6
4	2	16	4	8
5	4	25	16	20
15	10	55	26	37

Παράδειγμα

$$\beta = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

$$= \frac{37 - \frac{(15)(10)}{5}}{55 - \frac{(15)^2}{5}} = .70$$

$$\alpha = \bar{y} - \hat{\beta}_1 \bar{x} = 2 - (.70)(3) = -.10$$

$$\hat{y} = -.1 + .7x$$

Μοντέλο

Ο γενικότερος τύπος γραμμικού μοντέλου είναι :

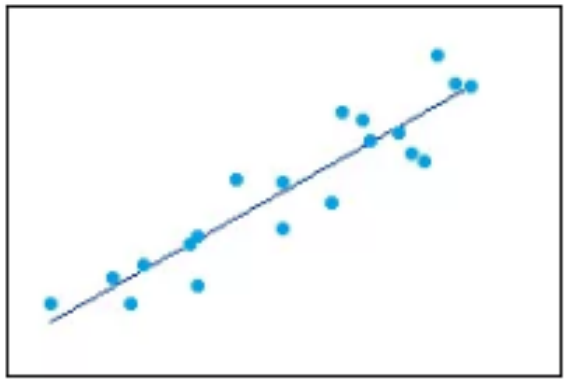
$Y=\beta_0+\beta_1 Z_1+\beta_2 Z_2+\beta_3 Z_3+....+\beta_p Z_p+\varepsilon$ (1),

όπου ε είναι η προσαύξηση μέσω της οποίας κάποια παρατήρηση Y πέφτει έξω από τη γραμμή της παλινδρόμησης.

- Αν p=1 και $Z_1=X$ η (1) γίνεται : $Y=\beta_0+\beta_1 X+\varepsilon$ ($1^{ης}$ τάξης)
- Αν p=2, $Z_1=X$ και $Z_2=X^2$ η (1) γίνεται : $Y=\beta_0+\beta_1 X+\beta_2 X^{\mathbf{2}}+\varepsilon$ ($2^{ης}$ τάξης)
- Αν p=3, $Z_1=X$, $Z_2=X^2$ και $Z_3=X^3$ η (1) γίνεται : $Y=\beta_0+\beta_1 X+\beta_2 X^{2+}\beta_3 X^{\mathbf{3}}+\varepsilon$ ($3^{ης}$ τάξης)

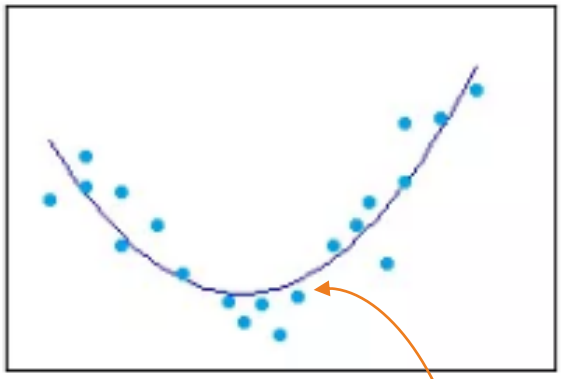
Μοντέλο

Linear



$Y=\beta_0+\beta_1X+\varepsilon$

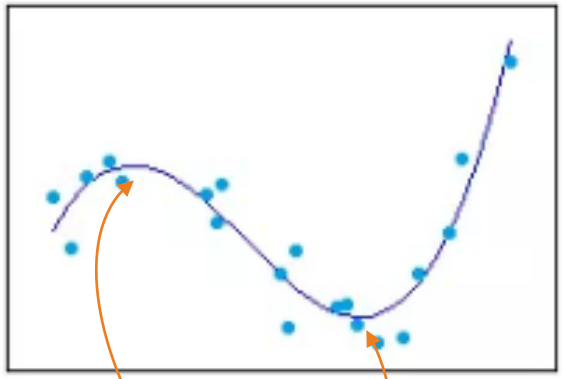
Quadratic



$Y=\beta_0+\beta_1X+\beta_2X^2+\varepsilon$

1 ακρότατο

Cubic



2 ακρότατα

$Y=\beta_0+\beta_1X+\beta_2X^{2+}\beta_3X^3+\varepsilon$

Μοντέλο

Ένας λογαριθμικός μετασχηματισμός επιτρέπει στο γραμμικό μοντέλο να προσαρμοστεί σε καμπύλες που διαφορετικά απαιτούν μη γραμμικά μοντέλα.

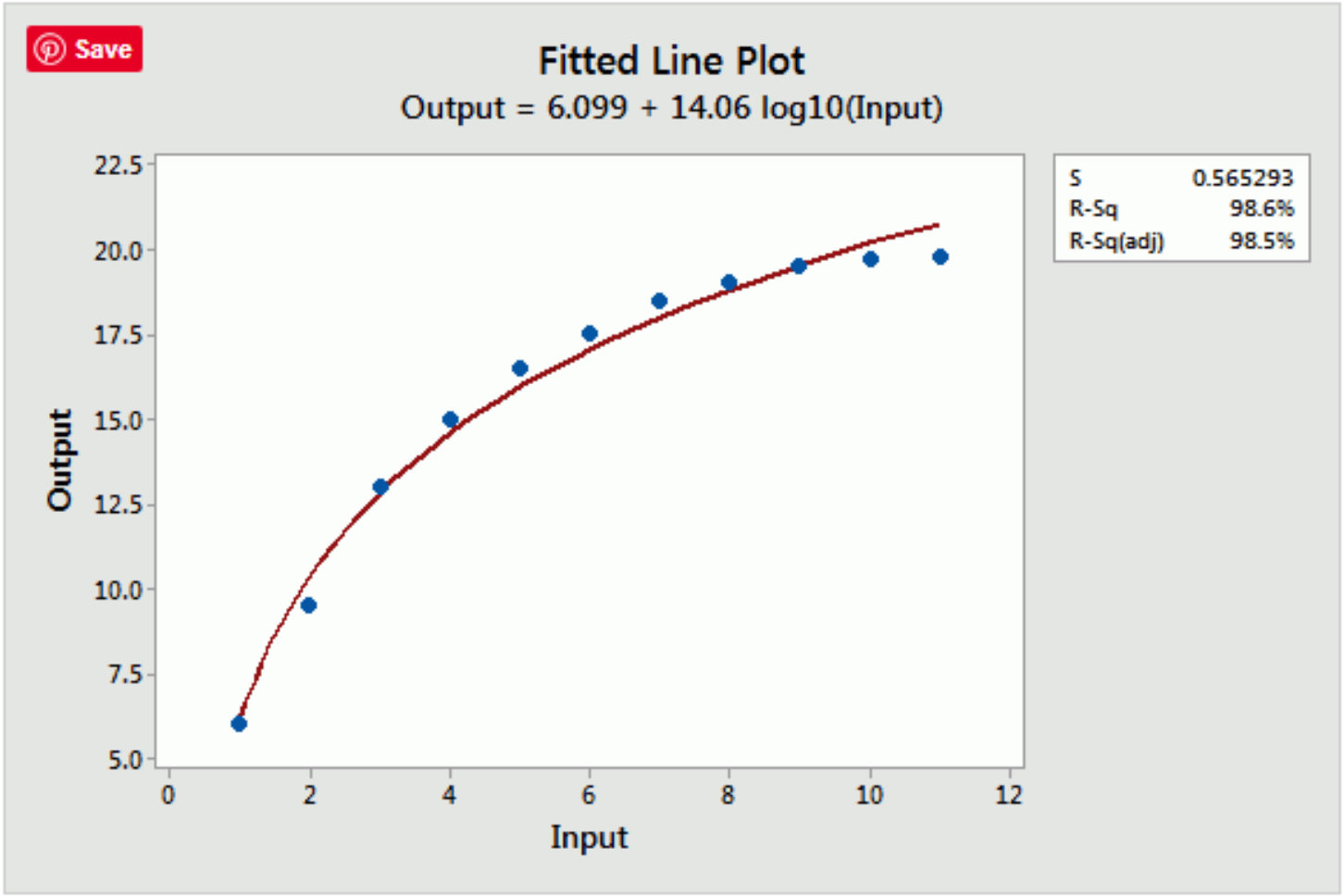
Για παράδειγμα :

$Y=\alpha X_1^{B1} X_2^{B2} \varepsilon \leftrightarrow$

$\ln Y = \ln\alpha + B_1 \ln X_1 + B_2 \ln X_2 + \ln\varepsilon$

Μοντέλο

Το πολλαπλασιαστικό μοντέλο



Μοντέλο

Το εκθετικό μοντέλο

$$Y=(e^{(\beta_0+\beta_1 X)})\cdot\varepsilon \leftrightarrow \ln Y= \beta_0+\beta_1 X+\ln\varepsilon$$

Ένα αντίστροφο μοντέλο

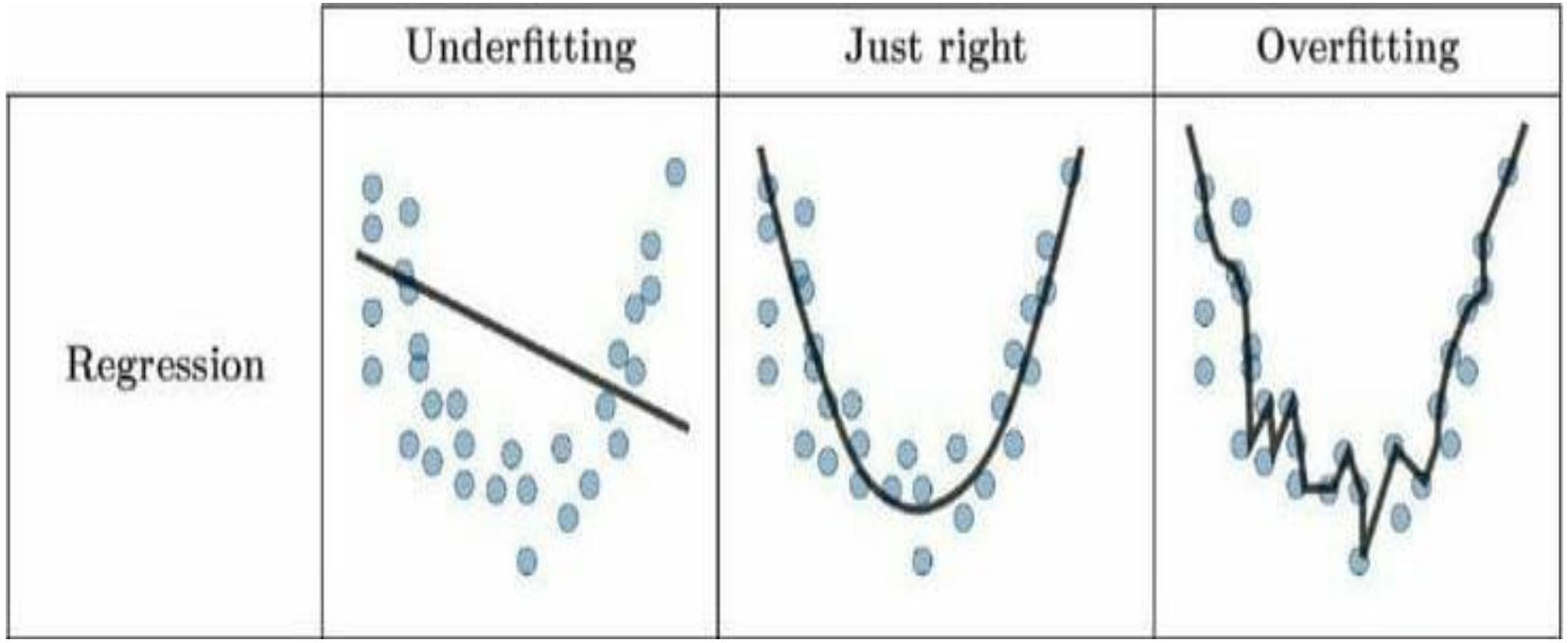
$$Y=1/(\beta_0+\beta_1 X+\varepsilon) \leftrightarrow 1/Y= \beta_0+\beta_1 X+\varepsilon$$

Σε αυτή την περίπτωση θα χρησιμοποιηθεί η αντίστροφη της εξαρτημένης μεταβλητής ως αποκριτική μεταβλητή

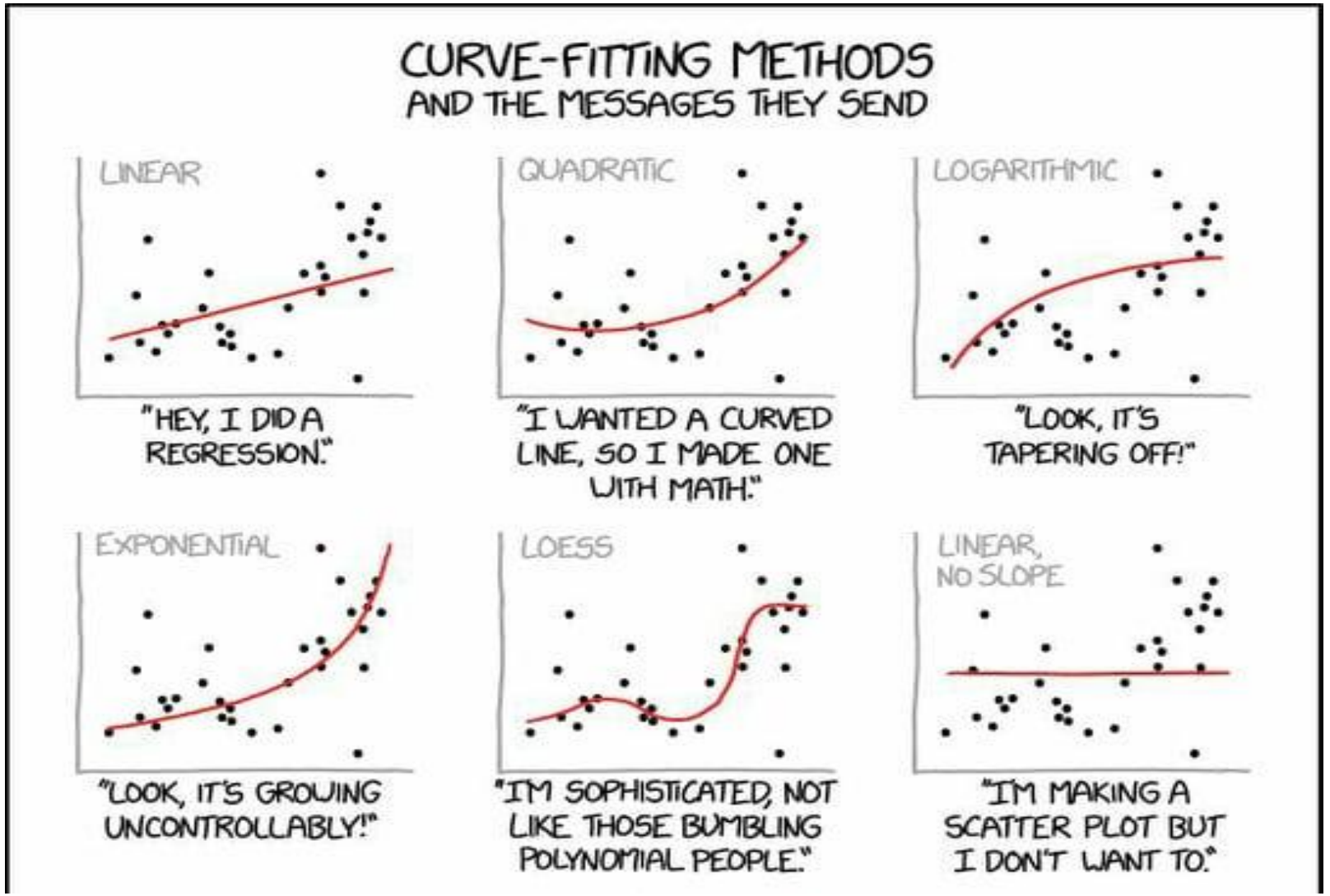
Ένα πολυπλοκότερο μοντέλο

$$Y=1/(e^{(\beta_0+\beta_1 X+\varepsilon)} \leftrightarrow \ln(1/Y-1)= \beta_0+\beta_1 X+\varepsilon$$

Μοντέλο



Μοντέλο


CURVE-FITTING METHODS
AND THE MESSAGES THEY SEND
LINEAR
"HEY, I DID A REGRESSION."
QUADRATIC
"I WANTED A CURVED LINE, SO I MADE ONE WITH MATH."
LOGARITHMIC
"LOOK, IT'S TAPERING OFF!"
EXPONENTIAL
"LOOK, IT'S GROWING UNCONTROLLABLY!"
LOESS
"I'M SOPHISTICATED, NOT LIKE THOSE BUMBLING POLYNOMIAL PEOPLE."
LINEAR, NO SLOPE
"I'M MAKING A SCATTER PLOT BUT I DON'T WANT TO."

Παράδειγμα

- Το πρώτο πράγμα που μπορούμε να διαπιστώσουμε είναι ότι
  - ο συντελεστής του X1 είναι αρνητικός, σημαίνει ότι η υψηλότερη τιμή οδηγεί σε λιγότερες πωλήσεις,
  - ο συντελεστής του X2 είναι θετικός, σημαίνει ότι το υψηλότερο εισόδημα σημαίνει μεγαλύτερες πωλήσεις.

$$Y = - 2{,}765 - 7{,}738\ X_1 + 12{,}286\ X_2$$

Μοντέλο

- Το R^2 μετρά το ποσοστό της απόκλισης της εξαρτημένης μεταβλητής το οποίο μπορεί να εξηγηθεί από την παλινδρόμηση.
- Η τιμή του R^2 κυμαίνεται πάντα από 0 έως 1 $R^2 = 1 - \frac{SSE}{SST}$

$$SSE = \Sigma \hat{u}_i^2 = \Sigma(Y_i - \hat{Y}_i)^2$$

SST: συνολική διακύμανση των Y από $\overline{Y}$

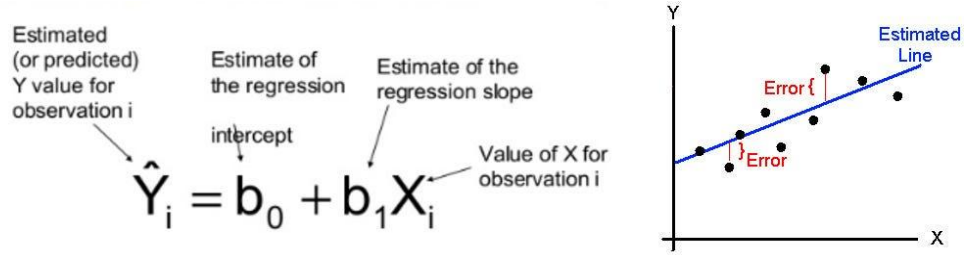
- Για να είναι αξιόπιστα τα αποτελέσματα της παλινδρόμησης θα πρέπει το πλήθος των παρατηρήσεων να είναι σημαντικά μεγαλύτερο από το πλήθος των συντελεστών που προσπαθούμε να εκτιμήσουμε.
- Έτσι, μερικές φορές συνιστάται ο υπολογισμός του προσαρμοσμένου R2:

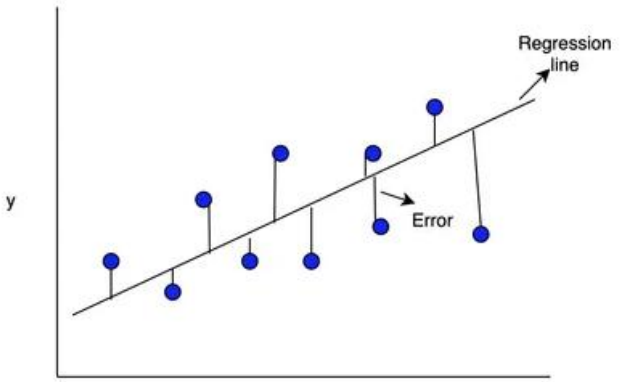
$$(\text{Προσαρμοσμένο})R^2 = 1 - \frac{(1-R^2)(n-1)}{n-m}$$

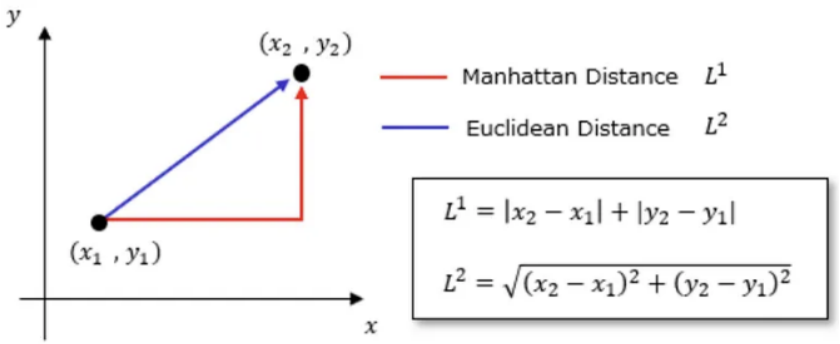
n = Το πλήθος παρατηρήσεων (x_i, y_i)
m= αριθμός των παραμέτρων

Μοντέλο

- Mean Squared Error (MSE) $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- Root Mean Squared Error (RMSE) $RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
- Mean Absolute Error (MAE) $MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$
- Mean Absolute Percentage Error (MAPE) $MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$
- Mean Absolute Scaled Error (MASE) $MASE = \frac{\frac{1}{h}\sum_{t=n+1}^{n+h}|y_t - \hat{y}_t|}{\frac{1}{n-1}\sum_{t=2}^{n}|y_t - y_{t-1}|}$







- From a geometric perspective, RMSE and MAE represent mean forms of the L2 and L1 norms, which correspond to the Euclidean distance and the Manhattan distance, respectively

Regularization

- λ (lambda): A hyperparameter that controls the strength of the penalty. A larger λ means more regularization

$$Loss = Original\ Loss + \lambda \times Regularization\ Term$$

- The most common types of regularization are L1 regularization and L2 regularization
- They differ in how they penalize the model's weights
- **L1 Regularization (Lasso Regression)**
  - Not all of features are important for predicting the output
  - L1 regularization can help the model to focus on the most significant features by reducing the weights of less important ones to zero

$$Loss = Original\ Loss + \lambda \sum_i |W_i|$$

- **L2 Regularization (Ridge Regression)**
  - L2 regularization helps ensure that the model doesn't assign too much importance to any one feature and considers all of them in a balanced way

$$Loss = Original\ Loss + \lambda \sum_i W_i^2$$

Regularization

- **Elastic Net Regression**
  - It is a combination of both L1 as well as L2 regularization
  - We add the absolute norm of the weights as well as the squared measure of the weights
  - With the help of an extra hyperparameter that controls the ratio of the L1 and L2 regularization

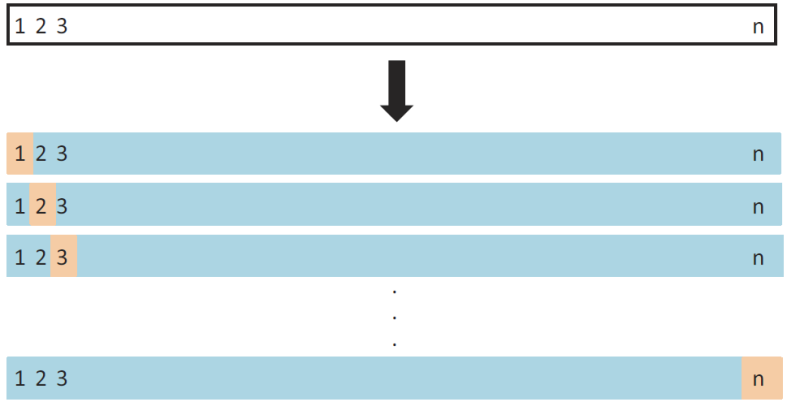
$$\text{Cost} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\left((1-\alpha)\sum_{i=1}^{m}|w_i| + \alpha\sum_{i=1}^{m}w_i^2\right)$$

Cross Validation

Cross Validation

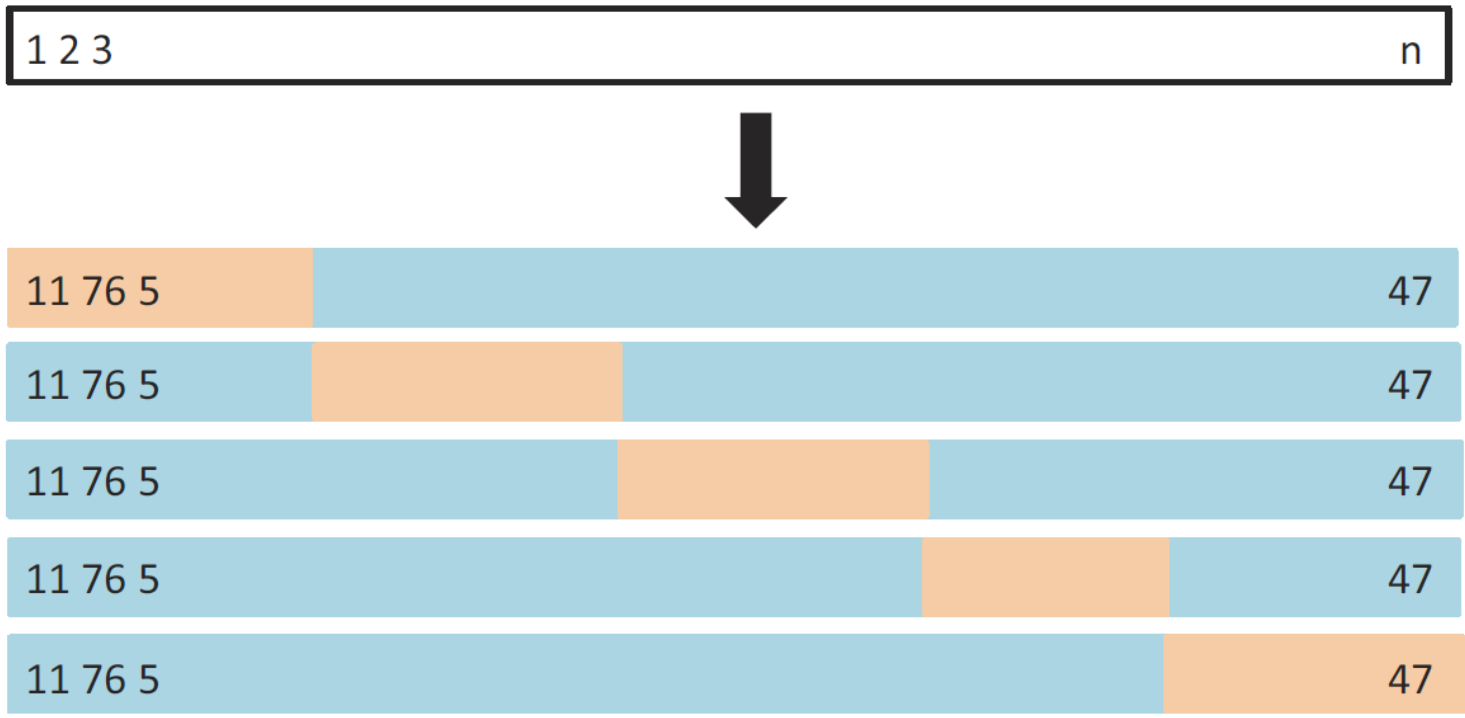
- Leave-One-Out Cross-Validation (LOOCV)
  - Like the validation set approach, LOOCV involves splitting the set of observations into two parts
  - However, instead of creating two subsets of comparable size, a single observation (x1, y1) is used for the validation set, and the remaining observations {(x2, y2), . . . , (xn, yn)} make up the training set
  - The statistical learning method is fit on the n − 1 training observations, and a prediction is made for the excluded observation, using its value x1
  - We can repeat the procedure by selecting (x2, y2) for the validation data, training the statistical learning procedure on the n − 1 observations and record the next MSE
  - The LOOCV estimate for the test MSE is the average of these n test error estimates

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{MSE}_i$$



Cross Validation

- A schematic display of 5-fold CV – A set of n observations is randomly split into five non-overlapping groups



Cross Validation

- k-Fold Cross-Validation
  - This approach involves randomly dividing the set of observations into k groups, or folds, of approximately equal size
  - The first fold is treated as a validation set, and the method is fit on the remaining k − 1 folds
  - This procedure is repeated k times; each time, a different group of observations is treated as a validation set
  - The most obvious advantage is computational

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{MSE}_i$$

Bayesian Learning

Εισαγωγή

- Ένας στατιστικός ταξινομητής: εκτελεί πιθανολογική πρόβλεψη, δηλ. προβλέπει πιθανότητες συμμετοχής στην κλάση
- Θεμέλιο: Βασισμένο στο θεώρημα του Bayes.
- Απόδοση: Ένας απλός ταξινομητής έχει συγκρίσιμη απόδοση με δέντρο αποφάσεων και επιλεγμένους ταξινομητές νευρωνικών δικτύων
- Αύξηση: Κάθε παράδειγμα εκπαίδευσης μπορεί σταδιακά να αυξήσει/μειώσει την πιθανότητα να είναι σωστή μια υπόθεση — η προηγούμενη γνώση μπορεί να συνδυαστεί με τα παρατηρούμενα δεδομένα
- Πρότυπο: Ακόμη και όταν οι μέθοδοι Bayes είναι υπολογιστικά δυσεπίλυτες, μπορούν να παρέχουν ένα πρότυπο βέλτιστης λήψης αποφάσεων έναντι του οποίου μπορούν να μετρηθούν άλλες μέθοδοι

Μοντέλο

- Μια καλή στρατηγική είναι να προβλέψετε:

$$\arg\max_{Y} P(Y|X_1, \ldots, X_n)$$

(για παράδειγμα: ποια είναι η πιθανότητα μια εικόνα να αντιπροσωπεύει τον αριθμό 5 με δεδομένα τα pixel της;)

Μοντέλο

- Θεώρημα συνολικής πιθανότητας: $P(B) = \sum_{i=1}^{M} P(B|A_i)P(A_i)$
- Θεώρημα Bayes: $P(H|\mathbf{X}) = \frac{P(\mathbf{X}|H)P(H)}{P(\mathbf{X})} = P(\mathbf{X}|H) \times P(H) / P(\mathbf{X})$
  - Έστω X ένα δείγμα δεδομένων («απόδειξη»): η ετικέτα κλάσης είναι άγνωστη
  - Έστω H μια υπόθεση ότι το X ανήκει στην κατηγορία C
  - Η κατηγοριοποίηση είναι για τον προσδιορισμό του P(H|X), (δηλαδή, η εκ των υστέρων πιθανότητα - posteriori probability): η πιθανότητα να ισχύει η υπόθεση δεδομένου του παρατηρούμενου δείγματος δεδομένων X
  - P(H) (prior probability - προηγούμενη πιθανότητα): η αρχική πιθανότητα
  - Για παράδειγμα, ο X θα αγοράσει υπολογιστή, ανεξαρτήτως ηλικίας, εισοδήματος,…
  - P(X): πιθανότητα να παρατηρηθούν δεδομένα δείγματος
  - P(X|H) (πιθανότητα): η πιθανότητα παρατήρησης του δείγματος X, δεδομένου ότι ισχύει η υπόθεση
  - Π.χ., δεδομένου ότι ο X θα αγοράσει υπολογιστή, η πιθανότητα. ότι το X είναι 31..40, μεσαίο εισόδημα

Μοντέλο

- Δεδομένων των δεδομένων εκπαίδευσης X, η εκ των υστέρων πιθανότητα μιας υπόθεσης H, P(H|X), ακολουθεί το θεώρημα του Bayes

$$P(H|\mathbf{X}) = \frac{P(\mathbf{X}|H)P(H)}{P(\mathbf{X})} = P(\mathbf{X}|H) \times P(H) / P(\mathbf{X})$$

- Ανεπίσημα, αυτό μπορεί να θεωρηθεί ως

posteriori = likelihood x prior/evidence

- Προβλέπει ότι το X ανήκει στο C_i αν η πιθανότητα $P(C_i|X)$ είναι η υψηλότερη μεταξύ όλων των $P(C_k|X)$ για όλες τις κλάσεις k
- Πρακτική δυσκολία: Απαιτεί αρχική γνώση πολλών πιθανοτήτων, που συνεπάγονται σημαντικό υπολογιστικό κόστος

Μοντέλο

- Έστω D ένα σύνολο πλειάδων εκπαίδευσης και των συσχετιζόμενων ετικετών κλάσης τους, και κάθε πλειάδα αντιπροσωπεύεται από ένα διάνυσμα χαρακτηριστικών n-D X = (x_1, x_2, ..., x_n)
- Ας υποθέσουμε ότι υπάρχουν m κατηγορίες C_1, C_2, …, C_m.
- Η ταξινόμηση είναι να εξαχθεί το μέγιστο εκ των υστέρων, δηλαδή το μέγιστο $P(C_i|X)$
- Αυτό μπορεί να εξαχθεί από το θεώρημα του Bayes

$$P(C_i|\mathbf{X}) = \frac{P(\mathbf{X}|C_i)P(C_i)}{P(\mathbf{X})}$$

- Επειδή το P(X) είναι σταθερό για όλες τις κλάσεις, πρέπει να μεγιστοποιηθεί μόνο το

$$P(C_i|\mathbf{X}) = P(\mathbf{X}|C_i)P(C_i)$$

Μοντέλο

- Μια απλοποιημένη υπόθεση: τα χαρακτηριστικά είναι υπό όρους ανεξάρτητα (δηλαδή, δεν υπάρχει σχέση εξάρτησης μεταξύ των χαρακτηριστικών):

$$P(\mathbf{X}\,|\,C_i) = \prod_{k=1}^{n} P(x_k\,|\,C_i) = P(x_1\,|\,C_i) \times P(x_2\,|\,C_i) \times \ldots \times P(x_n\,|\,C_i)$$

- Αυτό μειώνει σημαντικά το κόστος υπολογισμού: Μετρά μόνο την κατανομή κλάσης

- Εάν το A_k είναι κατηγορηματικό, το $P(x_k|C_i)$ είναι το # των πλειάδων στο C_i που έχουν τιμή x_k για το A_k διαιρούμενο με $|C_{i,\,D}|$ (# από τις πλειάδες του C_i στο D)
- Εάν το Ak έχει συνεχή τιμή, το $P(x_k|C_i)$ υπολογίζεται συνήθως με βάση την κατανομή Gauss με μέσο μ και τυπική απόκλιση και $P(x_k|C_i)$ είναι

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$P(\mathbf{X}\,|\,C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i})$$

Παράδειγμα

Class:

C1:buys_computer = ‘yes’

C2:buys_computer = ‘no’

Data to be classified:

X = (age <=30,

Income = medium,

Student = yes

Credit_rating = Fair)

age	income	student	credit_rating	buys_comp
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Παράδειγμα

age	income	student	credit_rating	buys_comp
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

$P(C_i)$: P(buys_computer = “yes”) = 9/14 = 0.643

P(buys_computer = “no”) = 5/14= 0.357

Compute $P(X|C_i)$ for each class

P(age = “<=30” | buys_computer = “yes”) = 2/9 = 0.222

P(age = “<= 30” | buys_computer = “no”) = 3/5 = 0.6

P(income = “medium” | buys_computer = “yes”) = 4/9 = 0.444

P(income = “medium” | buys_computer = “no”) = 2/5 = 0.4

P(student = “yes” | buys_computer = “yes) = 6/9 = 0.667

P(student = “yes” | buys_computer = “no”) = 1/5 = 0.2

P(credit_rating = “fair” | buys_computer = “yes”) = 6/9 = 0.667

P(credit_rating = “fair” | buys_computer = “no”) = 2/5 = 0.4

X = (age <= 30 , income = medium, student = yes, credit_rating = fair)

$P(X|C_i)$: P(X|buys_computer = “yes”) = 0.222 x 0.444 x 0.667 x 0.667 = 0.044

P(X|buys_computer = “no”) = 0.6 x 0.4 x 0.2 x 0.4 = 0.019

$P(X|C_i)*P(C_i)$: P(X|buys_computer = “yes”) * P(buys_computer = “yes”) = 0.028

P(X|buys_computer = “no”) * P(buys_computer = “no”) = 0.007

Therefore, X belongs to class (“buys_computer = yes”)

Υπολογισμοί

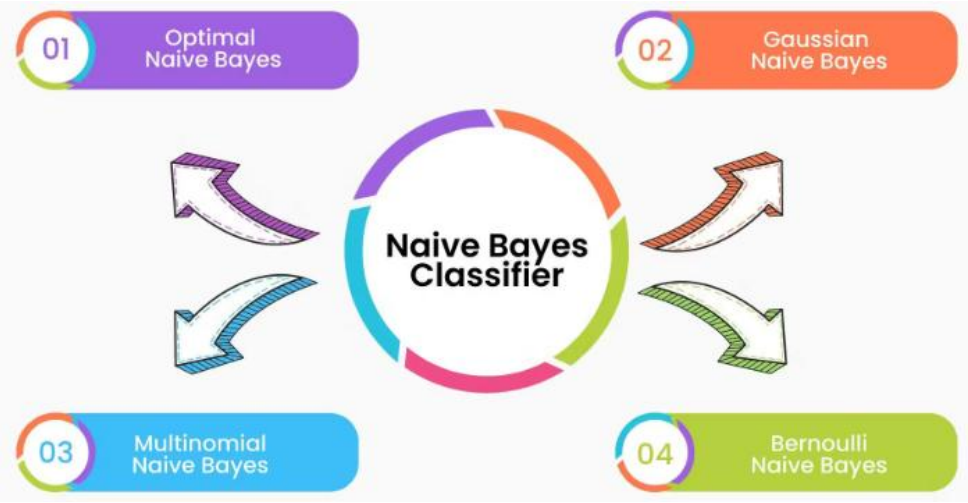
- Η αφελής μπεϋζιανή πρόβλεψη (Naïve Bayesian Classifier) απαιτεί η κάθε υπό όρους πιθανότητα να είναι μη μηδενική.
- Διαφορετικά, η προβλεπόμενη πιθανότητα θα είναι μηδέν
- Παράδειγμα: Ας υποθέσουμε ένα σύνολο δεδομένων με 1000 πλειάδες, income=low (0), income= medium (990), and income = high (10)
- Χρήση Λαπλασιανής διόρθωσης (ή Λαπλάσιας εκτίμησης - **Laplacian correction**)
- Προσθέτοντας 1 σε κάθε περίπτωση
  - **Prob(income = low) = 1/1003**
  - **Prob(income = medium) = 991/1003**
  - **Prob(income = high) = 11/1003**
- Η «διορθωμένη» πιθανότητα δίνει εκτιμήσεις που είναι κοντά στις «μη διορθωμένες» αντίστοιχές τους

Σχόλια

- Πλεονεκτήματα
    - Εύκολο στην εφαρμογή
    - Καλά αποτελέσματα που λαμβάνονται στις περισσότερες περιπτώσεις
- Μειονεκτήματα
    - Υπόθεση: ανεξαρτησία μεταξύ των κλάσεων, επομένως απώλεια ακρίβειας
    - Πρακτικά, υπάρχουν εξαρτήσεις μεταξύ των μεταβλητών
        - Π.χ. νοσοκομεία: ασθενείς: Προφίλ: ηλικία, οικογενειακό ιστορικό κ.λπ.
        - Συμπτώματα: πυρετός, βήχας κ.λπ.,
        - Ασθένεια: καρκίνος του πνεύμονα, διαβήτης κ.λπ.
    - Οι εξαρτήσεις μεταξύ αυτών δεν μπορούν να μοντελοποιηθούν από τον ταξινομητή Naïve Bayes

Κατηγορίες

- **Gaussian Naive Bayes**: Είναι ένας απλός αλγόριθμος που χρησιμοποιείται όταν τα χαρακτηριστικά είναι συνεχή. Τα χαρακτηριστικά που υπάρχουν στα δεδομένα πρέπει να ακολουθούν τον κανόνα της κατανομής Gauss. Επιταχύνει αξιοσημείωτα την αναζήτηση και υπό συνθήκες, το σφάλμα θα είναι δύο φορές μεγαλύτερο από το Optimal Naive Bayes.
- **Optimal Naive Bayes**: Ο Optimal Naive Bayes επιλέγει την τάξη που έχει τη μεγαλύτερη πιθανότητα να συμβεί. Σύμφωνα με το όνομα, είναι βέλτιστο. Αλλά θα περάσει από όλες τις δυνατότητες, κάτι που είναι πολύ αργό και χρονοβόρο.
- **Bernoulli Naive Bayes**: Ο Bernoulli Naive Bayes είναι ένας αλγόριθμος που είναι χρήσιμος για δεδομένα που έχουν δυαδικά ή δυαδικά χαρακτηριστικά.
- **Πολυωνυμικό Naive Bayes**: Τα χαρακτηριστικά που απαιτούνται για αυτόν τον τύπο είναι η συχνότητα των λέξεων που μετατρέπονται από το έγγραφο.



Παράδειγμα

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

	outlook	temperature	humidity	windy	play
1	sunny	hot	high	false	no
2	sunny	hot	high	true	no
3	overcast	hot	high	false	yes
4	rainy	mild	high	false	yes
5	rainy	cool	normal	false	yes
6	rainy	cool	normal	true	no
7	overcast	cool	normal	true	yes
8	sunny	mild	high	false	no
9	sunny	cool	normal	false	yes
10	rainy	mild	normal	false	yes
11	sunny	mild	normal	true	yes
12	overcast	mild	high	true	yes
13	overcast	hot	normal	false	yes
14	rainy	mild	high	true	no

(outlook=sunny), (temperature=cool), (humidity=high) and (windy=true)

P(outlook=sunny | play=yes) = 2/9
P(play=yes) = 9/14

$$\Pr[yes \mid E] = \Pr[Outlook = Sunny \mid yes] \times \Pr[Temperature = Cool \mid yes] \times \Pr[Humidity = High \mid yes] \times \Pr[Windy = True \mid yes] \times \frac{\Pr[yes]}{\Pr[E]}$$

Probability of class "yes"

$$= \frac{\frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14}}{\Pr[E]}$$

Likelihood of the two classes

For "yes" = 2/9 × 3/9 × 3/9 × 3/9 × 9/14 = 0.0053

For "no" = 3/5 × 1/5 × 4/5 × 3/5 × 5/14 = 0.0206

Conversion into a probability by normalization:

P("yes") = 0.0053 / (0.0053 + 0.0206) = 0.205

P("no") = 0.0206 / (0.0053 + 0.0206) = 0.795