

Στατιστική II

Γιώργος Τσιρογιάννης

Τμήμα Διοίκησης Επιχειρήσεων Αγροτικών
Προϊόντων και Τροφίμων,
Πανεπιστήμιο Πατρών



Διάλεξη 3η

Παλινδρόμηση



7^ο κεφάλαιο

Από την συσχέτιση στην παλινδρόμηση

- Όταν έχουμε εντοπίσει δύο μεταβλητές που σχετίζονται, μπορούμε να προβλέψουμε την μία από την άλλη;
- Πχ μπορούμε από τις ώρες διαβάσματος ενός φοιτητή, να προβλέψουμε την βαθμολογία του;
- Πχ μπορούμε από το ύψος στην παιδική ηλικία να προβλέψουμε το ύψος στην ενήλικη ζωή;

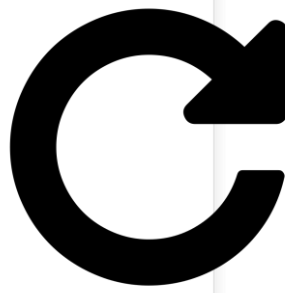
Το πρόβλημα των ευχετηρίων καρτών



Κάρτες

	Έστειλε	Έλαβε
Αντρι	5	10
Μαικ	7	12
Ντορις	13	14
Στιβ	9	18
Τζον	1	6

Παράδειγμα ερώτησης που ψάχνουμε απάντηση:
Αν η Έμα αποστείλει 11 κάρτες πόσες αναμένεται να λάβει ως απάντηση;
Ή αντίστροφα:
Πόσες κάρτες πρέπει να στείλει ώστε να λάβει 21;




Γενικές απαντήσεις μέσω των συσχετίσεων

$$SP_{xy} = \sum (X - \bar{X})(Y - \bar{Y}) = \sum XY - \frac{(\sum X)(\sum Y)}{n}$$


$$r = \frac{SP_{xy}}{\sqrt{SS_x SS_y}}$$

$$SS_x = \sum (X - \bar{X})^2 = \sum X^2 - \frac{(\sum X)^2}{n}$$

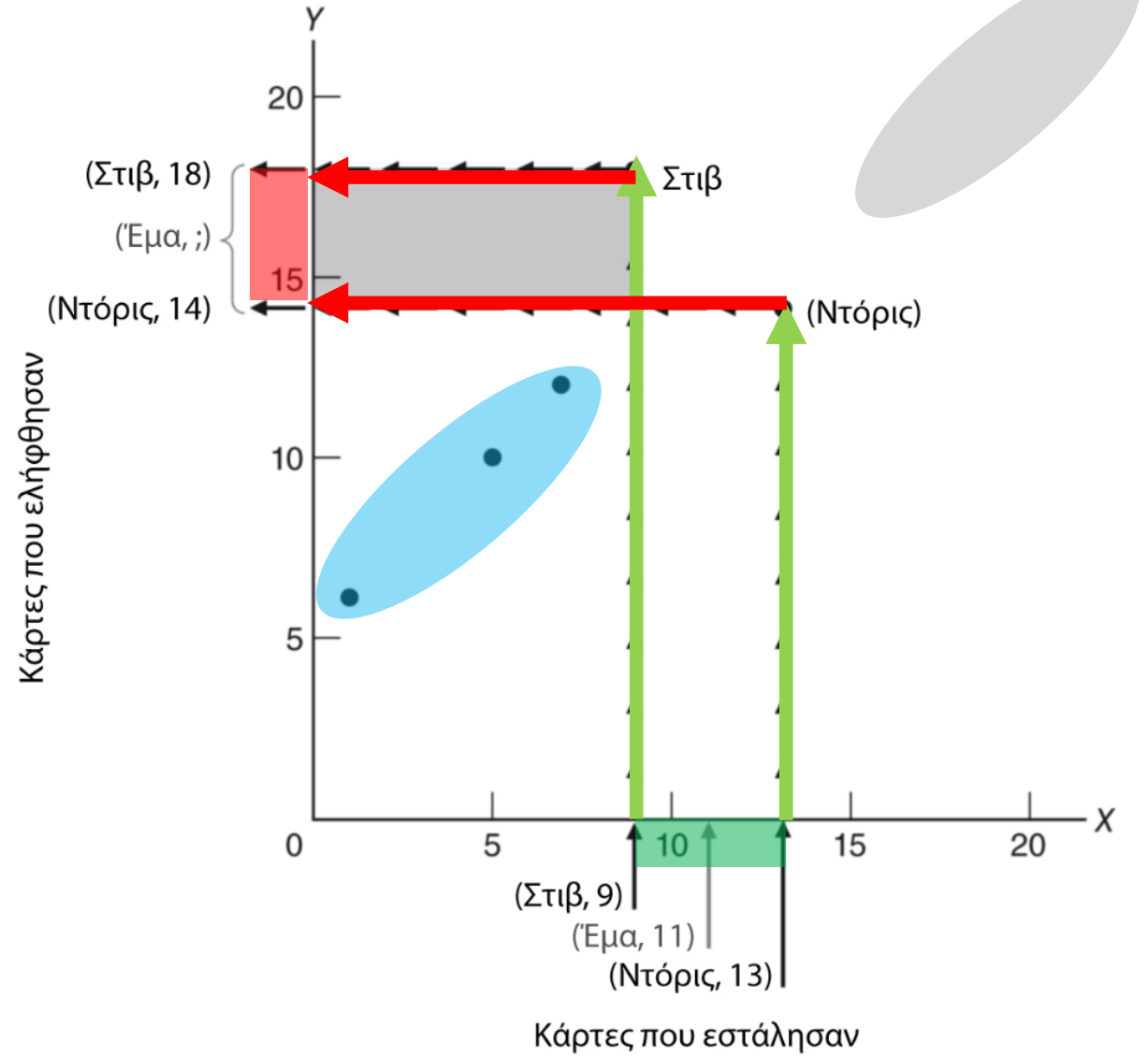
$$SS_y = \sum (Y - \bar{Y})^2 = \sum Y^2 - \frac{(\sum Y)^2}{n}$$


αποστολές

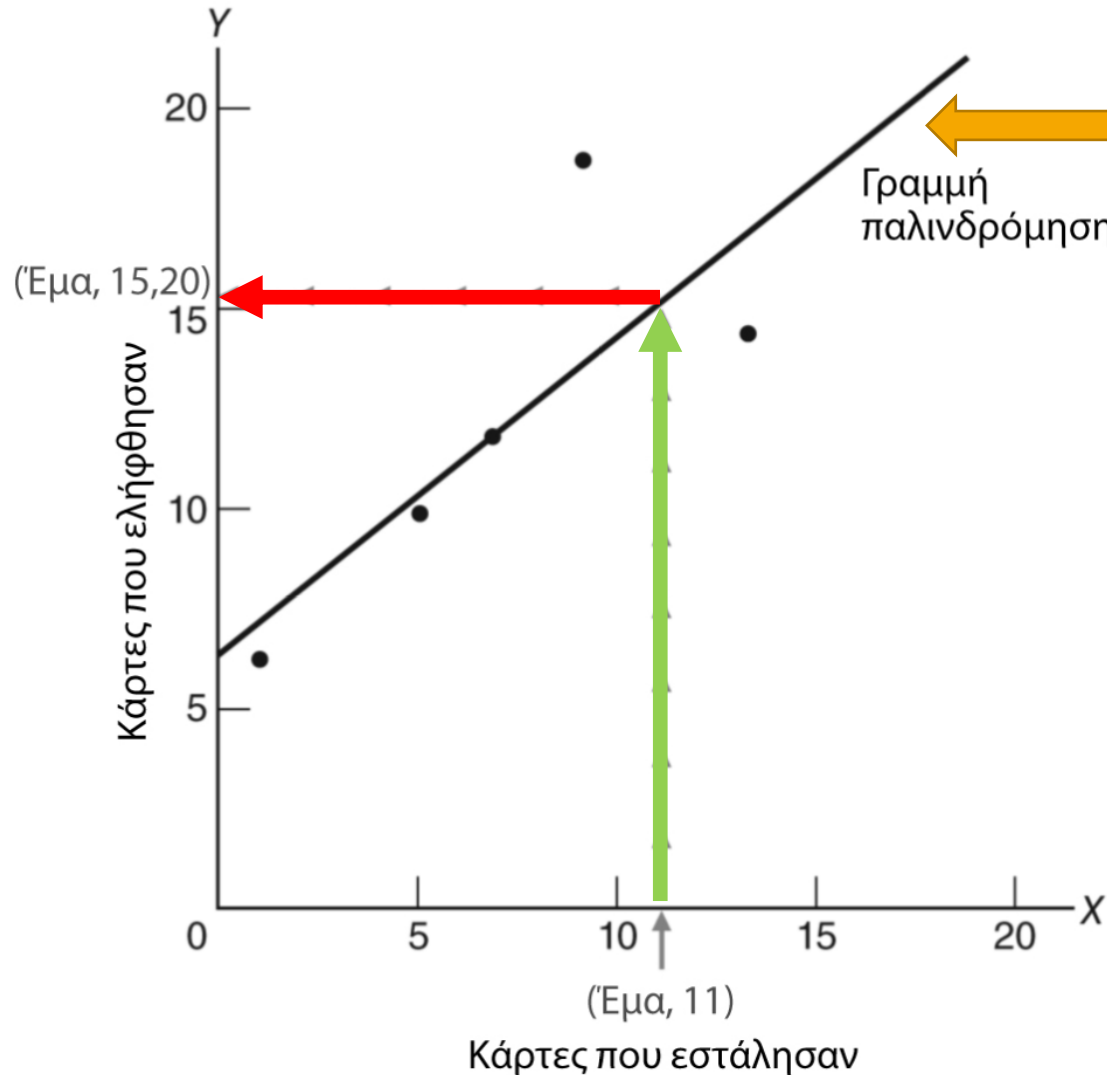

r=0.8


παραλαβές

Πρόβλεψη



Χρήση όλων των σημείων για την δημιουργία μοντέλου ανταλλαγής ευχετηρίων καρτών

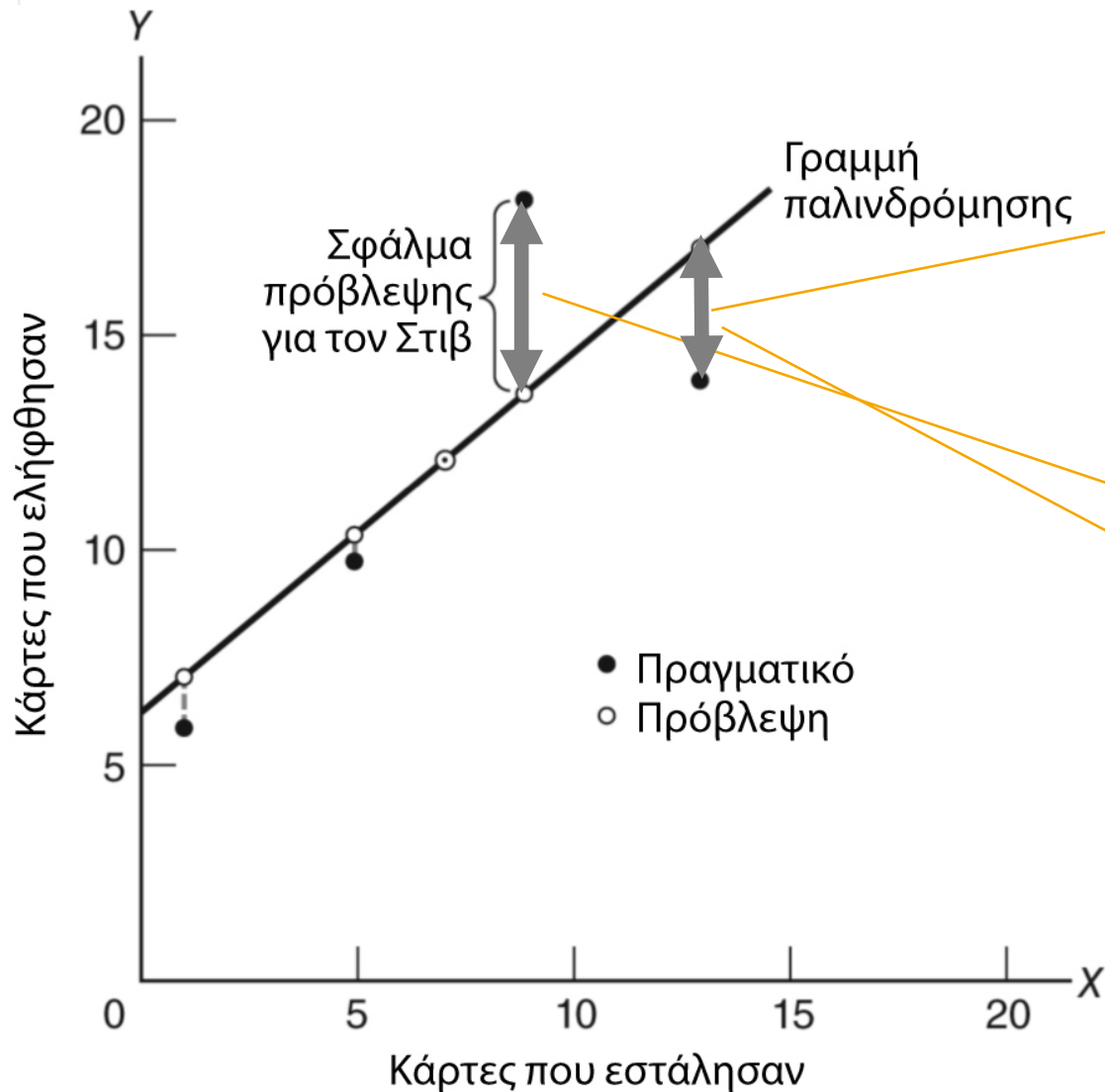


Κάρτες	Κάρτες	
	Έστειλε	Έλαβε
Αντρι	5	10
Μαικ	7	12
Ντορις	13	14
Στιβ	9	18
Τζον	1	6

Το φαινόμενο της ανταλλαγής καρτών που περιγράφεται μέσω των παρατηρήσεων, μπορεί να «απλοποιηθεί» με μια γραμμή!

Μπορούμε βάσει του μοντέλου αυτού να κάνουμε αντιστοιχίσεις και προβλέψεις

Σφάλμα της πρόβλεψης



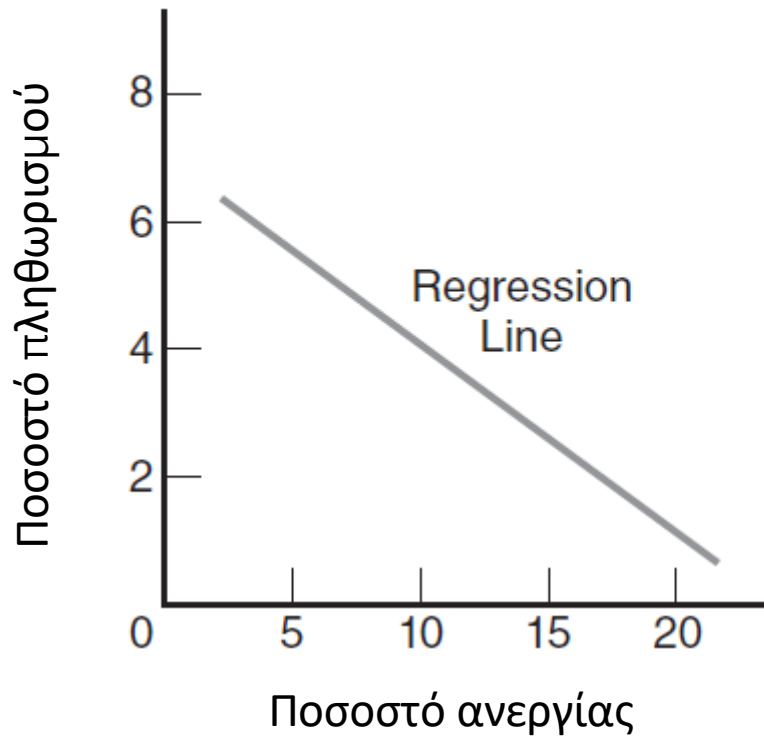
Σφάλμα: απόστασή μεταξύ άποψης του μοντέλου και πραγματικότητας

«Όλα τα μοντέλα είναι λάθος, αλλά κάποια είναι χρήσιμα» [George Box](#)

Υπό-εκτίμηση από το μοντέλο

Υπέρ-εκτίμηση από το μοντέλο

Παράδειγμα



- Προβλέψτε τον πληθωρισμό για ανεργία 5% και 16%;

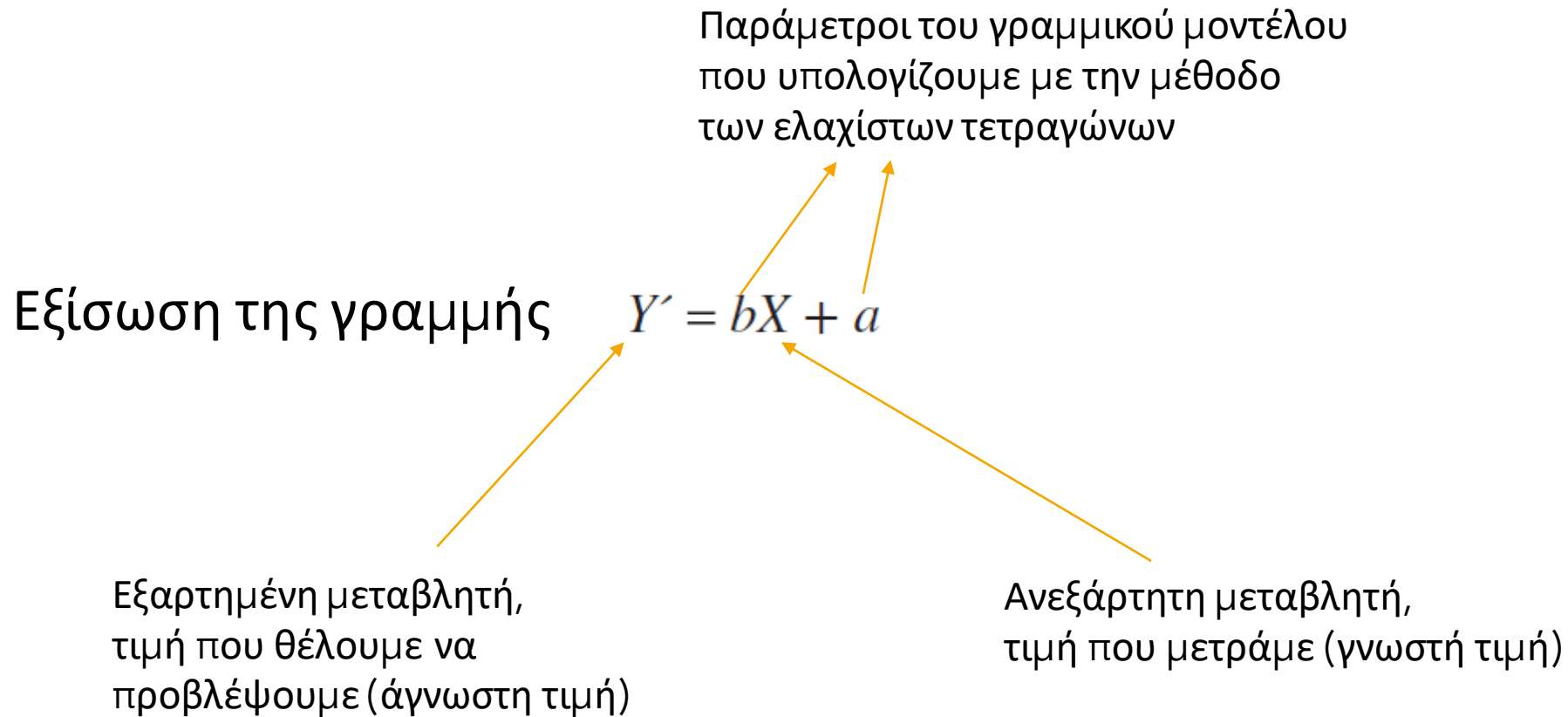
Συνολικό σφάλμα πρόβλεψης

- Αν επαναλάβουμε την διαδικασία και υπολογίσουμε την διαφορά «πραγματικής τιμής» - «πρόβλεψη μοντέλου» για όλα τα δεδομένα, λαμβάνουμε μια συνολική εικόνα του πόσο καλά το μοντέλο περιγράφει το προς εξέταση φαινόμενο και των δεδομένων του.
- Αρκεί να αθροίσω τις διαφορές;

Γραμμική παλινδρόμηση ελαχίστων τετραγώνων

- Τα διαφορετικά πρόσημα των σφαλμάτων οδηγούν σε αριθμητική απόκλιση στο μηδέν
- Ζητάμε μέθοδο που να μην επιτρέπει στα διαφορετικά πρόσημα να απαλείφουν τα σφάλματα
- Για ευκολία των υπολογισμών, καταφεύγουμε σε τετραγωνική μορφή του σφάλματος
- Ελαχιστοποιούμε το άθροισμα των τετραγώνων των επιμέρους σφαλμάτων

Υπολογισμός της γραμμής των ελαχίστων τετραγώνων



Υπολογισμός της γραμμής των ελαχίστων τετραγώνων

- Υπολογισμός σε δύο βήματα:

- Υπολογισμός της παραμέτρου που καθορίζει την κλίση

$$b = r \sqrt{\frac{SS_y}{SS_x}} \quad \begin{aligned} SS_x &= \sum (X - \bar{X})^2 = \sum X^2 - \frac{(\sum X)^2}{n} \\ SS_y &= \sum (Y - \bar{Y})^2 = \sum Y^2 - \frac{(\sum Y)^2}{n} \end{aligned}$$

- Υπολογισμός της παραμέτρου που καθορίζει την μετατόπιση

$$a = \bar{Y} - b\bar{X}$$

όπου \bar{Y}, \bar{X} οι δειγματικοί μέσοι.

Παράδειγμα



	Έστειλε	Έλαβε
Αντρι	5	10
Μαικ	7	12
Ντορις	13	14
Στιβ	9	18
Τζον	1	6

$$b = r \sqrt{\frac{SS_y}{SS_x}}$$

$$a = \bar{Y} - b\bar{X}$$

Υπολογισμός του r

$$r = \frac{SP_{xy}}{\sqrt{SS_x SS_y}}$$

$$SP_{xy} = \sum (X - \bar{X})(Y - \bar{Y}) = \sum XY - \frac{(\sum X)(\sum Y)}{n}$$

$$SS_x = \sum (X - \bar{X})^2 = \sum X^2 - \frac{(\sum X)^2}{n}$$

$$SS_y = \sum (Y - \bar{Y})^2 = \sum Y^2 - \frac{(\sum Y)^2}{n}$$

FRIEND	CARDS		4 XY	6 X ²	8 Y ²
	SENT, X	RECEIVED, Y			
Doris	13	14	182	169	196
Steve	9	18	162	81	324
Mike	7	12	84	49	144
Andrea	5	10	50	25	100
John	1	6	6	1	36

1 $n = 5$ 2 $\sum X = 35$ 3 $\sum Y = 60$ 5 $\sum XY = 484$ 7 $\sum X^2 = 325$ 9 $\sum Y^2 = 800$

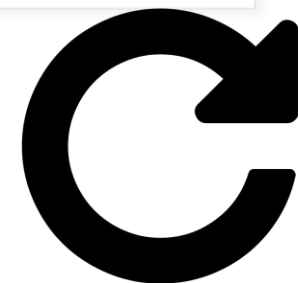
10 $SP_{xy} = \sum XY - \frac{(\sum X)(\sum Y)}{n} = 484 - \frac{(35)(60)}{5} = 484 - 420 = 64$

$$SS_x = \sum X^2 - \frac{(\sum X)^2}{n} = 325 - \frac{(35)^2}{5} = 325 - 245 = 80$$

$$SS_y = \sum Y^2 - \frac{(\sum Y)^2}{n} = 800 - \frac{(60)^2}{5} = 800 - 720 = 80$$

11 $r = \frac{SP_{xy}}{\sqrt{SS_x SS_y}} = \frac{64}{\sqrt{(80)(80)}} = \frac{64}{80} = .80$

Κοινά σημεία με τον υπολογισμό του r



Παράδειγμα

$$\begin{array}{l} 1 \quad SS_x = 80 \\ \quad \quad SS_y = 80 \\ \quad \quad r = .80 \\ \\ 2 \quad b = r \sqrt{\frac{SS_y}{SS_x}} = .80 \sqrt{\frac{80}{80}} = .80 \\ \\ 3 \quad \bar{X} = 7 \\ \quad \quad \bar{Y} = 12 \end{array}$$

$$4 \quad a = \bar{Y} - (b)(\bar{X}) = 12 - (.80)(7) = 12 - 5.60 = 6.40$$

$$\begin{array}{l} 5 \quad Y' = (b)(X) + a \\ \quad \quad = (.80)(X) + 6.40 \end{array}$$

Οι παράμετροι αυτοί ελαχιστοποιούν το άθροισμα των τετραγώνων του σφάλματος

Υπολογισμός της γραμμής των ελαχίστων τετραγώνων

- Υπολογισμός σε δύο βήματα:

- Υπολογισμός της παραμέτρου που καθορίζει την κλίση

$$b = r \sqrt{\frac{SS_y}{SS_x}} \quad \begin{array}{l} SS_x = \sum(x - \bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n} \\ SS_y = \sum(y - \bar{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n} \end{array}$$

- Υπολογισμός της παραμέτρου που καθορίζει την μετατόπιση

$$a = \bar{Y} - b\bar{X}$$

όπου \bar{Y}, \bar{X} οι δειγματικοί μέσοι.

Σχόλια

	Έστειλε	Έλαβε
Αντρι	5	10
Μαικ	7	12
Ντορις	13	14
Στιβ	9	18
Τζον	1	6

- Όταν με ένα μοντέλο κάνουμε πρόβλεψη για μια τιμή του X που δεν υπάρχουν δεδομένα ως προς την εξαρτημένη μεταβλητή, τότε ονομάζεται γνήσια πρόβλεψη (πχ αποστολή 11 καρτών).
- Για μηδενικές τιμές του X , ενδέχεται να λαμβάνουμε μη μηδενικές τιμές για την εξαρτημένη μεταβλητή λόγω της παρουσίας του b (πχ 6.4 κάρτες ακόμη και για κάποιον που δεν στέλνει καμία)
- Μια φυσική ερμηνεία του b είναι ως ρυθμός μεταβολής της εξαρτημένης μεταβλητής για κάθε μονάδα του X (πχ για κάθε κάρτα που στέλνουμε, σύμφωνα με το μοντέλο, αναμένουμε 0.8 παραλαβές)
- Δεν υπάρχει αποδεδειγμένη σχέση αιτίου-αιτιατού (πχ αν η Εμα αποστείλει κάρτες σε αγνώστους, πιθανότατα δεν θα παραλάβει)

Τυπικό σφάλμα εκτίμησης $s_{y|x}$

Άθροισμα τετραγώνων της διαφοράς

«πραγματική τιμή - πρόβλεψη»

(η πρόβλεψη εμπεριέχει το όρο «δεδομένου του x»
μέσω της γραμμικής εξίσωσης)

$$s_{y|x} = \sqrt{\frac{SS_{y|x}}{n-2}} = \sqrt{\frac{\sum(Y - Y')^2}{n-2}}$$

Διαβάζεται: τυπικό σφάλμα s υπό y δεδομένου x

Όταν έχουμε n σημεία, χάνουμε 2 βαθμούς
ελευθερίας γιατί μια ευθεία ορίζεται από
τον ίδιο αριθμό

Ή εναλλακτικά:

$$s_{y|x} = \sqrt{\frac{SS_y(1-r^2)}{n-2}}$$

Αποτελεί το ειδικό είδος της τυπικής απόκλισης
του σφάλματος πρόβλεψης του γραμμικού μοντέλου

Παράδειγμα

Τυπικό σφάλμα εκτίμησης $s_{y|x}$

$$s_{y|x} = \sqrt{\frac{SS_{y|x}}{n-2}} = \sqrt{\frac{\sum(Y-Y')^2}{n-2}}$$

Άθροισμα τετραγώνων της διαφοράς «πραγματική τιμή - πρόβλεψη» (η πρόβλεψη εμπεριέχει το όρο «δεδομένου του x» μέσω της γραμμικής εξίσωσης)

Διαβάζεται: τυπικό σφάλμα s υπό y δεδομένου x

Όταν έχουμε n σημεία, χάνουμε 2 βαθμούς ελευθερίας γιατί μια ευθεία ορίζεται από τον ίδιο αριθμό

Ή εναλλακτικά:
$$s_{y|x} = \sqrt{\frac{SS_y(1-r^2)}{n-2}}$$

Αποτελεί το ειδικό είδος της τυπικής απόκλισης του σφάλματος πρόβλεψης του γραμμικού μοντέλου

18



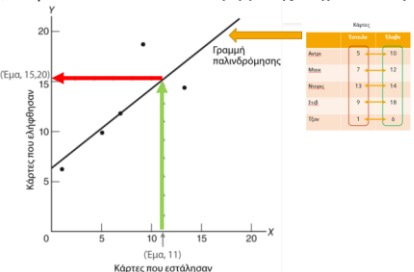
1 $SS_y = 80$

$r = .80$

2
$$s_{y|x} = \sqrt{\frac{SS_y(1-r^2)}{n-2}} = \sqrt{\frac{80(1-[.80]^2)}{5-2}} = \sqrt{\frac{80(.36)}{3}} = \sqrt{\frac{28.80}{3}} = \sqrt{9.60} = 3.10$$

Μια πιο ολοκληρωμένη απάντηση: η Εμα αναμένει 15.2 ± 3.1 κάρτες

Χρήση όλων των σημείων για την δημιουργία μοντέλου ανταλλαγής ευχετηρίων καρτών



Το φαινόμενο της ανταλλαγής καρτών που περιγράφεται μέσω των Παρατηρήσεων, μπορεί να «απλοποιηθεί» με μια γραμμή!

Μπορούμε βάσει του μοντέλου αυτό να κάνουμε αντιστοιχίες και προβλέψεις

Η σημασία του r

- Όταν το r είναι μεγάλο (κατ' απόλυτη τιμή), το τυπικό σφάλμα είναι μικρό
- Για ακραίες τιμές του r

$$SS_{y|x} = SS_y (1 - r^2) = SS_y [1 - (1)^2] = SS_y [1 - 1] = SS_y [0] = 0$$

$$SS_{y|x} = SS_y (1 - r^2) = SS_y [1 - (0)^2] = SS_y [1 - 0] = SS_y [1] = SS_y$$

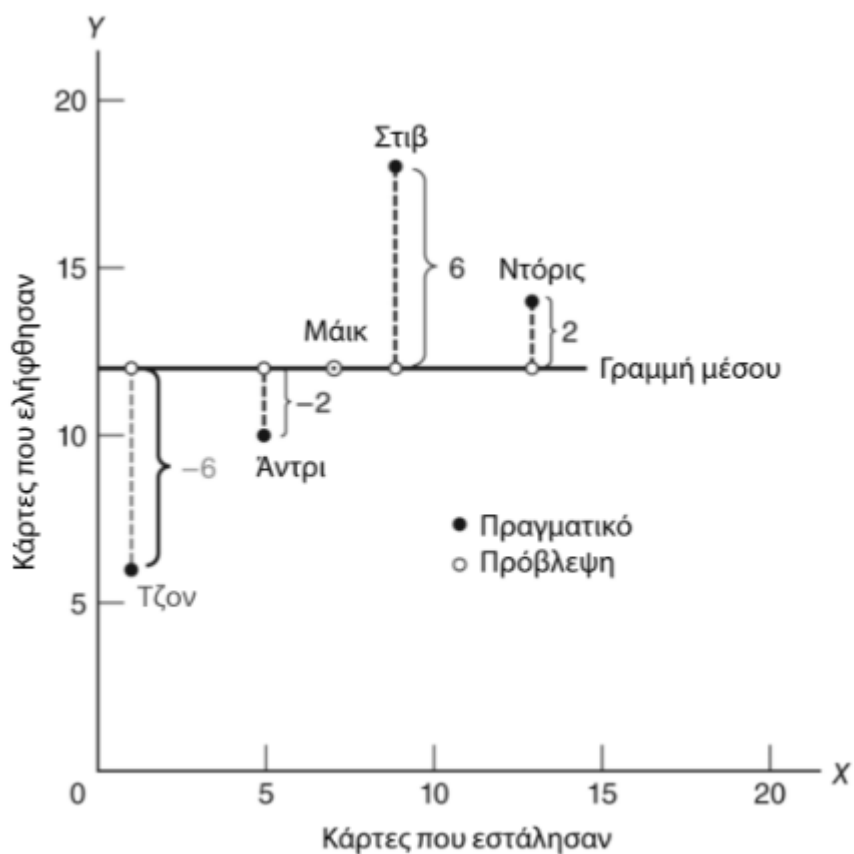
$$SS_y = \sum (Y - \bar{Y})^2 = \sum Y^2 - \frac{(\sum Y)^2}{n}$$

$$s_{y|x} = \sqrt{\frac{SS_y (1 - r^2)}{n - 2}}$$

Ερμηνεία του r^2

	Έστειλε	Έλαβε
Αντρι	5	10
Μαικ	7	12
Ντόρις	13	14
Στιβ	9	18
Τζον	1	6

Επαναλαμβανόμενη τιμή του μέσου

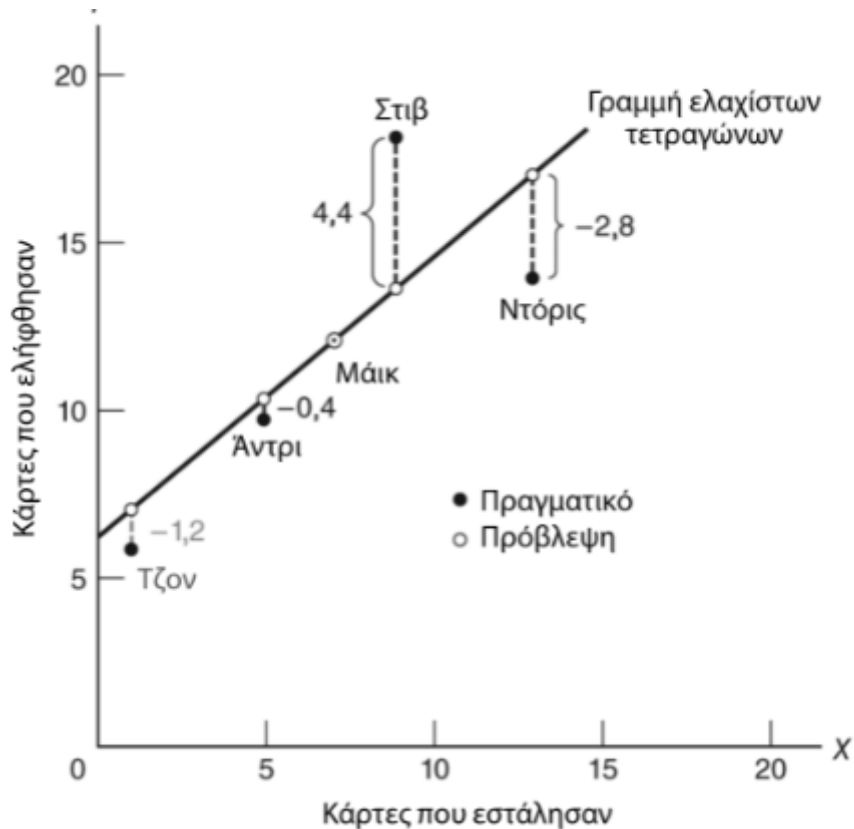


Έστω το εξαιρετικά απλό μοντέλο:
Για οποιαδήποτε τιμή του X ,
κάνει την ίδια πρόβλεψη ίση με την μέση τιμή

Το συνολικό σφάλμα είναι μηδέν

Είναι μέσο σύγκρισης με το μοντέλο που ελαχιστοποιεί το άθροισμα των τετραγώνων του σφάλματος (ελάχιστα τετράγωνα)

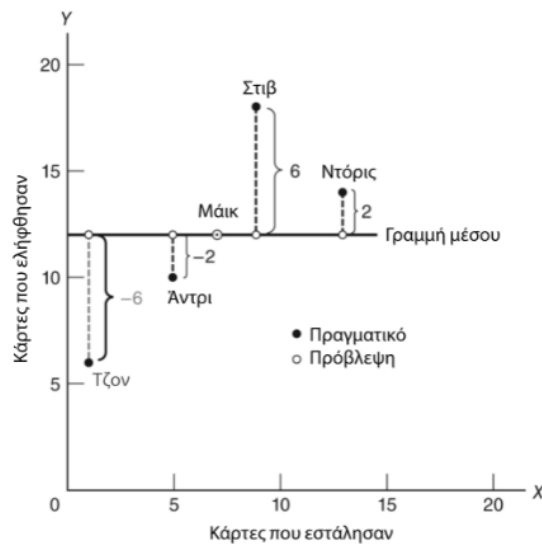
Ερμηνεία του r^2



Κάποιες τιμές του X , δίνουν υψηλότερο σφάλμα (ακόμη και από την επαναλαμβανόμενη τιμή του μέσου)
Γενικά μικρότερο γιατί αποφεύγεται η αριθμητική απόκλιση στο μηδέν

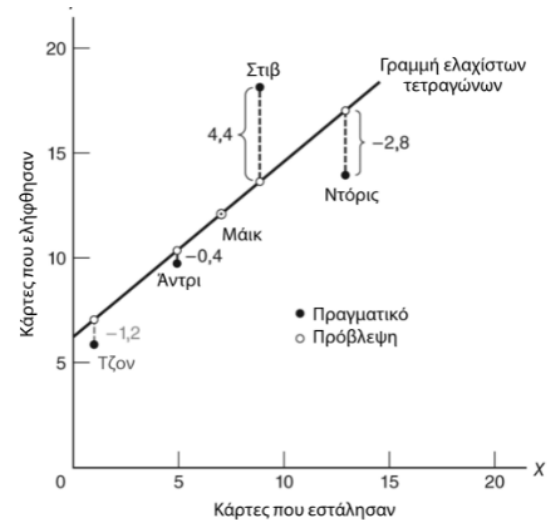
Ερμηνεία του r^2

- Σύγκριση των αριθμητικών αποτελεσμάτων



$$SS_y = \sum(Y - \bar{Y})^2$$

$$SS_y = [(-6)^2 + (-2)^2 + 0^2 + 6^2 + 2^2] = 80$$



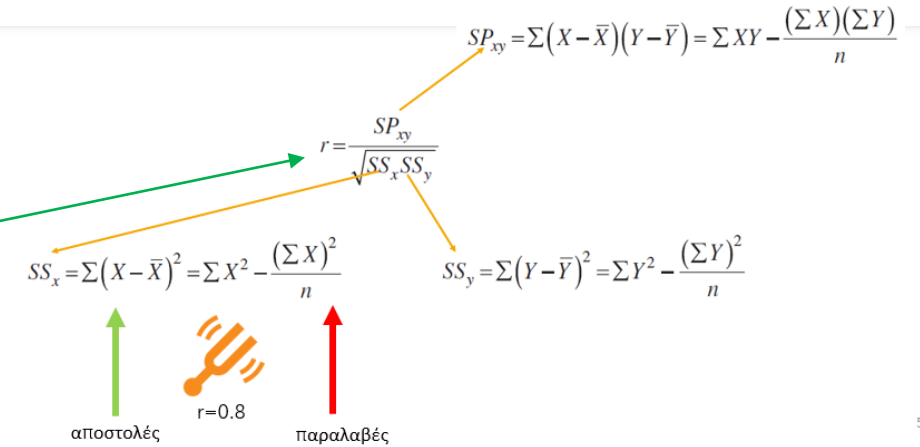
$$SS_{y|x} = \sum(Y - Y')^2$$

$$SS_{y|x} = [(-1.2)^2 + (-0.4)^2 + 0^2 + (4.4)^2 + (-2.8)^2] = 28.8$$



Ερμηνεία του r^2

$$r^2 = \frac{SS_{Y'}}{SS_Y} = \frac{SS_Y - SS_{Y|X}}{SS_Y}$$



Κανονικοποιημένη διαφορά του νέου μοντέλου σε σχέση με την επαναλαμβανόμενη τιμή του μέσου

Τυπικό σφάλμα εκτίμησης $s_{y|x}$

$$s_{y|x} = \sqrt{\frac{SS_{y|x}}{n-2}} = \sqrt{\frac{\sum (Y - Y')^2}{n-2}}$$

Άθροισμα τετραγώνων της διαφοράς «πραγματική τιμή - πρόβλεψη» (η πρόβλεψη περιέχει το όρο «δεδομένου του x» μέσω της γραμμικής εξίσωσης)

Διαβάζεται: τυπικό σφάλμα s υπό y δεδομένου x

Όταν έχουμε n σημεία, χάνουμε 2 βαθμούς ελευθερίας γιατί μια ευθεία ορίζεται από τον ίδιο αριθμό

Ή εναλλακτικά: $s_{y|x} = \sqrt{\frac{SS_y (1 - r^2)}{n - 2}}$

Αποτελεί το ειδικό είδος της τυπικής απόκλισης του σφάλματος πρόβλεψης του γραμμικού μοντέλου

Παράδειγμα



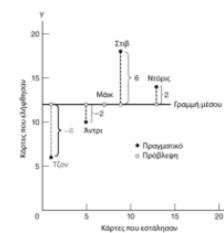
$$r^2 = \frac{SS_{Y'}}{SS_Y} = \frac{SS_Y - SS_{Y|X}}{SS_Y}$$

$$\frac{SS_y - SS_{y|x}}{SS_y} = \frac{80 - 28.8}{80} = \frac{51.2}{80} = .64$$

64% κέδρος αναλογίας στο συνολικό τετραγωνικό σφάλμα σε σχέση με την επαναλαμβανόμενη πρόβλεψη

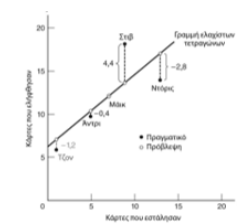
Ερμηνεία του r^2

- Σύγκριση των αριθμητικών αποτελεσμάτων



$$SS_y = \sum (Y - \bar{Y})^2$$

$$SS_y = [(-6)^2 + (-2)^2 + 0^2 + 6^2 + 2^2] = 80$$



$$SS_{y|x} = \sum (Y - \hat{Y})^2$$

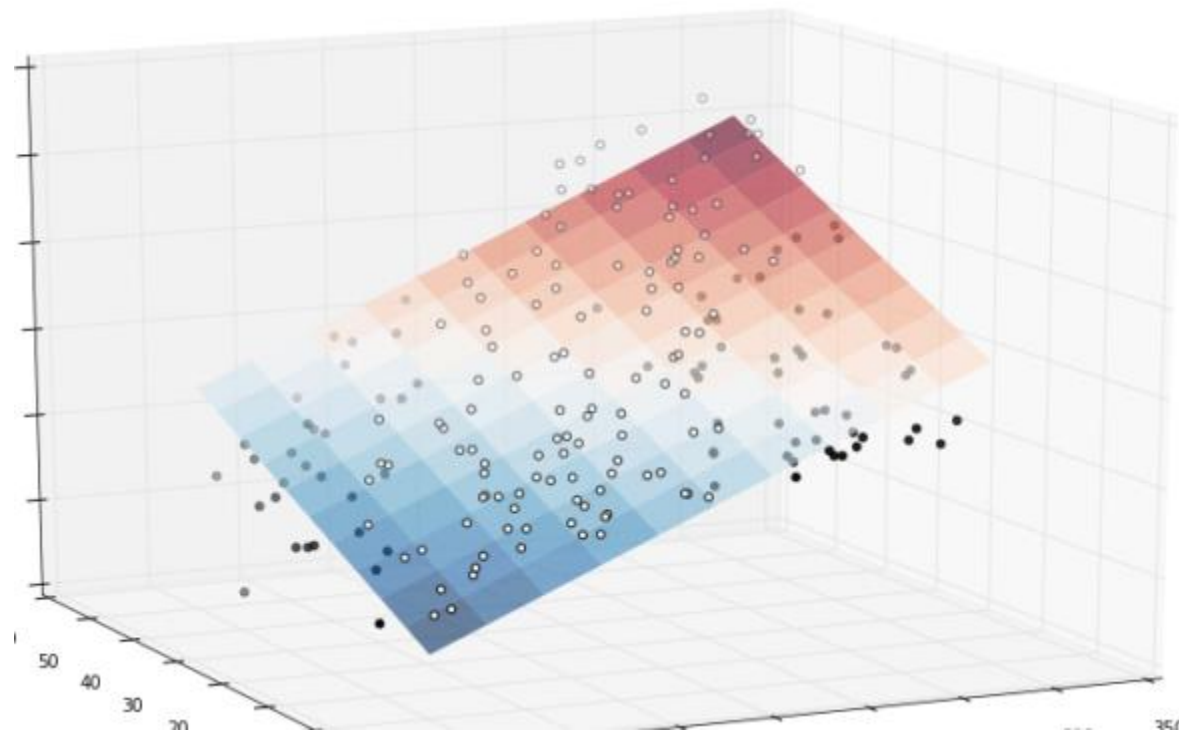
$$SS_{y|x} = [(-1.2)^2 + (-0.4)^2 + 0^2 + (4.4)^2 + (-2.8)^2] = 28.8$$

Σχόλια



- Η ερμηνεία του r^2 ισχύει για ολόκληρο το δείγμα, και όχι απαραίτητα σε μεμονωμένες περιπτώσεις
- Η τιμή του r^2 ποικίλει από εφαρμογή-σε-εφαρμογή ως προς την σημαντικότητά της
- Καμιά τιμή του r^2 δεν εξασφαλίζει σχέση αιτίου-αιτιατού

Παλινδρόμηση με πολλαπλές μεταβλητές



$$Y' = .410(X_1) + .005(X_2) + .001(X_3) + 1.03$$

On-line υλικό (μια μεταβλητή)

https://www.youtube.com/watch?v=nk2CQITm_eo

StatQuest: Linear Regression
(aka General Linear Models, part 1)

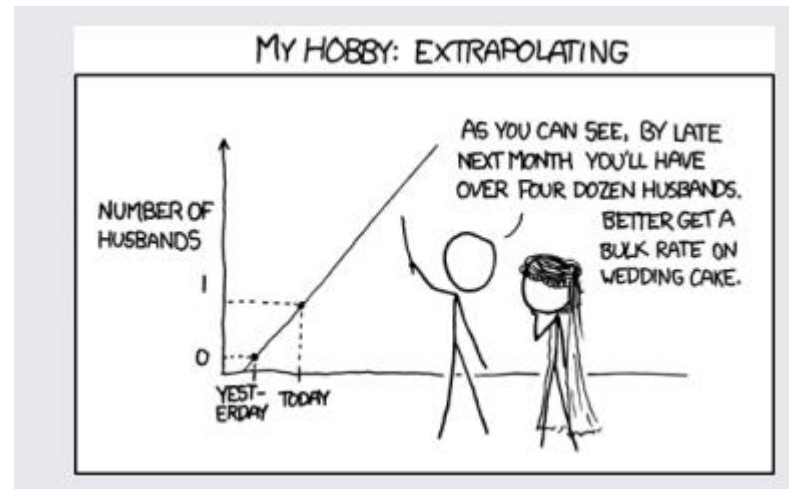
On-line υλικό (πολλαπλές μεταβλητές)

<https://www.youtube.com/watch?v=zITIFTsivN8>

StatQuest:
Multiple Regression...
Clearly explained!!!

Συγγράμματα (μια μεταβλητή)

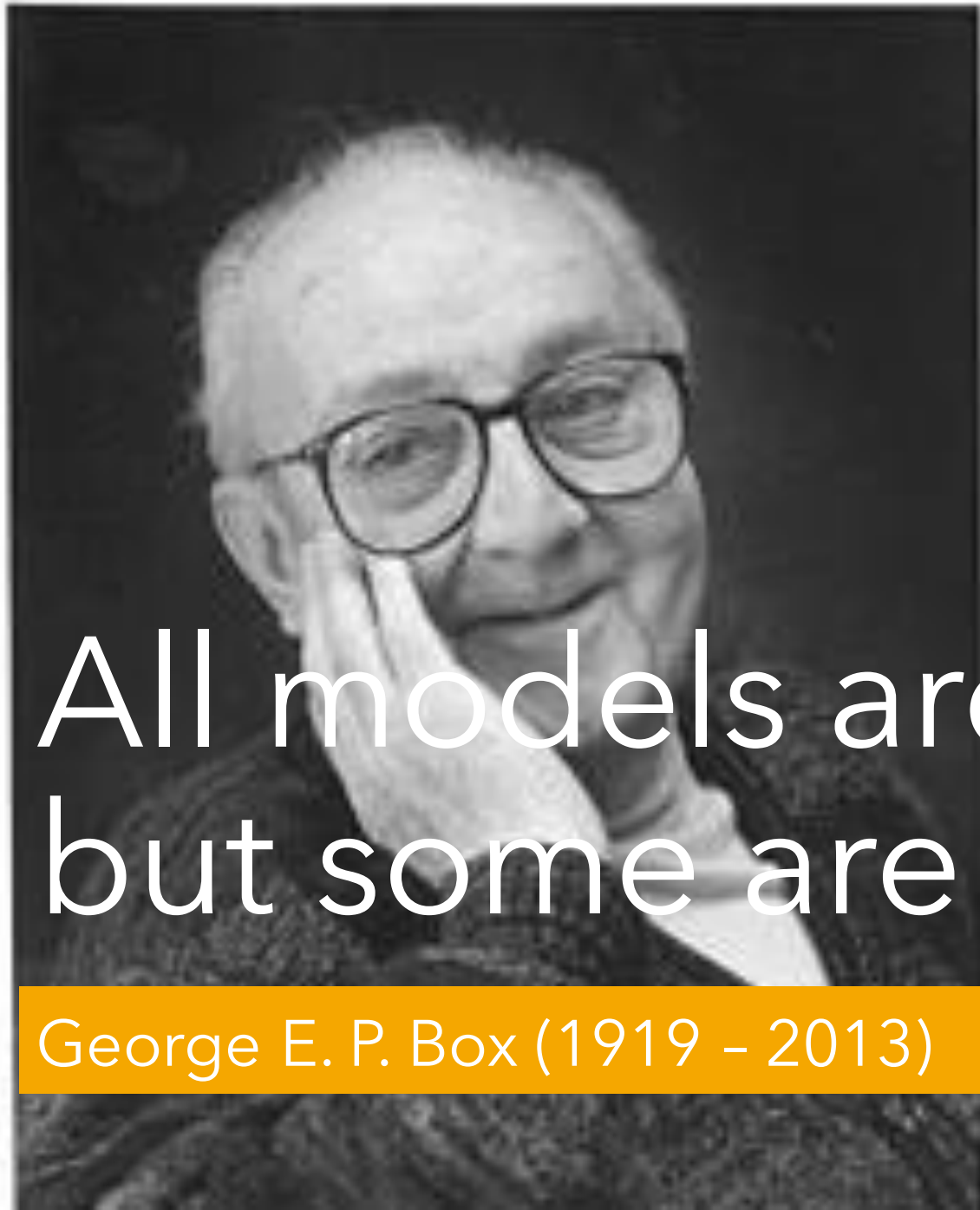
<http://www.mit.edu/~6.s085/notes/lecture3.pdf>



Συγγράμματα (πολλαπλές μεταβλητές)

https://www.stat.berkeley.edu/~brill/Stat131a/29_multi.pdf





All models are wrong,
but some are useful.

George E. P. Box (1919 - 2013)

