Στατιστική ΙΙ

Γιώργος Τσιρογιάννης

Τμήμα Διοίκησης Επιχειρήσεων Αγροτικών Προϊόντων και Τροφίμων,

Πανεπιστήμιο Πατρών

Ακαδημαϊκό έτος: 2020-2021

Διάλεξη 3η

Παλινδρόμηση

7ο 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

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

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

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

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

r=0.8

αποστολές

παραλαβές

Πως υπολογίζουμε το r;

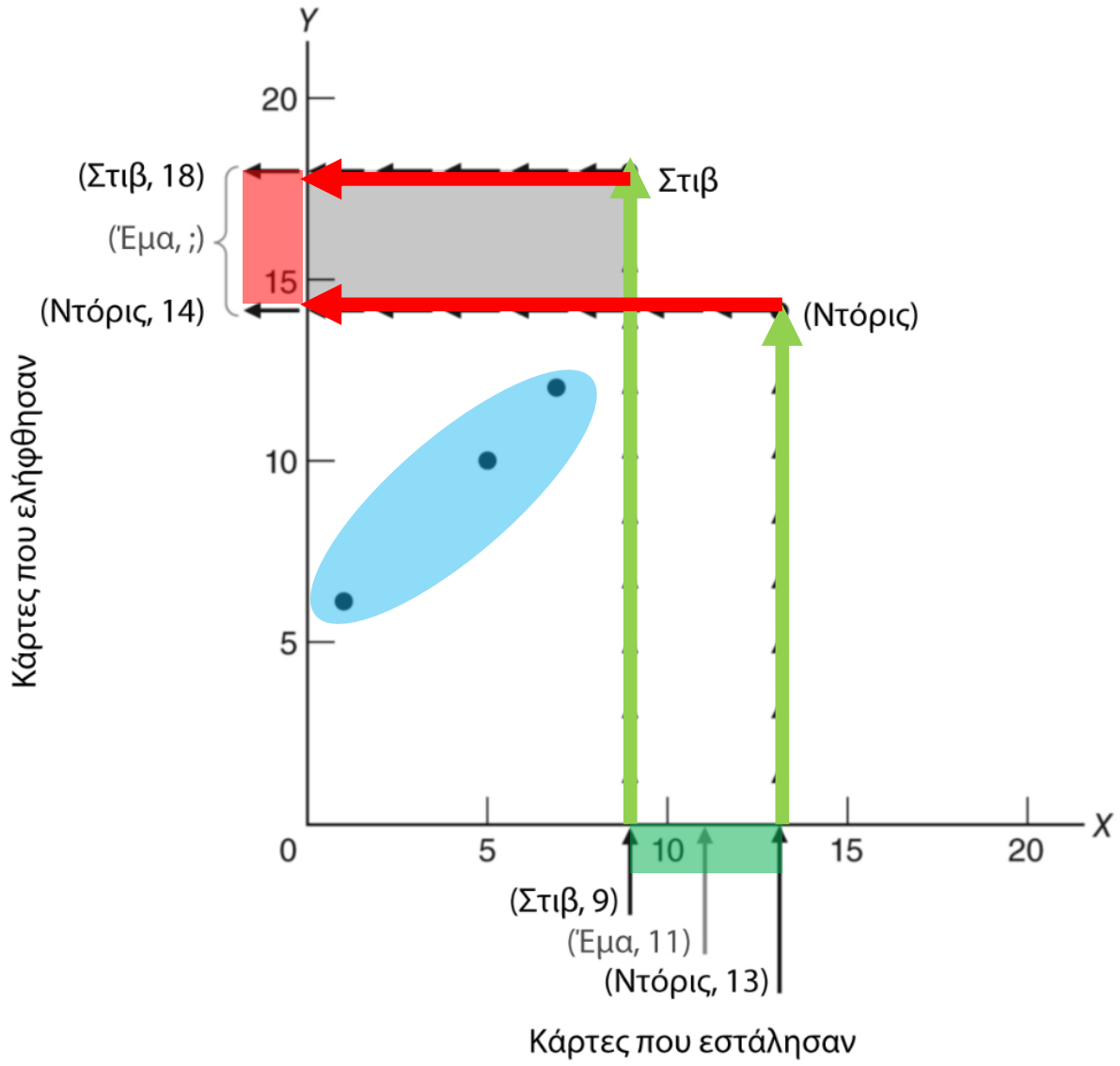
$$SP_{xy} = \sum \left(X - \bar{X}\right)\left(Y - \bar{Y}\right) = \sum XY - \frac{\left(\sum X\right)\left(\sum Y\right)}{n}$$

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

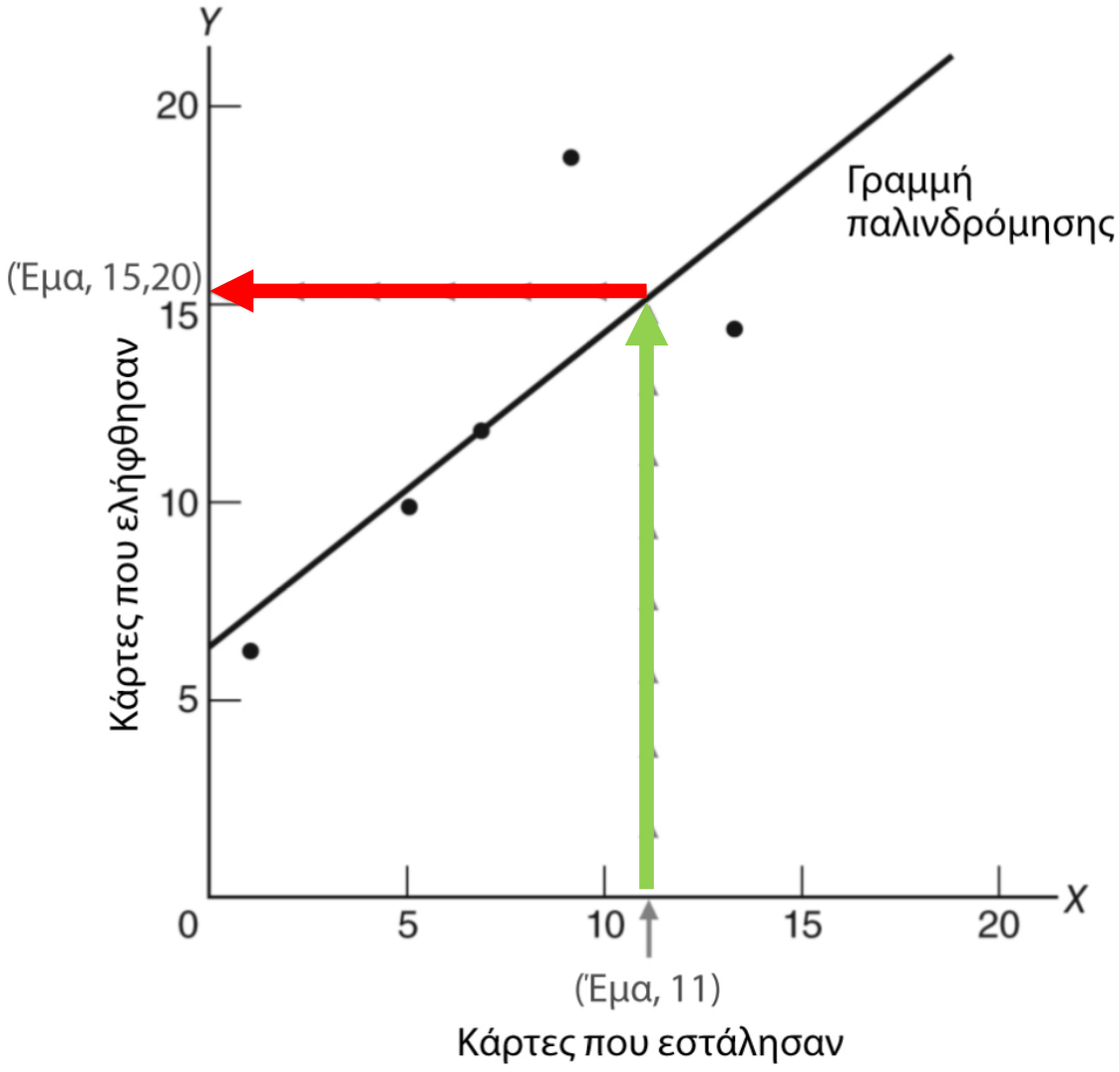
$$SS_x = \sum \left(X - \bar{X}\right)^2 = \sum X^2 - \frac{\left(\sum X\right)^2}{n}$$

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

Πρόβλεψη


Y
20
(Στιβ, 18)
(Έμα, ;)
15
(Ντόρις, 14)
Στιβ
(Ντόρις)
Κάρτες που ελήφθησαν
10
5
0
5
10
15
20
X
(Στιβ, 9)
(Έμα, 11)
(Ντόρις, 13)
Κάρτες που εστάλησαν

Χρήση όλων των σημείων για την δημιουργία μοντέλου ανταλλαγής ευχητηρίων καρτών



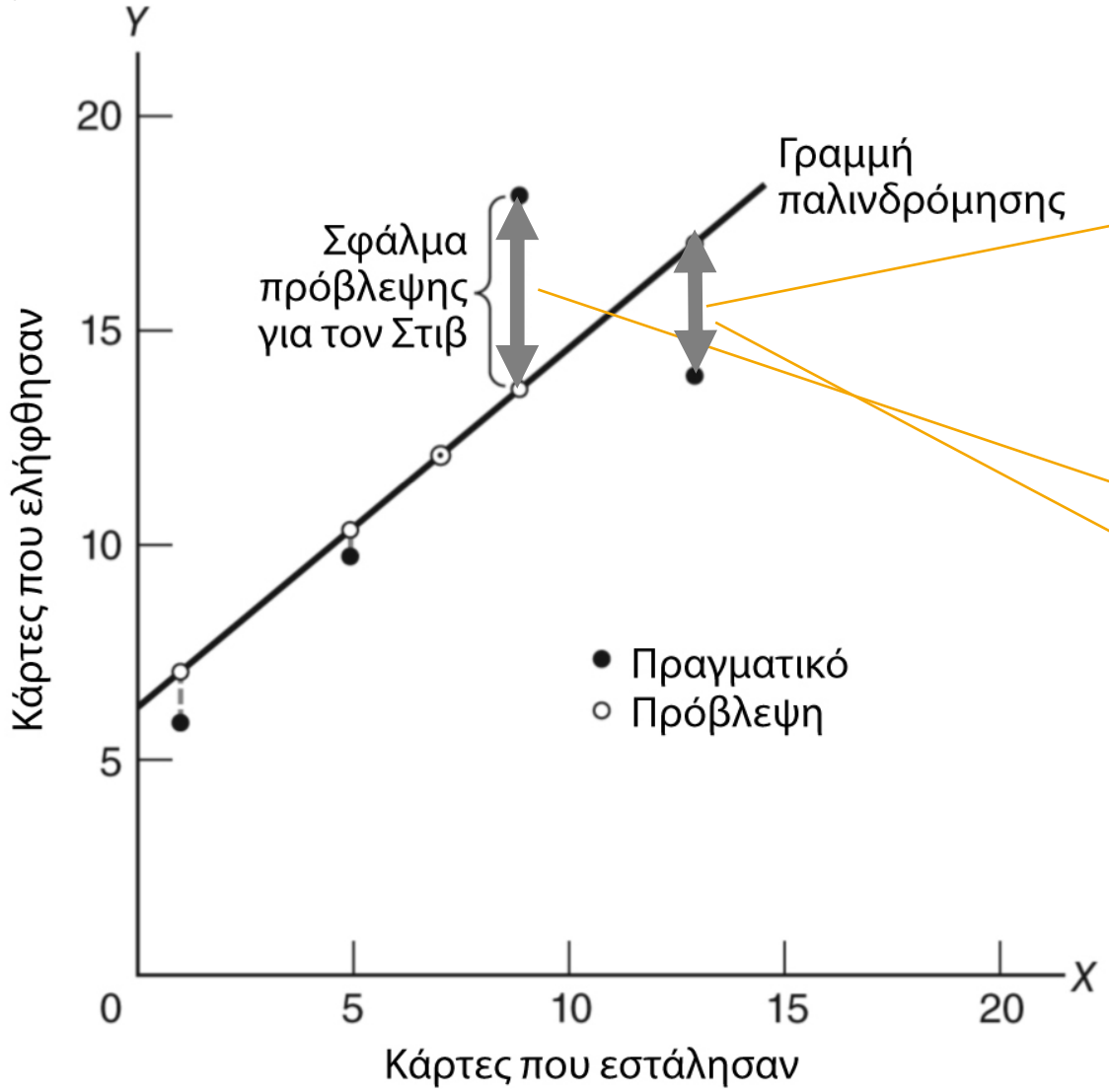
Κάρτες

	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

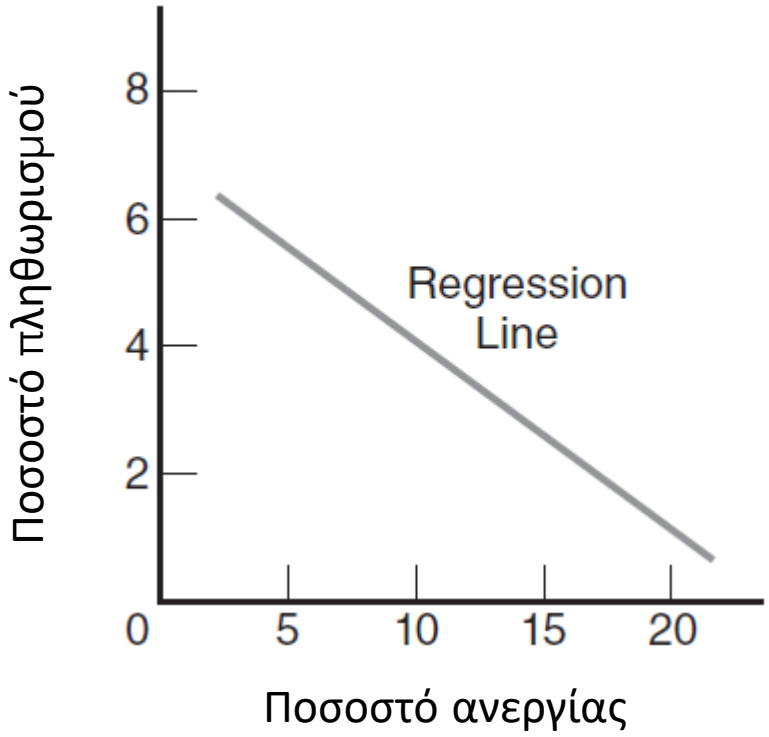
Υπό-εκτίμηση από το μοντέλο

Υπέρ-εκτίμηση από το μοντέλο

Συνολικό σφάλμα πρόβλεψης

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

Παράδειγμα



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

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

Παράμετροι του γραμμικού μοντέλου που υπολογίζουμε με την μέθοδο των ελαχίστων τετραγώνων

Εξίσωση της γραμμής $Y' = bX + a$

Εξαρτημένη μεταβλητή, τιμή που θέλουμε να προβλέψουμε (άγνωστη τιμή)

Ανεξάρτητη μεταβλητή, τιμή που μετράμε (γνωστή τιμή)

Υπολογισμός της γραμμής των ελαχίστων τετραγώνων

- Υπολογισμός σε δύο βήματα:
  - Υπολογισμός της παραμέτρου που καθορίζει την κλίση

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

  - 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

$$r = \frac{SP_{xy}}{\sqrt{SS_x SS_y}} \qquad SP_{xy} = \sum\left(X - \bar{X}\right)\left(Y - \bar{Y}\right) = \sum XY - \frac{(\sum X)(\sum Y)}{n}$$

$$SS_x = \sum\left(X - \bar{X}\right)^2 = \sum X^2 - \frac{(\sum X)^2}{n} \qquad SS_y = \sum\left(Y - \bar{Y}\right)^2 = \sum Y^2 - \frac{(\sum Y)^2}{n}$$

	CARDS		4	6	8
FRIEND	**SENT, *X***	**RECEIVED, *Y***	***XY***	***X*²**	***Y*²**
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 $\Sigma X = 35$ 3 $\Sigma Y = 60$ 5 $\Sigma XY = 484$ 7 $\Sigma X^2 = 325$ 9 $\Sigma Y^2 = 800$

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

Παράδειγμα

1 $SS_x = 80$
$SS_y = 80$
$r = .80$

2 $$b = r\sqrt{\frac{SS_Y}{SS_X}} = .80\sqrt{\frac{80}{80}} = .80$$

3 $\overline{X} = 7$
$\overline{Y} = 12$

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

5 $$Y' = (b)(X) + a$$
$$= (.80)(X) + 6.40$$

Οι παράμετροι αυτοί ελαχιστοποιούν το άθροισμα των τετραγώνων του σφάλματος

Σχόλια

- 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 a 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 $s_{y|x}$

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

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

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

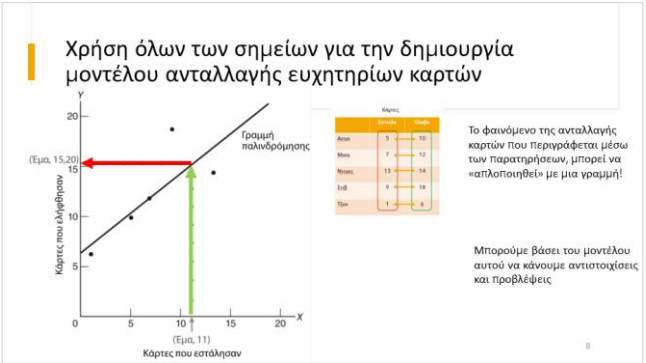
Παράδειγμα

1 $SS_y = 80$

$r = .80$

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

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



Η σημασία του r

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

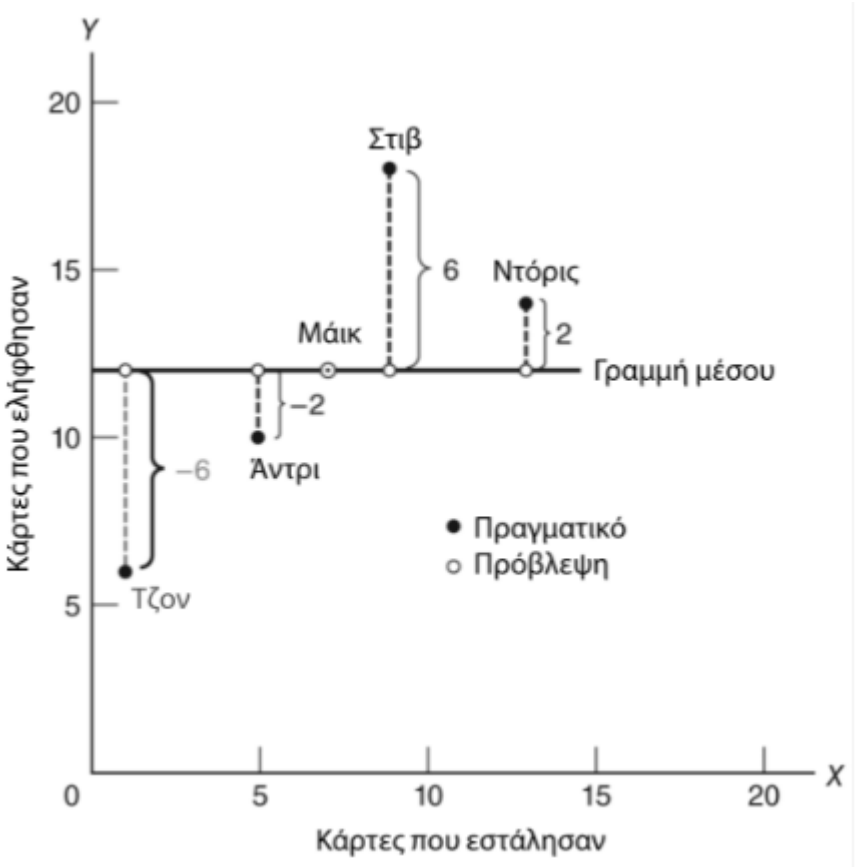
$$SS_{y|x} = SS_y(1-r^2) = SS_y[1-(1)^2] = SS_y[1-1] = SS_y[0] = 0$$

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

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

Ερμηνεία του r^2

Επαναλαμβανόμενη τιμή του μέσου

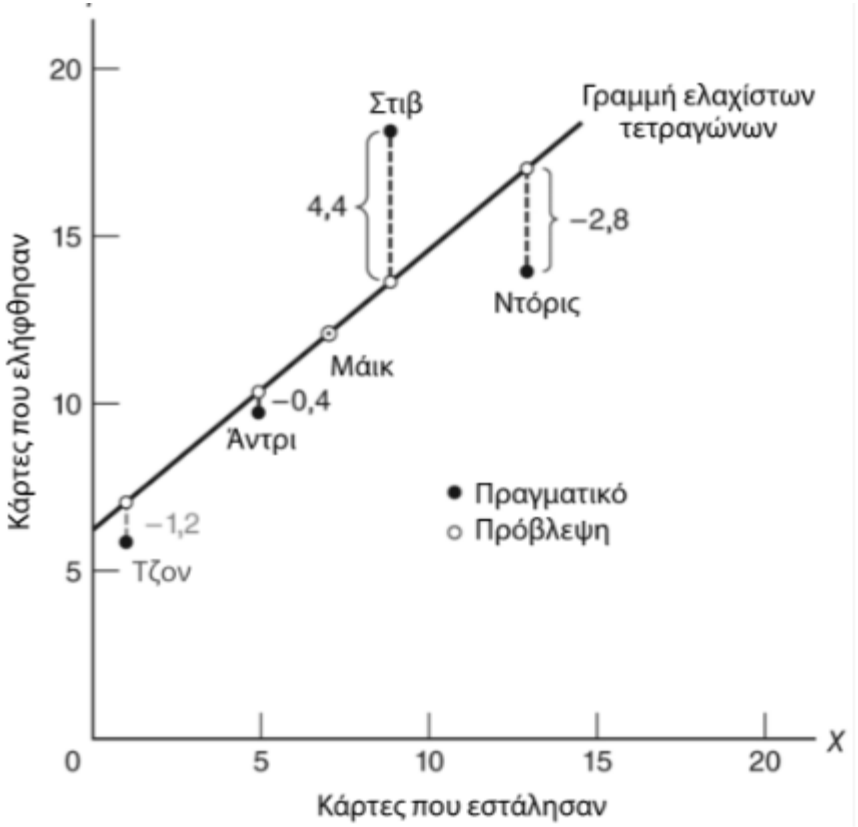


Ένα εξαιρετικά απλό μοντέλο:
Για οποιαδήποτε τιμή του X,
κάνει την ίδια πρόβλεψη ίση με την μέση τιμή

Το συνολικό σφάλμα είναι μηδέν

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

Ερμηνεία του r^2

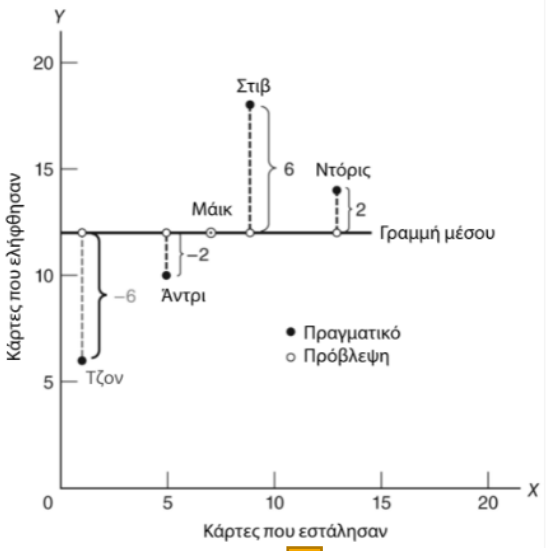


Κάποιες τιμές του X, δίνουν υψηλότερο σφάλμα (ακόμη και από την επαναλαμβανόμενη τιμή του μέσου)

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

Παράδειγμα

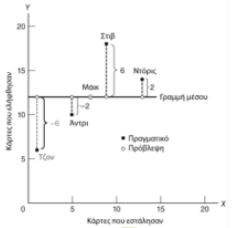
$$r^2 = \frac{SS_{Y'}}{SS_Y} = \frac{SS_Y - SS_{Y|X}}{SS_Y}$$

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

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

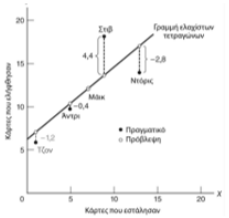
Ερμηνεία του r^2

- Σύγκριση των αριθμητικών αποτελεσμάτων



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

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



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

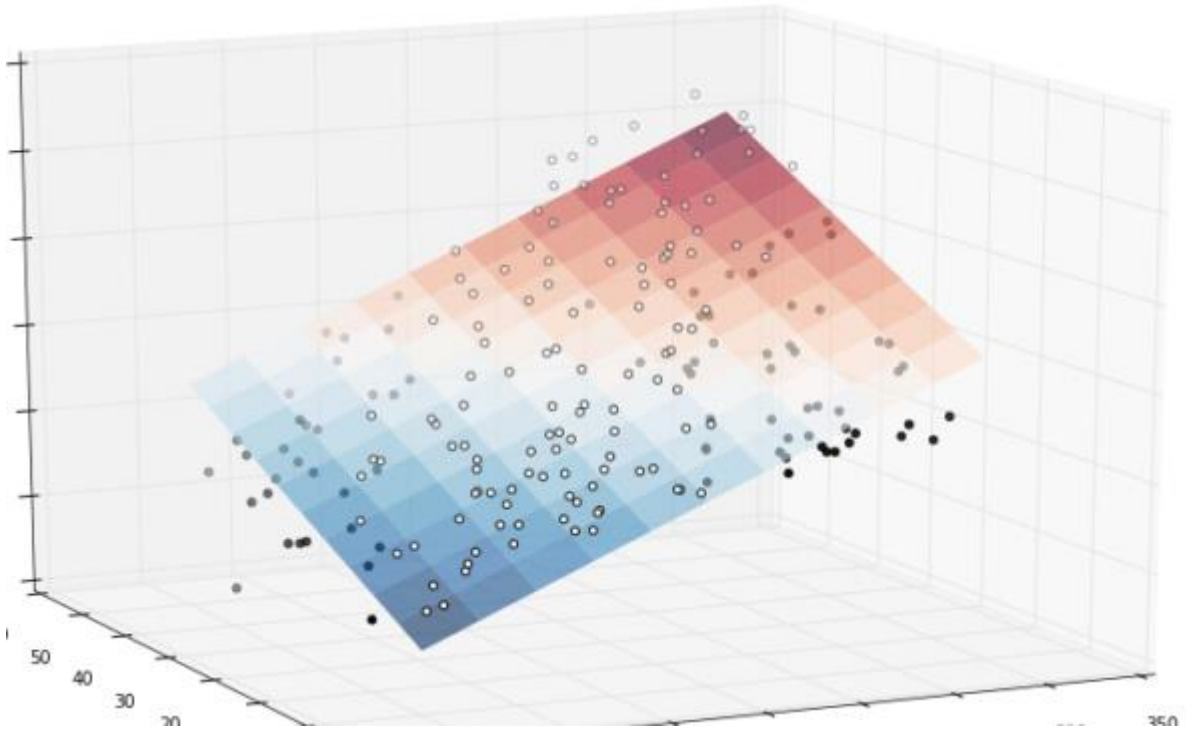
$$SS_{y|x} = [(-1.2)^2 + (-0.4)^2 + 0^2 + (4.4)^2 + (-2.8)^2] = 28.8$$

24

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 ώς προς την σημαντικότητά της
- Καμιά τιμή του r^2 δεν εξασφαλίζει σχέση αιτίου-αιτιατού

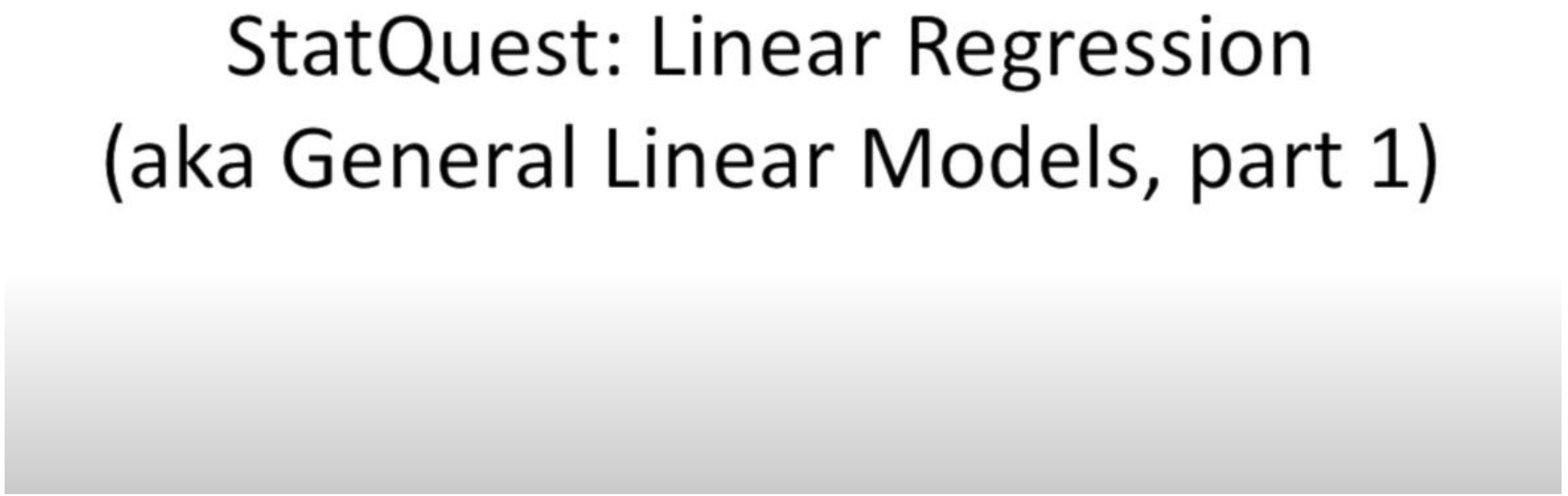
Παλινδρόμηση με πολλαπλές μεταβητές



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

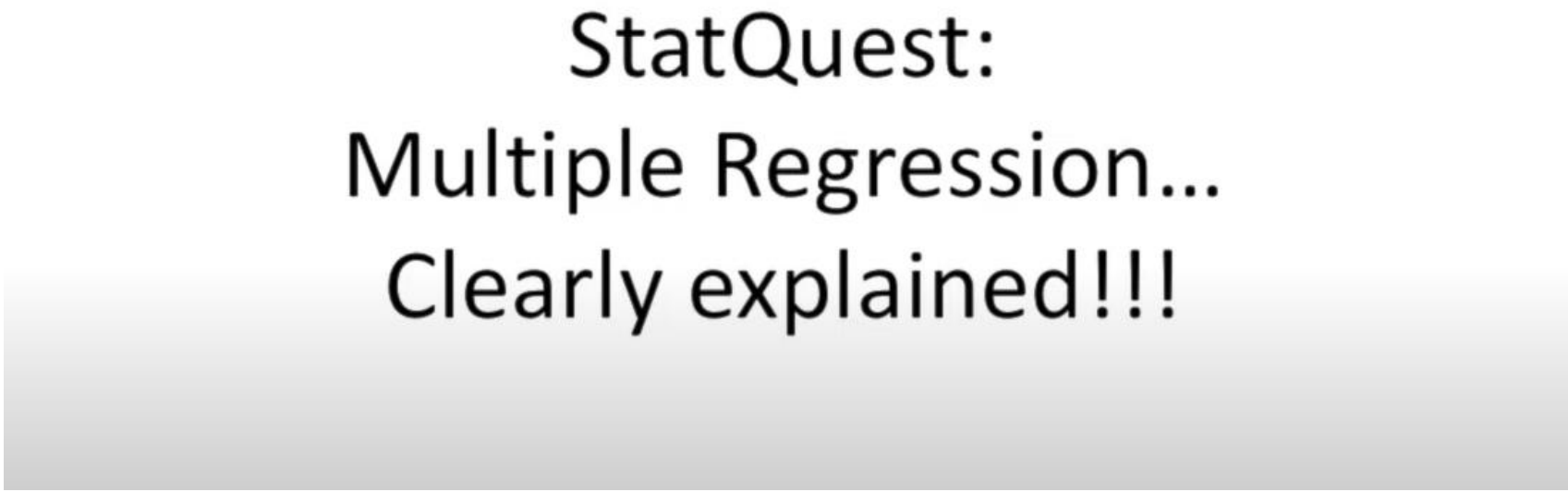
On-line υλικό (μια μεταβλητή)

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



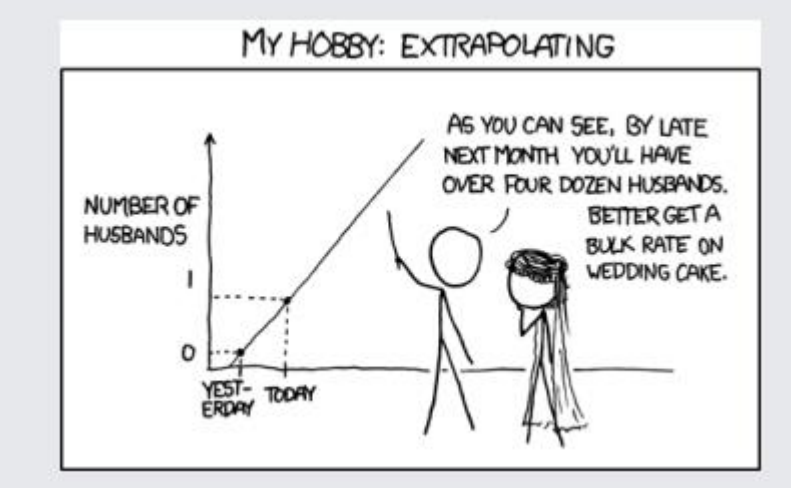
On-line υλικό (πολλαπλές μεταβλητές)

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



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