ΣΤΑΤΙΣΤΙΚΗ ΚΑΙ ΘΕΩΡΙΑ ΑΠΟΦΑΣΕΩΝ

ΔΙΑΛΕΞΗ: ΠΑΛΙΝΔΡΟΜΗΣΗ

Διδάσκουσα: Ε. Γάκη, Επίκ. Καθηγήτρια

Περιεχόμενα Ενότητας

- Ανάλυση Παλινδρόμησης
- Εκτίμηση Ευθείας παλινδρόμησης
- Εκτίμηση των συντελεστών της ευθείας παλινδρόμησης
- Στατιστική Συμπερασματολογία στην ανάλυση παλινδρόμησης
  - Διαστήματα Εμπιστοσύνης
  - Έλεγχοι Υποθέσεων
  - Συντελεστής Προσδιορισμού
- Πολλαπλή Παλινδρόμηση
- Μη γραμμικά Μοντέλα

Περιεχόμενα Μαθήματος

- Στατιστική Συμπερασματολογία στην ανάλυση παλινδρόμησης
  - Συντελεστής Προσδιορισμού
- Πολλαπλή Παλινδρόμηση
- Μη γραμμικά Μοντέλα

Συντελεστής Προσδιορισμού

Θέλουμε να δούμε πόσο καλά η ευθεία παλινδρόμησης εξηγεί τα δεδομένα του δείγματος. Για το σκοπό αυτό χρειάζεται ένα κριτήριο που θα καθορίζει το εάν μία γραμμή παλινδρόμησης προσαρμόζεται ικανοποιητικά στα δεδομένα.

Θέλουμε δηλαδή να απαντήσουμε στο ερώτημα: πόση είναι η μεταβλητότητα της Υ που εξηγείται από την παλινδρόμηση και πόση μένει ανερμήνευτη, δηλαδή οφείλεται σε τυχαίους παράγοντες;

Ο συντελεστής προσδιορισμού (Coefficient of Determination) χρησιμοποιείται, ουσιαστικά, ως κριτήριο καλής προσαρμογής των δεδομένων στο γραμμικό μοντέλο

$\sum\left(Y_i-\bar{Y}\right)^2 \equiv$ SST (συνολικό άθροισμα τετραγώνων, total sum of squares)

$\sum\left(Y_i-\hat{Y}\right)^2 \equiv$ SSE (ανερμήνευτο άθροισμα τετραγώνων ή άθροισμα των τετραγώνων των λαθών, unexplained sum of squares or error sum of squares)

$\sum\left(\hat{Y}_i-\bar{Y}\right)^2 \equiv$ SSR (ερμηνευόμενο άθροισμα τετραγώνων ή άθροισμα τετραγώνων παλινδρόμησης, explained sum of squares or regression sum of squares)

Συντελεστής Προσδιορισμού

SST = SSR + SSE

$$R^2 = \frac{SSR}{SST} = \frac{\sum\left(\hat{Y}_i - \bar{Y}\right)^2}{\sum\left(Y_i - \bar{Y}\right)^2} \qquad R^2 = 1 - \frac{SSE}{SST} \qquad 0 \leq R^2 \leq 1$$

Άλλοι τύποι υπολογισμού

$$R^2 = b^2 \frac{S_{XX}}{S_{YY}} \qquad R^2 = b^2 \frac{S_{XY}}{S_{YY}} \qquad R^2 = \frac{S_{XY}^2}{S_{XX} S_{YY}}$$

Συντελεστής συσχέτισης

- Ας υποθέσουμε τώρα ότι δεν είναι εύκολο να προκαθορίσουμε τις τιμές της μεταβλητής X, η X είναι δηλαδή και αυτή τυχαία μεταβλητή .
- Παρόλα αυτά είναι δυνατόν να μελετήσουμε μία παράμετρο της κατανομής αυτής που μπορεί να χρησιμοποιηθεί ως μέτρο της ισχυρότητας της γραμμικής εξάρτησης των X και Y.
- Η παράμετρος αυτή είναι ο γνωστός **συντελεστής συσχέτισης** (correlation coefficient) ο οποίος ορίζεται ως εξής
- Η εκτιμήτρια του ρ είναι $\rho = \dfrac{Cov(X,Y)}{\sigma_X \sigma_Y}$ $\quad 1 \leqslant \rho \leqslant 1$
- Για ένα συγκεκριμένο δείγμα η τιμή της εκτιμήτριας είναι $\hat{\rho} = \dfrac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2} \sqrt{(Y_i - \bar{Y})^2}}$

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2} \sqrt{\sum (Y_i - \bar{Y})^2}} = \frac{\sum (X_i Y_i - n\bar{X}\bar{Y})}{\sqrt{\left(\sum X_i^2 - n\bar{X}^2\right)} \sqrt{\left(\sum Y_i^2 - n\bar{Y}^2\right)}} = \frac{S_{XY}}{\sqrt{S_{XX}} \sqrt{S_{YY}}}$$

$r \leqslant 1$.

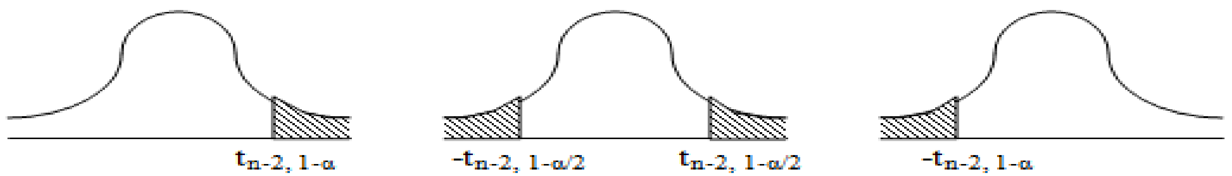
Συντελεστής συσχέτισης

$$T_o = r\sqrt{\frac{n-2}{1-r^2}} \sim t_{n-2}$$

$H_o: \rho = \rho_o$
$H_1: \rho > \rho_o$

$H_o: \rho = \rho_o$
$H_1: \rho \neq \rho_o$

$H_o: \rho = \rho_o$
$H_1: \rho < \rho_o$



Δηλαδή,

Απόρριψε την Ho αν

$t_o > t_{1-\alpha}$

Απόρριψε την Ho αν

$t_o > t_{1-\alpha/2}$
ή
$t_o < -t_{1-\alpha}$

Απόρριψε την H_o αν

$t_o < -t_{1-\alpha}$

όπου t_o είναι η τιμή της στατιστικής συνάρτησης ελέγχου T_o για το συγκεκριμένο δείγμα.

Να τονιστεί ότι το ακριβές κριτήριο ελέγχου χρησιμοποιείται μόνο για την περίπτωση που $\rho_o = 0$.

Πολλαπλή Παλινδρόμηση

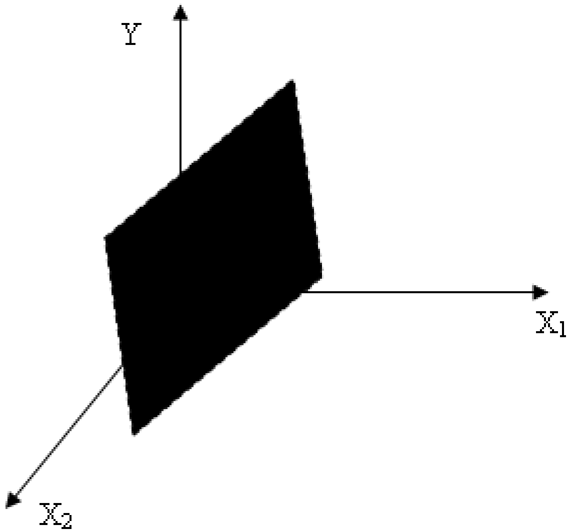
Πολλαπλή παλινδρόμηση

Το μοντέλο της απλής παλινδρόμησης που αναπτύξαμε στις προηγούμενες ενότητες αναφέρονταν σε σχέσεις που περιλάμβαναν μία μόνο ανεξάρτητη μεταβλητή. Οι περισσότερες όμως, κύρια οικονομικές, σχέσεις περιέχουν περισσότερες από μία ανεξάρτητες μεταβλητές.

Τα μοντέλα παλινδρόμησης που περιέχουν δύο ή περισσότερες ανεξάρτητες μεταβλητές ονομάζονται **μοντέλα πολλαπλής παλινδρόμησης** (multiple regression models). Υποθέτοντας ότι η συναρτησιακή σχέση που συνδέει την εξαρτημένη μεταβλητή Y με τις k ανεξάρτητες μεταβλητές είναι γραμμική, για ένα δείγμα n παρατηρήσεων θα έχουμε:

$$Y_i = a + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + e_i, \quad i = 1,2,\ldots,n$$

Πολλαπλή παλινδρόμηση



✓Το α αντιστοιχεί στο σημείο τομής του άξονα Y με το επίπεδο της παλινδρόμησης.

✓Το β1 εκφράζει τη μεταβολή της E(Y) όταν η X1 αυξάνεται κατά μία μονάδα ενώ η X2 παραμένει σταθερή.

✓Το β2 εκφράζει τη μεταβολή της E(Y) όταν η X2 αυξάνεται κατά μία μονάδα ενώ η X1 παραμένει σταθερή.

Πολλαπλή παλινδρόμηση

Έστω ότι το κατά κεφαλή εισόδημα (Y) είναι γραμμική συνάρτηση του ποσοστού του εργατικού δυναμικού που απασχολείται στην γεωργία (X_1) και του μέσου αριθμού ετών εκπαίδευσης του πληθυσμού άνω των 25 ετών (X_2). Δηλαδή η συνάρτηση είναι $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$. Τα δεδομένα προέρχονται από 15 αναπτυγμένα κράτη το 1981. Η μεταβλητή Y μετρείται σε χιλιάδες δολάρια.

Κράτος	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
y_i	6	8	8	7	7	12	9	8	9	10	10	11	9	10	11
x_{1i}	9	10	8	7	10	4	5	5	6	8	7	4	9	5	8
x_{2i}	8	13	11	10	12	16	10	10	12	14	12	16	14	10	12

Πολλαπλή παλινδρόμηση

$$\widehat{Y} = 6,26 - 0,38X_1 + 0,45X_2$$

Ερμηνεία του b1 = -0.38: Σε αύξηση του ποσοστού του εργατικού δυναμικού κατά μία μονάδα επέρχεται μείωση του κατά κεφαλή εισοδήματος κατά 0.38 μονάδες όταν διατηρείται σταθερός ο μέσος αριθμός ετών εκπαίδευσης.

Ερμηνεία του b2 = 0.45: Σε αύξηση του μέσου αριθμού ετών εκπαίδευσης κατά μία μονάδα επέρχεται αύξηση του κατά κεφαλή εισοδήματος κατά 0.38 μονάδες όταν διατηρείται σταθερό το ποσοστού του εργατικού δυναμικού που ασχολείται με τη γεωργία.

Μη Γραμμικά Μοντέλα

Μη Γραμμικά Μοντέλα

Στις εφαρμογές υπάρχουν περιπτώσεις κατά τις οποίες το μοντέλο παλινδρόμησης δεν είναι γραμμικό. Αυτό σημαίνει ότι η μέση τιμή δεν είναι γραμμικός συνδυασμός ως προς την ανεξάρτητη μεταβλητή αλλά μια οποιαδήποτε άλλη συνάρτηση.

Το μη γραμμικό μοντέλο με τη χρησιμοποίηση κατάλληλου μετασχηματισμού μετατρέπεται, τις περισσότερες φορές σε γραμμικό.

Στο μετασχηματισμένο μοντέλο εφαρμόζεται η μέθοδος των ελαχίστων τετραγώνων για την εκτίμηση των παραμέτρων του και στη συνέχεια με την αντίστροφη χρησιμοποίηση του μετασχηματισμού επιτυγχάνεται η εκτίμηση των παραμέτρων του μοντέλου.

Τέλος, μετά την εκτίμηση των παραμέτρων του αρχικού μοντέλου είναι δυνατή η χρησιμοποίηση του για προβλέψεις.

Μη Γραμμικά Μοντέλα

$$Y_i = \alpha + \beta \frac{1}{x_i} + \varepsilon_i$$

Εφαρμόζουμε το μετασχηματισμό $z_i = \frac{1}{x_i}$ οπότε το μοντέλο παίρνει τη μορφή:

$$Y_i = a + \beta z_i + \varepsilon.$$

Με εφαρμογή της μεθόδου των ελαχίστων τετραγώνων βρίσκουμε τις εκτιμήτριες ελαχίστων τετραγώνων του μετασχηματισμένου μοντέλου και κατά συνέπεια οδηγούμεθα στην εκτιμήτρια $\widehat{Y}$ της $E(Y_i / z_i)$ και μέσω αυτής, για κάθε συγκεκριμένο δείγμα $(z_1, y_1), (z_2, y_2), \ldots, (z_n, y_n)$, σε μία εκτίμηση $\widehat{y}$ της $E(Y_i / z_i)$. Δηλαδή,

$$\hat{Y} = \hat{\alpha} + \hat{\beta} z$$

και

$$\hat{Y} = a + b\, z$$

Μη Γραμμικά Μοντέλα

$$Y_i = \frac{1}{\alpha + \beta x_i + \varepsilon_i}$$

Το μοντέλο μπορεί να γραφτεί:

$$\frac{1}{Y_i} = \alpha + \beta\, x + \varepsilon_i$$

Εφαρμόζουμε το μετασχηματισμό $W_i = \frac{1}{Y_i}$ οπότε το μοντέλο παίρνει τη μορφή: $W_i = \alpha + \beta\, x_i + \varepsilon_i$. Με εφαρμογή της μεθόδου των ελαχίστων τετραγώνων βρίσκουμε τις εκτιμήτριες ελαχίστων τετραγώνων του μετασχηματισμένου υποδείγματος και κατά συνέπεια οδηγούμεθα στην εκτιμήτρια $\widehat{W}$ της $E(W_i / x_i)$ και μέσω αυτής, για κάθε συγκεκριμένο δείγμα $(x_1, w_1), (x_2, w_2), \ldots, (x_n, w_n)$, σε μία εκτίμηση $\widehat{W}$ της $E(W_i / x_i)$. Δηλαδή,

$$\widehat{W} = \widehat{a} + \widehat{\beta}\, x \quad \text{και}$$

$$\widehat{w} = \mathrm{a} + b\, x$$

Μη Γραμμικά Μοντέλα

$$Y_i = \alpha + \beta x_i^2 + \varepsilon_i$$

Εφαρμόζουμε το μετασχηματισμό $z_i = x^2$ οπότε το μοντέλο παίρνει τη μορφή: $Y_i = a + \beta z_i + \varepsilon$. Με εφαρμογή της μεθόδου των ελαχίστων τετραγώνων βρίσκουμε τις εκτιμήτριες ελαχίστων τετραγώνων του μετασχηματισμένου μοντέλου και κατά συνέπεια οδηγούμεθα στην εκτιμήτρια $\widehat{Y}$ της $E(Y_i / z_i)$ και μέσω αυτής, για κάθε συγκεκριμένο δείγμα $(z_1, y_1), (z_2, y_2), \ldots, (z_n, y_n)$, σε μία εκτίμηση $\widehat{y}$ της $E(Y_i / z_i)$.

Δηλαδή,

$$\widehat{Y} = \widehat{\alpha} + \widehat{\beta} z$$

και

$$\widehat{y} = a + b z$$

Μη Γραμμικά Μοντέλα

$$Y_i = \alpha + \beta\sqrt{x_i^{\ 2}} + \varepsilon_i$$

Εφαρμόζουμε το μετασχηματισμό $z_i = \sqrt{x_i^{\ 2}}$ οπότε το μοντέλο παίρνει τη μορφή: $Y_i = a + \beta z_i + \varepsilon$. Με εφαρμογή της μεθόδου των ελαχίστων τετραγώνων βρίσκουμε τις εκτιμήτριες ελαχίστων τετραγώνων του μετασχηματισμένου μοντέλου και κατά συνέπεια οδηγούμεθα στην εκτιμήτρια $\widehat{Y}$ της $E(Y_i / z_i)$ και μέσω αυτής, για κάθε συγκεκριμένο δείγμα $(z_1, y_1), (z_2, y_2), \ldots, (z_n, y_n)$, σε μία εκτίμηση $\widehat{y}$ της $E(Y_i / z_i)$. Δηλαδή,

$$\widehat{Y} = \widehat{a} + \widehat{\beta} z$$

και

$$\widehat{y} = a + b z$$

Παράδειγμα

Ο ακόλουθος πίνακας δίνει στοιχεία, για την ετήσια κατανάλωση κρέατος (σε τόνους) και τον πληθυσμό (σε χιλιάδες) για 9 διαφορετικά χωριά για ένα συγκεκριμένο έτος

Κρέας σε τόνους (Y)	Πληθυσμός σε χιλιάδες (X_1)
54	2
57	3
57	3
62	6
64	6
64	6
65	8
71	10
73	10

1. Να προσδιοριστεί το γραμμικό μοντέλο για την πρόβλεψη της ετήσιας κατανάλωσης κρέατος σ' ένα χωριό

   Για διευκόλυνση των υπολογισμών δίνεται ότι: Sxy=148
2. Σύμφωνα με έρευνα του Παγκόσμιου Οργανισμού Υγείας, αύξηση του πληθυσμού κατά 1000 άτομα, προκαλεί αύξηση της κατανάλωσης κρέατος κατά 2,5 τόνους. Να εξεταστεί, σε επίπεδο στατιστικής σημαντικότητας 5%, αν ισχύει η υπόθεση αυτή για τα υπό μελέτη χωριά.
3. Να κατασκευαστεί το 95% διάστημα εμπιστοσύνης για το σταθερό όρο του γραμμικού μοντέλου

Βοηθητικός Πίνακας

Y_i	X_i	$(X_i-\overline{X})$	$(Y_i-\overline{Y})$	$(X_i-\overline{X})^2$	$(Y_i-\overline{Y})^2$	$(X_i-\overline{X})Y_i$	$(X_i)^2$	$(Y_i)^2$	X_iY_i
54	2	-4	-9	16	81	-216	4	2916	108
57	3	-3	-6	9	36	-171	9	3249	171
57	3	-3	-6	9	36	-171	9	3249	171
62	6	0	-1	0	1	0	36	3844	372
64	6	0	1	0	1	0	36	4096	384
64	6	0	1	0	1	0	36	4096	384
65	8	2	2	4	4	130	64	4225	520
71	10	4	8	16	64	284	100	5041	710
73	10	4	10	16	100	292	100	5329	730
567	54			70	324	148	394	36045	3550
63	6			S_{XX}	S_{YY}	S_{XY}	ΣX_i^2	ΣY_i^2	$\Sigma X_i Y_i$

Υπολογισμοί

$\bar{X}=6$

$\bar{Y}=63$

$S_{XX}=\sum(X_i-\bar{X})^2=70$

$S_{YY}=\sum(Y_i-\bar{Y})^2=324$

$S_{XY}=S_{YX}\ \ 148$

Εναλλακτικά

$$S_{XX}=\sum X_i^2-\frac{\left(\sum X_i\right)^2}{n}=394-\frac{54*54}{9}=394-324=70$$

$$S_{YY}=\sum Y_i^2-\frac{\left(\sum Y_i\right)^2}{n}=36045-\frac{567\,^\wedge 2}{9}=324$$

$$S_{XY}=S_{YX}=\sum X_iY_i-\frac{\left(\sum X_i\right)\left(\sum Y_i\right)}{n}=148$$

Η ευθεία παλινδρόμησης της *Y* πάνω στη *X* είναι

$$a=\bar{Y}-\frac{S_{XY}}{S_{XX}}*\bar{X}=63-\frac{148}{70}*6=63-2{,}11*6=50{,}3$$

$$\beta=\frac{S_{XY}}{S_{XX}}=\frac{148}{70}\quad\Rightarrow\quad\beta=2{,}11$$

$$\hat{Y}=50.3+2{,}11X$$

$$R^2=\beta*\frac{S_{XY}}{S_{YY}}=\frac{148}{324}*2{,}11=0{,}96$$

Έλεγχος για το β

$H_o : \beta = 2.5$ versus **$H_1 : \beta \neq 2.5$**

Ως στατιστική συνάρτηση ελέγχου, κάτω από την H_o , χρησιμοποιείται η συνάρτηση

$$T_o = \frac{\widehat{\beta} - \beta_o}{S_{Y/X} \big/ \sqrt{S_{XX}}} \sim t_{n-2}$$

$$S^2_{Y/x} = \frac{1}{n-2}\left(S_{yy} - \frac{S^2_{xy}}{S_{xx}}\right) = \frac{1}{7}\left(324 - \frac{148\wedge 2}{70}\right) = 1{,}58$$

$S_{xx}=70$

$\beta o=2{,}11$

Άρα η τιμή της στατιστικής συνάρτησης ελέγχου είναι η :

$$to = \frac{2.5 - 2.11}{\sqrt{1{,}58} \big/ \sqrt{70}} = 2.59$$

Επειδή $t_{7,0.975}=2.365$

Απορρίπτουμε την Ho

Διάστημα Εμπιστοσύνης για το α

$$S_{\hat{\alpha}} = S_{Y/x}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}} = 1,25*\sqrt{\frac{1}{9}+\frac{36}{70}} = 1.25*0.79 = 0,995$$

$$\hat{\alpha} \pm t_{n-2,1-a/2} S_{\hat{\alpha}}$$

$t_{7,0.975} = 2.365$

Άρα το διάστημα εμπιστοσύνης για το α είναι

(50.3-2.365*0.995 , 50.3+2.365*0.995)= (47.94, 52.65)

Παράδειγμα

Ένας σύμβουλος επιχειρήσεων πιστεύει ότι η σχέση που συνδέει το ποσό των επιδοτήσεων (X) που λαμβάνουν οι εταιρείες από προγράμματα της Ε.Ε. και το περιθώριο κέρδους των εταιρειών (Y) είναι της μορφής $Y = \alpha + \beta x^2 + \varepsilon$. Προκειμένου να υποστηρίξει τον ισχυρισμό του πήρε δείγμα 5 εταιρειών και κατέγραψε το ποσό των επιδοτήσεων που έλαβαν (σε χρηματικές μονάδες) και το αντίστοιχο περιθώριο κέρδους (σε ποσοστό). Τα στοιχεία που συνέλεξε βρίσκονται στον πίνακα που ακολουθεί.

Πίνακας

Ποσό επιδότησης (X)	5	10	20	15	10
Περιθώριο κέρδους (Y)	1,5	2,0	2,5	3,0	2,5

Με βάση τα στοιχεία αυτά:

α. Να εκτιμηθούν οι συντελεστές του μοντέλου αυτού.

β. Αν γνωρίζουμε ότι το αντίστοιχο γραμμικό μοντέλο $Y = \alpha + \beta x + \varepsilon$ έχει συντελεστή προσδιορισμού R^2=0,53 να συγκριθεί η ερμηνευτική του ικανότητα με την αντίστοιχη του τετραγωνικού μοντέλου που χρησιμοποιήθηκε στο προηγούμενο ερώτημα.

Παράδειγμα

Υπολογισμοί για Μη γραμμικό Μοντέλο

z=sqr(X)	Y	zi-zems	yi-ymes	sqr(zi-zems)	sqr(yi-ymes)	(zi-zmes)yi
25	1,5	-145	-0,80	21025	0,64	-217,5
100	2	-70	-0,30	4900	0,09	-140
400	2,5	230	0,20	52900	0,04	575
225	3	55	0,70	3025	0,49	165
100	2,5	-70	0,20	4900	0,04	-175
850,00	11,50			86750	1,30	207,50
170	2,30			Szz	Syy	Szy

$$b = \frac{S_{ZY}}{S_{ZZ}} = \frac{207,5}{86750} = 0,0024$$

$$a = \bar{Y} - b\bar{Z} = 2,3 - 0,0024 * 170 = 1,892$$

$$\hat{y} = a + bz = 1,892 - 0,0024z$$

ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

διοικώ σημαίνει αποφασίζω

Παράδειγμα

Υπολογισμοί για Γραμμικό Μοντέλο

	X	Y	xi-xmes	yi-ymes	sqr(xi-xmes)	sqr(yi-ymes)	(xi-xmes)yi	sqr(xi)	sqr(yi)	xi*yi
	5	1,5	-7,00	-0,80	49	0,64	-10,5	25	2,25	7,5
	10	2	-2,00	-0,30	4	0,09	-4	100	4	20
	20	2,5	8,00	0,20	64	0,04	20	400	6,25	50
	15	3	3,00	0,70	9	0,49	9	225	9	45
	10	2,5	-2,00	0,20	4	0,04	-5	100	6,25	25
Sum	60,00	11,50			130,00	1,30	9,50	850,00	27,75	147,50
Average	12,00	2,30			Sxx	Syy	Sxy	Σxi^2	Σyi^2	Σxiyi

$$b = \frac{S_{XY}}{S_{XX}} = \frac{9,5}{130} = 0,073$$

$$a = \bar{Y} - b\bar{X} = 2,3 - 0,073 * 12 = 1,424$$

$$\hat{y} = a + bx = 1,424 - 0,073x$$

Παράδειγμα

Σύγκριση των R^2

Μη γραμμικό μοντέλο

$$R^2 = b\frac{S_{zy}}{S_{yy}} = 0{,}0024\frac{207{,}5}{1{,}3} = 0{,}38$$

Γραμμικό μοντέλο

$$R^2 = b\frac{S_{xy}}{S_{yy}} = 0{,}073\frac{9{,}5}{1{,}3} = 0{,}53$$

Άρα καλύτερο το γραμμικό μοντέλο