Ανάλυση

Παλινδρόμησης

Ανάλυση Παλινδρόμησης

Σε διάφορα προβλήματα της Στατιστικής το ενδιαφέρον μας εστιάζεται στην ταυτόχρονη μελέτη δύο ή περισσοτέρων μεταβλητών, για να προσδιορίσουμε με ποιο τρόπο οι μεταβλητές αυτές σχετίζονται μεταξύ τους.

Ο κλάδος της Στατιστικής που εξετάζει τη σχέση μεταξύ δύο ή περισσότερων μεταβλητών με σκοπό την πρόβλεψη της μίας εξ αυτών μέσω των άλλων, ονομάζεται **ανάλυση παλινδρόμησης** (regression analysis)

Απλή Παλινδρόμηση

Στην απλή παλινδρόμηση, χρησιμοποιούμε μόνο μια μεταβλητή Χ, και μια δεύτερη μεταβλητή Υ, η οποία μπορεί να προσεγγιστεί ικανοποιητικά από μία συνάρτηση του Χ.

Η Χ ονομάζεται **ανεξάρτητη μεταβλητή** (independent or input variable)

Η Υ ονομάζεται **εξαρτημένη μεταβλητή** (dependent or response variable)

Παράδειγμα

Η Υ εκφράζεται μέσω της Χ με την γραμμική σχέση $Y \approx 3X+5$

Απλή Παλινδρόμηση - Πολλαπλή Παλινδρόμηση

Η παλινδρόμηση στην οποία υπάρχει μόνο μια ανεξάρτητη μεταβλητή καλείται **απλή παλινδρόμηση** ενώ αν υπάρχουν περισσότερες από μία ανεξάρτητες μεταβλητές λέγεται **πολλαπλή παλινδρόμηση**.

Για την εύρεση του κατάλληλου μοντέλου για την περιγραφή της σχέσης μεταξύ δύο μεταβλητών που μας ενδιαφέρουν, συνήθως ξεκινάμε κατασκευάζοντας το **διάγραμμα διασποράς** (scatter plot) στο επίπεδο των παρατηρήσεων που διαθέτουμε. Σε ένα τέτοιο διάγραμμα οι τιμές της μεταβλητής X τοποθετούνται στον οριζόντιο άξονα και της μεταβλητής Y στον κατακόρυφο άξονα.

Η απλούστερη περίπτωση παλινδρόμησης είναι η **απλή γραμμική παλινδρόμηση** (simple linear regression), κατά την οποία χρησιμοποιούμε μόνο μια μεταβλητή X, και μια δεύτερη μεταβλητή Y η οποία μπορεί να προσεγγιστεί ικανοποιητικά από μία γραμμική συνάρτηση του X.

Αν υπάρχουν ν τιμές (X_i, Y_i), i=1,…,ν, τότε τίθεται το ερώτημα: Ποια είναι η ευθεία

$$Y = \beta_0 + \beta_1 X$$

που «ταιριάζει περισσότερο» στα ν σημεία του επιπέδου (X_i, Y_i), i=1,…,ν;

Μια μέθοδος που χρησιμοποιείται για την περιγραφή της στοχαστικής εξάρτησης των δύο μεταβλητών x (ανεξάρτητη) και y (εξαρτημένη) και προσδιορισμού των συντελεστών β_0 και β_1 της ευθείας $Y = \beta_0 + \beta_1 X$ είναι η μέθοδος των ελαχίστων τετραγώνων. Σύμφωνα με τη μέθοδο ελαχίστων τετραγώνων, η ευθεία που προσαρμόζεται καλύτερα στα

δεδομένα (ν σημεία στο επίπεδο) είναι αυτή που ελαχιστοποιεί το άθροισμα των τετραγώνων

$$\sum_{i=1}^{\nu}\left(Y_i-(\beta_0-\beta_1 X_i)\right)^2$$

Θεώρημα. *Οι εκτιμήτριες ελαχίστων τετραγώνων για τις παραμέτρους* β_0 *και* β_1 *της ευθείας* $Y = \beta_0 + \beta_1 X$ *με βάση τα ν ζεύγη* (X_i, Y_i), $i=1,\ldots,\nu$, *δίνονται από τους τύπους:*

$$\beta_1=\frac{\sum_{i=1}^{\nu}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{N}(X_i-\bar{X})^2}$$

$$\beta_0=\bar{Y}-\beta_1\bar{X}$$

όπου $\bar{X}=\frac{\sum_{i=1}^{\nu}X_i}{\nu}$ *και* $\bar{Y}=\frac{\sum_{i=1}^{\nu}Y_i}{\nu}$.

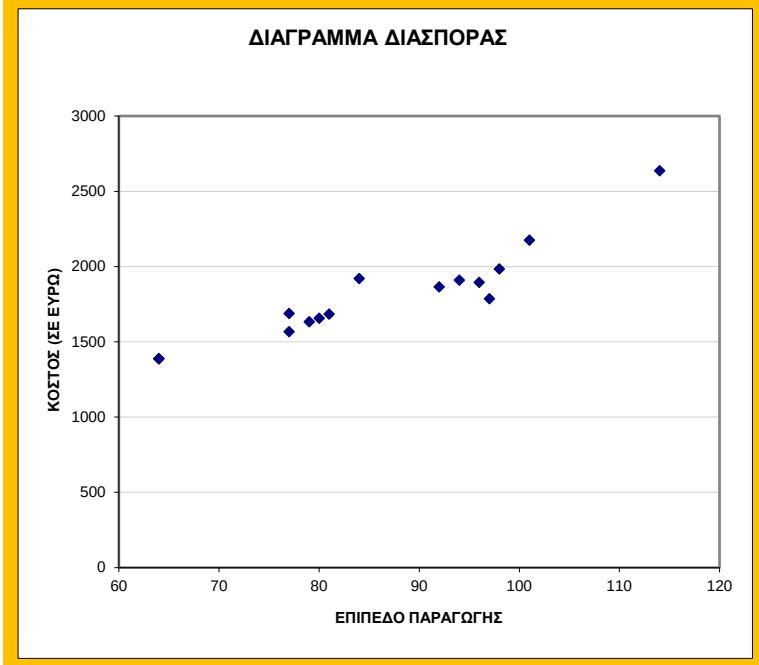
Ο ισοδύναμος τύπος υπολογισμού του συντελεστή β_1 *που δεν απαιτεί να εκφρασθεί το Y σε αποκλίσεις από το μέσο του είναι*

$$\beta_1=\frac{\sum_{i=1}^{\nu}X_iY_i-\nu\bar{X}\bar{Y}}{\sum_{i=1}^{\nu}X_i^2-\nu\bar{X}^2}$$

Παράδειγμα

Μια εταιρεία επιθυμεί να εκτιμήσει το αναμενόμενο κόστος παραγωγής όταν γνωρίζει το επίπεδο παραγωγής. Για το λόγο αυτό κατέγραψε από τα αρχεία της το επίπεδο παραγωγής και το αντίστοιχο κόστος παραγωγής ανά εβδομάδα, για 16 εβδομάδες που επιλέχθηκαν τυχαία από τους προηγούμενους έξι μήνες. Το κόστος (σε ευρώ) και το επίπεδο παραγωγής (σε τεμάχια) δίνονται στον διπλανό πίνακα.

i	Επίπεδα Παραγωγής (x_i)	Κόστος (σε χιλ. Ευρώ) (y_i)
1	114	2637
2	101	2177
3	84	1920
4	94	1910
5	98	1984
6	97	1787
7	77	1689
8	92	1866
9	96	1896
10	81	1684
11	79	1633
12	80	1657
13	77	1569
14	64	1390
15	64	1387
16	56	1289



Από το διάγραμμα προκύπτει μία θετική σχέση μεταξύ των δύο μεταβλητών. Επίσης, από το διάγραμμα φαίνεται ότι η σχέση των δυο μεταβλητών είναι γραμμική και ισχυρή.

i	Επίπεδα Παραγωγής (x_i)	Κόστος (σε χιλ. Ευρώ) (y_i)
1	114	2637
2	101	2177
3	84	1920
4	94	1910
5	98	1984
6	97	1787
7	77	1689
8	92	1866
9	96	1896
10	81	1684
11	79	1633
12	80	1657
13	77	1569
14	64	1390
15	64	1387
16	56	1289

Υπολογισμός των συντελεστών της γραμμικής εξίσωσης

$$Y = \beta_0 + \beta_1 X$$

με την μέθοδο των ελαχίστων τετραγώνων.

Α΄ Τρόπος (με αποκλίσεις από τους μέσους)

Σύμφωνα με τον πρώτο τρόπο πρέπει να χρησιμοποιήσουμε τους τύπους

$$\beta_1 = \frac{\sum_{i=1}^{\nu} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N} (X_i - \bar{X})^2}$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

Από τα στοιχεία μας μπορούμε να υπολογίσουμε τους αριθμητικούς μέσους των μεταβλητών X και Y οι οποίοι είναι

$$\bar{X} = \frac{\sum_{i=1}^{\nu} X_i}{\nu} = \frac{1354}{16} = 84{,}625$$

και

$$\bar{Y} = \frac{\sum_{i=1}^{\nu} Y_i}{\nu} = \frac{28475}{16} = 1779{,}6875.$$

Για τη διευκόλυνση των πράξεων σχηματίζουμε τον παρακάτω Πίνακα

i	X_i	Y_i	$X_i - \bar{X}$	$\left(X_i - \bar{X}\right)^2$	$\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$
1	114	2637	29,38	862,89	25183,55
2	101	2177	16,38	268,14	6505,99
3	84	1920	-0,63	0,39	-87,70
4	94	1910	9,38	87,89	1221,68
5	98	1984	13,38	178,89	2732,68
6	97	1787	12,38	153,14	90,49
7	77	1689	-7,63	58,14	691,49
8	92	1866	7,38	54,39	636,55
9	96	1896	11,38	129,39	1323,05
10	81	1684	-3,63	13,14	346,87
11	79	1633	-5,63	31,64	825,12
12	80	1657	-4,63	21,39	567,43
13	77	1569	-7,63	58,14	1606,49
14	64	1390	-20,63	425,39	8037,30
15	64	1387	-20,63	425,39	8099,18
16	56	1289	-28,63	819,39	14045,93
ΑΘΡΟΙΣΜΑ	**1354**	**28475**	**0,00**	**3587,75**	**71826,125**

Κατά συνέπεια έχουμε ότι:

ν=16

$\sum_{i=1}^{\nu}(X_i-\bar{X})(Y_i-\bar{Y})$= **71826,125**

$\sum_{i=1}^{\nu}(X_i-\bar{X})^2$ = **3587,75**

Επομένως,

$$\beta_1=\frac{\sum_{i=1}^{\nu}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{\nu}(X_i-\bar{X})^2}=\frac{71826,175}{3587,75}=20,02.$$

Για δε την εκτίμηση της σταθεράς έχουμε

$$\beta_0=\bar{Y}-\beta_1\bar{X}=1779,6875-20,02\cdot 84,625=85,5$$

Άρα η εξίσωση παλινδρόμησης είναι:

$$\hat{Y}=85,5+20X$$

Β΄ Τρόπος (χωρίς τις αποκλίσεις από τους μέσους)

Ο εναλλακτικός τύπος υπολογισμού του συντελεστή β_1 που δεν απαιτεί να εκφρασθεί το Υ σε αποκλίσεις από το μέσο του είναι

$$\beta_1 = \frac{\sum_{i=1}^{\nu} X_i Y_i - \nu \bar{X} \bar{Y}}{\sum_{i=1}^{\nu} X_i^2 - \nu \bar{X}^2}$$

Για τη διευκόλυνση των πράξεων σχηματίζουμε τον παρακάτω Πίνακα

i	X_i	Y_i	X_i^2	$X_i Y_i$
1	114	2637	12996	300618
2	101	2177	10201	219877
3	84	1920	7056	161280
4	94	1910	8836	179540
5	98	1984	9604	194432
6	97	1787	9409	173339
7	77	1689	5929	130053
8	92	1866	8464	171672
9	96	1896	9216	182016
10	81	1684	6561	136404
11	79	1633	6241	129007
12	80	1657	6400	132560
13	77	1569	5929	120813
14	64	1390	4096	88960
15	64	1387	4096	88768
16	56	1289	3136	72184
ΑΘΡΟΙΣΜΑ	**1354**	**28475**	**118170**	**2481523**

Κατά συνέπεια έχουμε ότι:

$$\beta_1 = \frac{2481523 - 16 \cdot 84,625 \cdot 1779,6875}{118170 - 16 \cdot (84,625)^2} = 20,02$$

και επομένως,

$$\beta_0 = \bar{Y} - \beta_1 \bar{X} = 1779,6875 - 20,02 \cdot 84,625 = 85,5$$

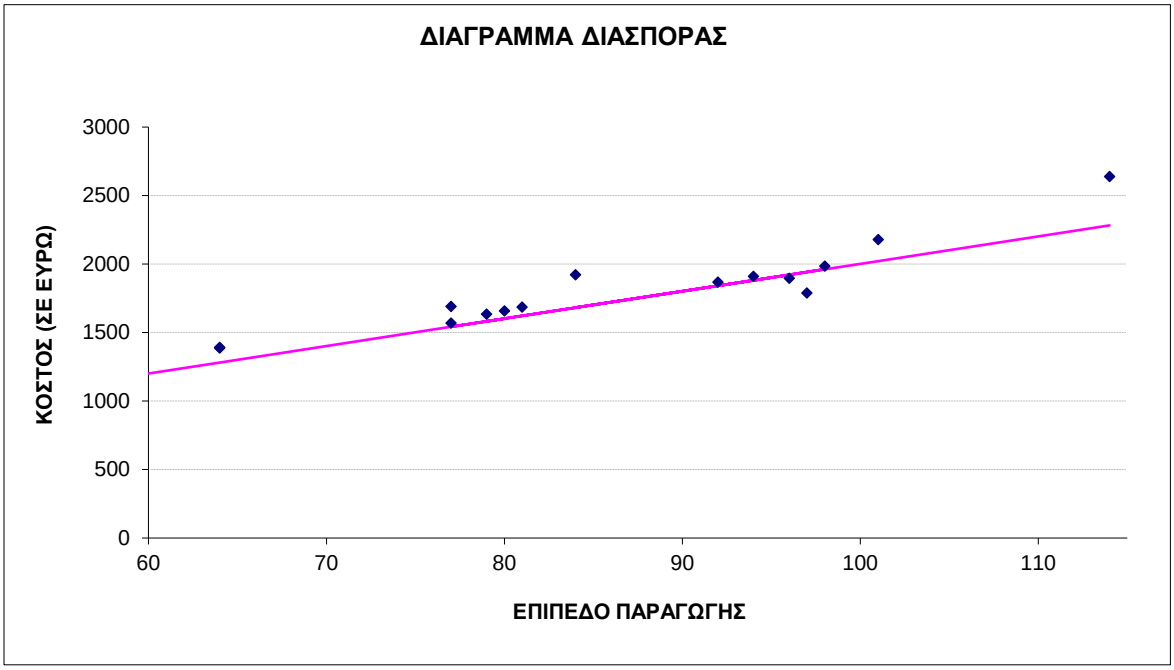
Άρα η εξίσωση παλινδρόμησης είναι:

$$\hat{Y} = 85,5 + 20X$$

Ερμηνεία των συντελεστών: Η σταθερά **β_0** εκφράζει την αναμενόμενη τιμή της Y όταν το X είναι μηδέν, δηλαδή μπορούμε να πούμε ότι όταν η επιχείρηση δεν παράγει προϊόντα τότε το σταθερό κόστος θα είναι περίπου ίσο με 85,5 ευρώ.

Ο συντελεστής κλίσης **β_1** εκφράζει την επίδραση στην αναμενόμενη τιμή της Y που προκαλεί η μεταβολή της X κατά μια μονάδα. Επομένως αν αυξηθεί η παραγωγή κατά 1 προϊόν τότε το κόστος παραγωγής θα αυξηθεί περίπου κατά 20 Ευρώ.

το Διάγραμμα Διασποράς εμπλουτισμένο με την Γραμμή Παλινδρόμησης που υπολογίσαμε



Έχοντας κατασκευάσει την Γραμμή Παλινδρόμησης, μπορούμε να προβλέψουμε το κόστος της παραγωγής για όποιον αριθμό προϊόντων θα μας ζητηθεί. Για παράδειγμα για την παραγωγή 50 προϊόντων το αναμενόμενο κόστος παραγωγής είναι ίσο με 1085,5 ευρώ, διότι:

$$\hat{Y}_{50} = 85{,}5 + 20 \cdot 50 = 1085{,}5$$

Ο **συντελεστής συσχέτισης** r δίνεται από τη σχέση:

$$r = \frac{\sum_{i=1}^{\nu}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{\nu}(X_i - \bar{X})^2 \cdot \sum_{i=1}^{\nu}(Y_i - \bar{Y})^2}}$$

Υπολογίζουμε ότι

$$\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = 1637459,4375$$

$$r = \frac{71826,125}{\sqrt{(3587,75)(1637459,4375)}} = 0,937$$

Η τιμή αυτή δηλώνει την ισχυρή γραμμική εξάρτηση (όπως αναμέναμε) μεταξύ του επιπέδου παραγωγής και του κόστους παραγωγής.

Ο **συντελεστής προσδιορισμού** R^2 ισούται με το τετράγωνο του συντελεστή συσχέτισης. Κατά συνέπεια, στην περίπτωσή μας θα έχουμε ότι:

$$R^2 = (0,937)^2 = 0,878.$$

Η τιμή αυτή δηλώνει ότι το 87,8% της μεταβλητότητας του κόστους (Y) ερμηνεύεται από το αριθμό των παραγόμενων προϊόντων δηλαδή το επίπεδο παραγωγής (X).