Προβλήματα

Φροντιστήριο #2

Απλή γραμμική παλινδρόμηση

Ενδιαφερόμαστε να μελετήσουμε τη σχέση μεταξύ δύο μεταβλητών X και Y.

- Αν για κάθε τιμή x της X καθορίζεται μονοσήμαντα η τιμή y της Y τότε ορίζεται μια συνάρτηση $y = f(x)$, όπου $f(x)$ είναι ο μαθηματικός τύπος υπολογισμού της y μέσω της x.
- Για κάθε τιμή x της X η ποσότητα Y μπορεί να είναι μια τ.μ.

Στην πρώτη περίπτωση λέμε ότι υπάρχει μία συναρτησιακή ή ντετερμινιστική (deterministic, functional) σχέση μεταξύ των μεταβλητών X και Y. Στη δεύτερη περίπτωση έχουμε μία στοχαστική ή στατιστική ή μη ντετερμινιστική (stochastic, statistic, non-deterministic) σχέση.

Για τις σχέσεις της δεύτερης μορφής, η μεταβλητή X θα καλείται ανεξάρτητη ή προβλέπουσα ή εξηγούσα ενώ η Y θα καλείται εξαρτημένη ή μεταβλητή απόκρισης.

Συνηθως οι ανεξάρτητες μεταβλητές είναι αυτές στις οποίες μπορούμε να δώσουμε μία τιμή (π.χ. τιμή πώλησης ενός προϊόντος, ποσότητα διαφήμισης για έναν αγρό κλπ) και οι εξαρτημένες μεταβλητές (π.χ. αριθμός πωλήσεων παραγωγή του αγρού) επηρεάζονται από μεταβολές των ανεξάρτητων μεταβλητών.

Το πιο απλό μοντέλο ανάλυσης παλινδρόμησης είναι το γραμμικό (στοχαστικό) μοντέλο για το οποίο η μορφή εξάρτησης των μεταβλητών X και Y περιγράφεται από τη σχέση:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

όπου x είναι η τιμή που πήρε η μη-τυχαία μεταβλητή X, Y είναι η τ.μ. που περιγράφει τη μεταβλητή απόρριψης για τη συγκεκριμένη τιμή x της X και ε είναι μια τ.μ. που εκφράζει την απόκλιση της Y από το γραμμικό όρο $\beta_0 + \beta_1 x$.

Έστω $x_1, x_2, \dots, x_n$ οι τιμές που δίνουμε στην X. Καταγράφουμε τις αντίστοιχες τιμές $y_1, y_2, \dots, y_n$ που παίρνει η μεταβλητή Y. Για τα ζεύγη (x_i, y_i), $i = 1, \dots, n$ μπορούμε να γράψουμε: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1, \dots, n$

Αν παραστήσουμε τα ζεύγη (x_i, y_i) σε ένα σύστημα ορθογωνίων αξόνων θα πάρουμε το διάγραμμα διασποράς. Στο διάγραμμα αυτό, η εξίσωση $y = \beta_0 + \beta_1 x$ θα παριστάνει μία ευθεία που προσεγγίζει τα δεδομένα μας. Η παράμετρος β_0 δίνει τη θέση όπου η ευθεία τέμνει τον άξονα $y'y$ και η β_1 δίνει το συντελεστή διευθύνσης της ευθείας.

Ερώτημα: Με ποιο τρόπο μπορούμε να καθορίσουμε τις άγνωστες παραμέτρους β_0 και β_1 ώστε η ευθεία που θα προκύψει να μας δίνει όσο το δυνατόν καλύτερη περιγραφή της σχέσης εξάρτησης των X και Y;

Ο καθορισμός των β_0, β_1 λέγεται εκτίμηση των παραμέτρων ενώ οι τιμές που προκύπτουν για αυτές με την υλοποίηση της διαδικασίας εκτίμησης λέγονται εκτιμήτριες. Η ευθεία που προσεγγίζει καλύτερα τα σημεία καλείται ευθεία παλινδρόμησης της Υ στη Χ.

Για το απλό γραμμικό μοντέλο παλινδρόμησης ισχύουν οι εξής υποθέσεις: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, i = 1, \dots, n$

1. Οι ποσότητες β_0, β_1 είναι οι άγνωστες παράμετροι
2. Τα x_i είναι γνωστοί αριθμοί που συνήθως καθορίζονται από τον ερευνητή. Είναι οι τιμές της ανεξάρτητης μεταβλητής Χ.
3. Τα Y_i είναι οι τιμές της εξαρτημένης μεταβλητής Υ. Το Y_i είναι τ.μ. και με y_i συμβολίζουμε την τιμή που λαμβάνει αυτή όταν $X = x_i$. Δηλ. είναι η παρατηρούμενη τιμή της Υ όταν $X = x_i$.
4. Τα $\varepsilon_i, i = 1, \dots, n$ είναι τυχαία σφάλματα με μέση τιμή 0 και διασπορά σ^2, δηλ. $E(\varepsilon_i) = 0$

$$Var(\varepsilon_i) = \sigma^2.$$

5. Τα σφάλματα $\varepsilon_i, \varepsilon_j$ που αντιστοιχούν σε διαφορετικές μετρήσεις i, j, $i \neq j$ θεωρούνται ασυσχέτιστα δηλ. $cov(\varepsilon_i, \varepsilon_j) = 0 \ \forall i \neq j$.

δηλ. $E(Y_i) = \beta_0 + \beta_1 x_i$, $Var(Y_i) = \sigma^2$

και $cov(Y_i, Y_j) = 0$

Μπορούμε να γράψουμε:

$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $Y = \beta_0 + \beta_1 x + \varepsilon$ ή

$E(Y_i) = \beta_0 + \beta_1 x_i$ ή $E(Y) = \beta_0 + \beta_1 x$

Δε γίνεται καμία υπόθεση για το είδος της κατανομής των Y_i ή των ε_i π.χ. αν είναι κανονική ή κάποια άλλη κατανομή πέραν του ότι αυτή έχει την ίδια διασπορά σ^2 για όλες τις τιμές x_i της ανεξάρτητης μεταβλητής X.

Το πλέον συνηθισμένο κριτήριο για την εκτίμηση των παραμέτρων β_0, β_1 είναι η ελαχιστοποίηση του αθροίσματος των τετραγώνων:

$$g(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

όπου $\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$, $i = 1, 2, \ldots, n$ είναι οι αποκλίσεις ή τα σφάλματα (errors, deviations).

Το άθροισμα $g(\beta_0, \beta_1)$ λέγεται άθροισμα τετραγώνων των σφαλμάτων (sum of squares of errors), ενώ οι τιμές των προκύπτων για τις παραμέτρους β_0, β_1 ελαχιστοποιώντας το άθροισμα αυτό καλούνται εκτιμήτριες ελαχίστων τετραγώνων, $\hat{\beta}_0, \hat{\beta}_1$ των β_0, β_1, αντίστοιχα.

Η συγκεκριμένη διαδικασία εκτίμησης των παραμέτρων β_0, β_1 είναι γνωστή ως μέθοδος των ελάχιστων τετραγώνων (Least Squares Method).

Ελαχιστοποιούμε την $g(\beta_0, \beta_1)$ ως προς β_0, β_1.

$$\frac{\partial g}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial g}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\Rightarrow \begin{cases} \sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i \quad (1) \\ \sum_{i=1}^{n} y_i x_i = \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 \end{cases}$$

Με μέθοδο οριζουσών

$$\hat{\beta}_1 = \frac{\begin{vmatrix} n & \sum y_i \\ \sum x_i & \sum y_i x_i \end{vmatrix}}{\begin{vmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{vmatrix}} = \frac{n\sum y_i x_i - \left(\sum y_i\right)^2}{n\sum x_i^2 - \left(\sum x_i\right)^2} \quad (2)$$

Από (1) $\Rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ (3)

Άρα $\hat{\beta}_0, \hat{\beta}_1$ δίνονται από (3) Λ (2).

Παραδείγματα (1) Έστω $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ η ευθεία ελαχίστων τετραγώνων που βρέθηκε με βάση τη ζεύξη σημείων (x_i, y_i), $i = 1, \ldots, n$. Έστω $\hat{y}_i$ οι εκτιμημένες τιμές $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, $i = 1, 2, \ldots, n$. Να δειχθεί ότι ο μέσος όρος των $\hat{y}_i$, $i = 1, \ldots, n$ είναι ίσος με $\bar{y}$.

Λύση Επειδή $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ έχουμε

$$\hat{y}_i = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x}), \quad i = 1, \ldots, n$$

Αθροίζοντας για κάθε $i = 1, 2, \ldots, n$ έχουμε:

$$\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} \left(\bar{y} + \hat{\beta}_1 (x_i - \bar{x})\right) = \sum_{i=1}^{n} \bar{y} + \hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x})$$

Όμως $\sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = \sum_{i=1}^{n} x_i - n\bar{x}$

$= n\bar{x} - n\bar{x} = 0$

Άρα $\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} \bar{y} = n\bar{y} = \sum_{i=1}^{n} y_i$

Αν διαιρέσουμε τη τελευταία σχέση με n, έχουμε:

$$\frac{1}{n} \sum_{i=1}^{n} \hat{y}_i = \frac{1}{n} \sum_{i=1}^{n} y_i \Leftrightarrow \frac{1}{n} \sum_{i=1}^{n} \hat{y}_i = \bar{y}$$

δηλαδή ο μέσος όρος των $\hat{y}_i, i=1,2,\ldots,n$ είναι ίσος με $\bar{y}$.

Παρατήρηση: Αν θέσουμε $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

μπορούμε να γράψουμε:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \bar{y} - \frac{S_{xy}}{S_{xx}} \bar{x}$$

Επίσης, μπορούμε να γράψουμε:

$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x}) y_i$ και για τη τ.μ. Y

$S_{xY} = \sum_{i=1}^{n} (x_i - \bar{x}) Y_i$. Άρα $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) Y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \sum_{i=1}^{n} c_i Y_i$

όπου $c_i = \dfrac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \dfrac{x_i-\bar{x}}{S_{xx}}$

Με αυτό τον μετασχηματισμό για το $\hat{\beta}_1$ μπορούμε να δείξουμε ότι $E(\hat{\beta}_1) = \beta_1$ δηλαδή ότι $\hat{\beta}$ είναι αμερόληπτη εκτιμήτρια της παραμέτρου β_1, διότι,

$$E(\hat{\beta}_1) = E\left(\sum_{i=1}^{n} c_i Y_i\right) = \sum_{i=1}^{n} c_i E(Y_i)$$

$$= \sum_{i=1}^{n} c_i(\beta_0 + \beta_1 x_i) = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i$$

Όμως $\sum_{i=1}^{n} c_i = \dfrac{\sum_{i=1}^{n}(x_i-\bar{x})}{S_{xx}} = \dfrac{n\bar{x} - n\bar{x}}{S_{xx}} = 0$

Επίσης

$$\sum_{i=1}^{n} c_i x_i = \frac{\sum_{i=1}^{n}(x_i-\bar{x})x_i}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x}+\bar{x})}{S_{xx}}$$

$$= \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{S_{xx}} + \frac{\bar{x}\sum_{i=1}^{n}(x_i-\bar{x})}{S_{xx}} = \frac{S_{xx}}{S_{xx}} + 0 = 1$$

Άρα $E(\hat{\beta}_1) = \beta_0 \cdot 0 + \beta_1 1 = \beta_1$

Αυόμη, $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{x}$ και $\hat{\beta}_1 = \sum_{i=1}^{n} c_i Y_i$

Άρα $\hat{\beta}_0 = \bar{Y} - \left(\sum_{i=1}^{n} c_i Y_i\right)\bar{x} = \sum_{i=1}^{n}\left(\frac{1}{n}\right)Y_i - \sum_{i=1}^{n} c_i \bar{x} Y_i$

$$= \sum_{i=1}^{n}\left(\frac{1}{n} - c_i\bar{x}\right)Y_i = \sum_{i=1}^{n} d_i Y_i \quad \text{όπου}$$

$$d_i = \left(\frac{1}{n} - c_i\bar{x}\right) = \frac{1}{n} - \frac{(x_i-\bar{x})\bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

Συνεπώς,

$$E(\hat{\beta}_0) = E(\bar{Y} - \hat{\beta}\bar{x}) = E(\bar{Y}) - \bar{x}E(\hat{\beta}) = \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} = \beta_0$$

Άρα, $\hat{\beta}_0$ αμερόληπτος εκτιμητής της παραμέτρου β_0. ⊠

Εφαρμογή Ένα υλικό συσκευάζεται από ένα εργοστάσιο σε μεγάλα κουτιά ο αριθμός των οποίων ποικίλει ανάλογα με την παραγγελία. Ο πίνακας δίνει τον αριθμό των κουτιών που συσκευάστηκαν ώστε να καλυφθούν οι παραγγελίες του εργοστασίου και οι εργατώρες που χρειάστηκαν για 10 πρόσφατες παραγγελίες που εκτελέστηκαν.

(α) Ποια από τις δύο μεταβλητές (αριθμός κουτιών, εργατώρες) μπορεί να θεωρηθεί ως ανεξάρτητη μεταβλητή X και ποια ως εξαρτημένη Y;

(β) Να σχεδιαστεί το αντίστοιχο διάγραμμα διασποράς των δεδομένων.

(γ) Να υπολογιστεί με τη μέθοδο ελαχίστων τετραγώνων η ευθεία γραμμικής παλινδρόμησης της Y πάνω στη X και να χαραχθεί στο αντίστοιχο διάγραμμα διασποράς.

(δ) Να δοθεί ερμηνεία της τιμής $\hat{\beta}_1$ και του σταθερού όρου $\hat{\beta}_0$ της ευθείας παλινδρόμησης.

#κουτιών	60	40	120	160	80	100	100	140	180	70
Εργατώρες	230	161	365	515	263	335	335	464	587	245

Λύση (α) Ως ανεξάρτητη μεταβλητή (X) θα θεωρήσουμε τον αριθμό των κουτιών ενώ ως εξαρτημένη (Y) τις εργατώρες που χρειάζονται για τη συσκευασία τους. Προφανώς, στόχος είναι να μπορέσουμε να κατασκευάσουμε ένα μοντέλο με το οποίο θα μπορούμε να προβλέπουμε

τις εργατώρες που θα χρειαστούν όταν δίνεται ο αριθμός των μονάδων που πρέπει να συσκευαστούν.

(β) Από το διάγραμμα διασποράς (βλέπε Σχήμα) μπορούμε να πούμε ότι η σχέση των μεταβλητών X και Y είναι κατά προσέγγιση γραμμική.

(γ) Από τα δεδομένα μας, για $X :=$ # μονάδων και $Y :=$ εργατώρες, βρίσκουμε:

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{10 \cdot 422250 - 1050 \cdot 3500}{10 \cdot 128500 - 1050^2} = 3$$

$$\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10} 3500 - 3\frac{1}{10} 1050$$

$$= 350 - 315 = 35.$$

Άρα, η ευθεία παλινδρόμησης της Y πάνω στη X είναι

η $y = 35 + 3x$ ή $\hat{Y} = 35 + 3x$.

Το σχήμα με τη ευθεία παλινδρόμησης μαζί με το διάγραμμα διασποράς δίνεται παρακάτω.

(δ) Ερμηνεία $\hat{\beta}_1$: Για κάθε επιπλέον μονάδα σε μια παραγγελία που φτάνει στο εργοστάσιο χρειάζονται 3 επιπλέον εργατώρες για τη συσκευασία των.

Ερμηνεία $\hat{\beta}_0$: Χρειάζονται επίσης 35 εργατώρες για προπαρασκευαστικές δουλειές πριν αρχίσει η συσκευασία των μονάδων που απαιτούνται για την κάλυψη των παραγγελιών. Αυτό προκύπτει από το γεγονός ότι για παραγγελία $x = 0$ μονάδων η ευθεία παλινδρόμησης δίνει $\hat{Y} = \hat{\beta}_0 = 35$.

Κάποια γενικά σχόλια:

① Επισημαίνουμε ότι ο όρος "γραμμικό" για το χαρακτηρισμό του μοντέλου αναφέρεται στις παραμέτρους και όχι στις μεταβλητές.

Για παράδειγμα, το μοντέλο $Y = \beta_0 + \beta_1 \frac{1}{X} + \varepsilon$ είναι επίσης γραμμικό και μπορούμε να υπολογίσουμε τους ε.ε.τ. των β_0, β_1:

$$\hat{\beta}_1 = \frac{n \sum_{i=1}^{n} \frac{y_i}{x_i} - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} \frac{1}{x_i}}{n \sum_{i=1}^{n} \frac{1}{x_i^2} - \left(\sum_{i=1}^{n} \frac{1}{x_i}\right)^2}$$

και $\hat{\beta}_0 = \bar{y} - \frac{\hat{\beta}_1}{n} \sum_{i=1}^{n} \frac{1}{x_i}$.

Επίσης, το μοντέλο $Y = e^{\beta_1 X + \varepsilon}$ είναι επίσης γραμμικό ως προς την παράμετρο β_1 γιατί μπορεί να γραφεί:

$\ln(Y) = \beta_1 X + \varepsilon$ και μπορούν και για αυτό το μοντέλο να υπολογιστούν τα ε.ε.τ της β_1.

② Στην απλή γραμμική παλινδρόμηση της Y στη X η ευθεία ελαχίστων τετραγώνων περνά από το σημείο $(\bar{x}, \bar{y})$ διότι $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ και άρα

$$\hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \bar{x} = \bar{y}.$$

③ Ισχύει ότι $\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$ διότι:

$$\sum_{i=1}^{n} \hat{\varepsilon}_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

$$= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\beta}_0 - \sum_{i=1}^{n} \hat{\beta}_1 x_i$$

$$= \sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i - n(\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 n\bar{x}$$

$$= \sum_{i=1}^{n} y_i - n\bar{y} = n\bar{y} - n\bar{y} = 0.$$