

4. ΠΑΛΙΝΔΡΟΜΗΣΗ ΤΑΞΗΣ ΜΕΓΕΘΟΥΣ (RANK REGRESSION)

4.1 Μονότονη Παλινδρόμηση (Monotonic Regression)

Από τη γραφική παράσταση των δεδομένων του προηγούμενου προβλήματος παρατηρούμε ότι τα ζευγάρια (X_i, Y_i) έχουν μια μονότονη σχέση. Η σχέση δηλαδή μεταξύ του X και του Y δεν είναι μεν γραμμική είναι όμως μονότονη. Στην πράξη πολλές φορές χρησιμοποιείται η μέθοδος της γραμμικής παλινδρόμησης-παρότι οι ενδείξεις οδηγούν στο συμπέρασμα ότι θα πρέπει να χρησιμοποιηθεί κάποια εναλλακτική μέθοδος ανάλυσης. Υπάρχει όμως μια πολύ απλή μη παραμετρική τεχνική που μας επιτρέπει να εφαρμόσουμε κάποια γραμμή στα δεδομένα η οποία να δίνει αξιόπιστες προβλέψεις. Η μέθοδος αυτή χρησιμοποιεί τεχνικές γραμμικής παλινδρόμησης και βασίζεται σε ένα μετασχηματισμό των δεδομένων σύμφωνα με την τάξη μεγέθους. Η μόνη υπόθεση που χρειαζόμαστε στην ανάλυση αυτή είναι η μονότονη σχέση. Ακριβώς λόγω της υποθέσης αυτής η μέθοδος ονομάζεται μέθοδος **Μονότονης Παλινδρόμησης** (*Monotonic regression*). Γενικότερα η μέθοδος παλινδρόμησης που εφαρμόζεται στις τάξεις μεγέθους των παρατηρήσεων, αντί στις ίδιες τις παρατηρήσεις ονομάζεται **παλινδρόμηση τάξης μεγέθους** (*rank regression*).

Η διαδικασία που χρησιμοποιούμε στη μέθοδο αυτή είναι η εξής:

1. Αντικαθιστούμε τις τιμές της ανεξάρτητης μεταβλητής X_i , $i=1,2,\dots,n$ με τις αντίστοιχες τάξεις μεγέθους (*Ranks*) από το ένα έως το n . Συμβολίζουμε αυτές τις τάξεις μεγέθους με R_{X_i} .

Αντικαθιστούμε τις τιμές των Y_i , $i=1,2,\dots,n$ με τις αντίστοιχες τάξεις μεγέθους από το 1 έως το n . Συμβολίζουμε τις τάξεις μεγέθους των Y με R_{Y_i} (Στην περίπτωση ίσων τιμών χρησιμοποιούμε την μέθοδο της μέσης τάξης μεγέθους).

2. Υπολογίζουμε τους συντελεστές ελαχίστων τετραγώνων α και β για τις τάξεις μεγέθους (*ranks*). Για τις τάξεις μεγέθους οι υπολογισμοί απλοποιούνται στους εξής τύπους:

$$\hat{\beta} = \frac{\sum R_{x_i} R_{y_i} - n(n+1)^2/4}{\sum R_{x_i^2} - n(n+1)^2/4}$$

$$\hat{\alpha} = \frac{n+1}{2} - \hat{\beta} \frac{n+1}{2}$$

3. Η εξίσωση παλινδρόμησης εκφράζεται μέσω των τάξεων μεγέθους όπως στην εξίσωση που ακολουθεί:

$$\hat{R}_{Y_i} = \hat{\alpha} + \hat{\beta} R_{X_i}$$

4.2 Προβλέψεις στη Μονότονη Παλινδρόμηση

Μέχρι τώρα στην μονότονη παλινδρόμηση εφαρμόστηκαν οι συνήθεις μέθοδοι γραμμικής παλινδρόμησης. Η μέθοδος διαφοροποιείται από τη μέθοδο της γραμμικής παλινδρόμησης όταν ο ερευνητής θέλει να προχωρήσει σε προβλέψεις. Αυτό γιατί η εξίσωση παλινδρόμησης είναι εκφρασμένη μέσω των τάξεων μεγέθους και επομένως τάξεις μεγέθους θα πρέπει να χρησιμοποιηθούν και πάλι αν μας ενδιαφέρει να προβλέψουμε την τιμή Y_0 της εξαρτημένης μεταβλητής για μία τιμή X_0 της ανεξάρτητης μεταβλητής. Είναι προφανές ότι για να χρησιμοποιήσουμε την μονότονη παλινδρόμηση θα πρέπει πρώτα να αντιστοιχίσουμε μια τάξη μεγέθους στην τιμή X_0 της ανεξάρτητης μεταβλητής για την οποία θέλουμε να κάνουμε πρόβλεψη.

Για να γίνει αυτό συγκρίνουμε την τιμή X_0 με τις παρατηρηθείσες τιμές X_i και αντιστοιχούμε μια τάξη μεγέθους στο X_0 σε σχέση με τις τιμές του δείγματος που είχαμε παρατηρήσει. Η τάξη μεγέθους που δίνεται στο X_0 δεν είναι απαραίτητο να είναι ακέραιος αριθμός.

Αν $X_0 \leq X_1$ τότε θεωρούμε ως τάξη μεγέθους του X_0 το R_{X_1} .

Αν $X_0 \geq X_n$ τότε στο X_0 δίνουμε ως τάξη μεγέθους το R_{X_n} .

Αν $X_0 = X_1$, $i = 2, \dots, n-1$ θεωρούμε ως τάξη μεγέθους του X_0 το R_{X_i} .

Αν $X_i < X_0 < X_{i+1}$ όπου X_i και X_{i+1} είναι δυο διαδοχικές τιμές του δείγματος τότε θεωρούμε ως τάξη μεγέθους του X_0 το R_{X_0} που υπολογίζεται με την μέθοδο της γραμμικής παρεμβολής και δίνεται από τον τύπο:

$$R_{X_0} = R_{X_i} + \left[\frac{X_0 - X_i}{X_{i+1} - X_i} \right] (R_{X_{i+1}} - R_{X_i})$$

Αφού καθορίσουμε την τιμή R_{X_0} (την τάξη μεγέθους της τιμής X_0) την αντικαθιστούμε στην εξίσωση παλινδρόμησης που αναφέρεται στις τάξεις μεγέθους και μέσω αυτής υπολογίζουμε την προβλεπόμενη τάξη μεγέθους για το Y_0 την οποία συμβολίζουμε με $\hat{R}_{Y_0}$. Η τιμή αυτή είναι η

$$\hat{R}_{Y_0} = \hat{\alpha} + \hat{\beta} R_{X_0}.$$

Το τελευταίο βήμα είναι η μετατροπή της προβλεφθείσας τάξης μεγέθους του Y_0 $\hat{R}_{Y_0}$ σε μια προβλεπόμενη τιμή του Y_0 , $\hat{Y}_0$. Αυτό γίνεται με σύγκριση της προβλεφθείσας τάξης μεγέθους $\hat{R}_{Y_0}$, (η οποία είναι δυνατόν να μην είναι ακέραιος αριθμός), με τις παρατηρηθείσες τάξεις μεγέθους των τιμών του Y του δείγματος και την εν συνεχεία επιλογή μιάς τιμής $\hat{Y}_0$ για το $\hat{R}_{Y_0}$ η οποία αναλογεί με τις παρατηρηθείσες τιμές του Y στο δείγμα και τις αντίστοιχες τάξεις μεγέθους.

Συγκεκριμένα:

$$\text{Αν } \hat{R}_{Y_0} \leq R_{Y_1} \text{ τότε } \hat{Y}_0 = Y_1.$$

$$\text{Αν } \hat{R}_{Y_0} \geq R_{Y_n} \text{ τότε } \hat{Y}_0 = Y_n.$$

$$\text{Αν } \hat{R}_{Y_0} = R_{Y_i} \text{ τότε } \hat{Y}_0 = Y_i.$$

Αν $R_{Y_i} < \hat{R}_{Y_0} < R_{Y_{i+1}}$, αν δηλαδή $\hat{R}_{Y_0}$ το είναι μεταξύ δύο διαδοχικών τάξεων μεγέθους των Y , τότε χρησιμοποιώντας την γραμμική παρεμβολή, θεωρούμε ως προβλεπόμενη τιμή $\hat{Y}_0$ για το X_0 την τιμή

$$\hat{Y}_0 = Y_i + \left[\frac{\hat{R}_{Y_0} - R_{Y_i}}{R_{Y_{i+1}} - R_{Y_i}} \right] (Y_{i+1} - Y_i)$$

Παράδειγμα: Στο προηγούμενο παράδειγμα των πωλήσεων και των εξόδων διαφήμισης προσδιορίζουμε την εξίσωση παλινδρόμησης που είναι βασισμένη στις τάξεις μεγέθους υπολογίζοντας τους συντελεστές της. Έχουμε

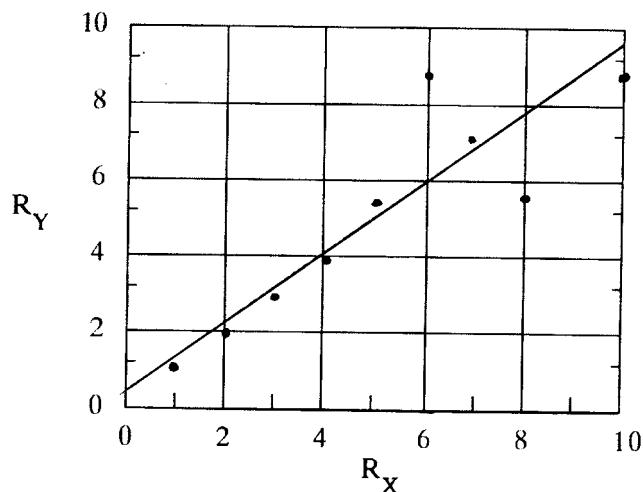
$$b = \frac{376.5 - 10(11)^2 / 4}{385 - 10(11)^2 / 4} = 0.897 \quad \text{και}$$

$$a = \frac{11}{2} - 0.897 \left[\frac{11}{2} \right] = 0.567$$

Επομένως, η εκτιμώμενη εξίσωση παλινδρόμησης των τάξεων μεγέθους είναι η:

$$\hat{R}_{Y_i} = 0.567 + 0.897 R_{X_i}$$

Αν κανείς κατασκευάσει το γράφημα των τάξεων μεγέθους του δείγματος για τις τιμές του X σε σχέση με τις τάξεις μεγέθους του δείγματος για τις τιμές του Y θα παρατηρήσει ότι οι τάξεις μεγέθους φαίνεται να ακολουθούν για γραμμική παλινδρόμηση παρότι κάτι τέτοιο δεν συμβαίνει για τις αρχικές τιμές του δείγματος



**Διάγραμμα διασποράς τάξεων μεγέθους πωλήσεων (Y)
σε σχέση με τάξεις μεγέθους διαφήμισης (X)
& η ευθεία ελαχίστων τετραγώνων**

Εστω ότι ένας απο τους διαφημιζόμενους στο συγκεκριμένο ραδιοφωνικό σταθμό αποφασίσει να αυξήσει κατά 25% τα έξοδα του για διαφήμιση στο σταθμό αυτό. Τι αύξηση θα πρέπει να περιμένει στις πωλήσεις του αν χρησιμοποιήσει την μέθοδο παλινδρόμησης που βασίζεται στις τάξεις μεγέθους;

Λύση:

1. Μετατροπή του $X_0 = 25$ σε τάξη μεγέθους.

Επειδή $X_0 = 25$ βρίσκεται μεταξύ των τιμών $X_i = 21$ (που έχει τάξη μεγέθους 5) και $X_{i+1} = 31$ (που έχει τάξη μεγέθους 6) η τάξη μεγέθους για το X_0 είναι:

$$R_{X_0} = 5 + \left(\frac{25 - 21}{31 - 21} \right) (6 - 5) = 5.4$$

2. Πρόβλεψη της τάξης μεγέθους του Y_0 .

Αντικαθιστώντας την τιμή του $R_{X_0} = 5.4$ στην εκτιμηθείσα εξίσωση παλινδρόμησης για τις τάξεις μεγέθους βρίσκουμε

$$\hat{R}_{Y_0} = 0.567 + .897(5.4) = 5.41$$

3. Μετατροπή του $\hat{R}_{Y_0}$ σε αντίστοιχη τιμή για το $\hat{Y}_0$.

Η τάξη μεγέθους 5.41 βρίσκεται μεταξύ των παρατηρηθεισών τιμών για τις τάξεις μεγέθους $R_{Y_i} = 4$ (που αντιστοιχεί στην τιμή $Y_i = 18$) και $R_{Y_{i+1}} = 5.5$ (που αντιστοιχεί στην τιμή $Y_{i+1} = 33$). Επομένως έχουμε ότι:

$$\hat{Y}_0 = 18 + \left[\frac{5.41 - 4}{5.5 - 4} \right] (33 - 18) = 32.1$$

Αυτή είναι η προβλεπόμενη τιμή για το Y δοθέντος ότι $X_0 = 25$. Επομένως, σε μια αύξηση κατά 25% του προϋπολογισμού για διαφήμιση μπορεί κανείς να αναμένει μια αύξηση στις πωλήσεις κατά 32.1%.

Οι υπολογισμοί που χρησιμοποιήθηκαν για τον καθορισμό της εξίσωσης παλινδρόμησης βασισμένης στις τάξεις μεγέθους φαίνονται στο πίνακα που ακολουθεί.

X_i	Y_i	R_{X_i}	R_{Y_i}	$R_{X_i}^2$	$R_{X_i} R_{Y_i}$
-11	-14	1	1.0	1	1.0
-1	-8	2	2.0	4	4.0
4	10	3	3.0	9	9.0
16	18	4	4.0	16	16.0
21	33	5	5.5	25	27.5
31	39	6	8.5	36	51.0
47	37	7	7.0	49	49.0
62	33	8	5.5	64	44.0
74	45	9	10.0	81	90.0
88	39	10	8.5	100	85.0
Σύνολα				385	376.5

Σημείωση: Είναι χρήσιμο να υπολογίσουμε τα κατάλοιπα που προκύπτουν από την χρησιμοποίηση της εξίσωσης παλινδρόμησης που είναι βασισμένη στις τάξεις μεγέθους και να τα συγκρίνουμε με τα κατάλοιπα που προέρχονται από τη χρήση της συνήθους μεθόδου της γραμμικής παλινδρόμησης. Ο υπολογισμός των καταλοίπων

αυτών θα μας επιτρέψει επίσης να προχωρήσουμε την ανάλυσή μας για να εξετάσουμε αν υφίσταται ισχυρή σειριακή συσχέτιση όπως είχαμε παρατηρήσει ότι συμβαίνει στο παράδειγμά μας με τη συνήθη γραμμική παλινδρόμηση. Οι υπολογισμοί φαίνονται στον πίνακα που ακολουθεί:

Πίνακας

(1) X_i	(2) Y_i	(3) R_{X_i}	(4) R_{Y_i}	(5) $\hat{R}_{Y_i}$	(6) $\hat{Y}_i$
-11	-14	1	1.0	1.46	-11.24
-1	-8	2	2.0	2.36	-1.52
4	10	3	3.0	3.26	12.08
16	18	4	4.0	4.15	19.50
21	33	5	5.5	5.05	28.50
31	39	6	8.5	5.95	34.20
47	37	7	7.0	6.85	36.30
62	33	8	5.5	7.74	37.99
74	45	9	10.0	8.64	39.56
88	39	10	8.5	9.54	43.16

Πίνακας (συνέχεια)

(7) $W_i = Y_i - \hat{Y}_i$	(8) W_{i+1}	(9) W_i^2	(10) W_{i+1}^2	(11) $W_i W_{i+1}$
-2.76	-6.48	7.62	41.99	17.88
-6.48	-2.08	41.99	4.33	13.48
-2.08	-1.50	4.33	2.25	3.12
-1.50	4.50	2.25	20.25	-6.75
4.50	-12.09	20.25	146.17	-54.41
-12.09	0.40	146.17	0.16	-4.84
0.40	0.24	0.16	0.06	0.10
0.24	5.44	0.06	29.59	1.31
5.44	-4.16	29.59	17.31	-22.63
-4.16 *				
-14.33	-15.73	252.41	262.10	-52.74

* Η τιμή αυτή δεν χρησιμοποιείται στο άθροισμα γιατί χρειαζόμαστε

$$\text{το } \sum_{i=1}^{n-1}.$$

Ο σειριακός συντελεστής συσχέτισης για τα κατάλοιπα αυτά υπολογίζεται, σύμφωνα με τα προαναφερθέντα, να είναι:

$$r_1 = \frac{-52.74 - (-14.33)(-15.73)/9}{\sqrt{[252.41 - (-14.33)^2/9][262.10 - (-15.73)^2/9]}}$$

Παρατηρούμε ότι ενώ ο σειριακός συντελεστής συσχέτισης για τα πραγματικά δεδομένα ήταν $r_1 = 0.565$ για τα μετασχηματισμένα δεδομένα που προκύπτουν από τη χρήση των τάξεων μεγέθους είναι $r_1=0.335$, τιμή που είναι σαφώς μικρότερη και αποτελεί μια επιβεβαίωση ότι το μοντέλο της μονότονης παλινδρόμησης προσφέρεται περισσότερο για την ανάλυση του συγκεκριμένου προβλήματος.

Παρατήρηση: Είναι χρήσιμο να επαναλάβουμε ότι η μέθοδος της μονότονης παλινδρόμησης είναι μια μη παραμετρική μέθοδος ανάλογη της μεθόδου των ελαχίστων τετραγώνων που όμως δεν χρειάζεται καμιά υπόθεση για ανεξαρτησία, κατανομές του πληθυσμού κλπ. Αυτό βέβαια έχει ως συνέπεια ότι δεν είναι δυνατόν να προχωρήσουμε σε συμπερασματολογία για ελέγχους υποθέσεων ή διαστήματα εμπιστοσύνης χρησιμοποιώντας την μέθοδο της μονότονης παλινδρόμησης.

Ενας τρόπος για να συγκρίνουμε τις δύο μεθόδους είναι να συγκρίνουμε τα μεγέθη των καταλοίπων. Αυτά ως γνωστόν εκφράζονται μέσω του αθροίσματος των τετραγώνων των λαθών (SSE) που ορίζεται από την σχέση

$$SSE = \sum (Y - \hat{Y}_i)^2.$$

Για το SSE που προκύπτει από την μέθοδο της παλινδρόμησης των τάξεων μεγέθους έχουμε

$$SSE = \sum_{i=1}^n W_i^2 = 269.72$$

Αντίστοιχα, το SSE που προκύπτει από την ευθεία γραμμικής παλινδρόμησης είναι:

$$SSE = \sum_{i=1}^n e_i^2 = 1179.42.$$

Η σύγκριση των δύο αυτών τιμών αποτελεί μια ακόμα ένδειξη ότι το μοντέλο της μονότονης παλινδρόμησης είναι πιο κατάλληλο για το πρόβλημά μας από το μοντέλο της απλής γραμμικής παλινδρόμησης (ένδειξη που είχαμε ήδη μετά τους υπολογισμούς των αντιστοίχων σειριακών συντελεστών συσχέτισης).

Θα πρέπει να τονισθεί ότι το μικρότερο SSE που αντιστοιχεί στην μονότονη παλινδρόμηση δεν συνεπάγεται απαραίτητα ότι το μοντέλο της μονότονης παλινδρόμησης είναι καλύτερο για να χρησιμοποιηθεί για πρόβλεψη τιμών του Y από ότι η αντίστοιχη γραμμική παλινδρόμηση αλλά ως ένδειξη ότι ίσως είναι καλύτερο.