

3. ΣΕΙΡΙΑΚΟΣ ΣΥΝΤΕΛΕΣΤΗΣ ΣΥΣΧΕΤΙΣΗΣ

Πρόβλημα: Ένας ραδιοφωνικός σταθμός ενδιαφέρεται να κάνει μια ανάλυση για τους πελάτες του που διαφημίζονται σ' αυτόν για να εξετάσει την ποσοστιαία μεταβολή των πωλήσεων Y τη χρονιά που πέρασε και την αντίστοιχη ποσοστιαία μεταβολή στις δαπάνες διαφήμισης X για τον ίδιο χρόνο. Τα σχετικά στοιχεία για τους δέκα πελάτες του συγκεκριμένου ραδιοφωνικού σταθμού παρατίθενται στον πίνακα που ακολουθεί.

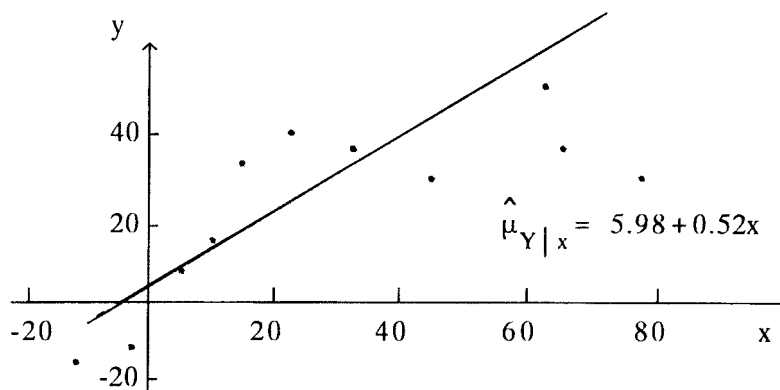
Ποσοστιαίες μεταβολές πωλήσεων (Y) και διαφήμισης (X) ενός έτους

Πελάτες	1	2	3	4	5	6	7	8	9	10
X (διαφήμιση)	4	62	31	-11	47	89	16	-1	74	21
Y (πωλήσεις)	10	33	39	-14	37	39	18	-8	45	33

Είναι εύκολο να επιβεβαιωθεί ότι η ευθεία παλινδρόμησης για τα συγκεκριμένα δεδομένα είναι η $\hat{\mu}_{Y|X} = 5.98 + 0.52x$.

Αν κανείς εξετάσει το διάγραμμα των παρατηρήσεων αυτών ή, εναλλακτικά, το διάγραμμα των καταλοίπων θα παρατηρήσει ότι μια καμπύλη θα εξέφραζε, ίσως, καλύτερα τα δεδομένα αυτά απ' ότι η ευθεία παλινδρόμησης που χρησιμοποιήθηκε. Αυτό ίσως οφείλεται σε ετεροσκεδαστικότητα των παρατηρήσεων ή, ισοδύναμα, των καταλοίπων, όπως φαίνεται στο σχήμα που ακολουθεί.

(Θα δώσουμε στη συνέχεια μία μέθοδο που βοηθά τον ερευνητή να αποφασίσει αν η ευθεία γραμμή εξηγεί επαρκώς ένα σύνολο δεδομένων).



Διάγραμμα διασποράς ποσοστιαίων μεταβολών πωλήσεων (Y) σε σχέση με ποσοστιαίες μεταβολές διαφήμισης (X) και η ευθεία ελαχίστων τετραγώνων

Ο ορισμός που ακολουθεί είναι χρήσιμος για την μελέτη της ετεροσκεδαστικότητας.

Ορισμός: Ορίζουμε ως *σειριακό συντελεστή συσχέτισης* R_1 (*serial correlation coefficient* R_1) ή *συντελεστή σειριακής συσχέτισης* για μια ακολουθία παρατηρήσεων $W_1, W_2, \dots, W_n$, τον *συντελεστή συσχέτισης* των W_i και W_{i+1} . Δηλαδή $R_1 = \rho(W_i, W_{i+1})$, ή, ισοδύναμα

$$R_1 = \frac{\sum_{i=1}^{n-1} W_i W_{i+1} - \left(\sum_{i=1}^{n-1} W_i \right) \left(\sum_{i=2}^n W_i \right) / (n-1)}{\sqrt{\left[\sum_{i=1}^{n-1} W_i^2 - \left(\sum_{i=1}^{n-1} W_i \right)^2 / n \right] \left[\sum_{i=2}^n W_i^2 - \left(\sum_{i=2}^n W_i \right)^2 / n \right]}}$$

Παρατήρηση: Είναι προφανές ότι ο σειριακός συντελεστής συσχέτισης R_1 , όπως ορίστηκε, είναι ο συντελεστής συσχέτισης κάθε παρατήρησης W_i με την παρατήρηση W_{i+1} που ακολουθεί. Για τον λόγο αυτό ο συντελεστής σειριακής συσχέτισης R_1 ονομάζεται πολλές

φορές **σειριακός συντελεστής συσχέτισης πρώτης τάξης** (first order serial correlation coefficient). Ονομάζεται επίσης σειριακός συντελεστής συσχέτισης με **υστέρηση 1** (serial correlation of lag 1). Αντίστοιχα, ορίζονται οι σειριακοί συντελεστές συσχέτισης με υστέρηση 2 και υστέρηση 3 (serial correlation of lag 2 and serial correlation of lag 3, αντίστοιχα) όταν θεωρούμε το συντελεστή συσχέτισης για τα ζεύγη W_i και W_{i+2} και για τα ζεύγη W_i και W_{i+3} , $i=1,2,\dots$. Ο ορισμός αυτός επεκτείνεται, με προφανή τρόπο, στον σειριακό συντελεστή συσχέτισης με υστέρηση v .

3.1 Σειριακή Συσχέτιση των Καταλοίπων

Σε πολλές περιπτώσεις, όταν η ευθεία παλινδρόμησης δεν είναι η καλύτερη μέθοδος για να εκφραστούν τα δεδομένα ενός προβλήματος, τα κατάλοιπα δεν έχουν ένα τυχαίο "άπλωμα". Αντίθετα, αρνητικά κατάλοιπα τείνουν να ακολουθούν άλλα αρνητικά κατάλοιπα ενώ θετικά κατάλοιπα τείνουν να ακολουθούν άλλα θετικά κατάλοιπα. Μια μέθοδος μέτρησης της συσχέτισης καταλοίπων που βρίσκονται κοντά το ένα στο άλλο (*adjacent residuals*) είναι αυτή που χρησιμοποιεί τον σειριακό συντελεστή συσχέτισης πρώτης τάξεως. Για να υπολογίσουμε το R_1 τοποθετούμε τις παρατηρήσεις της ανεξάρτητης μεταβλητής X κατά αύξουσα σειρά. Στη συνέχεια, υπολογίζουμε, με τον συνηθισμένο τρόπο το δειγματικό συντελεστή συσχέτισης (θεωρώντας ζεύγη καταλοίπων W_i με το διαδοχικό κατάλοιπο W_{i+1}) όπως ορίστηκε προηγουμένως.

Οι υπολογισμοί για τον προσδιορισμό του R_1 στο πρόβλημά μας εμφανίζονται στον πίνακα που ακολουθεί.

Πίνακας

Έξοδα Διαφήμισης	Πωλήσεις	Προβλεπόμενες Πωλήσεις	Κατάλοιπα	Υστέρηση 1
X_i	Y_i	$\hat{Y}_i = 5.98 + .52X_i$	$W_i = Y_i - \hat{Y}_i$	W_{i+1}
-11	-14	0.26	-14.26	-13.46
-1	-8	5.46	-13.46	1.94
4	10	8.06	1.94	3.69
16	18	14.31	3.69	16.09
21	33	16.91	16.09	16.89
31	39	22.11	16.89	6.57
47	37	30.43	6.57	-5.23
62	33	38.23	-5.23	0.53
74	45	44.47	0.53	-12.75
88	39	51.75	-12.75*	
Σύνολα			12.76	14.27

* Η τιμή αυτή δεν περιλαμβάνεται στο άθροισμα γιατί χρειάζεται το

$$\sum_{i=1}^{n-1}$$

Πίνακας (συνέχεια)

W_i^2	W_{i+1}^2	$W_i W_{i+1}$
203.35	181.17	191.94
181.17	3.76	-26.11
3.76	13.62	7.16
13.62	258.89	59.37
258.89	285.27	271.76
285.27	43.16	110.97
43.16	27.35	-34.36
27.35	0.28	-2.77
0.28	162.56	6.76
Σύνολα	1016.86	976.07

Από τα στοιχεία αυτά υπολογίζεται ο σειριακός συντελεστής συσχέτισης πρώτης τάξης

$$r_1 = \frac{571.1 - 12.76(14.29)/9}{\sqrt{[1016.86 - (17.76)^2 / 9][976.07 - (14.27)^2 / 9]}} = 0.565$$

Μεγάλες θετικές τιμές του R αποτελούν ένδειξη ότι το χρησιμοποιούμενο μοντέλο (στην περίπτωση μας η ευθεία παλινδρόμησης) δεν είναι ίσως το καλύτερο που μπορούμε να χρησιμοποιήσουμε (αφού αυτό αποτελεί ένδειξη ισχυρής συσχέτισης των καταλοίπων). Στην περίπτωση αυτή θα πρέπει να στραφούμε προς εναλλακτικά μοντέλα.

Σημείωση: Θα πρέπει τονιστεί ότι η συμπερασματολογία για την ανάλυση του συντελεστή γραμμικής συσχέτισης που αποδείχθηκε σε προηγούμενα κεφάλαια δεν εφαρμόζεται ακριβώς για τον σειριακό συντελεστή συσχέτισης. Παρόλα αυτά, χρησιμοποιούμε τις τιμές των πινάκων για τον συντελεστή συσχέτισης για **προσεγγιστικές** συγκρίσεις των παρατηρηθεισών τιμών.

3.2 Μοντέλο Αυτοπαλινδρόμησης Πρώτης Τάξης (First-Order Autoregressive Error Model)

Ο υπολογισμός του σειριακού συντελεστή συσχέτισης των καταλοίπων στον οποίο αναφερθήκαμε αποτελεί έναν τρόπο διαπίστωσης της σειριακής συσχέτισης των καταλοίπων που έχει ως αποτέλεσμα την μη ικανοποίηση της υπόθεσης της ανεξαρτησίας στο απλό μοντέλο γραμμικής παλινδρόμησης. Το πρόβλημα αυτό εμφανίζεται συχνά κυρίως σε προβλήματα στο χώρο της οικονομίας και των επιχειρήσεων που τα στοιχεία αποτελούν μια χρονολογική σειρά.

Όπως έχει ήδη αναφερθεί, η αυτοσυσχέτιση των καταλοίπων οφείλεται σε πολλούς λόγους. Ο κυριότερος από τους λόγους αυτούς είναι η παράλειψη μιας, ή περισσότερων, σημαντικών μεταβλητών από το μοντέλο παλινδρόμησης που εφαρμόσθηκε, όπως, πχ, η παράλειψη μιας μεταβλητής που αναφέρεται σε επιτόκια από ένα μοντέλο που χρησιμοποιείται για την πρόβλεψη του ύψους

τραπεζικών ταμείων. Στην περίπτωση αυτή, το ύψος των επιτοκίων παρουσιάζει διαφορές που έχουν κυκλική εμφάνιση και επηρεάζουν αντίστοιχα τις δανειοδοτήσεις. Ένας άλλος λόγος είναι ότι σημαντικοί τυχαίοι παράγοντες συχνά εμφανίζονται επίμονα για πολλές περιόδους. Κάτι τέτοιο συμβαίνει, για παράδειγμα, όταν η επίδραση μιας πολύ χαμηλής σοδειάς ενός αγροτικού προϊόντος έχει αντίκτυπο για πολλά χρόνια.

Το απλούστερο μοντέλο που χρησιμοποιείται για τη μελέτη του φαινομένου της αυτοσυσχέτισης των καταλοίπων είναι το *μοντέλο αυτοπαλινδρόμησης πρώτης τάξης (first-order autoregressive error model)*.

Το μοντέλο αυτό έχει τη μορφή

$$\varepsilon_t = \rho_1 \varepsilon_{t-1} + u_t$$

όπου ρ_1 είναι ο σειριακός συντελεστής συσχέτισης πρώτης τάξης των καταλοίπων του πληθυσμού $-1 < \rho_1 < 1$, και u_t είναι ανεξάρτητα τυχαία λάθη που ακολουθούν την κατανομή $N(0, \sigma^2)$.

Το μοντέλο αυτό υποθέτει ότι το κατάλοιπο ε_t για την περίοδο t περιέχει δύο συνιστώσες: μια συνιστώσα που προέρχεται από το κατάλοιπο ε_{t-1} για την προηγούμενη περίοδο (όταν $\rho_1 \neq 0$) και ένα όρο διαταραχής u_t (disturbance term) που δεν εξαρτάται από τις προηγούμενες χρονικές περιόδους.

Η παράμετρος ρ_1 , δηλαδή η παράμετρος αυτοσυσχέτισης (autocorrelation parameter) είναι ο σειριακός συντελεστής συσχέτισης της ακολουθίας των λαθών ε_t , όπως είδαμε προηγουμένως. Όπως ήδη ελέχθη, όταν το ρ_1 είναι θετικό, η αυτοσυσχέτιση είναι θετική. Όταν το ρ_1 είναι αρνητικό, η αυτοσυσχέτιση είναι αρνητική.

Στις περισσότερες χρονολογικές σειρές δεδομένων που αναφέρονται σε οικονομικά στοιχεία, ή στοιχεία επιχειρήσεων, τα οποία εμφανίζουν χρονικά συσχετισμένα κατάλοιπα, η αυτοσυσχέτιση είναι θετική. Για παράδειγμα, υψηλές πωλήσεις κάποιου μηνός που οφείλονται σε καλές συνθήκες κυκλικά εμφανιζόμενες, είναι πιθανό να συνεχίσουν να εμφανίζονται και τον επόμενο μήνα.

Στην ειδική περίπτωση όπου $\rho_1 = 0$, το μοντέλο αυτό έχει την απλή μορφή $\varepsilon_t = u_t$. Τότε, τα ε_t είναι ανεξάρτητα $N(0, \sigma^2)$,

ικανοποιούν δηλαδή την υπόθεση που απαιτείται στα απλά γραμμικά μοντέλα.

Οι συνέπειες που προκύπτουν όταν τα κατάλοιπα παρουσιάζουν αυτοσυσχέτιση πρώτης τάξης, όπως αυτή που προαναφέρθηκε και παρ'όλα αυτά χρησιμοποιείται η συνήθης μέθοδος των ελαχίστων τετραγώνων για την εφαρμογή ενός γραμμικού μοντέλου παλινδρόμησης είναι οι εξής:

α) Οι συντελεστές παλινδρόμησης ελαχίστων τετραγώνων εξακολουθούν να είναι αμερόληπτες εκτιμήτριες αλλά τείνουν να είναι σχετικά αναποτελεσματικές εκτιμήτριες.

β) Το μέσο τετραγωνικό σφάλμα MSE τείνει να υποτιμά σημαντικά την πραγματική διακύμανση των λαθών.

γ) Οι συνήθεις μέθοδοι για τα διαστήματα εμπιστοσύνης και τους ελέγχους υποθέσεων με τη χρήση των κατανομών t και F δεν έχουν πια καλή εφαρμογή.

3.3 Έλεγχος Durbin - Watson (Durbin - Watson Test)

Ο ευρύτερα χρησιμοποιούμενος έλεγχος για την διαπίστωση του κατά πόσο τα κατάλοιπα σε ένα μοντέλο γραμμικής παλινδρόμησης είναι αυτοσυσχετιζόμενα είναι ο έλεγχος *Durbin-Watson* (*Durbin-Watson Test*). Ο έλεγχος αυτός βασίζεται στο μοντέλο αυτοπαλινδρόμησης πρώτης τάξης που αναφέρθηκε προηγουμένως. Οι εναλλακτικές υποθέσεις του ελέγχου αυτού για θετική αυτοσυσχέτιση (το σύνηθες πρόβλημα σε εφαρμογές που αναφέρονται στα οικονομικά και τις επιχειρήσεις) είναι οι εξής:

$$H_0 : \rho_1 \leq 0$$

$$H_1 : \rho_1 > 0$$

Στον έλεγχο αυτό η μηδενική υπόθεση είναι ότι τα κατάλοιπα είναι ή ασυσχέτιστα ή αρνητικά συσχετισμένα, ενώ η εναλλακτική υπόθεση έχει την έννοια ότι τα κατάλοιπα είναι θετικά συσχετισμένα.

Η στατιστική συνάρτηση ελέγχου για τον έλεγχο *Durbin-Watson* συμβολίζεται συνήθως με D και βασίζεται στις διαφορές μεταξύ γειτονικών καταλοίπων, ως εξής:

$$D = \frac{\sum_{t=2}^n (\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^n \varepsilon_t^2}$$

όπου $\hat{\varepsilon}_t$ είναι η εκτιμήτρια του καταλοίπου της παλινδρόμησης για την περίοδο t και n είναι ο αριθμός των χρονικών περιόδων που χρησιμοποιούνται για την εφαρμογή του γραμμικού μοντέλου.

Η τιμή

$$d = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2}$$

αποτελεί την τιμή της στατιστικής συνάρτησης D για ένα συγκεκριμένο δείγμα με κατάλοιπα $e_1, e_2, \dots, e_n$ και χρησιμοποιείται για τον έλεγχο της υπόθεσης αυτοσυσχέτισης.

Πολλά στατιστικά προγράμματα, όπως π.χ. το SAS, περιέχουν την τιμή d της στατιστικής συνάρτησης D .

Όταν τα κατάλοιπα είναι θετικά αυτοσυσχετισμένα, τα γειτονικά κατάλοιπα θα τείνουν να έχουν παρόμοιο μέγεθος και έτσι οι διαφορές $e_t - e_{t-1}$ των γειτονικών καταλοίπων θα είναι πολύ μικρές. Επομένως, ο αριθμητής της τιμής d της στατιστικής συνάρτησης D θα είναι μικρός.

Αν τα κατάλοιπα δεν έχουν αυτοσυσχέτιση, ή είναι αρνητικά συσχετισμένα, τα γειτονικά κατάλοιπα θα τείνουν να διαφέρουν περισσότερο και, επομένως, ο αριθμητής της στατιστικής συνάρτησης ελέγχου θα έχει μεγαλύτερη τιμή.

Επομένως, μικρές τιμές της στατιστικής συνάρτησης ελέγχου D ενισχύουν την υπόθεση H_1 , ενώ μεγάλες τιμές της ενισχύουν την H_0 .

Τα ακριβή όρια για τον μονόπλευρο έλεγχο Durbin-Watson έχουν μια πολύπλοκη μορφή. Παρ' όλα αυτά, υπάρχει ένα κατώτερο όριο d_L και ένα ανώτερο όριο d_U ως όρια αποφάσεων. Οι τιμές των ορίων d_L και d_U για διάφορα μεγέθη δείγματος n , για διαφορετικούς αριθμούς ανεξαρτήτων μεταβλητών του γραμμικού μοντέλου $(p-1)$ και διαφορετικά επίπεδα εμπιστοσύνης δίνονται συνήθως σε έτοιμα κατασκευασμένους πίνακες. (Βλέπε παράρτημα).

Όταν η τιμή d της στατιστικής συνάρτησης ελέγχου είναι μικρότερη από το κατώτερο όριο d_L απορρίπτουμε τη μηδενική υπόθεση και δεχόμαστε ότι υπάρχει θετική αυτοσυσχέτιση. Όταν η τιμή d της στατιστικής συνάρτησης ελέγχου υπερβαίνει το ανώτερο όριο d_U συμπεραίνουμε ότι δεν υπάρχουν ενδείξεις για αυτοσυσχέτιση. Όταν, τέλος, η τιμή της στατιστικής συνάρτησης βρίσκεται μέσα στο διάστημα d_L και d_U δεν μπορούμε να είμαστε βέβαιοι για το ποια απόφαση να πάρουμε. Παρ' όλα αυτά, συνήθως συμπεριφερόμαστε συντηρητικά και θεωρούμε ότι ίσως υπάρχει θετική αυτοσυσχέτιση.

3.4 Έλεγχος Durbin-Watson με το MINITAB και το SAS

Για τον υπολογισμό της τιμής της στατιστικής συνάρτησης Durbin-Watson με το MINITAB χρησιμοποιούμε την υποεντολή
DW

μετά την εντολή Regress.

Στο πακέτο SAS καθορίζουμε το DW ως μια επιλογή (option) στην εντολή MODEL μετά την εντολή PROC REG (ή GLM).

Μέχρι στιγμής το MINITAB δε δίνει απευθείας καθορισμό των εκτιμητριών ελαχίστων τετραγώνων του μοντέλου αυτοπαλινδρόμησης πρώτης τάξης

$$\varepsilon_t = \rho_1 \varepsilon_{t-1} + u_t$$

στο οποίο αναφερθήκαμε.

Το SAS το επιτυγχάνει αυτό με την εντολή PROC AUTOREG. Μετά την εντολή PROC AUTOREG; και την εντολή που αναφέρεται στο μοντέλο (MODEL) χρησιμοποιούμε την εντολή

LAGLIST1;

Η εντολή αυτή αναφέρεται στο μοντέλο αυτοπαλινδρόμησης πρώτης τάξης με την έννοια ότι το λάθος (κατάλοιπο) στην περίοδο t καθορίζεται από το λάθος στην περίοδο $t-1$ (δηλαδή μία περίοδο προηγουμένως).

Στην περίπτωση που η τιμή του δειγματικού σειριακού συντελεστή συσχέτισης οδηγεί στο συμπέρασμα ότι η υπόθεση

γραμμικότητας πρέπει να απορριφθεί θα πρέπει να χρησιμοποιηθεί ένα εναλλακτικό μοντέλο παλινδρόμησης. Ένα τέτοιο μοντέλο που υιοθετείται συχνά είναι ένα μη παραμετρικό μοντέλο που θα δούμε στη συνέχεια. (Όπως είναι γνωστό, από τη Στατιστική Συμπερασματολογία, οι μη παραμετρικές μέθοδοι δεν απαιτούν κάποιες υποθέσεις για το μοντέλο του υπό μελέτη πληθυσμού).