Εφαρμογές Στατιστικών Μεθόδων σε Επιχειρηματικά Προβλήματα

Φροντιστήριο #9

Συντελεστές Μερικού Προσδιορισμού

Στην πολλαπλή παλινδρόμηση ενδιαφερόμαστε κυρίως για το πόσο σημαντική είναι η επίδραση κάθεμίας από τις ανεξάρτητες μεταβλητές $X_1, X_2, \ldots, X_k$ στην εξαρτημένη μεταβλητή Y. Έστω ότι έχουμε δύο ανεξάρτητες μεταβλητές X_1 και X_2.

Ένας τρόπος να το ελέγξουμε είναι να θεωρήσουμε ότι οι X_1 και X_2 είναι τ.μ. και να βρούμε τους δειγματικούς συντελεστές μερικού προσδιορισμού (ή τους δειγματικούς συντελεστές μερικής συσχέτισης) ανάμεσα στις μεταβλητές Y, X_1 και X_2.

Ας δούμε ένα παράδειγμα. Έστω:

$Y :=$ ζήτηση ενός προϊόντος,

$X_1 :=$ τιμή πώλησης και $X_2 :=$ τιμή πώλησης ανταγωνιστικού προϊόντος.

Έστω ότι οι εκτιμηθείσες εξισώσεις παλινδρόμησης της Y ως προς X_1 και X_2 και ως προς X_1 και X_2 μαζί δίνονται ως εξής μαζί με τους αντίστοιχους συντελεστές προσδιορισμού:

$$\hat{Y} = 55{,}871 - 0{,}220 X_1 \quad \text{με } R^2_{Y,X_1} = 0{,}76$$

$\hat{Y} = -15,814 + 0,228 X_2$ με $R^2_{Y,X_2} = 0,77$

και $\hat{Y} = 18,874 - 0,119 X_1 + 0,130 X_2$ με $R^2_{Y,(X_1,X_2)} = 0,85$.

Οι δύο μεταβλητές μαζί, οι X_1 και X_2, "εξηγούν" το 85% των μεταβολών της Y. Δηλαδή, ο συντελεστής πολλαπλού προσδιορισμού $R^2_{Y,(X_1,X_2)}$ δεν ισούται με το άθροισμα των δύο απλών συντελεστών προσδιορισμού R^2_{Y,X_1} και R^2_{Y,X_2}, όπως ενδεχομένως αναμενόταν.

Αυτό οφείλεται στο γεγονός ότι υπάρχει έντονη αρνητική γραμμική συσχέτιση μεταξύ των μεταβλητών X_1, X_2 με την τιμή του συντελεστή γραμμικής συσχέτισης $r_{X_1X_2}$ να ισούται με -0.801. Τι κάνουμε λοιπόν σε μία τέτοια περίπτωση;

Αυτό που μπορούμε να κάνουμε είναι να "μετρήσουμε" τι ποσοστό της ανερμήνευτης διασποράς της Y από τις επιδράσεις της X_1 μπορεί να ερμηνεύσει η νέα ανεξάρτητη μεταβλητή X_2! Όμως, η X_2 σχετίζεται με την Y και με την X_1.

Άρα, θα πρέπει να υπολογίσουμε τον λεγόμενο συντελεστή μερικού προσδιορισμού της X_2, $R^2_{Y,X_2|X_1}$ ο οποίος μετρά την επίδραση της X_2 στις μεταβολές της Y αφού πρώτα αφαιρέσουμε τις επιδράσεις της X_1 στην Y και στην X_2.

Πως γίνεται ο υπολογισμός του $R^2_{Y,X_2|X_1}$;

Ο $R^2_{Y,X_2|X_1}$ καλείται συντελεστής μερικού προσδιορισμού της X_2.

__1ο Βήμα:__ Εκτιμούμε την ευθεία παλινδρόμησης της Y πάνω στην X_1, $\hat{Y} = b_0 + b_1 X_1$ και **υπολογίζουμε** τα κατάλοιπα $Y - \hat{Y}$. Έτσι, "αφαιρούμε" τις επιδράσεις της X_1 στην Y.

__2ο Βήμα:__ Εκτιμούμε την ευθεία παλινδρόμησης της X_2 πάνω στην X_1, $\hat{X}_2 = c_0 + c_1 X_1$ και υπολογίζουμε τα κατάλοιπα $X_2 - \hat{X}_2$. Έτσι, "αφαιρούμε" τις επιδράσεις της X_1 στην X_2.

__3ο Βήμα:__ Ο συντελεστής __μερικού__ προσδιορισμού της X_2, $R^2_{Y,X_2|X_1}$ είναι ο __συντελεστής προσδιορισμού__ μεταξύ των μεταβλητών $Y - \hat{Y}$ και $X_2 - \hat{X}_2$.

Ο συντελεστής συσχέτισης μεταξύ των καταλοίπων $Y - \hat{Y}$ και $X_2 - \hat{X}_2$ είναι ο συντελεστής __μερικής συσχέτισης__ της X_2, $R_{Y,X_2|X_1}$ δηλαδή

$R_{Y,X_2|X_1} = r_{Y-\hat{Y},\, X_2-\hat{X}_2}$ και ο __συντελεστής μερικού προσδιορισμού__ της X_2 είναι το τετράγωνο

$R^2_{Y,X_2|X_1}$.

Αν π.χ. $R_{Y,X_2|X_1} = 0,61 \Rightarrow R^2_{Y,X_2|X_1} = 0,61^2 = 0,372$

Ερμηνεία: Μετά την αφαίρεση των επιδράσεων της τιμής πώλησης του προϊόντος (μεταβλητή X_1) στις πωλήσεις (Y) και στην τιμή των ανταγωνιστικών προϊόντων (μεταβλητή X_2), ποσοστό 37,2% της ανερμήνευτης μεταβλητότητας της ζήτησης του προϊόντος (Y) που έχει απομείνει ερμηνεύεται από τις μεταβολές των τιμών των ανταγωνιστικών προϊόντων (μεταβλητή X_2).

Όμοια, υπολογίζουμε τον συντελεστή μερικού προσδιορισμού της X_1, $R^2_{Y,X_1|X_2}$.

Έστω, $R_{Y,X_1} = r_{Y,X_1}$, $R_{Y,X_2} = r_{Y,X_2}$, $R_{X_1,X_2} = r_{X_1,X_2}$ συμβολίσουμε τους απλούς συντελεστές συσχέτισης (κατά Pearson) των μεταβλητών Y, X_1 και X_2. Τότε οι συντελεστές μερικού προσδιορισμού των X_1, X_2 είναι:

$$R^2_{Y,X_1|X_2} = \frac{\left(R_{Y,X_1} - R_{Y,X_2} R_{X_1,X_2}\right)^2}{\left[\left(1-R^2_{X_1,X_2}\right)\left(1-R^2_{Y,X_2}\right)\right]} \text{ και}$$

$$R^2_{Y,X_2|X_1} = \frac{\left(R_{Y,X_2} - R_{Y,X_1} R_{X_1,X_2}\right)^2}{\left[\left(1-R^2_{X_1,X_2}\right)\left(1-R^2_{Y,X_1}\right)\right]}$$

Συγγραμμικότητα

Έστω ότι έχουμε το μοντέλο με δύο ανεξάρτητες μεταβλητές X_1 και X_2.

Τι μπορεί να συμβαίνει όταν οι δύο μεταβλητές X_1, X_2 εμφανίζονται να μην επηρεάζουν στατιστικά σημαντικά την Y (τα t-tests δεν οδηγούν σε απόρριψη των $H_0: \beta_1 = 0$ και $H_0: \beta_2 = 0$) ενώ από μόνες τους ασκούν σημαντική επίδραση και το πιο παράδοξο είναι η επίδραση της πολλαπλής παλινδρόμησης στο σύνολό της να είναι στατιστικά σημαντική (το F-test να οδηγεί σε απόρριψη της H_0).

Σε αυτήν την περίπτωση το μυστικό είναι η σχέση των μεταβλητών X_1 και X_2.

Τότε ο $R_{X_1,X_2} = r_{X_1,X_2}$ είναι πολύ μεγάλος και αυτή είναι η αιτία που οι συντελεστές μερικής παλινδρόμησης είναι στατιστικά ασήμαντοι.

Μεγάλη τιμή του R_{X_1,X_2} οδηγεί σε μεγάλα τυπικά σφάλματα s_{b_1}, s_{b_2} και συνεπώς, τα πηλίκα $\frac{b_1}{s_{b_1}}$, $\frac{b_2}{s_{b_2}}$ είναι μικρά με αποτέλεσμα οι σ.σ. που εμπλέκονται στα t-tests να παίρνουν μικρές τιμές και να οδηγούν σε μη-απόρριψη της H_0.

Αν υπάρχει, λοιπόν, έντονη σχέση (συσχέτιση) μεταξύ δύο (ή περισσότερων) ανεξάρτητων μεταβλητών το πρόβλημα καλείται συγγραμμικότητα (ή πολυσυγγραμμικότητα). Τότε, η επίδραση της εξίσωσης πολλαπλής παλινδρόμησης στο σύνολό της είναι στατιστικά σημαντική και οι ατομικοί συντελεστές (μερικής) παλινδρόμησης είναι στατιστικά ασήμαντοι.

<u>Πώς καταλαβαίνουμε το πρόβλημα;</u>

<u>1η ένδειξη:</u> Αν το πρόσημο κάποιας b_i αλλάξει με την προσθήκη μίας νέας ανεξάρτητης μεταβλητής.

<u>2η ένδειξη:</u> Αν παρατηρήσουμε μεγάλη αλλαγή στην τιμή των συντελεστών μερικής παλινδρόμησης.

<u>3η ένδειξη:</u> Οι συντελεστές μερικής παλινδρόμησης αλλάζουν από στατιστικά σημαντικοί σε στατιστικά ασήμαντοι λόγω αύξησης των τιμών των τυπικών σφαλμάτων μετά την προσθήκη της νέας ανεξάρτητης μεταβλητής.

Συνήθως θα πρέπει να παραλείπουμε τις ανεξάρτητες μεταβλητές που συσχετίζονται έντονα με τις υπόλοιπες μεταβλητές του υποδείγματος και προκαλούν το πρόβλημα.

Αυτοσυσχέτιση (ή σειριακή συσχέτιση) των καταλοίπων

Όταν το μοντέλο γραμμικής 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. Σε τέτοιες περιπτώσεις, αν κάποιος εξετάσει το διάγραμμα των παρατηρήσεων ή εναλλακτικά, το διάγραμμα των καταλοίπων θα παρατηρήσει ότι μία καμπύλη θα εξέφραζε (ίσως καλύτερα) τα δεδομένα απ' ότι η ευθεία παλινδρόμησης που χρησιμοποιήθηκε.

Αυτό ίσως οφείλεται σε ετεροσκεδαστικότητα των παρατηρήσεων (ή ισοδύναμα των καταλοίπων). Άλλος κύριος λόγος στον οποίο μπορεί να οφείλεται η αυτοσυσχέτιση των καταλοίπων είναι η παράλειψη μίας ή περισσότερων σημαντικών μεταβλητών από το μοντέλο παλινδρόμησης που εφαρμόσθηκε όπως π.χ. η παράλειψη μίας μεταβλητής που αναφέρεται σε επιτόκια από ένα μοντέλο που χρησιμοποιείται για την πρόβλεψη του ύψους τραπεζικών ταμείων.

Το φαινόμενο της αυτοσυσχέτισης των καταλοίπων (στην ουσία, της μη-ικανοποίησης της υπόθεσης της ανεξαρτησίας των καταλοίπων εμφανίζεται συχνά κυρίως σε προβλήματα στο χώρο της οικονομίας και των επιχειρήσεων όπου τα δεδομένα συνιστούν μία χρονολογική σειρά. 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 μήνα που οφείλονται σε καλές οικονομικές συνθήκες (κυκλικά εμφανιζόμενες) είναι πιθανόν να συνεχίσουν να εμφανίζονται και τον επόμενο μήνα.

Οι κυριότερες συνέπειες της αυτοσυσχέτισης των καταλοίπων είναι οι ακόλουθες:

(α) Οι συντελεστές παλινδρόμησης με τη μέθοδο των ελαχίστων τετραγώνων εξακολουθούν να είναι αμερόληπτες εκτιμήτριες αλλά τείνουν να είναι σχετικά αναποτελεσματικές εκτιμήτριες.

(β) Το μέσο τετραγωνικό σφάλμα MSE υποτιμά σημαντικά την πραγματική διακύμανση των καταλοίπων.

(γ) Οι συνήθεις μέθοδοι για τα δ.ε. και τους ελέγχους υποθέσεων με τη χρήση των κατανομών t και F <u>δεν</u> έχουν πια <u>καλή εφαρμογή</u>.

Στόχος μας είναι να αναπτύξουμε ένα κριτήριο, μέσω του οποίου να μπορούμε να εξετάσουμε την παρουσία της αυτοσυσχετίσεως.

Θέλουμε να ελέγξουμε την:

$$H_0: \rho = 0$$

έναντι της

$$H_1: \rho \neq 0$$

ή $H_1: 0 < \rho < 1$ ή $H_1: -1 < \rho < 0$

όπου ρ είναι ο σειριακός συντελεστής συσχέτισης της ακολουθίας των καταλοίπων $\varepsilon_i, i=1,\dots,n$. Η σ.σ. ελέγχου αναφέρεται ως στατιστική των Durbin and Watson (D.W. = d) και ορίζεται ως εξής:

$$d = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2},$$ όπου e_i είναι τα κατάλοιπα της $\hat{Y}_i = b_0 + b_1 X_i, i=1,\dots,n.$

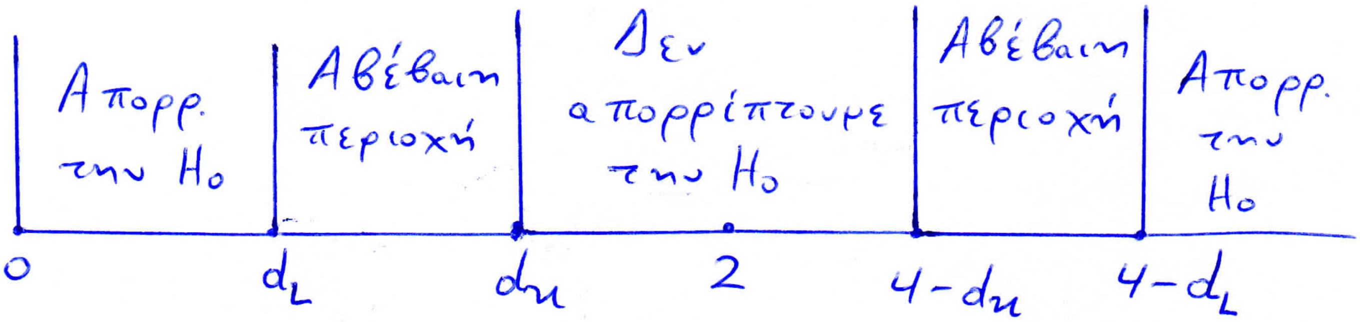
Αποδεικνύεται ότι: $d = 2(1-\rho)$

Αν $\rho = 0$ (δεν υπάρχει αυτοσυσχέτιση) υποδηλώνει $d \simeq 2.$

Αν $\rho = 1$ (θετική αυτοσυσχέτιση) υποδηλώνει $d \simeq 0.$

Αν $\rho = -1$ (αρνητική αυτοσυσχέτιση) υποδηλώνει $d \simeq 4.$

Για τον έλεγχο $H_0: \rho = 0$ vs $H_1: \rho \neq 0$ έχουμε πέντε περιοχές για την d. Αν με d_L και d_U το κάτω και το άνω φράγμα για την d, αντίστοιχα, έχουμε το διάγραμμα:



Συνεπώς, τιμές της d κοντά στο μηδέν και το 4 μας οδηγούν στην απόρριψη της H_0. Δεν απορρίπτουμε την $H_0: \rho = 0$ όταν η d παίρνει τιμές κοντά στο 2.

Οι τιμές d_L και d_U δίνονται σε κατάλληλους πίνακες τιμών (π.χ. βλέπε Πίνακα 7 με κριτικές τιμές της d για $\alpha = 0.01$ από Durdin and Watson). Ο πίνακας 7 περιέχεται στο τέλος του φρ. #9, στην επισυναπτόμενη φωτοτυπία, όπου παρουσιάζεται και ένα παράδειγμα εφαρμογής του ελέγχου D.W..

Παραδείγματα

① Δίνεται η παρακάτω εξίσωση πολλαπλής παλινδρόμησης: $\hat{Y}_i = 5837,52 - 53,22X_{1i} + 3,61X_{2i}$, $i = 1,2,\dots,20$. Δίνονται επίσης, $SSR = 33,47$, $SST = 52,09$, $D.W. = d = 1,98$, $R^2_{Y,X_1|X_2} = 0,6605$

$R^2_{Y,X_2|X_1} = 0,4728$.

(α) Να ελεγχθεί αν υπάρχει γραμμική σχέση μεταξύ εξαρτημένης και ερμηνευτικών μεταβλητών σε ε.σ.σ. $\alpha = 5\%$.

(β) Να ελεγχθεί αν υπάρχει αυτοσυσχέτιση των καταλοίπων στο υπόδειγμα για $\alpha = 0.01$. Ναι ή όχι και γιατί;

(γ) Να υπολογιστούν οι συντελεστές μερικής συσχέτισης.

(δ) Ποια από τις δύο ερμηνευτικές μεταβλητές θεωρείτε ότι συμβάλλει περισσότερο στην ερμηνευτική ικανότητα του μοντέλου και γιατί;

Λύση

(α) Ελέγχουμε σε ε.σ.σ. $\alpha = 5\%$ την

$$H_0: \beta_1 = \beta_2 = 0$$

vs

$$H_1: \exists \text{ ένα τουλάχιστον } i: \beta_i \neq 0.$$

Χρησιμοποιούμε την σ.σ. ελέγχου

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)}$$

Όμως $R^2 = \frac{SSR}{SST} = \frac{33,47}{52,09} = 0,6425 = 64,25\%$

και για $k=2$, $n=20$, $F = \frac{0,6425/2}{0,3575/(20-2-1)} = 15,279.$

Απορρίπτουμε την H_0 σε εσσ $\alpha = 5\%$, αν

$$F > F_{k,n-k-1,\alpha} = F_{2,17,0.05} = 19,44 \text{ (από πίνακες της } F)$$

Συνεπώς, η Ho <u>δεν</u> απορρίπτεται σε ε.σ.σ $\alpha = 5\%$ άρα, η εξίσωση παλινδρόμησης <u>δεν</u> εξηγεί στατιστικά σημαντικά τις μεταβολές της Y.

(β) Σε ε.σ.σ $\alpha = 0.01$, ελέγχουμε

$H_0: \rho = 0$ (δεν υπάρχει αυτοσυσχέτιση των καταλοίπων)

vs

$H_1: \rho \neq 0$ (υπάρχει αυτοσυσχέτιση των καταλοίπων).

Από τον Πίνακα 7, για $n = 20$, $k = 2$, $d_L = 0.86$, $d_U = 1.27$ για τη σ.σ. ελέγχου $D.W. = d$, όταν $\alpha = 0.01$.

Αφού $d = 1,98$ αυτό σημαίνει ότι η Ho <u>δεν</u> απορρ. και συνεπώς <u>δεν</u> υπάρχει αυτοσυσχέτιση των καταλοίπων.

(γ) Οι συντελεστές μερικής συσχέτισης είναι:

$$R_{Y,X_1|X_2} = r_{Y,X_1|X_2} = \sqrt{R^2_{Y,X_1|X_2}} =$$

$$= \sqrt{0,6605} = 0,8127$$

Επειδή $b_1 = -53,22 \Rightarrow r_{Y,X_1|X_2} = -0.8127$ δηλαδή ο συντελεστής μερικής συσχέτισης διατηρεί το πρόσημο του συντελεστή b_1.

Ομοίως $R_{Y,X_2|X_1} = r_{Y,X_2|X_1} = \sqrt{R^2_{Y,X_2|X_1}}$

$$= \sqrt{0,4728} = 0,6876.$$

Επειδή $b_2 = +3,61 \Rightarrow r_{Y,X_2|X_1} = +0,6876$

δηλαδή ο συντελεστής μερικής συσχέτισης διατηρεί το πρόσημο του συντελεστή b_2.

(δ) Επειδή, σε απόλυτους όρους, $r_{Y,X_1|X_2} > r_{Y,X_2|X_1}$ μπορούμε να πούμε ότι η μεταβλητή X_1 συμβάλλει περισσότερο από τη μεταβλητή X_2 στην ερμηνευτική ικανότητα του μοντέλου (δηλαδή συμβάλλει περισσότερο στο συνολικό ποσοστό της μεταβλητότητας της Y).

—

(2) Ως Παράδειγμα 2, επισυνάπτουμε την Ασκ. 1, από το βιβλίο του κ. Ι. Χαλικιά, στο τέλος του Κεφ. 10.

Πίνακας 7: Κριτικές τιμές της d-στατιστικές για α=.01

n	κ=1		κ=2		κ=3		κ=4		κ=5	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	.81	1.07	.70	1.25	.59	1.46	.49	1.70	.39	1.96
16	.84	1.09	.74	1.25	.63	1.44	.53	1.66	.44	1.90
17	.87	1.10	.77	1.25	.67	1.43	.57	1.63	.48	1.85
18	.90	1.12	.80	1.26	.71	1.42	.61	1.60	.52	1.80
19	.93	1.13	.83	1.26	.74	1.41	.65	1.58	.56	1.77
20	.95	1.15	.86	1.27	.77	1.41	.68	1.57	.60	1.74
21	.97	1.16	.89	1.27	.80	1.41	.72	1.55	.63	1.71
22	1.00	1.17	.91	1.28	.83	1.40	.75	1.54	.66	1.69
23	1.02	1.19	.94	1.29	.86	1.40	.77	1.53	.70	1.67
24	1.04	1.20	.96	1.30	.88	1.41	.80	1.53	.72	1.66
25	1.05	1.21	.98	1.30	.90	1.41	.83	1.52	.75	1.65
26	1.07	1.22	1.00	1.31	.93	1.41	.85	1.52	.78	1.64
27	1.09	1.23	1.02	1.32	.95	1.41	.88	1.51	.81	1.63
28	1.10	1.24	1.04	1.32	.97	1.41	.90	1.51	.83	1.62
29	1.12	1.25	1.05	1.33	.99	1.42	.92	1.51	.85	1.61
30	1.13	1.26	1.07	1.34	1.01	1.42	.94	1.51	.88	1.61
31	1.15	1.27	1.08	1.34	1.02	1.42	.96	1.51	.90	1.60
32	1.16	1.28	1.10	1.35	1.04	1.43	.98	1.51	.92	1.60
33	1.17	1.29	1.11	1.36	1.05	1.43	1.00	1.51	.94	1.59
34	1.18	1.30	1.13	1.36	1.07	1.43	1.01	1.51	.95	1.59
35	1.19	1.31	1.14	1.37	1.08	1.44	1.03	1.51	.97	1.59
36	1.21	1.32	1.15	1.38	1.10	1.44	1.04	1.51	.99	1.59
37	1.22	1.32	1.16	1.38	1.11	1.45	1.06	1.51	1.00	1.59
38	1.23	1.33	1.18	1.39	1.12	1.45	1.07	1.52	1.02	1.58
39	1.24	1.34	1.19	1.39	1.14	1.45	1.09	1.52	1.03	1.58
40	1.25	1.34	1.20	1.40	1.15	1.46	1.10	1.52	1.05	1.58
45	1.29	1.38	1.24	1.42	1.20	1.48	1.16	1.53	1.11	1.58
50	1.32	1.40	1.28	1.45	1.24	1.49	1.20	1.54	1.16	1.59
55	1.36	1.43	1.32	1.47	1.28	1.51	1.25	1.55	1.21	1.59
60	1.38	1.45	1.35	1.48	1.32	1.52	1.28	1.56	1.25	1.60
65	1.41	1.47	1.38	1.50	1.35	1.53	1.31	1.57	1.28	1.61
70	1.43	1.49	1.40	1.52	1.37	1.55	1.34	1.58	1.31	1.61
75	1.45	1.50	1.42	1.53	1.39	1.56	1.37	1.59	1.34	1.62
80	1.47	1.52	1.44	1.54	1.42	1.57	1.39	1.60	1.36	1.62
85	1.48	1.53	1.46	1.55	1.43	1.58	1.41	1.60	1.39	1.63
90	1.50	1.54	1.47	1.56	1.45	1.59	1.43	1.61	1.41	1.64
95	1.51	1.55	1.49	1.57	1.47	1.60	1.45	1.62	1.42	1.64
100	1.52	1.56	1.50	1.58	1.48	1.60	1.46	1.63	1.44	1.65

Πηγή: J. Durbin, G.S. Watson, "Testing for serial correlation in least squares regression, II." Biometrika, 1951,30

Για να βρούμε τις τιμές των d_L και d_U, πρέπει να γνωρίζουμε:

- το επίπεδο σημαντικότητας,
- το μέγεθος του δείγματος n και
- τον αριθμό των ερμηνευτικών μεταβλητών (κ), χωρίς να συμπεριλαμβάνεται η σταθερά.

Ετσι, για $\alpha=.01$, $n=30$, και $\kappa=1$, το κάτω φράγμα d_L είναι ίσο με 1.13 και το άνω φράγμα d_U με 1.26. Αν λοιπόν, η d είναι μικρότερη του 1.13 , $d<1.13$, απορρίπτουμε την H_0 και δεχόμαστε την H_1.

Παράδειγμα 7: Με τα δεδομένα του ακόλουθου πίνακα θα εξετάσουμε αν τα κατάλοιπα του υποδείγματος $Y_i = \beta_0+\beta_1 X_i+u_i$, ($i=1,2,...,15$) εμφανίζουν αυτοσυσχέτιση.

X_i	Y_i
1	2
2	2
3	2
4	1
5	3
6	5
7	6
8	6
9	10
10	10
11	10
12	12
13	15
14	10
15	11

Θα πρέπει να βρούμε τα κατάλοιπα, γι'αυτό πρώτα πρέπει να υπολογίσουμε τα $\hat{\beta}_0$ και $\hat{\beta}_1$. Από τα δεδομένα έχουμε:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{15} x_i y_i}{\sum_{i=1}^{15} x_i^2} = \frac{255}{280} = .91$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} = -.28$$

Ετσι, η εξίσωση παλινδρομήσεως είναι:

$$\hat{Y}_i = -.28 + .91X_i \quad i=1,...,15 \qquad (232)$$

οπότε τα κατάλοιπα δίνονται από τη σχέση:

$$\hat{u}_i = Y_i - \hat{Y}_i = Y_i - (-.28 + .91X_i) \qquad (233)$$

Από τη (233) υπολογίζουμε τα $\hat{u}_i^2$ και τα $(\hat{u}_i - \hat{u}_{i-1})^2$, για να τα χρησιμοποιήσουμε στον υπολογισμό της στατιστικής των Durbin-Watson - βλέπε (229). Οι υπολογισμοί των (232), $\hat{u}_i^2$ και $(\hat{u}_i - \hat{u}_{i-1})^2$ δίνονται στον πίνακα 8:

Πίνακας 8: Υπολογισμοί των $\hat{Y}_i$, $\hat{u}_i^2$ και $(\hat{u}_i - \hat{u}_{i-1})^2$

$\hat{Y}_i$	$\hat{u}_i^2$	$(\hat{u}_i - \hat{u}_{i-1})^2$
.63	1.876	----
1.54	.211	.828
2.45	.203	.828
3.36	5.570	3.648
4.27	1.612	1.188
5.18	.032	1.188
6.09	.008	.008
7	1	.828
7.91	4.368	9.548
8.82	1.392	.828
9.73	.073	.828
10.64	1.850	1.188
11.55	11.903	4.369
12.46	6.052	34.928
13.37	5.617	.008
	$\sum_{i=1}^{15} \hat{u}_i^2 = 41.767$	$\sum_{i=1}^{15} (\hat{u}_i - \hat{u}_{i-1})^2 = 60.213$

Συνεπώς:

$$d = \frac{\sum_{i=2}^{15} (\hat{u}_i - \hat{u}_{i-1})^2}{\sum_{i=1}^{15} \hat{u}_i^2} = \frac{60.213}{41.767} = 1.44$$

Από τον πίνακα των κριτικών τιμών της d-στατιστικής έχουμε:

$\alpha = .01$
$n = 15$
$\kappa = 1$
$d_L = .81$
$d_\upsilon = 1.07$

Ετσι, προκύπτουν τα διαστήματα

(α) $(0, d_L) = (0, .81)$
(β) $(d_L, 4-d_\upsilon) = (.81, 1.07)$
(γ) $(d_\upsilon, 4-d_\upsilon) = (1.07, 2.93)$
(δ) $(1-d_\upsilon, 4-d_L) = (2.93, 3.19)$
(ε) $(1-d_L, 4) = (3.19, 4)$

Η υπολογισθείσα τιμή $d=1.44$ ανήκει στο διάστημα (γ), συνεπώς δεχόμαστε την $H_0: \varrho=0$. Αρα, τα κατάλοιπα δεν αυτοσυσχετίζονται.

Θα εξετάσουμε στη συνέχεια την περίπτωση που η στατιστική των Durbin-Watson θα δείξει ότι παρουσιάζεται το πρόβλημα της αυτοσυσχετίσεως. Τι μπορούμε να κάνουμε τότε για την εκτίμηση του υποδείγματος; Το υπόδειγμά μας είναι:

$$Y_i = \beta_0 + \beta_1 X_i + u_i \quad i=1,...,n \qquad (234)$$

$$\text{με } u_i = \varrho u_{i-1} + v_i \quad |\varrho| < 1 \qquad (235)$$

και v_i ικανοποιεί τις (210) και (211).

Κεφάλαιο 10
Πολλαπλή Παλινδρόμηση

1. Ενα μεσιτικό γραφείο ενδιαφέρεται να εκτιμήσει ένα υπόδειγμα που θα επιτρέπει την πρόβλεψη της τιμής πώλησης ανεξαρτήτων διαμερισμάτων (μαιζονέτες) σε προάστιο της Αθήνας. Από δείγμα 15 διαμερισμάτων, που μεταβιβάστηκαν πρόσφατα, προέκυψαν τα εξής στοιχεία:

Διαμέρισμα	Τιμή Πώλησης (χιλ. ευρώ) (Y)	Εμβαδόν (τ.μ.) (X1)	Ηλικία (έτη) (X2)
1	422,0	200	4
2	387,0	171	12
3	378,5	145	8
4	429,5	176	1
5	395,5	193	7
6	352,0	120	32
7	379,0	155	16
8	429,5	193	2
9	392,5	159	2
10	396,0	150	3
11	433,5	190	1
12	396,5	139	1
13	372,5	154	13
14	419,0	189	3
15	384,0	159	7

1. Αναφέρατε το υπόδειγμα πολλαπλής παλινδρόμησης που προτίθεστε να εκτιμήσετε.
2. Εκτιμήστε το υπόδειγμα σας και ερμηνεύστε τις τιμές των συντελεστών μερικής παλινδρόμησης.
3. Ελέγξτε τη στατιστική σημαντικότητα των συντελεστών μερικής παλινδρόμησης σε $\alpha = 5\%$.
4. Προσδιορίστε την τιμή του συντελεστή πολλαπλού προσδιορισμού και ερμηνεύστε τον σε σχέση με το υπόδειγμα σας.
5. Εκτιμήστε τους συντελεστές μερικού προσδιορισμού και ερμηνεύστε τους.
6. Ποια είναι η προβλεπόμενη αξία πώλησης μιας μαιζονέτας 174 τ.μ. που κατασκευάστηκε πριν από μία δεκαετία;

Απαντήσεις:

1. Το υπόδειγμα πολλαπλής παλινδρόμησης που θα εκτιμηθεί είναι:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

όπου

Y = τιμή πώλησης κατοικίας (σε χιλιάδες €)
X_1 = Εμβαδόν κατοικίας (σε τ.μ.)
X_2 = Ηλικία κατοικίας (σε έτη)

2. Η εκτίμηση του υποδείγματος με τη μέθοδο των ελαχίστων τετραγώνων συμβολίζεται με

$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$ και είναι η εξής (αποτελέσματα Excel):

Regression Statistics	
Multiple R	0,902
R Square	0,814
Adjusted R Square	0,783
Standard Error	11,221
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	6617,433	3308,716	26,278	0,00004
Residual	12	1510,967	125,914		
Total	14	8128,400			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	317,998	28,264	11,251	0,000	256,416	379,580
Εμβαδόν (τ.μ.)	0,543	0,157	3,466	0,005	0,202	0,885
Ηλικία (έτη)	-1,405	0,444	-3,161	0,008	-2,373	-0,437

Δηλαδή, $\hat{Y} = 317{,}998 + 0{,}543\, X_1 - 1{,}405\, X_2$

Εάν ακολουθήσετε την κλασσική μέθοδο με το σύστημα των τριών εξισώσεων, δίνονται:

$\Sigma Y=5.967$ $\Sigma X_1=2.493$ $\Sigma X_2=112$ $\Sigma YX_1=998.159{,}5$,
$\Sigma YX_2=42.335{,}5$ $\Sigma X_1X_2=17.024{,}0$ $\Sigma X_1^2=422.085{,}0$ $\Sigma X_2^2=1.800{,}0$

Από την εξίσωση παλινδρόμησης προκύπτει ότι εάν το εμβαδόν αυξηθεί κατά ένα τ.μ. η συνολική τιμή αυξάνεται κατά μεσο όρο κατά 0,543 χιλ. € (ή 543 €), ενώ για κάθε έτος παλαιότητας η συνολική τιμή μειώνεται κατά 1,405 χιλ. € (ή 1.405 €).

3. Με βάση τα παραπάνω αποτελέσματα, όλοι οι συντελεστές μερικής παλινδρόμησης είναι στατιστικά σημαντικοί σε $\alpha = 5\%$.

4. Ο συντελεστής R^2 ισούται με 0,814 (και ο διορθωμένος με 0,783) που σημαίνει ότι, περίπου, κατά 80% η αξία πώλησης μιας μαιζονέτας εξαρτάται από το εμβαδόν και την ηλικία του ακινήτου.

5. Οι συντελεστές μερικού προσδιορισμού θα εκτιμηθούν ως εξής: Από τα δεδομένα έχουμε

$R_{Y,X_1} = 0{,}812$, $R_{Y,X_2} = -0{,}793$, και $R_{X_1,X_2} = -0{,}582$

$$\begin{aligned} R^2_{Y,X_1/X_2} &= (R_{Y,X_1} - R_{Y,X_2}\, R_{X_1,X_2})^2 / [(1-R^2_{X_1,X_2})\cdot(1-R^2_{Y,X_2})] = \\ &= [0{,}812 - (-0{,}793)\cdot(-0{,}582)]^2 / [(1-(-0{,}582)^2)\cdot(1-(-0{,}793)^2)] = \\ &= 0{,}500 \end{aligned}$$

$$\begin{aligned} R^2_{Y,X_2/X_1} &= (R_{Y,X_2} - R_{Y,X_1}\, R_{X_1,X_2})^2 / [(1-R^2_{X_1,X_2})\cdot(1-R^2_{Y,X_1})] = \\ &= [-0{,}793 - (0{,}812)\cdot(-0{,}582)]^2 / [(1-(-0{,}582)^2)\cdot(1-(0{,}812)^2)] = \\ &= 0{,}454 \end{aligned}$$

Η ερμηνεία τους είναι ότι εάν αφαιρέσουμε την επίδραση της ηλικίας του ακινήτου (X_2), ποσοστό 50% ($R^2_{Y,X_1/X_2}$) της ανερμήνευτης μεταβλητότητας της αξίας των ακινήτων ερμηνεύεται από το εμβαδόν τους (X_1). Ενώ, εάν αφαιρέσουμε την επίδραση του εμβαδού του ακινήτου (X_1), ποσοστό 45,4% ($R^2_{Y,X_2/X_1}$) της ανερμήνευτης μεταβλητότητας της αξίας των ακινήτων ερμηνεύεται από την ηλικία τους (X_2).

6. Η προβλεπόμενη αξία πώλησης μιας μαιζονέτας 174 τ.μ. που κατασκευάστηκε πριν από μία δεκαετία θα προκύψει από το υπόδειγμα πολαπλής παλινδρόμησης που εκτιμήθηκε, με αντικατάσταση των συγκεκριμένων τιμών των ανεξάρτητων μεταβλητών.

$$\begin{aligned} \hat{Y} &= 317{,}998 + 0{,}543\, X_1 - 1{,}405\, X_2 = 317{,}998 + 0{,}543 \cdot 174 - 1{,}405 \cdot 10 = \\ &= 317{,}998 + 94{,}482 - 14{,}05 \\ &= 398{,}43 \text{ χιλιάδες €} \end{aligned}$$