Εφαρμογές Στατιστικών Μεθόδων σε Επιχειρηματικά Προβλήματα

Φροντιστήριο #8

Συντελεστής Πολλαπλού Προσδιορισμού

Στην πολλαπλή παλινδρόμηση ανάλογος του συντελεστή προσδιορισμού της απλής παλινδρόμησης είναι ο συντελεστής πολλαπλού προσδιορισμού. Αυτός μετράει το ποσοστό της μεταβλητότητας της εξαρτημένης μεταβλητής Y που οφείλεται στις επιδράσεις όλων μαζί των ανεξάρτητων μεταβλητών $X_1, X_2, \ldots, X_k$. Δηλαδή, μετράει τη συνολική επίδραση που δέχεται η Y από τις $X_1, X_2, \ldots, X_k$.

Ορίζεται ως: $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$, όπου

$$SSR = \sum(\hat{Y}-\bar{Y})^2, \quad SST = \sum(Y-\bar{Y})^2 \text{ και}$$

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

Όπως και στην απλή γραμμική παλινδρόμηση, ισχύει ότι: $SST = SSR + SSE$.

Η προσθήκη μίας νέας ανεξάρτητης μεταβλητής στο υπόδειγμα οδηγεί σε αύξηση της τιμής του συντελεστή R^2. Η αύξηση της τιμής του R^2 δεν θα έχει αξία όταν μάλιστα ο αριθμός k των ανεξάρτητων μεταβλητών είναι υψηλός σε σχέση με την τιμή του μεγέθους δείγματος n. Για να αντιμετωπίσουμε αυτό το πρόβλημα, χρησιμοποιούμε τον "διορθωμένο" (ή προσαρμοσμένο) συντελεστή πολλαπλού προσδιορισμού R_a^2, όπου

$$R_a^2 = 1-(1-R^2)\left(\frac{n-1}{n-k-1}\right).$$

Ο R_a^2 έχει σημαντικό ρόλο στις περιπτώσεις όπου ο αριθμός κ είναι μεγάλος σε σχέση με το μέγεθος του δείγματος. Για μεγάλο n σε σχέση με το κ, ο R_a^2 διαφέρει ελάχιστα από τον R^2. Εάν η συνεισφορά της νέας μεταβλητής που προστίθεται στο μοντέλο είναι αμελητέα, ο διορθωμένος συντελεστής θα μειωθεί (αντί να αυξηθεί).

Για τον υπολογισμό του R_a^2 χρήσιμοι είναι οι ακόλουθοι τύποι: (για το μοντέλο με κ=2 ανεξάρτητες μεταβλητές)

$$SSE = \sum(Y-\hat{Y})^2 = \sum\left[Y-(b_0+b_1X_1+b_2X_2)\right]^2$$

$$= \sum Y^2 - b_0\sum Y - b_1\sum YX_1 - b_2\sum YX_2$$

Επίσης,

$$SST = \sum Y^2 - \frac{(\sum Y)^2}{n}$$

$$SSR = SST - SSE$$

<u>Παρατηρήσεις:</u> 1. Ο R^2 δίνει την αναλογία της διακύμανσης της Y που ερμηνεύεται από τη γραμμική επίδραση σε αυτήν των ανεξάρτητων X_i. Δηλαδή ο R^2 ελέγχει το βαθμό προσαρμογής της Y στο νέφος των δειγματικών σημείων $(X_{1i}, X_{2i}, \ldots, X_{ki}, Y_i)$, $i=1,\ldots,n$.

2. Προφανώς, $R_a^2 < R^2$ διότι:

$$1-R_a^2 = (1-R^2)\frac{(n-1)}{(n-k-1)} \Rightarrow 1-R_a^2 > 1-R^2 \Rightarrow \boxed{R_a^2 < R^2}$$

Η διαφορά τους είναι σημαντική σε μικρά δείγματα.

3. Ο R_a^2 προτιμάται από τον R^2 όταν θέλουμε να ελέγξουμε κατά πόσον η εισαγωγή μίας νέας X_i βελτιώνει την "ερμηνευτική ικανότητα" της εξίσωσης. Και αυτό, γιατί με κάθε νέα εισαγωγή ενώ ο R^2 αυξάνει πάντοτε (επειδή μειώνεται το $\sum e_i^2$, ο $\bar{R}_a$ που μπορεί να γραφεί ως:

$$R_a^2 = 1 - \frac{\sum e_i^2 / (n-k-1)}{\sum (y_i - \bar{y})^2 / (n-1)} = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2} \cdot \frac{(n-1)}{(n-k-1)} \quad (*)$$

είναι δυνατόν να **μην** αυξηθεί **σημαντικά** ή/και να μειωθεί, γιατί παράλληλα μειώνεται η ποσότητα $(n-k-1)$ κατά μία μονάδα. Αν με την εισαγωγή μίας νέας X_i, ο R_a^2 μειωθεί ή αυξηθεί ελάχιστα η νέα X_i <u>δεν</u> θεωρείται αναγκαία στο υπόδειγμα.

Πως προκύπτει η σχέση (*);

Έστω $y_i, i=1,\dots,n$ οι παρατηρούμενες τιμές της Y και $\hat{y}_i$ οι εκτιμώμενες τιμές της Y από το μοντέλο παλινδρόμησης. Για κάθε $i=1,\dots,n$ μπορούμε να γράψουμε:

$$y_i - \bar{y} = y_i - \hat{y}_i + \hat{y}_i - \bar{y}, \quad i=1,\dots,n, \quad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}.$$

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$SST = SSE + SSR \Rightarrow 1 = \frac{SSE}{SST} + R^2$$

$$1-R^2 = \frac{SSE}{SST} = \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2}$$. Με αντικατάσταση της

παραπάνω έκφρασης στην (*) προκύπτει η σχέση.

Έλεγχοι Στατιστικής Σημαντικότητας

Ο έλεγχος της στατιστικής σημαντικότητας της εξίσωσης παλινδρόμησης ταυτίζεται με τον έλεγχο της στατιστικής σημαντικότητας του συντελεστή πολλαπλού προσδιορισμού R^2.

Ελέγχουμε:

H_0: η εξίσωση παλινδρόμησης δεν εξηγεί καθόλου τις μεταβολές της Y (το ποσοστό της ερμηνευόμενης διασποράς της Y είναι μηδέν)

Ισοδύναμα, $B_1 = B_2 = \ldots = B_K = 0$,

έναντι της:

H_1: η εξίσωση παλινδρόμησης εξηγεί ένα μέρος των μεταβολών της Y (το ποσοστό της ερμηνευόμενης διασποράς της Y είναι μεγαλύτερο του μηδενός) και τουλάχιστον ένας συντελεστής B_i για κάποιο i είναι διάφορος του μηδενός, $B_i \neq 0$.

Ο πίνακας ανάλυσης διακύμανσης (ANOVA) για την πολλαπλή παλινδρόμηση διαμορφώνεται ως εξής:

Πηγή Μεταβλητότητας	Άθροισμα Τετραγώνων	β.ε.	Μέσα Τετράγωνα	$F_{K,n-K-1}$
Παλινδρόμηση	SSR	K	SSR/K	$\frac{SSR/K}{SSE/(n-K-1)}$
Κατάλοιπο	SSE	$n-K-1$	$SSE/(n-K-1)$	
Σύνολο	SST	$n-1$		

Ο έλεγχος ανάλυσης διακύμανσης γίνεται με τη σ.σ. ελέγχου $F_{k,n-k-1}$, όπου

$$F_{k,n-k-1} = \frac{SSR/k}{MSE}, \quad MSE = \text{μέσο τετραγωνικό σφάλμα}$$

($k := \#$ανεξάρτητων μεταβλητών) $= SSE/(n-k-1)$

Απορρίπτουμε την H_0 σε εσσ α, αν

$$F_{k,n-k-1} > F_{k,n-k-1;\alpha}$$ όπου με $F_{k,n-k-1,\alpha}$ συμβολίζουμε το α-ποσοστιαίο σημείο της κατανομής F με $n_1 = k$ και $n_2 = n-k-1$ β.ε.

<u>Έλεγχος σημαντικότητας συντελεστών μερικής παλινδρόμησης</u>

Ελέγχουμε: $H_0 : B_i = B_i^*$ δηλ. ο συντελεστής μερικής παλινδρόμησης B_i του πληθυσμού ισούται με B_i^*

έναντι:

$H_1 : B_i \neq B_i^*$ δηλ. ο συντελεστής μερικής παλινδρόμησης B_i του πληθυσμού είναι διάφορος του B_i^*. Ο έλεγχος γίνεται με το κριτήριο t με $n-k-1$ β.ε. δηλ.

$$T_{n-k-1} = \frac{|b_i - B_i^*|}{S_{b_i}},$$ όπου S_{b_i} είναι το τυπικό σφάλμα εκτίμησης του συντελεστή B_i.

Αν $|T_{n-k-1}| > t_{n-k-1,\alpha/2}$ η H_0 απορρίπτεται σε ε.σ.σ. α. Οι παραπάνω υποθέσεις για τον έλεγχο της στατιστικής σημαντικότητας των B_i μπορούν να διατυπωθούν ως εξής:

$H_0: B_i = B_i^*$ (δεδομένου ότι όλες οι ανεξάρτητες μεταβλητές περιλαμβάνονται στο υπόδειγμα)

έναντι:

$H_1: B_i \neq B_i^*$ (δεδομένου ότι όλες οι ανεξάρτητες μεταβλητές περιλαμβάνονται στο υπόδειγμα)

Το $100(1-\alpha)\%$ δ.ε. για τον συντελεστή παλινδρόμησης B_i είναι:

$$\left(b_i - s_{b_i} t_{n-k-1,\alpha/2},\ b_i + s_{b_i} t_{n-k-1,\alpha/2} \right),$$ όπου

$t_{n-k-1,\alpha/2}$ είναι το $\alpha/2$-ποσοστιαίο σημείο της t-student με $n-k-1$ β.ε.

Παράδειγμα (Συνέχεια στο Παράδειγμα 1 του Φρ. #7).

(iii) Να υπολογισθεί ο συντελεστής πολλαπλού προσδιορισμού.

(iv) Να κατασκευασθεί ο πίνακας ανάλυσης διακύμανσης (ANOVA) για τα δεδομένα μας.

(v) Να υπολογισθεί το τυπικό σφάλμα εκτίμησης της εξαρτημένης μεταβλητής στο μοντέλο της πολλαπλής παλινδρόμησης.

(vi) Να εκτιμηθούν οι διακυμάνσεις των συντελεστών μερικής παλινδρόμησης και να υπολογισθούν τα τυπικά σφάλματα των εκτιμήσεων αυτών.

(vii) Να ελεγχθεί σε ε.σ.σ. $\alpha = 5\%$, η στατιστική σημαντικότητα των παραμέτρων β_1, β_2 του πληθυσμού.

(viii) Να εκτιμηθεί με ένα 95% δ.ε. η παράμετρος β_1 και η παράμετρος β_2 του πληθυσμού.

(ix) Να υπολογισθεί ο διορθωμένος συντελεστής πολλαπλού προσδιορισμού R_a^2.

(x) Να ελεγχθεί σε ε.σ.σ. $\alpha = 5\%$, η στατιστική σημαντικότητα της εξίσωσης πολλαπλής παλινδρόμησης.

<u>Λύση</u> (iii) Έχουμε: $R^2 = \dfrac{SSR}{SST}$, όπου

$$SST = \sum y^2 = \left(-\tfrac{5}{4}\right)^2 + \left(-\tfrac{1}{4}\right)^2 + \left(-\tfrac{1}{4}\right)^2 + \left(\tfrac{7}{4}\right)^2$$

$$= 4.75 \text{ και } SSR = \sum(\hat{Y} - \bar{Y})^2 = \sum \hat{y}^2$$

αν θέσουμε $\hat{y} = \hat{Y} - \bar{Y}$.

Όμως για κάθε i, $\hat{y}_i = b_1 x_{1i} + b_2 x_{2i}, i = 1, \dots, n$

$$\sum \hat{y}_i^2 = \sum (b_1 x_{1i} + b_2 x_{2i})^2$$

$$= b_1^2 \sum x_{1i}^2 + 2 b_1 b_2 \sum x_{1i} x_{2i} + b_2^2 \sum x_{2i}^2$$

$$= b_1 \left(b_1 \sum x_{1i}^2 + b_2 \sum x_{1i} x_{2i}\right) + b_2 \left(b_2 \sum x_{2i}^2 + b_1 \sum x_{1i} x_{2i}\right)$$

$$= b_1 \sum x_{1i} y_i + b_2 \sum x_{2i} y_i$$ (από τις κανονικές εξισώσεις, Φρ. #7)

Άρα

$$\sum \hat{y}^2 = b_1 \sum x_1 y + b_2 \sum x_2 y$$

$$= 1.235 \cdot 1.5 + 0.529 \cdot 5.25 = 4.63$$

(από πίνακα
της άσκησης 1, φρ. #7)

$R^2 = \dfrac{4.63}{4.75} = 0.975$ δηλ. το 97.5% της συνολικής μεταβλητότητας των μηνιαίων δαπανών του πληθυσμού των νοικοκυριών ερμηνεύεται (επεξηγείται) από τους μισθούς X_1 και από τα άλλα εισοδήματα X_2.

(iv) Πίνακας Ανάλυσης Διακύμανσης (ANOVA)

Πηγή Μεταβλητότητας	Αθροίσματα τετραγώνων	β.ε	Μέσα Τετράγωνα	$f_{2,1}$
Παλινδρόμηση	4.63	$k=2$	2.315	19,29
Κατάλοιπα	0.12	1	0.12	
Άθροισμα	4.75	3		

(v) Έχουμε ήδη υπολογίσει στον πίνακα ανάλυσης διακύμανσης το μέσο τετραγωνικό σφάλμα MSE

$$MSE = \frac{SSE}{(n-k-1)} = \frac{0.12}{(4-2-1)} = 0.12,$$ το οποίο

ισούται με την διακύμανση των σφαλμάτων

$s_e^2 = MSE$. Άρα, το τυπικό σφάλμα εκτιμήσεως της εξαρτημένης μεταβλητής είναι:

$$S = \sqrt{S_e^2} = \sqrt{MSE} = \sqrt{0.12} = 0.346$$

δηλαδή η αναμενόμενη (κατά μέσο όρο) τυπική απόκλιση των τιμών της Y γύρω από τη εξίσωση πολλαπλής παλινδρόμησης είναι (περίπου) 0.346 χιλιάδες ευρώ.

(vi) Για να εκτιμήσουμε τη διακύμανση των συντελεστών μερικής παλινδρόμησης $\hat{B}_1$ και $\hat{B}_2$, $\sigma^2_{\hat{\beta}_1}$, $\sigma^2_{\hat{\beta}_2}$ αρκεί να υπολογίσουμε τις εκτιμήτριές τους $S^2_{b_1}$ και $S^2_{b_2}$ (ή $S^2_{\hat{\beta}_1}$, $S^2_{\hat{\beta}_2}$) αντίστοιχα. Αυτές υπολογίζονται από τους τύπους:

$$S^2_{\hat{\beta}_1} = S^2_{b_1} = S_e^2 \frac{\sum x_2^2}{\sum x_1^2 \sum x_2^2 - \left(\sum x_1 x_2\right)^2}$$

$$= 0.12 \cdot \frac{8.75}{1 \cdot 8.75 - (0.5)^2} = 0.1235 \cong 0.124$$

$$S^2_{\hat{\beta}_2} = S^2_{b_2} = S_e^2 \frac{\sum x_1^2}{\sum x_1^2 \sum x_2^2 - \left(\sum x_1 x_2\right)^2}$$

$$= 0.12 \quad \frac{1}{1 \cdot 8.75 - (0.5)^2} = 0.0141$$

Άρα, τα τυπικά σφάλματα εκτιμήσεως των β_1, β_2, είναι: $S_{\hat{\beta}_1} = S_{b_1} = \sqrt{0.124} = 0.352$

και $S_{\hat{\beta}_2} = S_{b_2} = \sqrt{0.014} = 0.1183.$

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

$$H_0: B_1 = 0 \quad vs \quad H_1: B_1 \neq 0$$

Χρησιμοποιούμε την σ.σ. ελέγχου:

$$T_{n-k-1} = \frac{|b_1 - 0|}{s_{b_1}} = \frac{|1.235|}{\sqrt{0.124}} = 3.51$$

$n-k-1 = 4-2-1 = 1$. Από τους πίνακες της t βρίσκουμε ότι $t_{1,0.025} = 12.706$. Άρα, η H_0 <u>δεν</u> απορρίπτεται σε εσσ $\alpha = 5\%$ που σημαίνει ότι η μεταβλητή X_1 <u>δεν</u> φαίνεται να ασκεί στατιστικά σημαντική επίδραση στην διαμόρφωση των τιμών της Y.

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

$$H_0: B_2 = 0 \quad vs \quad H_1: B_2 \neq 0$$

Χρησιμοποιούμε την σ.σ. ελέγχου:

$$T_{n-k-1} = \frac{|b_2 - 0|}{s_{b_2}} = \frac{|0.529|}{\sqrt{0.014}} = 4.47$$

< 12.706, άρα <u>δεν</u> απορρίπτουμε την H_0 σε ε.σ.σ. $\alpha = 5\%$, που σημαίνει ότι η μεταβλητή X_2 <u>δεν</u> φαίνεται να ασκεί στατιστικά σημαντική επίδραση στη διαμόρφωση των τιμών της Y.

(viii) Το 95% δ.ε. για την παράμετρο B_1 είναι

$$\left(b_1 - s_{b_1} \cdot t_{n-k-1,\alpha/2},\ b_1 + s_{b_1} \cdot t_{n-k-1,\alpha/2}\right)$$

$$= (1.235 - 0.352 \cdot 12.706,\ 1.235 + 0.352 \cdot 12.706)$$

$$= (-3.2375,\ 5.707512)$$

δηλ. με πιθανότητα 95% η πραγματική, άλλα με άγνωστη τιμή, παράμετρος του πληθυσμού B_1 αναμένεται να βρίσκεται μεταξύ των τιμών -3.2375 και 5.707512. Στο δ.ε. που βρήκαμε περιέχεται η τιμή 0 και όπως είδαμε σε ε.σ. $\alpha = 5\%$ η B_1 δεν προέκυψε στατιστικά σημαντική.

Με το δ.ε. κάνουμε έναν έμμεσο έλεγχο της $H_0: B_1 = 0$ vs $H_1: B_1 \neq 0$ για το ίδιο α.

Ομοίως, το 95% δ.ε. για την παράμετρο B_2 είναι:

$$\left(b_2 - s_{b_2} t_{n-k-1,\alpha/2},\ b_2 + s_{b_2} t_{n-k-1,\alpha/2}\right)$$

$$= (0.529 - 0.1183 \cdot 12.706,\ 0.529 + 0.1183 \cdot 12.706)$$

$= (-0.974,\ 2.03211)$, δηλ. με πιθανότητα 95% η πραγματική, αλλά με άγνωστη τιμή, παράμετρους του πληθυσμού B_2 αναμένεται να βρίσκεται μεταξύ των τιμών -0.974 και 2.03211. Μπορούμε να κάνουμε τον ίδιο έμμεσο έλεγχο όπως και για την παράμετρο B_1.

(ix) Ο διορθωμένος συντελεστής πολλαπλού προσδιορισμού είναι:

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

$$= 1-(1-0.975)\frac{3}{1} = 1-3(1-0.975)$$

$$= 0.925.$$

δηλ. το 92.5% της συνολικής μεταβλητότητας της εξαρτημένης μεταβλητής Υ ερμηνεύεται από τις μεταβολές των ερμηνευτικών (ανεξάρτητων) μεταβλητών X_1 και X_2. Το γεγονός ότι ο διορθωμένος συντελεστής πολλαπλού προσδιορισμού R_a^2 είναι μειωμένος σε σχέση με τον συντελεστή πολλαπλού προσδιορισμού R^2 είναι σαφέστατη ένδειξη ότι οι μεταβλητές X_1 και X_2 που εμπεριέχονται στο μοντέλο δεν (φαίνεται να) είναι στατιστικά σημαντικές για να ερμηνεύσουν τις μεταβολές της Υ. Κάτι τέτοιο άλλωστε το διαπιστώσαμε και μέσω του ερωτήματος (vii).

(x) Από τον πίνακα ανάλυσης διακύμανσης (ANOVA) έχουμε $F_{2,1} = 19,29$. Ελέγχουμε σε ε.σ.σ. $\alpha=5\%$ την $H_0: \beta_1=\beta_2=0$

vs

$H_1: \exists$ ένα τουλάχιστον $i: \beta_i \neq 0$

Χρησιμοποιούμε την σ.σ. ελέγχου

$$F_{\kappa,n-\kappa-1} = \frac{SSR/\kappa}{SSE/(n-\kappa-1)} \overset{(*)}{=} \frac{R^2/\kappa}{(1-R^2)/(n-\kappa-1)}$$

(*)
Για την απόδειξη της (*) βλέπε Σημειώσεις Αϋφαντή, 2η απόδειξη, σελ. 82.

Για τα δεδομένα μας, $F_{\kappa,n-\kappa-1} = 19,29$. Απορρίπτουμε την H_0 σε εσσ $\alpha = 5\%$ αν $F_{\kappa,n-\kappa-1} > F_{\kappa,n-\kappa-1,\alpha}$

Όμως $F_{\kappa,n-\kappa-1,\alpha} = F_{2,1,0.05} = 200$, άρα η H_0 δεν απορρίπτεται σε ε.σ.σ. $\alpha = 5\%$ και μπορούμε να πούμε ότι η εξίσωση παλινδρόμησης που μελετάμε δεν είναι στατιστικά σημαντική και συνεπώς οι μεταβλητές X_1, X_2 δεν φαίνεται να ασκούν (μαζί) στατιστικά σημαντική επίδραση στις μεταβολές των τιμών της εξαρτημένης μεταβλητής Y.