Εφαρμογές Στατιστικών Μεθόδων σε Επιχειρηματικά Προβλήματα

Φροντιστήριο #3

Συντελεστής Προσδιορισμού

Έστω $\hat{y}_i$ οι τιμές που προκύπτουν για τη μεταβλητή απόκρισης Y χρησιμοποιώντας την ευθεία παλινδρόμησης $y=\beta_0+\beta_1 x$, δηλαδή $\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1 x_i$, $i=1,\dots,n$

Αυτές οι ποσότητες είναι οι εκτιμήσεις μας για την τιμή της μεταβλητής απόκρισης στις θέσεις x_i, $i=1,\dots,n$ και καλούνται εκτιμημένες τιμές (ή εκτιμώμενες τιμές) της Y στις θέσεις x_i, $i=1,\dots,n$.

Οι διαφορές $\hat{\varepsilon}_i=y_i-\hat{y}_i=y_i-(\hat{\beta}_0+\hat{\beta}_1 x_i)$, $i=1,\dots,n$

καλούνται εκτιμημένα ή εκτιμώμενα σφάλματα και εκφράζουν τις αποκλίσεις των παρατηρηθεισών τιμών y_i της Y (για την τιμή x_i της X) από τις εκτιμημένες τιμές $\hat{y}_i$.

Η ποσότητα

$$g(\hat{\beta}_0,\hat{\beta}_1)=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}\hat{\varepsilon}_i^2=SSE$$

είναι η ελάχιστη τιμή του αθροίσματος $g(\beta_0,\beta_1)$ καλείται άθροισμα τετραγώνων των (εκτιμημένων) σφαλμάτων.

SSE: sum of squares of Errors.

Είναι φανερό ότι:

- αν SSE μικρό η ευθεία θα περνάει "κοντά" από τα σημεία (x_i,y_i), $i=1,\dots,n$
- αν SSE μεγάλο, η ευθεία <u>δεν</u> θα βρίσκεται "κοντά"

σε όλα τα σημεία (x_i, y_i), $i=1,\ldots,n$.

Ένα μέτρο μεταβλητότητας των τιμών της Y γύρω από τη μέση τιμή τους $\bar{y}$ είναι το συνολικό άθροισμα των τετραγώνων SST, όπου $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$

Μπορούμε να γράψουμε:

$$SST = SSR + SSE$$

όπου $SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ είναι το άθροισμα τετραγώνων της παλινδρόμησης (Regression Sum of Squares).

Έχουμε:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 x_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x})$$

Άρα

$$SSR = \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2 = \sum_{i=1}^{n} [\bar{y} + \hat{\beta}_1 (x_i - \bar{x}) - \bar{y}]^2$$

$$= \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Συνεπώς, για το SSR, έχουμε: $\boxed{SSR = \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}$

Επίσης:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

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

$$= \sum_{i=1}^{n} \left(y_i^2 - \hat{\beta}_0 y_i - \hat{\beta}_1 x_i y_i - \hat{\beta}_0 y_i + \hat{\beta}_0^2 + \hat{\beta}_0 \hat{\beta}_1 x_i - \hat{\beta}_1 x_i y_i + \hat{\beta}_1 \hat{\beta}_0 x_i + \hat{\beta}_1^2 x_i^2\right)$$

$$= \sum_{i=1}^{n} y_i^2 - \hat{\beta}_0 \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i y_i$$

$$- \hat{\beta}_0 \sum_{i=1}^{n} y_i + n \hat{\beta}_0^2 + \hat{\beta}_0 \hat{\beta}_1 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i y_i + \hat{\beta}_1 \hat{\beta}_0 \sum_{i=1}^{n} x_i$$

$$+ \hat{\beta}_1^2 \sum_{i=1}^{n} x_i^2 .$$

Όμως

$$- \hat{\beta}_0 \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i y_i + 2 \hat{\beta}_1 \hat{\beta}_0 \sum_{i=1}^{n} x_i + n \hat{\beta}_0^2 + \hat{\beta}_1^2 \sum_{i=1}^{n} x_i^2$$

$$= - \hat{\beta}_0 \left(n \hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i \right) - \hat{\beta}_1 \sum_{i=1}^{n} x_i y_i + 2 \hat{\beta}_1 \hat{\beta}_0 \sum_{i=1}^{n} x_i + n \hat{\beta}_0^2$$

$$+ \hat{\beta}_1^2 \sum_{i=1}^{n} x_i^2$$

$$= \cancel{- n \hat{\beta}_0^2} - \hat{\beta}_0 \hat{\beta}_1 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i y_i + 2 \hat{\beta}_1 \hat{\beta}_0 \sum_{i=1}^{n} x_i$$

$$+ \cancel{n \hat{\beta}_0^2} + \hat{\beta}_1^2 \sum_{i=1}^{n} x_i^2$$

$$= \hat{\beta}_0 \hat{\beta}_1 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i y_i + \hat{\beta}_1^2 \sum_{i=1}^{n} x_i^2$$

$$= \hat{\beta}_0 \hat{\beta}_1 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \left(\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \right) + \hat{\beta}_1^2 \sum_{i=1}^{n} x_i^2$$

$$= \hat{\beta}_0 \hat{\beta}_1 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1^2 \sum_{i=1}^{n} x_i^2 + \hat{\beta}_1^2 \sum_{i=1}^{n} x_i^2 = 0 .$$

Άρα,

$$\boxed{SSE = \sum_{i=1}^{n} y_i^2 - \hat{\beta}_0 \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i y_i}$$

Η ποσότητα: $s^2 = MSE = \dfrac{SSE}{n-2} = \dfrac{1}{n-2} \sum_{i=1}^{n} \hat{\varepsilon}_i^2$

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

είναι αμερόληπτη εκτιμήτρια της παραμέτρου σ^2 δηλ. $E(s^2) = \sigma^2 = var(Y_i) = var(\varepsilon_i)$ λέγεται μέσο άθροισμα τετραγώνων των υπολοίπων ή μέσο τετραγωνικό υπόλοιπο ή μέσο τετραγωνικό σφάλμα (error mean square, residual mean square).

Χρησιμοποιούμε τον λόγο $\frac{SSR}{SST}$ ως ένα δείκτη ποιότητας του μοντέλου γραμμικής παλινδρόμησης $y = \hat{\beta}_0 + \hat{\beta}_1 x$.

Η ποσότητα $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$ καλείται συντελεστής προσδιορισμού (coefficient of determination) του γραμμικού μοντέλου. Συχνά εκφράζεται ως ποσοστό, παίρνει τιμές μεταξύ 0 και 1.

Τιμές $R^2 \to 1$: η ευθεία παλινδρόμησης περνάει πολύ κοντά από τα περισσότερα σημεία, ενώ

$R^2 \to 0$ πολλά σημεία βρίσκονται μακριά από την ευθεία και θα πρέπει να αναζητηθεί μία άλλη σχέση της ανεξάρτητης και της εξαρτημένης μεταβλητής (ενδεχομένως, μη-γραμμική).

Ο συντελεστής R^2 εκφράζει το ποσοστό της συνολικής διασποράς των τιμών της εξαρτημένης μεταβλητής Υ η οποία ερμηνεύεται (εξηγείται) από τη ανεξάρτητη μεταβλητή Χ (μέσω της ευθείας παλινδρόμησης).

Μόνο για το απλό γραμμικό μοντέλο ισχύει ότι

$\boxed{R^2 = r^2}$ όπου r είναι ο συντελεστής γραμμικής συσχέτισης των Χ και Υ.

Έλεγχος στατιστικής σημαντικότητας του Συντελεστή R^2.

Ελέγχουμε:

$H_0: \beta_0 = \beta_1 = 0$ vs H_1: υπάρχει τουλάχιστον ένας συντελεστής παλινδρόμησης εκ των β_0, β_1 που είναι στατιστικά σημαντικά διάφορος του μηδενός
$\exists$ ένα τουλ. i : $\beta_i \neq 0, i=0,1$

Εναλλακτικά,
γράφουμε:

H_0: η εξίσωση παλινδρόμησης <u>δεν</u> εξηγεί καθόλου τις μεταβολές της Υ ή το ποσοστό της διασποράς που εξηγείται για την Υ είναι μηδέν vs.

H_1: η εξίσωση παλινδρόμησης εξηγεί ένα μέρος των μεταβολών της Υ ή το ποσοστό της ερμηνευόμενης διασποράς της Υ είναι μεγαλύτερο του μηδέν.

Ο έλεγχος καλείται έλεγχος ανάλυσης διακύμανσης - analysis of variance test (ANOVA test) και για τη διενέργεια του μπορούμε αρχικά να κατασκευάσουμε τον ακόλουθο πίνακα που καλείται πίνακας ανάλυσης διακύμανσης (ANOVA table).

Πηγή Μεταβλητότητας	Άθροισμα Τετραγώνων	Β.Ε.	Μέσα Τετράγωνα	Λόγος $F_{1,n-2}$
Παλινδρόμηση	SSR	1	$SSR/1 = MSR$	$\frac{MSR}{MSE} = F_{1,n-2}$
Κατάλοιπα	SSE	$n-2$	$SSE/n-2 = MSE$	
Σύνολο	SST	$n-1$		

Υπολογίζουμε τον λόγο $F_{1,n-2}$.

Απορρίπτουμε την H_0 σε ΕΣΣ α αν

$F_{1,n-2} > \mathcal{F}_{1,n-2,\alpha}$ όπου $\mathcal{F}_{1,n-2,\alpha}$ είναι το

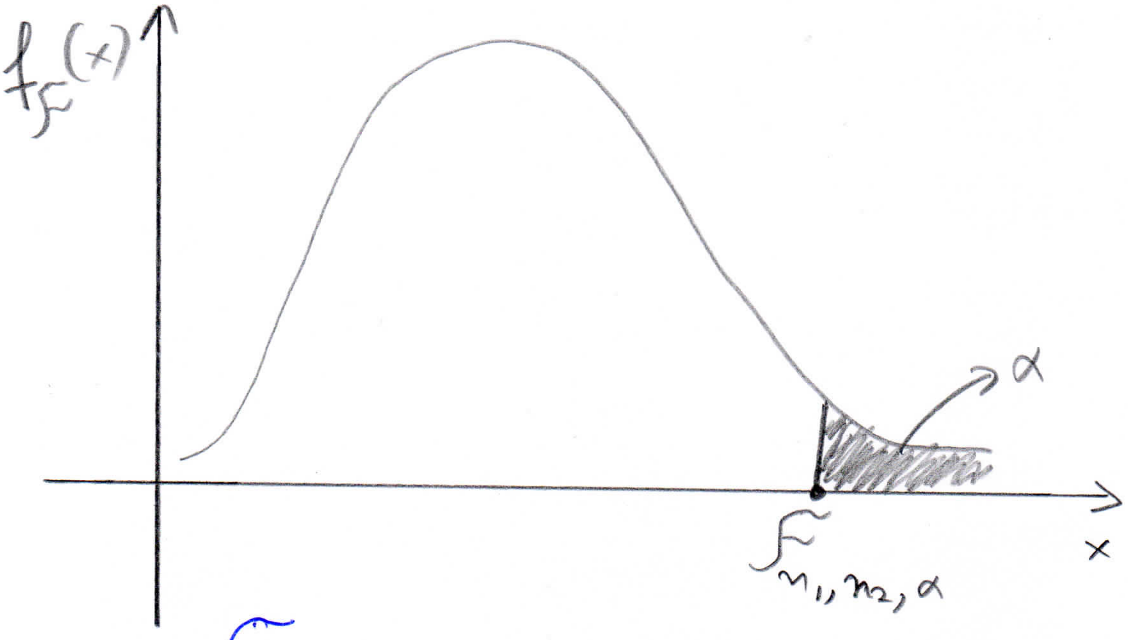
α-ποσοστιαίο σημείο της κατανομής $\mathcal{F}_{1,n-2}$ με

$n_1 = 1$ και $n_2 = n-2$ β.ε.

Λίγα στοιχεία για την κατανομή $\mathcal{F}_{n_1,n_2}$

Η γραφική της παράσταση έχει την ακόλουθη μορφή (του σχήματος) για διάφορες τιμές των n_1, n_2. Στο σχήμα καταγράφουμε και το σημείο

$F_{n_1,n_2,\alpha}$:



Το σημείο $F_{n_1,n_2,\alpha}$ είναι τ.ω. $P(F > F_{n_1,n_2,\alpha}) = \alpha$.

Παραδείγματα

① Μία τράπεζα ενδιαφέρεται να μελετήσει την αποταμιευτική συμπεριφορά των πελατών της. Θεωρούμε ότι η ετήσια αποταμίευση των πελατών της εξαρτάται από το ετήσιο εισόδημά τους. Σε δείγμα 10 πελατών έχουμε τα ακόλουθα στοιχεία (τιμές σε χιλιάδες €).

Ετήσια αποταμίευση (Y)	5	3	5	4	6	1	2	8	2	3
Ετήσια εισοδήματα (X)	50	31	28	45	50	32	36	55	26	47

(α) Να εκτιμήσετε και να ερμηνεύσετε τους συντελεστές του γραμμικού υποδείγματος.

(β) Ποια η τιμή του συντελεστή προσδιορισμού και ποια η ερμηνεία του.

(γ) Να πραγματοποιηθεί έλεγχος για ε.σ. $\alpha=5\%$ για τη στατιστική σημαντικότητα του συντελεστή προσδιορισμού (ή ολόκληρης της εξίσωσης παλινδρόμησης).

Δίνονται: $\sum_{i=1}^{10} x_i = 400$, $\sum_{i=1}^{10} y_i = 39$, $\sum_{i=1}^{10} x_i^2 = 17000$

$\sum_{i=1}^{10} y_i^2 = 193$, $\sum_{i=1}^{10} x_i y_i = 1700$.

<u>Λύση</u>

(α) Οι συντελεστές του γραμμικού υποδείγματος $Y = \beta_0 + \beta_1 X + \varepsilon$ δίνονται:

$$\hat{\beta}_1 = \frac{n \sum_{i=1}^{10} x_i y_i - \sum_{i=1}^{10} x_i \sum_{i=1}^{10} y_i}{n \sum_{i=1}^{10} x_i^2 - \left(\sum_{i=1}^{10} x_i\right)^2} = \frac{10 \cdot 1700 - 400 \cdot 39}{10 \cdot 17000 - (400)^2} = 0{,}14$$

και $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{39}{10} - 0{,}14 \cdot \frac{400}{10} = -1{,}7$.

Συνεπώς, η εκτιμώμενη ευθεία παλινδρόμησης της Y πάνω στη X θα είναι:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X = -1{,}7 + 0{,}14X.$$

Ερμηνεία των συντελεστών:

<u>$\hat{\beta}_1 = 0{,}14$</u> Για κάθε αύξηση κατά 1000 € των ετήσιων εισοδημάτων του πελάτη της τράπεζας

οι ετήσιες αποταμιεύσεις του αναμένεται να έχουν μία αύξηση περίπου ίση με 140 ευρώ (0,14·1000 €).

<u>$\hat{\beta}_0 = -1,7$</u> Αν υποθέσουμε ότι ένας πελάτης δεν έχει ετήσια εισοδήματα, οι ετήσιες αποταμιεύσεις του αναμένεται να μειωθούν κατά περίπου 1700 ευρώ (−1,7·1000 €).

(β) Έχουμε $R^2 = \dfrac{SSR}{SST}$

Όμως, $SSR = \hat{\beta}_1^2 \cdot \left(\sum_{i=1}^{n=10} x_i^2 - n\,\bar{x}^2 \right)$

$$= 0,14^2 \left(17000 - 10 \cdot 40^2\right) = 19,6$$

$$SST = \sum_{i=1}^{n=10} y_i^2 - n\,\bar{y}^2 = 193 - 10\left(\frac{39}{10}\right)^2$$

$$= 193 - 152,1 = 40,9$$

$$SSE = SST - SSR = 40,9 - 19,6 = 21,3$$

Άρα, $R^2 = \dfrac{SSR}{SST} = \dfrac{19,6}{40,9} = 0,479 = 47,9\%$

<u>Ερμηνεία R^2</u>: Το 48% περίπου της συνολικής μεταβλητότητας των ετήσιων αποταμιεύσεων των πελατών της τράπεζας οφείλονται στα ετήσια εισοδήματα τους και το υπόλοιπο 52% οφείλεται σε άλλους παράγοντες (ή/και σε μεταβλητές που

δεν έχουν προβλεφθεί από το μοντέλο).

(γ) Μπορούμε να φτιάξουμε τον πίνακα ανάλυσης διακύμανσης.

Πηγή Μεταβλητότητας	Αθροίσματα Τετραγώνων	Β.Ε	Μέσα Τετράγωνα	$F_{1,8}$
Παλινδρόμηση	19,6	1	19,6	$F_{1,8}=\frac{19,6}{2,66}$
Κατάλοιπα	21,3	8	2,66	$=7,36$
Σύνολο	40,9	9		

Ελέγχουμε $H_0: \beta_0=\beta_1=0$

$H_1: \exists$ τουλ. ένα $i: \beta_i \neq 0, \; i=1,2.$

Απορρίπτουμε την H_0 σε ε.σ.σ. $\alpha=5\%$ αν

$$F_{1,8}=7,36 > F_{1,8,0.05}=5,32$$

ισχύει η ανισότητα άρα τα δεδομένα παρέχουν ισχυρές ενδείξεις απόρριψης της H_0 συνεπώς φαίνεται ότι τα ετήσια εισοδήματα (μαζί με τη σταθερά παλινδρόμησης) επηρεάζουν στατιστικά σημαντικά τις μεταβολές (τη μεταβλητότητα) των ετήσιων αποταμιεύσεων των πελατών της τράπεζας.

(δ) (Extra ερώτημα)
Να υπολογιστεί το μέσο τετραγωνικό σφάλμα της εξίσωσης παλινδρόμησης.

Έχουμε: $s^2 = MSE = \dfrac{SSE}{n-2} = \dfrac{21,3}{8} = 2,6625$

Το τυπικό σφάλμα των καταλοίπων είναι:

$s = \sqrt{s^2} = \sqrt{2,6625} = 1,631.$

<u>Παρατήρηση</u> (για τον έλεγχο F)
Το πηλίκο $F_{1,n-2} = \dfrac{MSR}{MSE}$ μπορεί να γραφεί

εναλλακτικά:

$$F_{1,n-2} = \frac{R^2/1}{(1-R^2)/(n-2)} = \frac{SSR/1}{SSE/(n-2)} = \frac{MSR}{MSE}$$

$= \dfrac{R^2(n-2)}{(1-R^2)}$, αν γνωρίζουμε την τιμή του

συντελεστή προσδιορισμού R^2.
