Εφαρμογές Στατιστικών Μεθόδων σε Επιχειρηματικά Προβλήματα

Φροντιστήριο #10

Επαναληπτικές Ασκήσεις

#1 Έστω ότι έχουμε τις κατά κεφαλήν εκπομπές διοξειδίου του άνθρακα (CO_2/c), το κατά κεφαλήν Ακαθάριστο Εθνικό Προϊόν (GNP/c) και το ποσοστό βιομηχανοποίησης ($Indust$) για 12 χώρες, όπως φαίνεται στον παρακάτω πίνακα:

Χώρες	1	2	3	4	5	6	7	8
Y (CO_2/c)	2.59	2.86	2.97	2.97	2.98	2.98	3.31	3.35
X_1 (GNP/c)	390	730	930	940	960	960	2021	2240
X_2 ($INDUST$)	16.9	26.2	20	19.7	29.4	43.9	39	32.4

Χώρες	9	10	11	12
Y	3.42	3.78	3.9	4.02
X_1	2629	6.010	7992	10420
X_2	50	35	37.4	33.3

(α) Βρείτε τους εκτιμητές των παραμέτρων της εξίσωσης παλινδρόμησης. Γράψτε την εξίσωση παλινδρόμησης και ερμηνεύστε τους συντελεστές της.

(β) Εκτιμήστε το τυπικό σφάλμα εκτίμησης της παλι-

νδρόμησης καθώς και τα τυπικά σφάλματα των εκτιμηθέντων συντελεστών b_1 και b_2. Βρείτε τον απλό καθώς και το διορθωμένο συντελεστή προσδιορισμού.

(γ) Ελέγξτε σε επίπεδο σ.σ. $\alpha = 5\%$ τη στατιστική σημαντικότητα των συντελεστών παλινδρόμησης.

(δ) Υπολογίστε το μερικό συντελεστή συσχέτισης μεταξύ των X_1, X_2 και Y και σχολιάστε ποια από τις ερμηνευτικές μεταβλητές επηρεάζει περισσότερο την εξαρτημένη μεταβλητή.

(ε) Κατασκευάστε ένα 95% δ.ε για τους συντελεστές της παλινδρόμησης β_1, β_2.

(ζ) Ελέγξτε τη συνολική στατιστική σημαντικότητα του πολλαπλού υποδείγματος.

Λύση (α) Προσαρμόζουμε το μοντέλο πολλαπλής γραμμικής παλινδρόμησης: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$

$i = 1, 2, \dots, 12$.

Για τον υπολογισμό των εκτιμητών ελαχίστων τετραγώνων των παραμέτρων $\beta_1, \beta_2, \beta_0$ Θέτουμε:

$$y = Y_i - \bar{Y}, \quad x_1 = X_{1i} - \bar{X}_1, \quad x_2 = X_{2i} - \bar{X}_2$$

και για κάθε i υπολογίζουμε τις ποσότητες:

$$y, x_1, x_2, x_1^2, x_2^2, x_2 y, x_1 y, x_1 x_2$$

Οι εκτιμητές ελαχίστων τετραγώνων είναι:

$$b_1 = \hat{\beta}_1 = \frac{(\Sigma x_1 y)(\Sigma x_2^2) - (\Sigma x_2 y)(\Sigma x_1 x_2)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2}$$

Οι υπολογισμοί μας δίνουν:

$\sum x_1^2 = 119522842, \quad \sum x_2^2 = 1119.267$

$\sum x_2 y = 26.19066, \quad \sum x_1 y = 15213.19$

$\sum x_1 x_2 = 115595.6 \quad \sum y^2 = 2.227692$

Άρα $b_1 = 0,000116$

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

$= 0,0115$

$b_0 = \bar{Y} - b_1 \bar{X}_1 - b_2 \bar{X}_2 = 2,544$ (μετά από πράξεις).

Η εκτιμώμενη εξίσωση παλινδρόμησης είναι:

$$\hat{Y}_i = 2,544 + 0,000116 X_{1i} + 0,0115 X_{2i}$$

Η εξαρτημένη μεταβλητή είναι θετικά συσχετισμένη με τις ανεξάρτητες μεταβλητές X_1, X_2. Ο εκτιμώμενος συντελεστής b_1 υποδηλώνει ότι μία μοναδιαία αύξηση στην μεταβλητή X_1 συνδυάζεται με μία αύξηση στην Υ κατά περίπου ίση με 0,00012 μονάδες μέτρησης ρύπων διατηρώντας τη μεταβλητή X_2 σταθερή. Ομοίως, ο εκτιμώμενος συντελεστής b_2 υποδηλώνει ότι μία μοναδιαία αύξηση στην μεταβλητή X_2 συνδυάζεται με μία αύξηση στην Υ ίση κατά περίπου με 0,0115 μονάδες μέτρησης ρύπων

διατηρώντας τη μεταβλητή X_1 σταθερή. Αν $X_1 = X_2 = 0$ τότε $b_0 = 2,544$.

(β) Τα τυπικά σφάλματα των εκτιμητών δίνονται από τους παρακάτω τύπους:

$$s_{\hat{\beta}_1} = s_{b_1} = s_e \sqrt{\frac{\sum x_2^2}{\sum x_1^2 \sum x_2^2 - \left(\sum x_1 x_2\right)^2}},$$

όπου $s_e = \sqrt{\frac{\sum e_i^2}{n-k-1}} = \sqrt{\frac{\sum (Y_i - \hat{Y}_i)^2}{n-k-1}}$

$k :=$ #ανεξάρτητων μεταβλητών, $\sum e_i^2 = 0.15995$

άρα, $s_e = \sqrt{\frac{0.15995}{12-2-1}} = \sqrt{\frac{0.15995}{9}}$

$= 0,1334$

και

$$s_{b_1} = 0,1334 \sqrt{\frac{1118,267}{(11952842)(1118,267) - (11559 5,6)^2}}$$

$= 0,000013$

και

$$s_{\hat{\beta}_2} = s_{b_2} = s_e \sqrt{\frac{\sum x_1^2}{\sum x_1^2 \sum x_2^2 - \left(\sum x_1 x_2\right)^2}}$$

$$= 0,1334 \sqrt{\frac{11952842}{(11952842)(1118,267) - (11559 5,6)^2}}$$

$= 0,0042.$

Ο συντελεστής προσδιορισμού μπορεί να υπολογιστεί από τον τύπο:

$$R^2 = \frac{b_1 \sum yx_1 + b_2 \sum yx_2}{\sum y^2} =$$

$$= \frac{0,000116(15219,19) + 0,0115(26,1507)}{2,228} = 0,9273$$

Ο διορθωμένος συντελεστής προσδιορισμού είναι:

$$R^2_{adj} = R^2_a = 1-(1-R^2)\left(\frac{n-1}{n-k-1}\right) =$$

$$= 1-(1-0,9273)\left(\frac{12-1}{12-2-1}\right) = 0,9111.$$

(γ) Ελέγχουμε σε επίπεδο $\alpha = 5\%$ την $H_0: \beta_1 = 0$ έναντι την $H_1: \beta_1 \neq 0$.

Χρησιμοποιούμε την ελεγχοσυνάρτηση:

$$T_{n-k-1} = \frac{b_1 - 0}{s_{b_1}} = \frac{0,000116 - 0}{0,000013} \simeq 9$$

$$\| \quad T$$

Βρίσκουμε την κρίσιμη τιμή από τους πίνακες της $t_{n-k-1} = t_{12-2-1} = t_9$ δηλ. της t-student με $n=9$ β.ε. Για $\alpha/2 = 0.025$, $t_{9,0.025} = 2,262$.

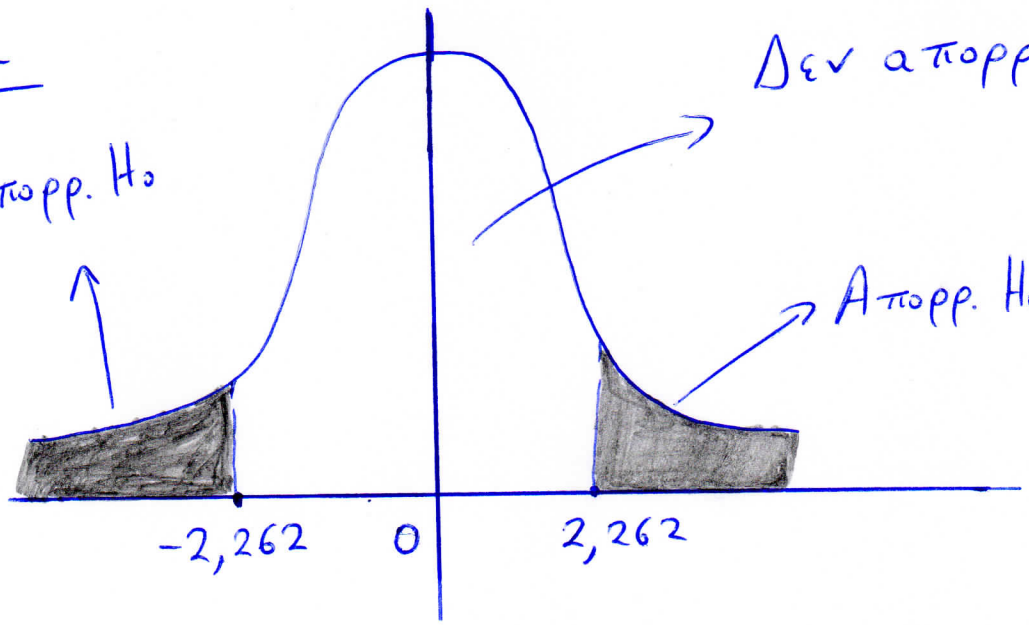
Απορρίπτουμε την H_0 σε επίπεδο $\alpha = 5\%$, αν $|T| > 2,262$

Άρα, η H_0 απορρίπτεται σε εσφ $\alpha = 5\%$, συνεπώς -6- τα δεδομένα παρέχουν ισχυρές ενδείξεις ότι η μεταβλητή X_1 είναι στατιστικά σημαντική για την ερμηνεία των μεταβολών της Y.

Ομοίως σε εσφ $\alpha = 5\%$ ελέγχουμε την $H_0: \beta_2 = 0$ έναντι της $H_1: \beta_2 \neq 0$. Η τιμή της ελεγχοσυνάρτησης είναι: $$T = T_{n-k-1} = T_9 = \frac{b_2}{S_{b_2}} = \frac{0,0115}{0,0042}$$

$= 2,74 > t_{9,0.025} = 2,262$. Άρα, η H_0 απορρίπτεται σε εσφ $\alpha = 5\%$ συνεπώς τα δεδομένα παρέχουν ενδείξεις ότι η μεταβλητή X_2 είναι στατιστικά σημαντική για την ερμηνεία των μεταβολών της Y.

Σχήμα



(δ) Βρίσκουμε αρχικά τους μερικούς συντελεστές συσχέτισης $r_{YX_1}, r_{YX_2}, r_{X_1X_2}$.

$$r_{YX_1} = \frac{\Sigma x_1 y}{\sqrt{\Sigma x_1^2}\sqrt{\Sigma y^2}} = 0,9323$$

$$r_{YX_2} = \frac{\Sigma x_2 y}{\sqrt{\Sigma x_2^2}\sqrt{\Sigma y^2}} = 0,5247$$

$$r_{X_1X_2} = \frac{\Sigma x_1 x_2}{\sqrt{\Sigma x_1^2}\sqrt{\Sigma x_2^2}} = 0,3162$$

Από τους τύπους του μερικού συντελεστή συσχέτισης για την επίδραση της X_1 στην Y έχουμε:

$$R_{Y,X_1|X_2} = \frac{r_{YX_1} - r_{YX_1} r_{X_1X_2}}{\sqrt{1-r_{X_1X_2}^2}\sqrt{1-r_{YX_2}^2}}$$

$$= \frac{0,9323 - 0,5247\cdot(0,3162)}{(0,9487)(0,8513)} = 0,949.$$

Η επίδραση της X_2 στην Y, βρίσκεται από τον τύπο:

$$R_{Y,X_2|X_1} = \frac{r_{YX_2} - r_{YX_1} r_{X_1X_2}}{\sqrt{1-r_{X_1X_2}^2}\sqrt{1-r_{YX_1}^2}}$$

$$= \frac{0,5247 - (0,93236)(0,3162)}{\sqrt{1-(0,3162)^2}\sqrt{1-(0,9323)^2}} = 0,67$$

Όπως φαίνεται από τις τιμές των συντελεστών μερικής συσχέτισης, η μεταβλητή X_2 (σε σχέση με τη X_1) φαίνεται να ασκεί πιο σημαντική επίδραση στην

ερμηνεία των μεταβολών της Υ.

(ε) Τα διαστήματα εμπιστοσύνης για τις πληθυσμιακές παραμέτρους β_1, β_2 με βάση τους εκτιμητές b_1, b_2 είναι:

$$b_1 - 2{,}262\, s_{b_1} \le \beta_1 \le b_1 + 2{,}262\, s_{b_1}$$

$\Rightarrow 0{,}000087 \le \beta_1 \le 0{,}000145$ (με συντελεστή εμπιστοσύνης 95%).

Ομοίως για το β_2: $0{,}002 \le \beta_2 \le 0{,}021$ από:

$b_2 - 2{,}262 s_{b_2} \le \beta_2 \le b_2 + 2{,}262 s_{b_2}$ (με συντελεστή εμπιστοσύνης 95%). Άρα, με εμπιστοσύνη 95%, η πληθυσμιακή παράμετρος β_1 θα βρίσκεται μεταξύ 0,000087 και 0,000145 και η πληθυσμιακή παράμετρος β_2 θα βρίσκεται μεταξύ 0,002 και 0,021.

(ζ) Για να ελέγξουμε τη συνολική στατιστική σημαντικότητα της εξίσωσης παλινδρόμησης:

$H_0: \beta_1 = \beta_2 = 0$

H_1: τουλάχιστον ένας εκ των συντελεστών β_i, $i=1,2$ είναι διάφορος του μηδενός

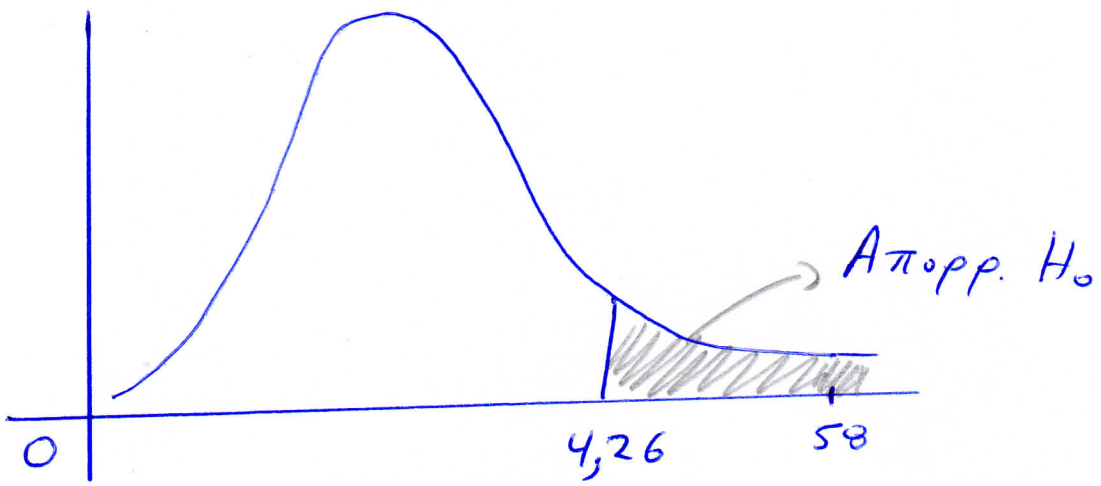
σε ε.σ.σ. $\alpha = 5\%$, χρησιμοποιούμε την ελεγχοσυνάρτηση:

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)} = \frac{\frac{0{,}9273}{2}}{\frac{(1-0{,}9273)}{(12-2-1)}} \cong 58$$

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

$F > f_{k,n-k-1;\alpha} = 4,26$ (από πίνακες της f).

Συνεπώς, υπάρχουν ισχυρές ενδείξεις απόρριψης της H_0 σε επ. $\alpha = 5\%$. Αυτό σημαίνει ότι υπάρχει τουλάχιστον μία ερμηνευτική μεταβλητή που είναι στατιστικά σημαντική.



---

② Σε ένα τ.δ. 12 παρατηρήσεων εφαρμόζουμε υπόδειγμα πολλαπλής παλινδρόμησης με ανεξάρτητες μεταβλητές X_1, X_2 και εξαρτημένη μεταβλητή την Υ. Υπολογίζεται ο παρακάτω πίνακας ανάλυσης διακύμανσης.

Πηγή Μεταβλητότητας	Άθροισμα Τετραγώνων	β.ε	Μέσα Τετράγωνα
Παλινδρόμηση	SSR = 868,56	2	434,28
Κατάλοιπα	SSE = 134,44	9	14,937
Σύνολο	SST = 1003	11	

(α) Να συμπληρωθεί ο πίνακας ανάλυσης διακύμανσης.

Είναι β.ε(SSR) = k = #ανεξάρτητων μεταβλητών = 2

$\frac{SSR}{2} = 434,28 \Rightarrow \boxed{SSR = 868,56}$

$β.ε(SST) = n-1 = 11$

$n = 12$

$β.ε(SSE) = n-k-1 = 12-2-1 = 9$

$MSE = \frac{SSE}{9} = \frac{134,44}{9} = 14,937.$

(β) Να υπολογιστεί ο συντελεστής προσδιορισμού.

Είναι $R^2 = \frac{SSR}{SST} = \frac{868,56}{1003} = 0,8659$ ή 86,59%. Και οι δύο μεταβλητές μαζί (X_1, X_2) εξηγούν (ερμηνεύουν) το 86,59% των συνολικών μεταβολών (της συνολικής μεταβλητότητας) της ανεξάρτητης μεταβλητής Υ.

(γ) Να διεξάγετε στατιστικό έλεγχο σημαντικότητας της εξίσωσης παλινδρόμησης σε επίπεδο $α = 5\%$. Σε ποιο συμπέρασμα καταλήγετε;

Ελέγχουμε σε επίπεδο $α = 5\%$ τις ακόλουθες υποθέσεις:

H_0: η εξίσωση παλινδρόμησης δεν εξηγεί καθόλου τις μεταβολές της Υ (το ποσοστό της εξηγούμενης διασποράς της Υ είναι μηδέν), $β_1 = β_2 = 0$

vs

H_1: η εξίσωση παλινδρόμησης εξηγεί ένα μέρος των μεταβολών της Υ (το ποσοστό της εξηγούμενης διασποράς της Υ είναι μεγαλύτερο του μηδενός) ή ένας τουλάχιστον συντελεστής διάφορος του μηδενός, $β_1$ ή $β_2 \neq 0$.

$F_{k,n-k-1} = \frac{434,28}{14,937} = 29,0741$. Από τους πίνακες της $f_{k,n-k-1,\alpha=0.05}$ έχουμε: $f_{2,9,0.05} = 19,37$.

Άρα, η H_0 απορρίπτεται σε επίπεδο $\alpha = 5\%$. Συνεπώς, υπάρχει τουλάχιστον ένας συντελεστής παλινδρόμησης στατιστικά σημαντικά διάφορος του μηδενός.

(δ) Σε έναν έλεγχο F του πολλαπλού γραμμικού μοντέλου υπολογίζουμε το p-value και το βρίσκουμε ότι είναι μεγαλύτερο του επιπέδου σημαντικότητας α. Τι σημαίνει αυτό;

Αυτό σημαίνει ότι η H_0 <u>δεν</u> απορρίπτεται δηλαδή η εξίσωση παλινδρόμησης <u>δεν</u> φαίνεται να εξηγεί καθόλου τις μεταβολές της Υ (το ποσοστό της εξηγούμενης διασποράς της Υ είναι μηδέν). Επομένως, κανένας εκ των συντελεστών παλινδρόμησης <u>δεν</u> φαίνεται να είναι στατιστικά σημαντικός για το μοντέλο.

(ε) Αν επιπλέον γνωρίζετε ότι ο συντελεστής συσχέτισης Υ με την X_1 είναι 0,84, ο συντελεστής συσχέτισης της Υ με την X_2 είναι 0,81 και ο συντελεστής συσχέτισης της X_1 με την X_2 είναι 0,58 να απαντήσετε στα ακόλουθα ερωτήματα:

(i) Ποιο από τα ακόλουθα υποδείγματα:

$Y = \beta_0 + \beta_1 X_1 + \varepsilon$ (1), $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ (2)

Θεωρείτε ότι πρέπει να προτιμηθεί; Να δικαιολογήσετε την απάντησή σας.

Λύση Είναι $r_{Y,X_1} = 0,84 \Rightarrow r^2_{Y,X_1} = R^2_{(1)} = 0,7056$ δηλ. για το μοντέλο (1) που περιέχει μόνο την X_1, από μόνη της η X_1 "εξηγεί" το 70,56% της συνολικής μεταβλητότητας της Υ.

Για το μοντέλο (2) όπως είδαμε, ο συντελεστής προσδιορισμού προέκυψε ίσος με 86,59% δηλαδή μαζί οι X_1, X_2 εξηγούν το 86,59% των μεταβολών της Υ. Επειδή επίσης ο $r_{X_1,X_2} = 0,58$ δεν είναι πάρα πολύ μεγάλος (ώστε να δημιουργεί μεγάλα προβλήματα συγγραμμικότητας) και επειδή όπως θα δούμε στο (ii) ο συντελεστής μερικού προσδιορισμού για την X_2 προκύπτει ίσος με 53,3% Θεωρούμε ότι πρέπει να προτιμηθεί το μοντέλο (2) διότι εξηγεί σε μεγαλύτερο ποσοστό τη συνολική μεταβλητότητα της Υ (εξηγεί μεγαλύτερο "μέρος" των μεταβολών της Υ σε σχέση με το μοντέλο (1)).

(ii) Να υπολογίσετε τον συντελεστή μερικού προσδιορισμού της Υ με τη μεταβλητή X_2, όταν η X_1 βρίσκεται ήδη στο υπόδειγμα και να δώσετε την ερμηνεία.

Μπορούμε να βρούμε τον $R^2_{Y,X_2|X_1}$, συντελεστή μερικού προσδιορισμού της μεταβλητής X_2. Έτσι "μετράμε" την επίδραση της X_2 στις μεταβολές της Υ αφού πρώτα "αφαιρέσουμε" τις επιδράσεις της X_1 στην Υ και στην X_2.

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

όπου $R_{Y,X_2} = r_{Y,X_2}$, $R_{Y,X_1} = r_{Y,X_1}$, $R_{X_1,X_2} = r_{X_1,X_2}$

Βρίσκουμε:

$$R^2_{Y,X_2|X_1} = \frac{(0,81 - 0,84 \cdot 0,58)^2}{(1-0,58^2)(1-0,84^2)} = 0,5330$$

που σημαίνει ότι μετά την αφαίρεση των επιδράσεων της X_1 στην Υ και στην X_2 ποσοστό 53,3% της ανερμήνευτης μεταβλητότητας της Υ ερμηνεύεται από την μεταβλητή X_2.

ξ