Φροντιστήριο #6

Άλλα είδη συσχέτισης

Ένα εναλλακτικό μοντέλο της απλής γραμμικής παλινδρόμησης που χρησιμοποιείται συχνά είναι το ακόλουθο μοντέλο:

$$Y = \beta_0 X^{\beta_1} e^{u} \quad \text{όπου } u \text{ είναι το σφάλμα}$$

ή κατάλοιπο για το οποίο: $u \sim N(0, \sigma_u^2)$. Ο συντελεστής ελαστικότητας για το μοντέλο αυτό είναι σταθερός για όλες τις τιμές των μεταβλητών X και Y, διότι:

$$\eta_{Y|X} = \frac{dY}{dX} \cdot \frac{X}{\hat{Y}} = \beta_1 \beta_0 X^{\beta_1 - 1} \frac{X}{\hat{Y}}$$

$$= \beta_1 \beta_0 X^{\beta_1} \frac{1}{\beta_0 X^{\beta_1}} = \beta_1 .$$

Η μορφή της καμπυλόγραμμης σχέσης που αντιστοιχεί στο μοντέλο εξαρτάται από την τιμή του συντελεστή β_1 (βλέπε Σχήματα 9.1, 9.2, βιβλίο Χαλικιά).

Μετά από κατάλληλο μετασχηματισμό του μοντέλου μπορούμε να γράψουμε:

$$\ln(Y) = \ln(\beta_0 X^{\beta_1} e^{u})$$

$$\Rightarrow \ln(Y) = \ln(\beta_0) + \beta_1 \ln(X) + u$$

Οι εκτιμώμενες τιμές της Y δίνονται από:

$$\hat{Y} = \hat{\beta}_0 X^{\beta_1} \quad \text{ή} \quad \ln(\hat{Y}) = \ln(\hat{\beta}_0) + \hat{\beta}_1 \ln(X)$$

Για τα σφάλματα u_i μπορούμε να γράψουμε, $i=1,\dots,n$:

$$u_i = \ln(y_i) - \beta_1 \ln(x_i) - \ln(\beta_0)$$

και

$$g(\ln(\beta_0), \beta_1) = \sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} \left(\ln(y_i) - \beta_1 \ln(x_i) - \ln(\beta_0)\right)^2$$

Όμως

$$\frac{\partial g(\ln(\beta_0), \beta_1)}{\partial(\ln(\beta_0))} = 2\sum_{i=1}^{n} \left(\ln(y_i) - \beta_1 \ln(x_i) - \ln(\beta_0)\right)(-1) = 0$$

$$\Rightarrow \sum_{i=1}^{n} \ln(y_i) = n \ln(\beta_0) + \beta_1 \sum_{i=1}^{n} \ln(x_i)$$

και

$$\frac{\partial g(\ln(\beta_0), \beta_1)}{\partial \beta_1} = 2\sum_{i=1}^{n} \left(\ln(y_i) - \beta_1 \ln(x_i) - \ln(\beta_0)\right)(-\ln(x_i)) = 0$$

$$\Rightarrow \sum_{i=1}^{n} \ln(y_i)\ln(x_i) = \ln(\beta_0) \sum_{i=1}^{n} \ln(x_i) + \beta_1 \sum_{i=1}^{n} (\ln(x_i))^2$$

Άρα, έχουμε τις εξισώσεις:

$$\sum_{i=1}^{n} \ln(y_i) = n \ln(\beta_0) + \beta_1 \sum_{i=1}^{n} \ln(x_i)$$

$$\sum_{i=1}^{n} \ln(y_i) \ln(x_i) = \ln(\beta_0) \sum_{i=1}^{n} \ln(x_i) + \beta_1 \sum_{i=1}^{n} \left(\ln(x_i)\right)^2$$

Συνεπώς, με τη μέθοδο των οριζουσών οι ε.ε.τ. είναι $\hat{\beta}_1$, $\widehat{\ln(\beta_0)}$ και δίνονται από τις εξισώσεις:

$$\hat{\beta}_1 = \frac{\begin{vmatrix} n & \sum_{i=1}^{n} \ln(y_i) \\ \sum_{i=1}^{n} \ln(x_i) & \sum_{i=1}^{n} \ln(x_i)\ln(y_i) \end{vmatrix}}{\begin{vmatrix} n & \sum_{i=1}^{n} \ln(x_i) \\ \sum_{i=1}^{n} \ln(x_i) & \sum_{i=1}^{n} \left(\ln(x_i)\right)^2 \end{vmatrix}}$$

$$= \frac{n \sum_{i=1}^{n} \ln(x_i)\ln(y_i) - \sum_{i=1}^{n} \ln(x_i) \sum_{i=1}^{n} \ln(y_i)}{n \sum_{i=1}^{n} \left(\ln(x_i)\right)^2 - \left(\sum_{i=1}^{n} \ln(x_i)\right)^2}$$

και

$$\widehat{\ln(\beta_0)} = \frac{\sum_{i=1}^{n} \ln(y_i)}{n} - \frac{\hat{\beta}_1 \sum_{i=1}^{n} \ln(x_i)}{n}$$

Ο συντελεστής γραμμικής συσχέτισης για το μοντέλο διαμορφώνεται ως εξής:

$$r = \frac{n\sum_{i=1}^{n} \ln(x_i)\ln(y_i) - \sum_{i=1}^{n}\ln(x_i)\sum_{i=1}^{n}\ln(y_i)}{\sqrt{n\sum_{i=1}^{n}(\ln(x_i))^2 - \left(\sum_{i=1}^{n}\ln(x_i)\right)^2}\sqrt{n\sum_{i=1}^{n}(\ln(y_i))^2 - \left(\sum_{i=1}^{n}\ln(y_i)\right)^2}}$$

Ανάλυση συσχέτισης κατά τάξεις

Ανάλυση συσχέτισης και παλινδρόμησης εφαρμόσαμε μέχρι τώρα στην περίπτωση όπου οι μεταβλητές είναι ποσοτικές ή μετριούνται σε κλίμακα διαστήματος. Η συσχέτιση κατά τάξεις χρησιμοποιείται όταν οι μεταβλητές είναι ιεραρχημένες (ή μπορούν να ιεραρχηθούν). Τότε οι τιμές των μεταβλητών μπορούν μόνο να ιεραρχηθούν από τη μικρότερη στη μεγαλύτερη τιμή αλλά δεν μπορούν να αφαιρεθούν ή να διαιρεθούν μεταξύ τους οπότε σε κάθε μία από αυτές τις περιπτώσεις έχουμε κλίμακα διαστήματος ή κλίμακα λόγου.

Ο κατάλληλος συντελεστής συσχέτισης για τέτοιου τύπου δεδομένα ιεραρχικής κλίμακας είναι ο συντελεστής συσχέτισης κατά τάξεις του Spearman.

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

όπου, $n :=$ μέγεθος δείγματος

και $d_i = R_X - R_Y$ η διαφορά μεταξύ των τάξεων μεγέθους για κάθε ζεύγος παρατηρήσεων

Π.Χ. αν $Y :=$ βαθμολογία

$X :=$ οικογενειακό εισόδημα

και $r_s = 0.776$ (έχουμε έντονη θετική συσχέτιση μεταξύ των τάξεων μεγέθους του οικογενειακού εισοδήματος και της βαθμολογίας).

Δηλ. τα παιδιά των οικονομικά πιο εύρωστων οικογενειών επιτυγχάνουν υψηλότερη μέση βαθμολογία.

Έλεγχος στατιστικής σημαντικότητας του r_s

Ελέγχουμε σε ε.σ.σ. α την $H_0 : \rho_s = 0$ vs

$H_1 : \rho_s \neq 0$

όπου ρ_s είναι ο πληθυσμιακός συντελεστής συσχέτισης κατά τάξεις μεταξύ των μεταβλητών X και Y. Χρησιμοποιούμε την σ.σ. ελέγχου:

$$T_{n-2} = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}} \overset{H_0}{\sim} t_{n-2}$$

Απορρίπτουμε την H_0 σε ε.σ.σ. α αν

$|T_{n-2}| > t_{n-2,\alpha/2}$. Η τάξη μεγέθους για δύο ίδιες τιμές είναι ο μέσος όρος των δύο τάξεων:

π.χ. $\frac{9+10}{2}=9.5$

αν συμπίπτουν 3 τιμές $\frac{4+5+6}{3}=5$

Ισχύει ότι $-1 \leq r_s \leq +1$, με ανάλογες ερμηνείες για τις τιμές τους όπως αυτές που είδαμε για τον συντελεστή γραμμικής συσχέτισης r, $-1 \leq r \leq +1$.

Παραδείγματα

① 9.1 (Χαλικιάς) Η επιθεώρηση εργασίας διεξάγει έρευνα για να διαπιστώσει αν οι δαπάνες των επιχειρήσεων για την ασφάλεια των εργαζόμενων συμβάλλουν στον περιορισμό των εργατικών ατυχημάτων. Από δείγμα 10 μεγάλων επιχειρήσεων του κατασκευαστικού κλάδου έχουμε τα στοιχεία

Εταιρεία	(X) Ετήσια δαπάνη ανά εργαζόμενο (€)	(Y) ετήσιος αριθμός ατυχημάτων	R(X)	R(Y)
Α	65	2	1	9.5
Β	36	8	7	2.5
Γ	33	7	8	4
Δ	23	10	10	1
Ε	27	6	9	5
Ζ	45	5	5	6
Η	42	8	6	2.5
Θ	61	2	3	9.5
Ι	52	4	4	7
Κ	63	3	2	8

Χρησιμοποιώντας τον r_s, σε $\alpha=5\%$, υπάρχει στατιστικά σημαντική συσχέτιση ανάμεσα στις δαπάνες και τον αριθμό των εργατικών ατυχημάτων;

d_i^2
72.25
20.25
16
81
16
1
12.25
42.25
9
36
$\sum_{i=1}^{10} d_i^2 = 306$

Λύση

Ο συντελεστής συσχέτισης κατά τάξεις είναι:

$$r_s = 1 - \frac{6\sum_{i=1}^{10} d_i^2}{n(n^2-1)}$$

$$= 1 - \frac{6\cdot 306}{10(10^2-1)} = -0.855$$

Έλεγχος στατιστικής σημαντικότητας του r_s. Χρησιμοποιούμε την ελεγχοσυνάρτηση:

$$T_{n-2} = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}}$$

$H_0: \rho_s = 0$

$H_1: \rho_s \neq 0$

$$= -\frac{0.855\sqrt{8}}{\sqrt{1-0,855^2}} = -4.66$$

Από τους πίνακες της t_8 βρίσκουμε για $\alpha = 5\%$

ότι $t_{8,\alpha/2} = t_{8,0.025} = 2,306$

Άρα $|T_{n-2}| = 4.66 > 2,306$ άρα η H_0 απορρ. σε εσσ $\alpha = 5\%$ άρα τα δεδομένα παρέχου ισχυρές ενδείξεις ότι υπάρχει έντονη σημαντική στατιστική και μάλιστα οι ενδείξεις δείχνουν αρνητική συσχέτιση μεταξύ των δαπανών και του αριθμού των εργατικών ατυχημάτων.

(2) (9.2 Χαλικιάς) Ο υπουργός Εργασίας ισχυρίζεται ότι μία αύξηση από τις επιχειρήσεις κατά 10% των δαπανών για την ασφάλεια των εργαζομένων προκαλεί μείωση του αριθμού των ατυχημάτων κατά 15%. Ποιο υπόδειγμα πρέπει να προσαρμόσετε στα δεδομένα του παραδείγματος (1) για να ελέγξετε σε εσσ $\alpha=5\%$ τον ισχυρισμό του υπουργού;

Λύση Ο ισχυρισμός του υπουργού υποδηλώνει ότι η ελαστικότητα του αριθμού των εργατικών ατυχημάτων ως προς τις δαπάνες για τη ασφάλεια των εργαζομένων είναι σταθερή και άρα το υπόδειγμα που περιγράφει τη σχέση μεταξύ των δύο μεταβλητών είναι της μορφής: $\hat{Y}=\hat{\beta}_0 X^{\hat{\beta}_1}$ ή

$$\ln(\hat{Y})=\widehat{\ln(\beta_0)}+\hat{\beta}_1 \ln(X)$$ όπου

Y: = ετήσιος αριθμός εργατικών ατυχημάτων

X: = ετήσια δαπάνη ανά εργαζόμενο (σε €).

Για τον υπολογισμό των $\hat{\beta}_1$, $\widehat{\ln(\beta_0)}$ χρειαζόμαστε τις στήλες $\ln(X)$, $\ln(Y)$ από τις οποίες παίρνουμε: $$\sum_{i=1}^{10} \ln(x_i)=37,435$$

$$\sum_{i=1}^{10} \ln(y_i)=15,680$$

$$\sum_{i=1}^{10} \ln(x_i)\ln(y_i) = 57{,}052, \quad \sum_{i=1}^{10}(\ln(x_i))^2 = 141{,}325$$

$$\sum_{i=1}^{10}(\ln(y_i))^2 = 27{,}627$$

$\ln(X)$	$\ln(Y)$
4,174	0,693
3,584	2,079
3,497	1,946
3,135	2,303
3,296	1,792
3,807	1,609
3,738	2,079
4,111	0,693
3,951	1,386
4,143	1,099

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n}\ln(x_i)\ln(y_i) - \sum_{i=1}^{n}\ln(x_i)\sum_{i=1}^{n}\ln(y_i)}{n\sum_{i=1}^{n}(\ln(x_i))^2 - \left[\sum_{i=1}^{n}\ln(x_i)\right]^2}$$

$$\overset{n=10}{=} \frac{10\cdot(57{,}052) - (37{,}435)(15{,}680)}{10\cdot(141{,}325) - (37{,}435)^2} = -1{,}390$$

$$\widehat{\ln(\beta_0)} = \frac{\sum_{i=1}^{n}\ln(y_i)}{n} - \hat{\beta}_1 \frac{\sum_{i=1}^{n}\ln(x_i)}{n}$$

$$= \frac{15{,}680}{10} - (-1{,}390)\frac{37{,}435}{10} = 6{,}771$$

Άρα, το εκτιμώμενο μοντέλο παλινδρόμησης θα είναι:

$$\boxed{\ln(\hat{Y}) = 6{,}771 - 1{,}390\ln(X)}$$

Εάν ο ισχυρισμός του Υπουργού είναι σωστός η ελαστικότητα των εργατικών ατυχημάτων ως προς τις δαπάνες για την ασφάλεια των εργαζομένων ισούται με −1,5 (ενώ η εκτίμηση από τα δεδομένα των δείγματος βρέθηκε ίση με -1.39). Θα ελέγξουμε σε εσσ $\alpha=5\%$ αν $H_0: \beta_1=-1.5$ vs $H_1: \beta_1 \neq -1.5$

Χρησιμοποιούμε τη σ.σ. ελέγχου:

$$T_{n-2}=\frac{|\beta_1-\beta_1^*|}{S_{\hat{\beta}_1}}, \quad \beta_1^*=-1.5$$

$$S_{\hat{\beta}_1}=\frac{s}{\sqrt{\sum_{i=1}^{n}(\ln(x_i))^2-n\left(\frac{\sum_{i=1}^{10}\ln(x_i)}{10}\right)}}, \quad (n=10)$$

όπου $s^2=\frac{1}{8}(SST-SSR)$

$$SST=\sum_{i=1}^{10}(\ln(y_i))^2-n\left(\frac{\sum_{i=1}^{n}\ln(y_i)}{n}\right)^2$$

$$=27,627-\frac{1}{10}(15,680)^2=27,627-24,58624$$
$$=3,04076$$

$$SSR=\hat{\beta}_1^2\left(\sum_{i=1}^{10}(\ln(x_i))^2-n\left(\frac{\sum_{i=1}^{n}\ln(x_i)}{n}\right)^2\right)$$

$$=1,9321\left(141,325-\frac{1}{10}(37,435)^2\right)$$

$$=1,9321\left(141,325-140,1379\right)$$

$$=2,29359$$

$$s^2=\frac{1}{8}(3,04076-2,29359)=0,0933$$

$$\Rightarrow \boxed{s=0,305}$$

Άρα $s_{\hat{\beta}_1} = \dfrac{0,305}{1,089} \cong 0,28$

Συνεπώς, $|T_{n-2}| = \dfrac{|\hat{\beta}_1 - \beta_1^*|}{s_{\hat{\beta}_1}} = \dfrac{|-1.39-(-1.5)|}{0,28}$

$= 0,39$

Από πίνακες της t_8, έχουμε $t_{8,0.025} = 2,306$

δηλ. $|T_{n-2}| < 2,306$ άρα η $H_0: \beta_1 = -1.5$ δεν απορρίπτεται σε εσσ $\alpha = 5\%$, και τα δεδομένα παρέχουν ενδείξεις υπέρ του ισχυρισμού του υπουργού.

Παρατήρηση Αν έχουμε ισοβαθμίες μπορούμε να "βελτιώσουμε" ("διορθώσουμε") την τιμή του r_s χρησιμοποιώντας τον **εναλλακτικό** τύπο:

$$r_s = \frac{\sum x^2 + \sum y^2 - \sum d_i^2}{2\sqrt{\sum x^2}\sqrt{\sum y^2}}$$

όπου

$$\sum x^2 = \frac{n^3-n}{12} - \frac{1}{12}\sum_j z_j(z_j^2-1) \text{ για την } X$$

και

$$\sum y^2 = \frac{n^3-n}{12} - \frac{1}{12}\sum_h z_h(z_h^2-1) \text{ για την } Y$$

όπου z_j, z_h: μέγεθος ισοψηφιών.

π.χ. $X :=$ # παιδιών ανά οικογένεια

$Y :=$ IQ του νεότερου παιδιού.

Υπολογίστε τον κατάλληλο συντελεστή συσχέτισης για τις X και Y, στο δείγμα των $n=6$ μετρήσεων.

X	Y	R_X	R_Y	$d_i = R_X - R_Y$	d_i^2
2	110	1	6	−5	25
3	100	3	4.5	−1.5	2.25
3	100	3	4.5	−1.5	2.25
3	80	3	2.5	0.5	0.25
4	80	5	2.5	2.5	6.25
5	70	6	1	5	25

$$61 = \sum_{i=1}^{6} d_i^2$$

$n=6$

$$\sum x^2 = \frac{6^3-6}{12} - \frac{1}{12}\,3(3^2-1)$$ (μία ισοβαθμία μεγέθους 1)

$$= 17{,}5 - 2 = 15{,}5$$

(δύο ισοβαθμίες μεγέθους 2)

$$\sum y^2 = \frac{6^3-6}{12} - \frac{1}{12}\left(2(2^2-1)+2(2^2-1)\right)$$

$$= 17{,}5 - \frac{1}{12}(6+6) = 16{,}5$$

$$r_s = \frac{15{,}5+16{,}5-61}{2\sqrt{15{,}5}\sqrt{16{,}5}} = -0{,}906698$$

$H_0: \rho_s = 0$

$H_1: \rho_s < 0$

$$T_{n-2} = r_s \frac{\sqrt{n-2}}{\sqrt{1-r_s^2}}$$

$$= -0{,}9066\,\frac{\sqrt{4}}{\sqrt{1-0{,}821}} = -4{,}2856$$

Απορρίπτουμε H_0 σε εσσ $\alpha = 1\%$ αν

$$T_{n-2} < -t_{4,0.01} = -3,747$$

η H_0 απορρίπτεται άρα φαίνεται να υπάρχει στατιστικά σημαντική αρνητική διατακτική συσχέτιση μεταξύ των X και Y.