ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ
ΜΕΡΟΣ Β

Δημήτρης Κουγιουμτζής
e-mail: ***dkugiu@auth.gr***

Ιστοσελίδα αυτού του τμήματος του μαθήματος:
http://users.auth.gr/~dkugiu/Teach/CivilTransport/index.html

Εφαρμοσμένη Στατιστική:

- **Συντελεστής συσχέτισης**
- **Παλινδρόμηση**
  απλή γραμμική, πολλαπλή γραμμική
- **Χρονικές σειρές**
  στασιμότητα, αυτοσυσχέτιση,
  μοντέλα αυτοπαλινδρόμησης

Παραδείγματα

τιμής εισιτηρίου ~
χώρου στάθμευσης

αριθμός σηματοδοτημένων διασταυρώσεων ~
χρόνος κάλυψης μιας διαδρομής

κατανάλωση ηλεκτρικής ενέργειας κατοικίας ~
κατανάλωση νερού ~
μέγεθος κατοικίας

Εξαρτάται η μια τ.μ. από την άλλη?

Εξαρτιούνται και οι δύο από κάποια άλλη?

Συσχέτιση και Παλινδρόμηση

Δύο τ.μ.: X με διασπορά σ_X^2 , Y με σ_Y^2

συνδιασπορά $\sigma_{XY} \equiv \mathrm{Cov}[X,Y]$

$$\sigma_{XY} = \mathrm{E}[(X-\mu_X)(Y-\mu_Y)] = \mathrm{E}[X\,Y] - \mu_X\,\mu_Y$$

- εκφράζει τη γραμμική συσχέτιση δύο τ.μ. δηλαδή την αναλογική μεταβολή (αύξηση ή μείωση) της μιας τ.μ. που αντιστοιχεί σε μεταβολή της άλλης μεταβλητής
- εξαρτάται από τις μονάδες μέτρησης των δύο τ.μ.

συντελεστής συσχέτισης $\rho = \dfrac{\sigma_{XY}}{\sigma_X\,\sigma_Y}$

- ρ δεν εξαρτάται από τη μονάδα μέτρησης των X και Y
- ρ είναι συμμετρικός ως προς τις X και Y.
- $\rho \in$ [-1,1]

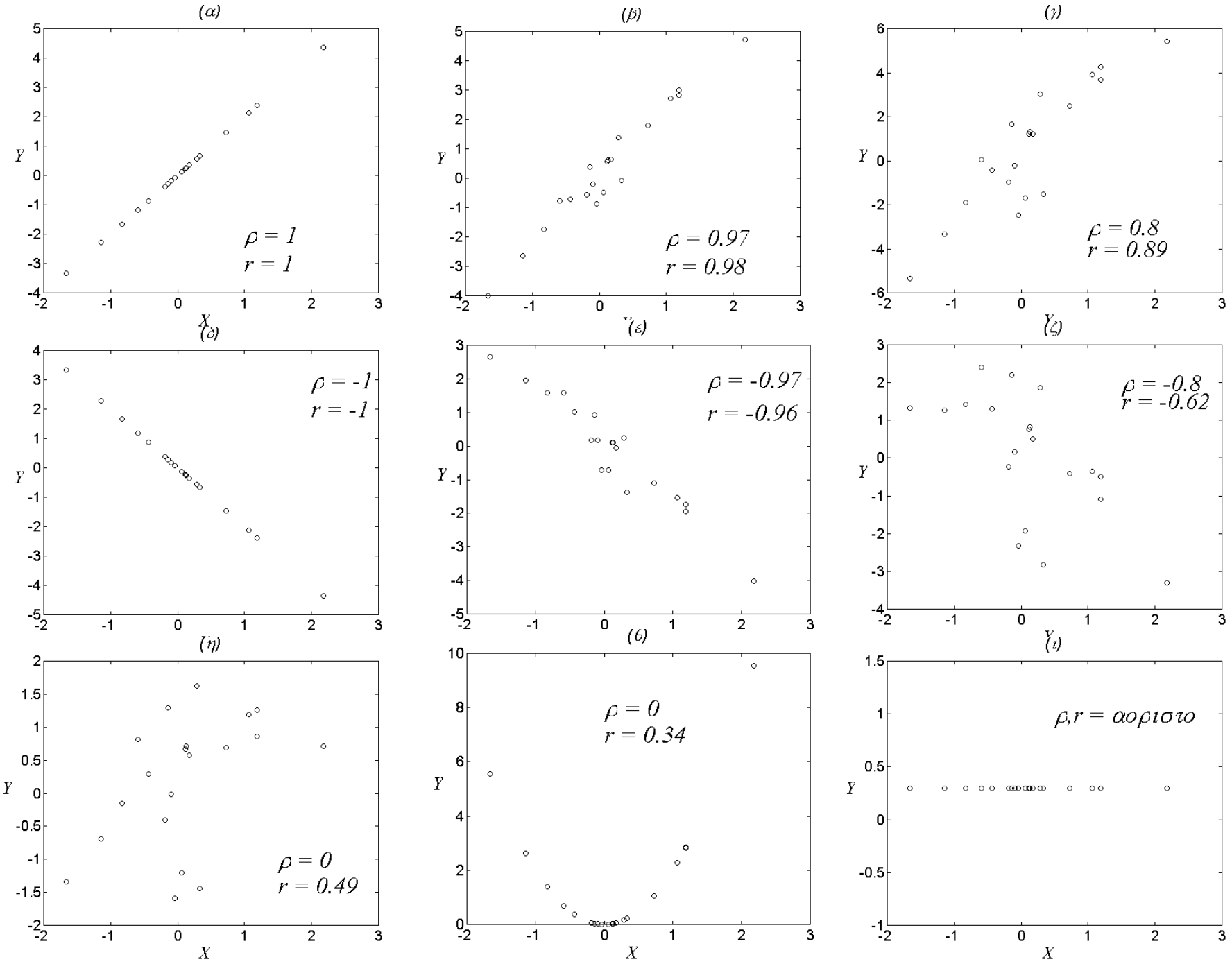
Ερμηνεία της τιμής του συντελεστή συσχέτισης

- *ρ=1*: υπάρχει *τέλεια θετική* σχέση μεταξύ των *X* και *Y*.
- *ρ=0*: δεν υπάρχει καμιά (γραμμική) σχέση μεταξύ των *X* και *Y*.
- *ρ=-1*: υπάρχει *τέλεια αρνητική* σχέση μεταξύ των *X* και *Y*.
- "*ρ κοντά στο 1* "
  ➔ η γραμμική συσχέτιση των δύο τ.μ. είναι θετική και ισχυρή
- "*ρ κοντά στο -1* "
  ➔ η γραμμική συσχέτιση των δύο τ.μ. είναι αρνητική και ισχυρή
- "*ρ κοντά στο 0* " ➔ οι τ.μ. είναι πρακτικά ασυσχέτιστες

Παρατηρήσεις των δύο τ.μ. X και Y κατά ζεύγη

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

για ποιοτική εκτίμηση της συσχέτισης ➔ ***διάγραμμα διασποράς***


(α)
ρ = 1
r = 1
(β)
ρ = 0.97
r = 0.98
(γ)
ρ = 0.8
r = 0.89
(δ)
ρ = -1
r = -1
(ε)
ρ = -0.97
r = -0.96
(ζ)
ρ = -0.8
r = -0.62
(η)
ρ = 0
r = 0.49
(θ)
ρ = 0
r = 0.34
(ι)
ρ,r = αοριστο
X
Y

Σημειακή εκτίμηση του συντελεστή συσχέτισης

Σημειακή εκτίμηση του ρ από το δείγμα των n ζευγαρωτών παρατηρήσεων των X και Y

$$\hat{\rho} \equiv r = \frac{s_{XY}}{s_X s_Y}$$

αμερόληπτη εκτιμήτρια s_{XY}

$$\sigma_{XY} = \mathrm{E}[(X-\mu_X)(Y-\mu_Y)] = \mathrm{E}[X\,Y] - \mu_X \mu_Y$$

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}\right)$$

αμερόληπτες εκτιμήτριες s_X και s_Y
είναι οι τετραγωνικές ρίζες των δειγματικών διασπορών

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$$

Άρα

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

συντελεστής συσχέτισης Pearson

συντελεστής προσδιορισμού $100r^2\%$

γνωρίζοντας τη μια μεταβλητή μπορούμε να προσδιορίσουμε το $100r^2\%$ της μεταβλητότητας της άλλης μεταβλητής

Διάστημα εμπιστοσύνης για ρ ?

Ναι, αν X και Y ακολουθούν τη δι-μεταβλητή κανονική κατανομή

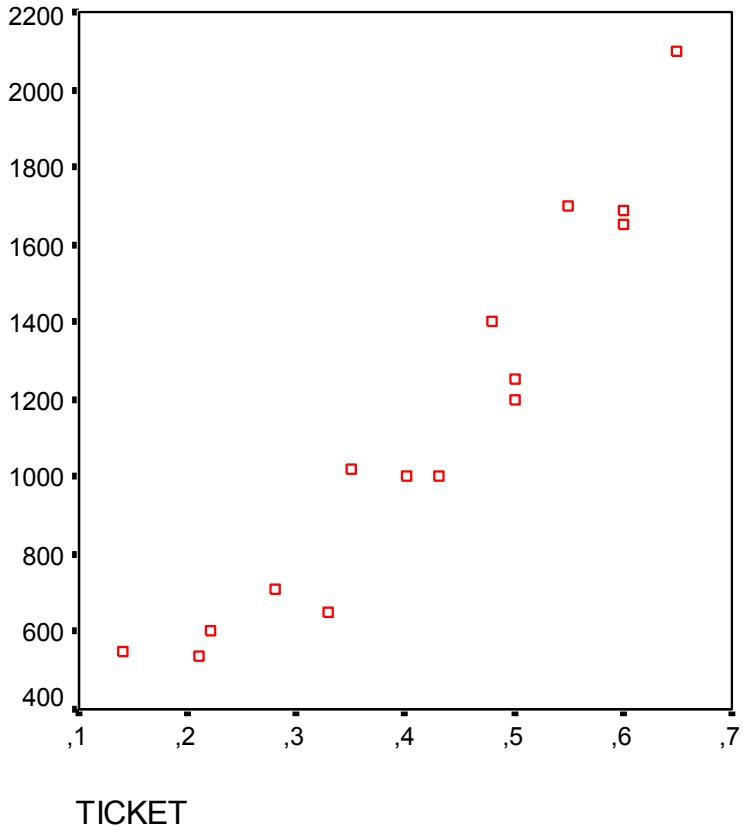
Έλεγχος υπόθεσης για ρ ?

Ναι, αν X και Y ακολουθούν τη δι-μεταβλητή κανονική κατανομή

$H_0 : \rho = 0$ ⟹ εξετάζει αν τα X και Y είναι γραμμικά ανεξάρτητα

Παράδειγμα

Θέλουμε να ελέγξουμε αν το αντίτιμο του εισιτηρίου αστικού λεωφορείου σε μια πόλη συσχετίζεται με τη χρήση χώρων στάθμευσης στο κέντρο της πόλης.



εισιτήριο λεωφορείου	*χώροι στάθμευσης*
,14	550,00
,21	540,00
,22	600,00
,28	710,00
,33	650,00
,35	1020,00
,40	1000,00
,43	1000,00
,48	1400,00
,50	1200,00
,50	1250,00
,55	1700,00
,60	1690,00
,60	1650,00
,65	2100,00

<u>Ποιοτική εκτίμηση</u>

συσχέτιση της τιμής εισιτηρίου και χώρου στάθμευσης είναι θετική και ισχυρή

Υπολογισμός του *r*

$$\bar{x} = 0.416 \qquad \bar{y} = 1137.3$$

$$\sum_{i=1}^{20} x_i^2 = 2.942 \qquad \sum_{i=1}^{20} y_i^2 = 22762200 \qquad \sum_{i=1}^{20} x_i y_i = 8123.7$$

$$r = \frac{8123.7 - 15 \cdot 0.416 \cdot 1137.3}{\sqrt{\left(2.942 - 15 \cdot 0.416^2\right) \cdot \left(22762200 - 15 \cdot 1137.3^2\right)}} = 0.952$$

Η μεταβλητότητα της μιας τ.μ. (εισιτήριο ή χώρος στάθμευσης) μπορεί να εξηγηθεί σε μεγάλο ποσοστό από τη συσχέτιση της με την άλλη

Συμπέρασμα

Η γνώση της μιας τ.μ. μας επιτρέπει να προσδιορίσουμε την άλλη με μεγάλη ακρίβεια

... και συγκεκριμένα ...

γνωρίζοντας την τιμή του εισιτηρίου μπορούμε να καθορίσουμε ~90% της μεταβλητότητας του χώρου στάθμευσης

Correlations

		TICKET	PARK1
TICKET	Pearson Correlation	1	,952**
	Sig. (2-tailed)	.	,000
	N	15	15
PARK1	Pearson Correlation	,952**	1
	Sig. (2-tailed)	,000	.
	N	15	15

**. Correlation is significant at the 0.01 level

Παλινδρόμηση

Επίδραση του αριθμού φαναριών στο χρόνο μιας διαδρομής?

Ζητάμε να εκτιμήσουμε την (γραμμική) εξάρτηση του αριθμού φαναριών στο χρόνο μιας διαδρομής.

εξαρτημένη μεταβλητή Y: χρόνος διαδρομής

ανεξάρτητη μεταβλητή X: αριθμός φαναριών

Μελέτη της μεταβλητότητα μιας τ.μ. *Y* χρησιμοποιώντας την πληροφορία από κάποια άλλη μεταβλητή *X*

➔ ***ανάλυση παλινδρόμησης***

παλινδρόμηση:

απλή: σχέση εξάρτησης μόνο ως προς μια ανεξάρτητη μεταβλητή
πολλαπλή: σχέση εξάρτησης ως προς περισσότερες από μια ανεξάρτητες μεταβλητές
γραμμική: η πιο απλή σχέση εξάρτησης (αναλογική)
μη-γραμμική:

Απλή γραμμική παλινδρόμηση

Υπόθεση απλής γραμμικής παλινδρόμησης:

$$\mathrm{E}[Y \mid X = x] = \alpha + \beta x$$

μέση τιμή της τ.μ. Y για κάθε τιμή x της X:

Άλλες υποθέσεις:

$$\mathrm{Var}[Y \mid X = x] \equiv \sigma^2_{Y|X} = \sigma^2$$

$$Y \mid X = x \;\sim\; \mathrm{N}\left(\alpha + \beta x, \sigma^2\right)$$

$$\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$$

για κάποια τιμή x_i της X μπορεί να αντιστοιχούν διαφορετικές τιμές y_i της Y
➔ y_i είναι τ.μ.

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

ε_i: ***σφάλμα παλινδρόμησης***

$$\varepsilon_i = y_i - \mathrm{E}[Y \mid X = x_i]$$

Εκτίμηση παραμέτρων απλής γραμμικής παλινδρόμησης

Εκτίμηση των τριών παραμέτρων της παλινδρόμησης:

- της **διαφοράς ύψους (σταθερός όρος)** της ευθείας παλινδρόμησης α
- της **κλίσης** ή του **συντελεστή** της ευθείας παλινδρόμησης β
- της **διασποράς σφάλματος** της παλινδρόμησης σ^2

από το δείγμα $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$

Εκτίμηση των α και β

Η εκτίμηση των παραμέτρων *α* και *β* γίνεται με τη μέθοδο των ***ελαχίστων τετραγώνων***
βρίσκει την ευθεία παλινδρόμησης με παραμέτρους *α* και *β* έτσι ώστε το άθροισμα των τετραγώνων των κατακόρυφων αποστάσεων των σημείων από την ευθεία να είναι το ελάχιστο.

Οι εκτιμήσεις των α και β δίνονται από την ελαχιστοποίηση του αθροίσματος των τετραγώνων των σφαλμάτων

$$b = \frac{s_{XY}}{s_X^2}$$

και

$$a = \bar{y} - b\,\bar{x}$$

s_{XY}
δειγματική συνδιασπορά των X και Y

s_X^2
δειγματική διασπορά της X

Τα a και b ορίζουν την ευθεία

$$\hat{y} = a + b\,x$$

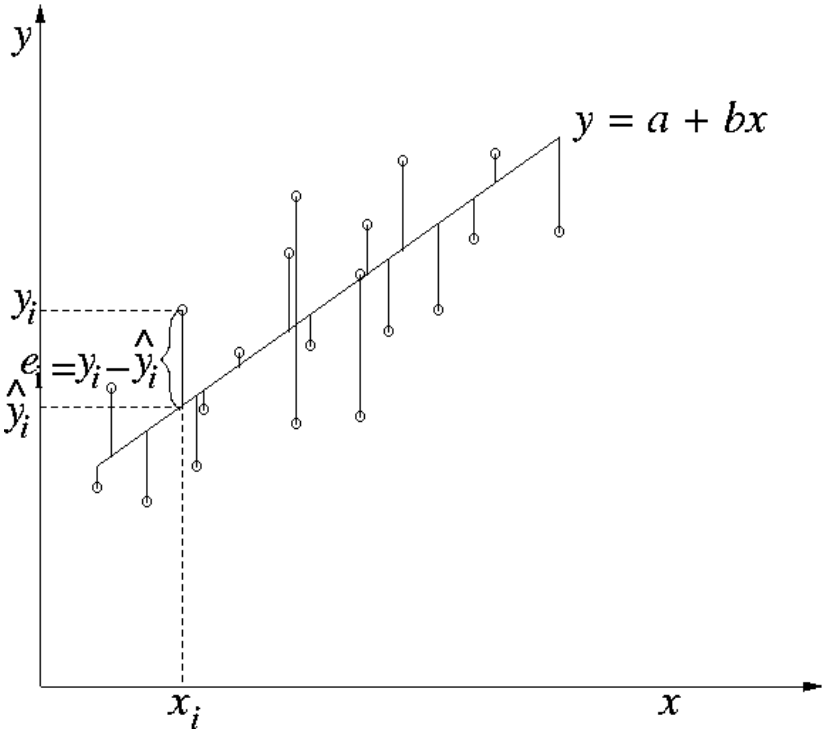
ευθεία ελαχίστων τετραγώνων

Εκτίμηση του σ^2

Για κάθε x_i ➔ $\hat{y}_i = a + b\,x_i$

υπόλοιπο ή σφάλμα ελαχίστων τετραγώνων

$$e_i = y_i - \hat{y}_i = y_i - a - b\,x_i$$



$$\varepsilon_i = y_i - \alpha - \beta\,x_i$$

e_i είναι η εκτίμηση του ε_i,

Εκτίμηση της διασποράς του σφάλματος σ^2

δειγματική διασπορά s^2 των υπολοίπων e_i

$$s^2 \equiv s_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

$$s^2 = \frac{n-1}{n-2}\left(s_Y^2 - \frac{s_{XY}^2}{s_X^2}\right) = \frac{n-1}{n-2}\left(s_Y^2 - b^2\,s_X^2\right)$$

Διαστήματα εμπιστοσύνης για τις παραμέτρους α και β

$$a \pm t_{n-2,1-\alpha/2} \times s_a \qquad b \pm t_{n-2,1-\alpha/2} \times s_b$$

εκτιμήσεις των τυπικών σφαλμάτων των a και b

Έλεγχος υποθέσεων για τα α και β

H_0: $\beta = 0$

αν η κλίση β της ευθείας παλινδρόμησης είναι στατιστικά ασήμαντη

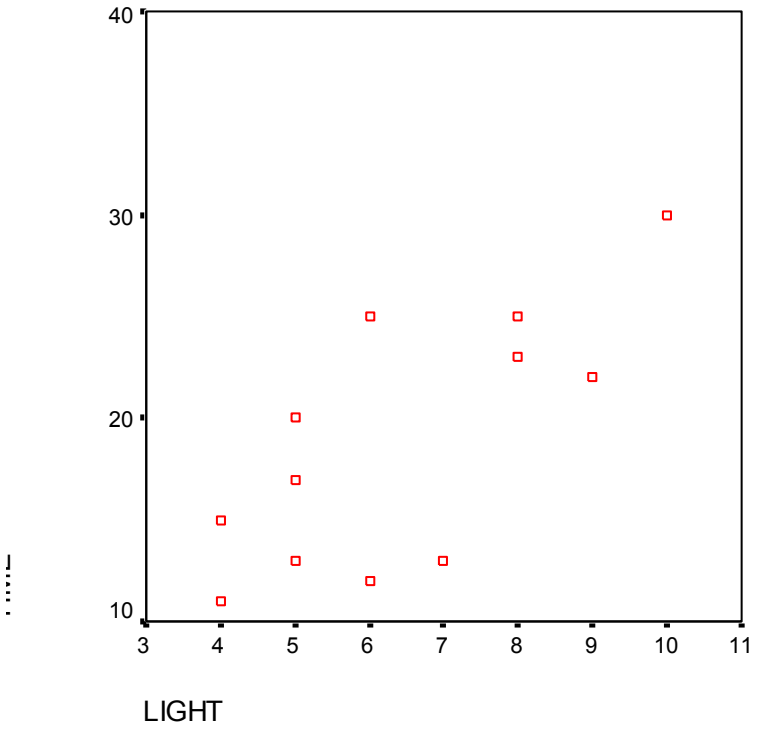
η τ.μ. Y δεν εξαρτάται από τη μεταβλητή X

Παράδειγμα

Θέλουμε να διερευνήσουμε την επίδραση του αριθμού των σηματοδοτημένων διασταυρώσεων στο χρόνο κάλυψης μιας διαδρομής

Υποθέτουμε πως ο χρόνος διαδρομής εξαρτάται γραμμικά από τον αριθμό των σηματοδοτημένων διασταυρώσεων

Σωστή υπόθεση?



αριθμός φαναριών	χρόνος διαδρομής
4	11,00
4	15,00
5	13,00
5	20,00
5	17,00
6	12,00
6	25,00
7	13,00
8	25,00
8	23,00
9	22,00
10	30,00

Εκτίμηση και πρόβλεψη της εξαρτημένης μεταβλητής

Για κάθε x_0 η σημειακή πρόβλεψη είναι ➔ $\hat{y}_0 = a + b\,x_0$

$(1-\alpha)\%$ διάστημα εμπιστοσύνης για $\hat{y}_0$

$$(a+bx_0) \pm t_{n-2,1-\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_X^2}}$$

$(1-\alpha)\%$ διάστημα πρόβλεψης για $\hat{y}_0$

$$(a+bx_0) \pm t_{n-2,1-\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_X^2}}$$

Πολλαπλή Γραμμική Παλινδρόμηση

Η τ.μ. Y εξαρτάται γραμμικά από κάποιες μεταβλητές $X_1, X_2, \ldots, X_k$

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i$$

Άλλες υποθέσεις: $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$

Γενικά το πρόβλημα και η εκτίμηση της πολλαπλής γραμμικής παλινδρόμησης δε διαφέρει ουσιαστικά από αυτό της απλής γραμμικής παλινδρόμησης.

Χρειάζονται όλες οι ανεξάρτητες μεταβλητές στο μοντέλο;

Επιλογή των ανεξάρτητων μεταβλητών

τεχνική της βηματικής παλινδρόμησης

Αρχίζουμε με το απλό σταθερό μοντέλο $y_i = \beta_0 + \varepsilon_i$

Σε κάθε βήμα προστίθεται μια μεταβλητή στο μοντέλο

μόνο αν αυτή η μεταβλητή προσφέρει σημαντική πληροφορία για τη *Y*, επιπρόσθετα στην πληροφορία που παρέχουν οι ανεξάρτητες μεταβλητές που είναι ήδη στο μοντέλο από το προηγούμενο βήμα

Σε κάθε βήμα γίνονται δύο στατιστικοί έλεγχοι:

1. Αν κάποια από τις μεταβλητές που δεν ήταν στο μοντέλο πρέπει να προστεθεί σ' αυτές που ήδη έχουν συμπεριληφθεί.

2. Αν κάποια από τις μεταβλητές που ήταν στο μοντέλο του προηγούμενου βήματος πρέπει να παραμείνει στο νέο μοντέλο, στο οποίο έχει συμπεριληφθεί μια νέα μεταβλητή.

Παράδειγμα

Θέλουμε να διερευνήσουμε την επίδραση στο χρόνο κάλυψης μιας διαδρομής:
1. του αριθμού των σηματοδοτημένων διασταυρώσεων
2. της ποιότητας του οδοστρώματος
3. του αριθμού βιομηχανικών μονάδων
4. του ποσοστού των φορτηγών στο πλήθος των οχημάτων

Χρονικές Σειρές

$x_1, x_2, \ldots, x_n$ σύνολο παρατηρήσεων ενός μεγέθους με το χρόνο με κάποιο χρονικό βήμα

πολλαπλή παλινδρόμηση: Y εξαρτάται από $X_1, X_2, \ldots, X_k$

χρονοσειρές: X_t εξαρτάται από $X_{t-1}, X_{t-2}, \ldots$ για κάθε *t*

αυτοπαλινδρόμηση

κλασικό μοντέλο $x_t = \mu_t + s_t + y_t$

στοιχείο τάσης

στοιχείο περιοδικότητας ή εποχικότητας

μη ομαλό στοιχείο

Στάσιμη χρονοσειρά:

όταν οι στατιστικές ιδιότητες της χρονοσειράς δεν αλλάζουν με το χρόνο

Στοιχείο τάσης

$$x_t = \mu_t + y_t$$

μ_t συνάρτηση του χρόνου

π.χ. $\mu_t = f(t) = a_0 + a_1 t$

Απαλοιφή στοιχείου τάσης

$$y_t = x_t - \hat{\mu}_t$$

$$y_t = \nabla x_t = x_t - x_{t-1}$$

Στοιχείο περιοδικότητας

$$x_t = s_t + y_t$$

s_t συνάρτηση του χρόνου

π.χ. $\mu_t = f(t) = \sin(2\pi t / k)$

Απαλοιφή στοιχείου τάσης

$$y_t = x_t - \hat{s}_t$$

Αυτοσυσχέτιση

συνάρτηση αυτοδιασποράς $\gamma_X(\tau) = \mathrm{Cov}\left[X_t, X_{t-\tau}\right]$

συνάρτηση αυτοσυσχέτισης $\rho_X(\tau) = \dfrac{\gamma_X(\tau)}{\sigma_X^2}$

εκτίμηση αυτοδιασποράς:

$$\hat{\gamma}_X(\tau) = \frac{1}{n-\tau}\sum_{t=\tau+1}^{n}\left(x_t - \bar{x}\right)\left(x_{t-\tau} - \bar{x}\right) = \frac{1}{n-\tau}\left(\sum_{t=\tau+1}^{n} x_t x_{t-\tau} - n\bar{x}^2\right)$$

εκτίμηση αυτοσυσχέτισης:

$$r_X(\tau) = \hat{\rho}_X(\tau) = \frac{\hat{\gamma}_X(\tau)}{s_X^2}$$

$$-1 \le r_X(\tau) \le 1$$

$$r_X(-\tau) = r_X(\tau)$$

Έτσι αν η χρονοσειρά είναι λευκός θόρυβος,
η αυτοσυσχέτιση μηδενίζεται για κάθε υστέρηση $\tau \neq 0$

Πρακτικά όρια για λευκό θόρυβο

$$\pm z_{1-\alpha/2}\frac{1}{\sqrt{n}}$$

Μοντέλα αυτοπαλινδρόμησης

ε_t ανεξάρτητα μεταξύ τους και έχουν πανομοιότυπη κατανομή

Γενικό μοντέλο

$$X_t = \phi_1 X_{t-1} + \cdots \phi_p X_{t-p} + \theta_0 \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

αυτοπαλινδρομούμενο μέρος

μέρος του κινούμενου μέσου

σφάλμα ή θόρυβο στη χρονική στιγμή *t* $\theta_0 = 1$

αυτοπαλινδρομούμενο μοντέλο κινούμενου μέσου ARMA(*p*,*q*)

αυτοπαλινδρομούμενο μοντέλο $X_t = \phi_1 X_{t-1} + \cdots \phi_p X_{t-p} + \varepsilon_t \qquad \varepsilon_t \sim \mathrm{N}(0, \sigma_\varepsilon^2)$

όπως για πολλαπλή γραμμική παλινδρόμηση

παράμετροι $\phi_1, \ldots, \phi_p, \sigma_\varepsilon^2$