ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ

2η Διάλεξη

Ελένη Κανδηλώρου (Αναπλ. Καθηγήτρια)
Οικονομικό Πανεπιστήμιο Αθηνών
Τμήμα Στατιστικής

24-2-2017

Ύλη 1ης Εβδομάδας

Γραμμική Παλινδρόμηση-*Έννοια Παλινδρόμισης*

1. *Σχέση μεταξύ μεταβλητών*
2. *Παραδείγματα*
3. *Ανάλυση παλινδρόμησης*
4. *Απλή & πολλαπλή παλινδρόμιση*
5. *Παράδειγμα*
6. Ανεξάρτητες και εξαρτημένες μεταβλητές

Σχέσεις Μεταβλητών

Σε διάφορα προβλήματα της Στατιστικής το ενδιαφέρον του ερευνητή εντοπίζεται στην ταυτόχρονη μελέτη δύο ή περισσότερων μεταβλητών, για να εντοπιστεί ο τρόπος με τον οποίο σχετίζονται οι μεταβλητές αυτές μεταξύ τους.

Σχέση μεταξύ Μεταβλητών Παραδείγματα

1. Η ηλικία (X) και η αρτηριακή πίεση αίματος (Y) ενός ενήλικα έχουν κάποια σχέση εξάρτησης μεταξύ τους (όσο μεγαλύτερη η ηλικία τόσο υψηλότερη η αρτηριακή πίεση).

Ηλικία (X)	36	38	42	42	47	49	55	56	60	63	68	72
Πίεση αίματος (Y)	118	115	125	140	128	145	150	147	155	149	152	160

(εδώ $(X_1,Y_1) = (36, 118)$, $(X_2,Y_2) = (38, 115)$, κ.ο.κ.)

Παραδείγματα*συνέχεια*

2. Η συνολική παραγωγή ενός κτήματος εξαρτάται από το ύψος της θερμοκρασίας, την ποσότητα λιπάσματος, το ύψος της υγρασίας, το ύψος της βροχόπτωσης, το είδος λιπάσματος, κ.λ.π.

Νερό (*X*)	12	18	24	30	36	42	48
Σοδειά (*Y*)	5.27	5.68	6.25	7.21	8.05	8.71	8.42

5

Παραδείγματα*συνέχεια*

3. Η επίδοση ενός φοιτητή επηρεάζεται από το φύλο, τις ώρες μελέτης, τις ώρες παρακολούθησης στο αμφιθέτρο, κ.λ.π.

% Παρακ/σης (*X*)	,5	,2	,77	,01	,35	,4	,30	,90	1	,48
Βαθμός (*Y*)	6	3	8,5	2	3,5	4	3,5	9,5	10	5,5

6

Σχέση Μεταξύ Μεταβλητών

Σε διάφορα προβλήματα της Στατιστικής το ενδιαφέρον του αναλυτή εντοπίζεται στην ταυτόχρονη μελέτη δύο ή περισσοτέρων μεταβλητών, ώστε να προσδιρίσει τον τρόπο με τον οποίο σχετίζονται μεταξύ τους.

7

Ανάλυση Παλινδρόμισης

Η διαδικασία που εξετάζει τη σχέση ανάμεσα σε δύο (***Y, X***) ή περισσότερες μεταβλητές (***Y, X_1, X_2, ..., $X_κ$***) με στόχο την **πρόβλεψη μιας από αυτές** (*Y*) μέσω των υπολοίπων λέγεται **ανάλυση παλινδρόμησης** (regression analysis).

8

Χρήση του όρου «Regression»:

….έγινε το 1885 από τον Galton (Άγγλο ανθρωπολόγο).

Το θέμα της μελέτης εκείνης είχε να κάνει με το ύψος των παιδιών σε σχέση με το ύψος των γονέων τους....

Απλή Παλινδρόμηση

Στην Απλή Παλινδρόμηση, χρησιμοποιούνται δύο μεταβλητές. Η X και η Y.

Η Y μπορεί να προσεγγιστεί από μια συνάρτηση του X ($\hat{Y}_i \approx 2 + 0{,}90X$).

X : ανεξάρτητη μεταβλητή (independent/input var.)

Y : εξαρτημένη μεταβλητή (dependent/response var.)

συνέχεια

Απλή: Παρατηρούμε ότι όσο αυξάνεται η X (Ηλικία) τόσο αυξάνεται και η Y (πίεση αίματος). Μάλιστα φαίνεται ότι τα σημεία (X_i,Y_i) συγκεντρώνται κοντά σε μία ευθεία, π.χ. την $Y_i=\beta_0+\beta_1 X_i$, $i = 1, 2, \ldots, n$, για κάποιες σταθερές β_0, β_1. Οι αποκλίσεις $(Y_i-(\beta_0+\beta_1 X_i))$, $i = 1, 2,\ldots,n$ των σημείων (X_i,Y_i) από την ευθεία αυτή φαίνονται τυχαίες. Αν ονομάσουμε ε_i, $i = 1, 2,\ldots,n$ τις διαφορές αυτές,τότε προκύπτει το γνωστό ***Α***πλό ***Γ***ραμμικό ***Υ***πόδειγμα

Απλή & Πολλαπλή Παλινδρόμηση Παραδείγματα

Απλή Παλινδρόμηση (το προαναφερόμενο παράδειγμα, σελ 4)

Ηλικία (X)	36	38	42	42	47	49
Πίεση αίματος (Y)	118	115	125	140	128	145
Ηλικία (X)	55	56	60	63	68	72
Πίεση αίματος (Y)	150	147	155	149	152	160

Συνέχεια από την Προηγούμενη Σελίδα

Πολλαπλή Παλινδρόμηση

Στην Πολλαπλή Παλινδρόμηση, χρησιμοποιούνται περισσότερες από δύο μεταβλητές. Η Y και δύο ή περισσότερες X. Δηλαδή,

$$Y_i=\beta_0+\beta_1 X_{1i}+\beta_{2i}+\varepsilon_i$$

Παράδειγμα Πολλαπ. Παλινδ.

X_1	X_2	Y	X_1	X_2	Y
36	Α	118	55	Γ	150
38	Γ	115	56	Α	147
42	Γ	125	60	Γ	155
42	Α	140	63	Α	149
47	Γ	128	68	Γ	152
49	Γ	145	72	Α	160

X_1=ηλικία

X_2=φύλο

Y=πίεση

Ανεξάρτητες & Εξαρτημένες Μεταβλητές

Στις Κλινικές Μελέτες, ο ερευνητής προκαθορίζει τις δόσεις ενός φαρμάκου (ανεξάρτητη μεταβλητή) που παρέχει στους ασθενείς του και μετρά την αντίδρασή τους στο φάρμακο (εξαρτημένη μεταβλητή).

Με την παλινδρόμηση προσδιορίζεται η σχέση δόσης-αντίδρασης, στην περίπτωση του συγκεκριμένου φαρμάκου. Δηλαδή, για δεδομένη δόση να προβλέπεται η αντίδραση!

Ποιά Μεταβλητή είναι Ανεξάρτητη & ποιά Εξαρτημένη;

- Έστω ότι ένας ερευνητής θέλει να μελετήσει τη σχέση μεταξύ ύψους και βάρους των εφήβων της Ελλάδας. *(πρόβλημα)*
- *Πληθυσμός: το σύνολο των εφήβων της χώρας*
- *Ζητούμενο:* προσδιορισμός της σχέσης μεταξύ του ύψους - βάρους
- Δείγμα: τυχαίο, μεγέθους *n για τις μεταβλητές μας (*ύψος και βάρος). Ποιό το X & ποιό το Y;

συνέχεια

- *Στατιστική Ανάλυση: βασίζεται στο δείγμα*
  - *Ανάλυση Απλής Παλινδρόμησης*
  - *Γενίκευση Αποτελεσμάτων (από το δείγμα στον πληθυσμό-Συμπερασματολογία)*

17

Τι Παρατηρούμε στο Προηγούμενο Παράδειγμα (βλ.σελ.11);

1. Παρατηρούμε ότι όσο αυξάνεται η X (*Age*) τόσο αυξάνεται και η Y (*Pressure*). Μάλιστα φαίνεται ότι τα σημεία (X_i,Y_i) βρίσκονται «κοντά» σε μία ευθεία, π.χ. την

$$\hat{Y}_i \approx b_0 + b_1 X_i$$

18

1. (Xi,Yi) από την ευθεία αυτή φαίνονται τυχαίες. Αν ονομάσουμε εi, $i = 1, 2,\ldots,n$ τις διαφορές αυτές τότε προκύπτει φυσιολογικά το γνωστό ως απλό γραμμικό υπόδειγμα που θα περιγράψουμε στη συνέχεια.
2. , $i = 1, 2, \ldots, n$ για κάποιες σταθερές β_0, β_1. Οι αποκλίσεις $Y_i - \beta_0 + \beta_1 X_i$, $i = 1, 2,\ldots,n$ των σημείων

19

ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ

3η Διάλεξη

Ελένη Κανδηλώρου (Αναπλ. Καθηγήτρια)
Οικονομικό Πανεπιστήμιο Αθηνών
Τμήμα Στατιστικής

28-2-2017

20

Ύλη 2ης Εβδομάδας (3η Διάλεξη)

Απλό Γραμμικό Υπόδειγμα:

1. Διάγραμμα διασποράς
2. *Απλό γραμμικό υπόδειγμα*
3. *Ανεξάρτητες τυχαίες μεταβλητές*
4. Y_i, ε_i - *Τι Σχέση Έχουν;*
5. Υπολογισμός των $E(Y_i)$ & $V(Y_i)$

Παράδειγμα

Η ηλικία (***X***) και η αρτηριακή πίεση αίματος (***Y***) ενός ενήλικα έχουν κάποια σχέση εξάρτησης μεταξύ τους (όσο μεγαλύτερη η ηλικία τόσο υψηλότερη η αρτηριακή πίεση).

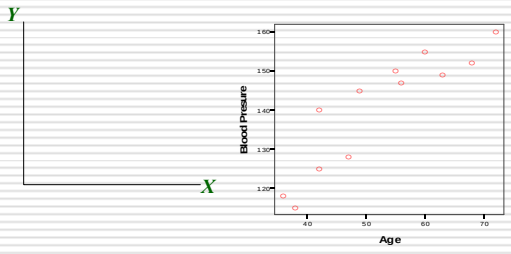
Ηλικία (*X*)	36	38	42	42	47	49	55	56	60	63	68	72
Πίεση αίματος (*Y*)	118	115	125	140	128	145	150	147	155	149	152	160

Και τώρα τι;

Για να επιλέξουμε το κατάλληλο υπόδειγμα για να περιγράψουμε τη σχέση ανάμεσα σε 2 μεταβλητές, είθιστε να ξεκινάμε από τη γραφική απεικόνισή τους, δηλαδή, απο το **διάγραμμα διασποράς** (*scatter plot*). Το διάγραμμα διασποράς μεταξύ της ηλικίας και της πίεσης αποκαλύπτει τη θετική σχέση μεταξύ των 2 αυτών μεταβλητών.

Το πρώτο πράγμα που μπορούμε να κάνουμε είναι να δούμε με τη χρήση του **SPSS** τη «σχέση» των συγκεκριμένων μεταβλητών: Εκτελούμε **Graphs/ Scatterplot / Simple**/ **Y** Axis: *Pressure*, **X** Axis: *Age* λαμβάνοντας το ακόλουθο γράφημα

συνέχεια



Τι Παρατηρούμε στο Προηγούμενο Παράδειγμα;

Παρατηρούμε ότι όσο αυξάνεται η ***X*** (***Age***) τόσο αυξάνεται και η ***Y*** (***Pressure***). Μάλιστα φαίνεται ότι τα σημεία (X_i, Y_i) βρίσκονται «κοντά» σε μία ευθεία, π.χ. την

$$Y_i \approx \beta_0 + \beta_1 X_i$$

Όπου, $i = 1, 2, \ldots, n$ για κάποιες σταθερές β_0, β_1.

συνέχεια

Οι αποκλίσεις $Y_i - (\beta_0 + \beta_1 X_i)$, $i = 1, 2, \ldots, n$ των σημείων (X_i, Y_i) από την ευθεία αυτή φαίνονται τυχαίες. Αν ονομάσουμε ε_i, $i = 1, 2, \ldots, n$ τις διαφορές αυτές τότε προκύπτει φυσιολογικά το γνωστό μας απλό γραμμικό υπόδειγμα.

Απλό Γραμμικό Υπόδειγμα

Γενικά, θέλουμε να προσδιορίσουμε τη σχέση μεταξύ των μεταβλητών ***X***, ***Y***. Το πιο απλό υπόδειγμα που θα μπορούσε να κάνει αυτό είναι το **απλό γραμμικό υπόδειγμα**.

Σύμφωνα με το **υπόδειγμα** αυτό θεωρούμε ότι τα X_i, Y_i συνδέονται με τη σχέση

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \; i = 1, 2, \ldots, n$$

Ανεξάρτητες Τυχαίες Μεταβλητές

όπου β_0, β_1 είναι δύο άγνωστες σταθερές (και καλούνται *τεταγμένη* ή *intercept*

και *κλίση* ή *slope* αντίστοιχα),

ενώ οι

$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ είναι *ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν κανονική κατανομή* $N(0, \sigma^2)$ (σ^2 άγνωστο) και συνήθως καλούνται «σφάλματα» των μετρήσεων.

Y_i , ε_i - Τι Σχέση Έχουν;

Μπορεί να θεωρηθεί ότι τα σφάλματα $\varepsilon_1, \varepsilon_2,\ldots,\varepsilon_n$ εμπεριέχουν όλους τους άλλους παράγοντες (εκτός της X) που επηρεάζουν την τιμή της μεταβλητής Y.

Υπογραμμίζεται και πάλι ότι οι τιμές $X_1, X_2,\ldots, X_n$ δεν είναι τυχαίες, αντίθετα με τις $Y_1, Y_2, \ldots, Y_n$ οι οποίες προφανώς

είναι ***τυχαίες*** και μάλιστα θα ακολουθούν κανονική κατανομή (αφού είναι γραμμικές συναρτήσεις των κανονικών τ.μ. ε_i) με παραμέτρους ... ⟶

29

Υπολογισμός των $E(Y_i)$ & $V(Y_i)$

$E(Y_i)$, και $V(Y_i)$

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

$$V = \sigma^2$$

30

συνέχεια

Δηλαδή,

$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2).$$

Επίσης οι τ.μ. $Y_1, Y_2, \ldots, Y_n$ είναι ανεξάρτητες αφού τα σφάλματα $\varepsilon_1, \varepsilon_2,\ldots,\varepsilon_n$ είναι ανεξάρτητα.

Αρα, θα πρέπει με βάση τα (X_i, Y_i), i=1,2,…,n, να εκτιμήσουμε τις παραμέτρους

β_0, β_1 και

σ^2

31

Ύλη Επόμενης Εβδομάδας (4η Διάλεξη)

1. Εκτίμηση των παραμέτρων β_0, β_1 και σ^2
2. Έλεγχοι υποθέσεων και Δ.Ε. για τις παραμέτρους του υποδείγματος

32

ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ

4η Διάλεξη

Ελένη Κανδηλώρου (Αναπλ. Καθηγήτρια)
Οικονομικό Πανεπιστήμιο Αθηνών
Τμήμα Στατιστικής

(3-3-2017 ανεβλήθει)
7-3-2017

ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ

5η Διάλεξη

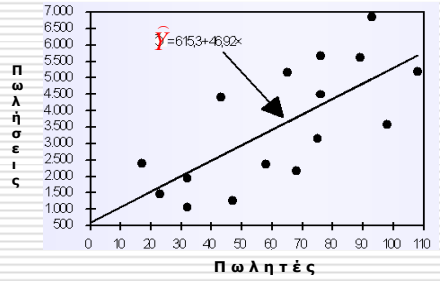
Ελένη Κανδηλώρου (Αναπλ. Καθηγήτρια)
Οικονομικό Πανεπιστήμιο Αθηνών
Τμήμα Στατιστικής

10-3-2017

Ύλη 5ης Διάλεξης

1. Διάγραμμα: Διασποράς & Γραμμής Παλινδρόμησης
2. Ιδιότητες Εκτιμητών
3. Συντελεστής Προσδιορισμού
4. Η εκτίμηση του Συντελεστή Προσδιορισμού
5. Έλεγχος του Συντελεστή Προσδιορισμού

Διάγραμμα Διασποράς Μεταξύ Πωλήσεων & Πωλητών & Γραμμή Παλινδρόμησης



Ιδιότητες Εκτιμητών

Οι εκτιμητές $\hat{\beta}_0$ και $\hat{\beta}_1$ είναι Best Linear Unbiased Estimators (BLUE).

Τι σημαίνουν τα αρχικά BLUE?

1. **Εκτιμητής:** Οι

$\hat{\beta}_0$ $\hat{\beta}_1$

είναι εκτιμητές της πραγματικής τιμής των

β_0 β_1

37

Ιδιότητες Εκτιμητών*συνέχεια*

2.Unbiased-Αμερόληπτος: Κατά μέσο όρο, οι τιμές των $\hat{\beta}_0$ και $\hat{\beta}_1$

θα είναι ίσες με τις πραγματικές τιμές των

β_0 και β_1 .

Δηλαδή,

$$E\left(\hat{\beta}_0\right) = \beta_0 \quad \text{και} \quad E\left(\hat{\beta}_1\right) = \beta_1$$

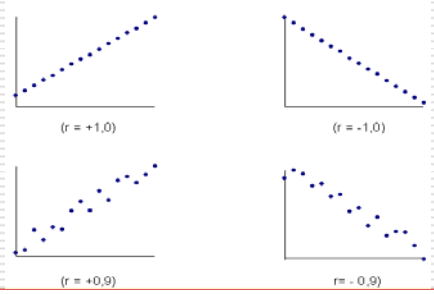
38

Ιδιότητες Εκτιμητών*συνέχεια*

3. ***Best-Καλύτερος***: σημαίνει ότι οι εκτιμητές της μεθόδου ΕΤ (OLS) έχουν την μικρότερη διακύμανση σε σχέση με όλους τους αμερόληπτους εκτιμητές (***Αποτελεσματικός***).

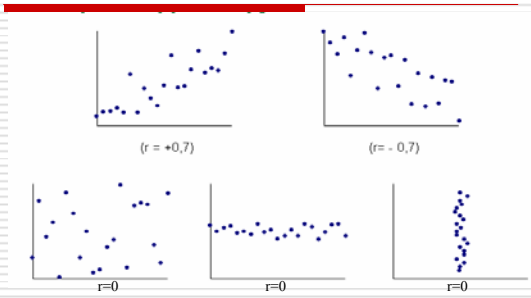
39

Είδη Συσχέτισης



40

Είδη Συσχέτισης



41

Υπάρχει Σχέση Μεταξύ *X* & *Y*

Ο πιο απλός τρόπος για να διαπιστώσουμε αν υπάρχει συσχέτιση μεταξύ δύο μεταβλητών είναι η κατασκευή του **διαγράμματος διασποράς** (όπως έχουμε αναφέρει).

Το διάγραμμα διασποράς μεταξύ ***πωλήσεων*** και ***διαφήμισης*** αποκαλύπτει τη **θετική** σχέση μεταξύ των 2 μεταβλητών.

Το διάγραμμα διασποράς μεταξύ **πωλητών** και **διαφήμισης** αποκαλύπτει τη ***θετική*** σχέση μεταξύ των 2 μεταβλητών, η οποία δεν είναι τόσο έντονη όσο στη προηγούμενη περίπτωση

42

Συντελεστής Συσχέτισης

Η ποσοτική μέτρηση της γραμμικής σχέσης μεταξύ δύο μεταβλητών ονομάζεται **συντελεστής συσχέτισης (correlation coefficient)** ο οποίος παίρνει τιμές μεταξύ του ***-1*** και του ***1***.

Η εκτίμηση του συντελεστή συσχέτισης, που συμβολίζεται με ***r*** προκύπτει ως:

43

Συντελεστής Συσχέτισης

Μετρά το βαθμό της γραμμικής συσχέτισης 2 τ.μ. *X* & *Y* με διασπορά σ^2_X & σ^2_Y αντίστοιχα & συνδιακύμανση , *Cov.(X,Y) = E(X, Y)-E(X)E(Y).*

$$r = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\sum (X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum \left((X-\overline{X})^2\right)\left(\sum (Y-\overline{Y})^2\right)}}$$

$$= \frac{n\sum XY - \sum X \sum Y}{\sqrt{\left(n\left(\sum X^2\right) - \left(\sum X\right)^2\right)\left(n\left(\sum Y^2\right) - \left(\sum Y\right)^2\right)}}$$

44

Συντελεστής Συσχέτισης

Η συσχέτιση μεταξύ (***Y, X***) μετράει τον βαθμό αλληλεξάρτησής τους. Αν υποτεθεί ότι ***Y*** και ***X*** συσχετίζονται, τότε:

- υπάρχουν ενδείξεις γραμμικής σχέσης μεταξύ των δύο μεταβλητών. Οι μεταβολές των δύο μεταβλητών, κατά μέσο όρο, συνδέονται με το συντελεστή συσχέτισης.

Συντελεστής Συσχέτισης Παράδειγμα

Έτος	Πωλήσεις	Διαφήμιση	Πωλητές	Έτος	Πωλήσεις	Διαφήμιση	Πωλητές
1985	1050	162	32	1993	3570	720	98
1986	1260	285	47	1994	4410	1140	43
1987	1470	540	23	1995	4500	1395	76
1988	2160	261	68	1996	5610	1560	89
1989	1950	360	32	1997	5190	1380	108
1990	2400	690	17	1998	5670	1260	76
1991	2370	495	58	1999	5160	1710	65
1992	3150	948	75	2000	6840	1860	93

Συντελεστής Συσχέτισης υπολογισμός

Πωλήσεις & Διαφήμιση

$$r = \frac{16\,(66.883.410) - (56.760)\,(14.766)}{\sqrt{\left(16\,(18.289.944) - (14.766)^2\right)\left(16\,(251.029.800) - (56.760)^2\right)}} = 0.95$$

Πωλήσεις & Πωλητές

$$r = \frac{16\,(4.094.160) - (56.760)\,(1.000)}{\sqrt{\left(16\,(74.152) - (1.000)^2\right)\left(16\,(251.029.800) - (56.760)^2\right)}} = 0.72$$

Έλεγχος Στατιστικής Σημαντικότητας του r

Αυτό που ενδιαφέρει περισσότερο είναι εάν η γραμμική σχέση μεταξύ των μεταβλητών ***X*** και ***Y*** είναι στατιστικά σημαντική.

Ο έλεγχος περιγράφεται ως:

$$H_o : \rho = 0$$

$$H_1 : \rho \neq 0$$

Έλεγχος για το $r_{συνέχεια}$

Ο έλεγχος γίνεται με βάση την κατανομή ***t*** και η κριτική τιμή υπολογίζεται ως:

$$t_{n-2} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

Αν $|t_{n-2}| > |t_{n-2,\,1-\alpha/2}| \Rightarrow \not{H}_0$

Έλεγχος για το $r_{συνέχεια}$

$$t_{n-2} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0,72}{\sqrt{\frac{1-0,72^2}{16-2}}} = 11,38$$

$$|t_{n-2\,;1-\alpha/2}| = |t_{n-2\,;0,975}| = 2,145 \quad \acute{\alpha}\rho\alpha \quad \not{H}_0$$

Σχολιασμός του r

Το πιο ουσιαστικό ερώτημα που θα πρέπει να απαντήσουμε πριν χρησιμοποιήσουμε την εξίσωση παλινδρόμησης είναι:

ποια είναι η προβλεπτική ικανότητα της εξίσωσης ή τι ποσοστό των μεταβολών της εξαρτημένης μεταβλητής Y οφείλεται στις επιδράσεις της X.

Σχολιασμός του r

Όλη η ανάλυση που θα ακολουθήσουμε βασίζεται στα κατάλοιπα (e_i). Όσο μεγαλύτερη είναι η επίδραση της ***X*** επί της ***Y*** τόσο μικρότερα είναι τα κατάλοιπα και αντίστροφα.

Πρώτη δουλειά μας είναι να εκτιμήσουμε τη συνολική διασπορά γύρω από τη γραμμή της παλινδρόμησης (δηλαδή τι?)

Σχολιασμός του r

... το άθροισμα των τετραγώνων των αποκλίσεων των πραγματικών τιμών της ***Y*** από τις αντίστοιχες τιμές της.

Δηλαδή, θα υπολογίσουμε το

$$\sum_{i}^{n} e_i^2 = \sum_{i}^{n} (Y_i - \hat{Y}_i)^2$$

που επίσης ονομάζεται και άθροισμα των τετραγώνων των σφαλμάτων (sum of squared errors) και συμβολίζεται με SSE.

Σχολιασμός του r

$$\sum \epsilon^2 = \sum \left(Y - \hat{Y}\right)^2 = \sum \left(Y - \hat{\beta}_0 - \hat{\beta}_1 X\right)^2 = \sum Y^2 - \hat{\beta}_0 \sum Y - \hat{\beta}_1 \sum YX$$

Άθροισμα Τετραγωνικών Σφαλμάτων

$\hat{Y} = 615,3,4 + 46,92 \times X$

Έτος	Πωλήσεις (Y)	Πωλητές (X)	$\hat{Y}$	$SSE = (Y - \hat{Y})^2$
1985	1050	32	2117	1137584
1986	1260	47	2820	2434564
1987	1470	23	1694	50327
1988	2160	68	3806	2707787
1989	1950	32	2117	27747
1990	2400	17	1413	974480
1991	2370	58	3336	933890
1992	3150	75	4134	968147
1993	3570	98	5213	2699456
1994	4410	43	2633	3158985
1995	4500	76	4181	101850
1996	5610	89	4791	671151
1997	5190	108	5682	242219
1998	5670	76	4181	2217538
1999	5160	65	3665	2235656
2000	6840	93	4978	3465464
Άθροισμα				**24026845**

ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ

6η Διάλεξη

Ελένη Κανδηλώρου (Αναπλ. Καθηγήτρια)
Οικονομικό Πανεπιστήμιο Αθηνών
Τμήμα Στατιστικής

14-3-2017

Ύλη 6ης Διάλεξης

1. Παράδειγμα για:
   1. Εκτιμήσεις παραμέτρων
   2. Υπολογισμός των $\widehat{Y}_i$
   3. Υπολογισμός εκτιμημένων σφαλμάτων
   4. Πρόβλεψη τιμών της ***Y***
2. Εκτίμηση Διακύμανσης του Σφάλματος

Δεδομένα & Υπολογισμοί

Yi	Xi	Xi^2	Yi^2	XiYi n=16
1050	*32*	1024	1102500	33600
1260	*47*	2209	1587600	59220
1470	*23*	529	2160900	33810
2160	*68*	4624	4665600	146880
1950	*32*	1024	3802500	62400
2400	*17*	289	5760000	40800
2370	*58*	3364	5616900	137460
3150	*75*	5625	9922500	236250
3570	*98*	9604	12744900	349860
4410	*43*	1849	19448100	189630
4500	*76*	5776	20250000	342000
5610	*89*	7921	31472100	499290
5190	*108*	11664	26936100	560520
5670	*76*	5776	32148900	430920
5160	*65*	4225	26625600	335400
6840	*93*	8649	46785600	636120
56760	**1000**	**74152**	**251029800**	**4094160**

Ζητούμενο

1. Εκτιμήσεις των παραμέτρων β_1 και β_2:

$$\widehat{\beta}_1 = ?$$

$$\widehat{\beta}_0 = ?$$

2. Υπολογισμός του: $\widehat{Y}_i = ?$ $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$

$\widehat{e}_i = Y_i - \widehat{Y}_i$ (εκτιμημένα σφάλματα)

Ζητούμενο$_{\text{συνέχεια}}$

3. Υπολογισμός προβλέψεων των τιμών ***της εξαρτημένης μεταβλητής***.

   Με δεδομένη την τιμή της ***X*** μπορούμε να προβλέψουμε την τιμή της ***Y***. Αν ***X=10***, πόσο είναι το ***Y***?

Συνολική Μεταβλητότητα του Υποδείγματος

Yi	Xi	$\hat{Y}_i$	$(Y_i - \bar{Y})$	$(Yi - \bar{Y})^2$
1050	32	2116,5757	-1066,5757	1137583,724
1260	47	2820,30896	-1560,30896	2434564,051
1470	23	1694,33574	-224,33574	50326,52424
2160	68	3805,53553	-1645,53553	2707787,18
1950	32	2116,5757	-166,5757	27747,46383
2400	17	1412,84243	987,15757	974480,068
2370	58	3336,38002	-966,38002	933890,3431
3150	75	4133,94439	-983,94439	968146,5626
3570	98	5213,00206	-1643,00206	2699455,769
4410	43	2632,64676	1777,35324	3158984,54
4500	76	4180,85994	319,14006	101850,3779
5610	89	4790,7621	819,2379	671150,7368
5190	108	5682,15757	-492,15757	242219,0737
5670	76	4180,85994	1489,14006	2217538,118
5160	65	3664,78888	1495,21112	2235656,293
6840	93	4978,4243	1861,5757	3465464,087
				24026844,91

Απαντήσεις- Εκτιμήσεις Παραμέτρων

1. $$\hat{\beta}_1 = \frac{n\sum Y_i X_i - \sum Y_i \sum X_i}{n\sum X_i^2 - (\sum (X_i))^2} = 46,916$$

$$\hat{\beta}_0 = \bar{Y} - \beta_1 \bar{X} = 615,278$$

Ερμηνεία των παραπάνω αποτελεσμάτων

Υπολογισμός Εκτιμημένων Σφαλμάτων (e_i^2)

2	Yi	Xi	$\hat{Y}_i$	$(Yi - \bar{Y})$	$(Yi - \bar{Y})^2$
	1050	32	2116,5757	-1066,5757	1137583,724
	1260	47	2820,30896	-1560,30896	2434564,051
	1470	23	1694,33574	-224,33574	50326,52424
	2160	68	3805,53553	-1645,53553	2707787,18
	1950	32	2116,5757	-166,5757	27747,46383
	2400	17	1412,84243	987,15757	974480,068
	2370	58	3336,38002	-966,38002	933890,3431
	3150	75	4133,94439	-983,94439	968146,5626
	3570	98	5213,00206	-1643,00206	2699455,769
	4410	43	2632,64676	1777,35324	3158984,54
	4500	76	4180,85994	319,14006	101850,3779
	5610	89	4790,7621	819,2379	671150,7368
	5190	108	5682,15757	-492,15757	242219,0737
	5670	76	4180,85994	1489,14006	2217538,118
	5160	65	3664,78888	1495,21112	2235656,293
	6840	93	4978,4243	1861,5757	3465464,087
					24026844,91

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Αποτελέσματα (e_i^2)

- Τα αποτελέσματα δίνονται στον προηγούμενο Πίνακα, στην 4η στήλη.

Πρόβλεψη του *Y με δεδομένο X=100*

3.

Y= 615,278 + 46,913*(100) = 5306,578

65

Εκτίμηση Διακύμανση του Σφάλματος (*SSE*)

Η πληθυσμιακή διακύμανση του σφάλματος σ_ε^2 είναι η παράμετρος που καθορίζει την ένταση της εξάρτησης της ***Y*** από την ***X***. Η εκτίμηση της θα βασισθεί στο άθροισμα των τετραγώνων των σφαλμάτων γύρω από τη γραμμή παλινδρόμησης, δηλαδή το ***SSE***. Συμβολίζοντας την εκτίμηση του σ_ε^2 με s_ε^2, τότε

$$s_\epsilon^2 = \frac{\sum \left(Y - \hat{Y}\right)^2}{n-2} = \frac{SSE}{n-2}$$

66

Εκτίμηση Διακύμανση του Σφάλματος*συνέχεια*

όπου $n-2$ είναι οι ***BE (df).*** Χάνουμε 2 βαθμούς ελευθερίας διότι η εκτίμησή του βασίζεται στην εκτίμηση 2 παραμέτρων.

67

Συνολικό Άθροισμα Τετραγώνων

Η συνολική μεταβλητότητα της εξαρτημένης μεταβλητής ***Y*** ονομάζεται **συνολικό άθροισμα τετραγώνων (Total Sum of Squares)** και συμβολίζεται ως:

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

SSE: το μέρος της συνολικής μεταβλητικότητας της ***Y*** που δεν εξηγείται από την παλινδ.

(***SST-SSE = SSR***): το μέρος της διασποράς της ***SST*** που φείλεται στις επιδράσεις της ***X***.

68

Συνολικό Άθροισμα Τετραγώνων *συνέχεια*

Δηλαδή, η συνολική μεταβλητότητα της ***Y*** χωρίζεται σε 2 μέρη:

1. Σε εκείνη που ***εξηγείτε*** από την εξίσωση παλινδρόμησης ***SSR***, και
2. Σε εκείνη που ***δεν εξηγείτε*** (ανεξήγητη), δηλαδή, εκείνη που οφείλεται στην επίδραση όλων των άλλων παραγόντων ***SSE εκτός της X.***

69

Συνολικό Άθροισμα Τετραγώνων *συνέχεια*

Η απόκλιση $(Y_i - \bar{Y})$ διακρίνεται σε:

$$(Y_i - \bar{Y}) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$$

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

70

Συνολικό Άθροισμα Τετραγώνων *συνέχεια*

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

$$SST = SSE + SSR$$

71

SST, SSR, SSE

1. SST: εκφράζει τη συνολική παρατηρούμενη μεταβλητότητα των Y_i

2. SSR: εκφράζει τη μεταβλητότητα των προσαρμοσμένων τιμών διότι:

$$\bar{\hat{Y}} = \frac{1}{n}\sum_{i=1}^{n}\hat{Y} = \frac{1}{n}\sum_{i=1}^{n}Y_i = \bar{Y}$$

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

72

SST, SSR, SSE$_{\text{συνέχεια}}$

3. SSE εκφράζει τη μεταβλητότητα των Y_i σε σχέση με τις αντίστοιχες προσαρμοσμένες τιμές $\hat{Y}_i$:

Η μεταβλητότητα αυτή οφείλεται στην διασπορά σ^2 των σφαλμάτων ε_i τα οποία όπως είπαμε μπορεί να θεωρηθεί ότι «περιέχουν» όλους τους άλλους παράγοντες που πηρεάζουν την τιμή των Y_i (και δεν υπάρχουν στο υπόδειγμα).

Συντελεστής Προσδιορισμού

Άρα, η συνολική παρατηρούμενη μεταβλητότητα των Y_i (***SST***) μπορεί να χωριστεί στα **δύο** στην μεταβλητότητα :

α) που ερμηνεύεται από το υπόδειγμα (***SSR***)

β) που οφείλεται σε παράγοντες που δεν έχουν περιληφθεί στο υπόδειγμα.

Άρα, το πηλίκο (***συντελεστής προσδιορισμού***)

Συντελεστής Προσδιορισμού

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST}$$

μπορεί να θεωρηθεί ότι εκφράζει το ποσοστό της μεταβλητότητας των παρατηρήσεων που ερμηνεύεται από το υπόδειγμα.

Συντελεστής Προσδιορισμού Παράδειγμα$_{\text{συνέχεια}}$

Είναι προφανές ότι όσο μεγαλύτερο (πιο «κοντά» στην μονάδα) είναι το R^2 τόσο καλύτερο είναι το υπόδειγμα που έχουμε θεωρήσει διότι ερμηνεύει μεγαλύτερο μέρος της παρατηρούμενης μεταβλητότητας.

Παράδειγμα$_{\text{συνέχεια}}$

$R^2=(SSR)/(SST) = 0,908$

Ερμηνεία ???

Γραμμικά Μοντέλα

10η Παράδοση

31-3-2017

Ελένη Κανδηλώρου
Αναπλ. Καθηγήτρια

Εξέταση της ορθότητας του υποδείγματος

Ό,τι έχουμε συζητήσει μέχρι τώρα στηρίζεται στις υποθέσεις του γραμμικού υποδείγματος:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, i = 1,2,...,n$$

όπου τα σφάλματα $\varepsilon_1, \varepsilon_2,\ldots,\varepsilon_n$ είναι ανεξάρτητα και κατανέμονται $N(0,\sigma^2)$.

Πρέπει όμως να βεβαιωθούμε ότι οι υποθέσεις που έχουμε κάνει για να εφαρμόσουμε τη ΜΕΤ (που οδηγεί σε BLUE εκτιμητές) ικανοποιούνται!

συνέχεια

Αν διαπιστώσουμε ότι αυτό δεν συμβαίνει τότε τροποποιούμε κατάλληλα το υπόδειγμά μας.

Αν δηλαδή, δεν ισχύουν οι παρακάτω υποθέσεις στις οποίες έχουμε αναφερθεί, θα πρέπει να μετασχηματίσουμε τα δεδομένα μας.

Υποθέσεις

1. **Γραμμικότητα** (Linearity)
2. **Ομοσκεδαστικότητα**-(Homoscedasticity)-$Var(\varepsilon_i)=\sigma^2$, $i = 1,2,...n$
3. **Ανεξαρτησία** (Independence)-$Cov.(\varepsilon_i, \varepsilon_j)=0$ για κάθε $i \neq j$, $i,j = 1,2,...n$
4. **Κανονικότητα** (Normality)-$\varepsilon_i \sim N(0,\sigma^2)$.

Πώς διαπιστώνουμε ότι οι υποθέσεις ισχύουν;

Η κατανομή της Y έχει, για τα διάφορα επίπεδα X_i, $i = 1,2,...,n$ της X, μέση τιμή

$E(Y) = \beta_0+\beta_1 X$, όπου, β_0 και β_1 παράμετροι που εκτιμώνται από το δείγμα (X_i , Y_i), $i = 1,2,...,n$.

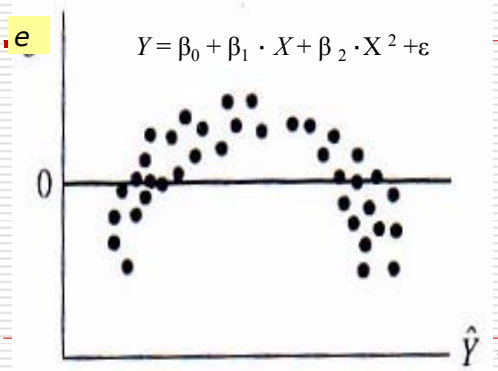
Δηλαδή, υποθέτουμε ότι οι μέσες τιμές της Y, για τα διάφορα επίπεδα της X, είναι γραμμικές συναρτήσεις της X (ότι βρίσκονται δηλαδή σε ευθεία γραμμή). Σημειώνουμε ότι στο υπόδειγμα $Y = \beta_0+\beta_1 X+\varepsilon$, τυχαίες μεταβλητές είναι μόνο οι Y και ε.

81

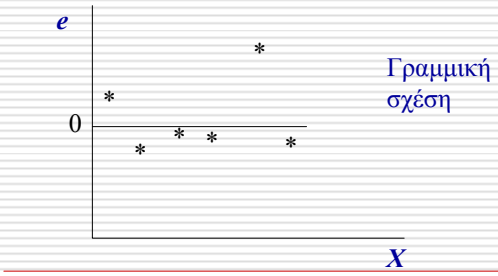
Γραμμικότητα (1η υπόθεση)

Το πρώτο που μπορούμε να κάνουμε είναι να δούμε τη «σχέση» των συγκεκριμένων μεταβλητών, χρησιμοποιώντας το SPSS & εκτελώντας την εντολή: ***Graphs/ Scatterplot / Simple/ Y Axis: Pressure, X Axis: Age***. *παίρνοντας το παρακάτω διάγραμμα*

82

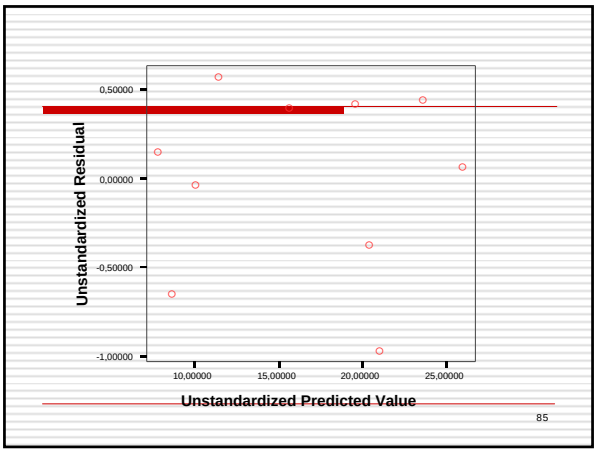


83



Γραμμική σχέση

84

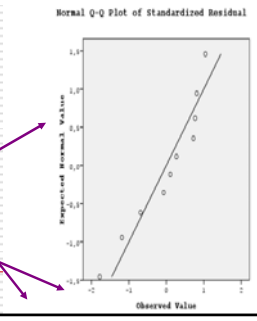


85

Κανονικότητα (4η υπόθεση)

Εξετάζουμε αν τα τυποποιημένα κατάλοιπα ακολουθούν πράγματι κανονική κατανομή, χρησιμοποιωτας

- ιστόγραμμα ή
- Q-Q plot ή
- P-P plot ή
- K-S τεστ ή
- χ^2 *test*.



Συνέχεια: SPSS

- Q-Q plot: Analyze/ Descriptive Statistics/ Q-Q plot.
- P-P plot: Analyze/ Regression/ Linear Regression Plot/ Standardized Residual plots/ P-P Normal Probability Plot
- Histogram: Analyze/ Regression/ Linear Regression Plot/ Standardized Residual/ Histogram.
- Kolmogorov-Smirnov Test. Μη παραμετρικός έλεγχος: Για να ελέγξουμε αν η κατανομή μιας μεταβλητής είναι συμβατή με την κανονική

εφαρμόζουμε το test Kolmogorov-Smirnov.

87

συνέχεια

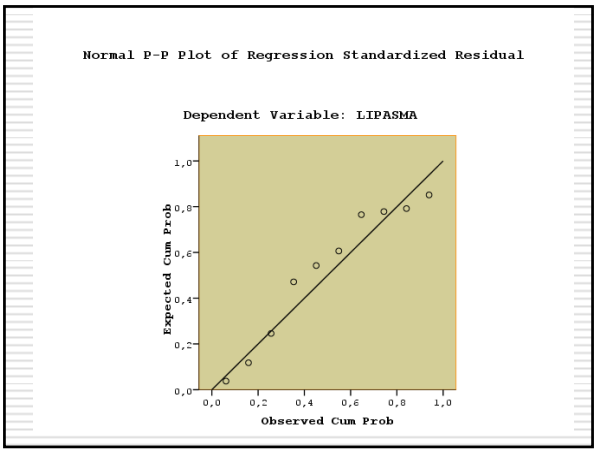
H_0: Η υπό έλεγχο κατανομή, δε διαφέρει από την κανονική κατανομή.

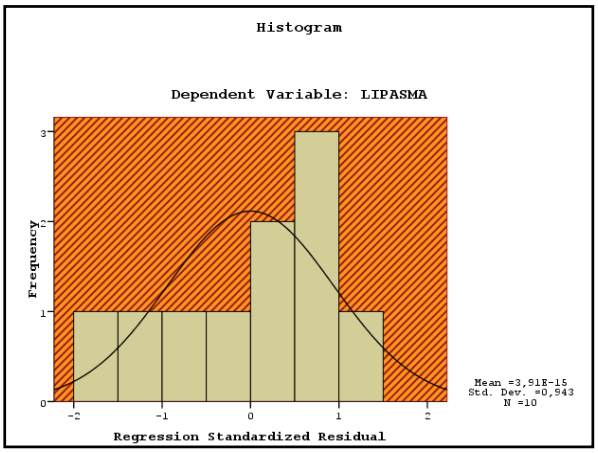
H_1: Η υπό έλεγχο κατανομή, διαφέρει από την κανονική κατανομή.

SPSS:

Analyze → Nonparametric tests → One sample K-S → Test variable list: βάζουμε τη μεταβλητή που θέλουμε να ελέγξουμε την κανονικότητα της (***τυποποιημένα κατάλοιπα = Res_1 =*** e_i), **Test distribution: Normal → Ok**

88





One-Sample Kolmogorov-Smirnov Test

		LIPASMA	Standardized Residual	Zscore(LIPASMA)
N		10	10	10
Normal Parameters[a,b]	Mean	16,4000	,0000000	,0000000
	Std. Deviation	6,58618	,94280904	1,000000
Most Extreme Differences	Absolute	,208	,179	,208
	Positive	,148	,134	,148
	Negative	-,208	-,179	-,208
Kolmogorov-Smirnov Z		,657	,565	,657
Asymp. Sig. (2-tailed)		,782	,907	,782

Δεδομένου ότι Sig = 0,907, η H_0 δεν απορρίπτεται! Δηλαδή, τα τυποποιημένα κατάλοιπα ακολουθούν πράγματι κανονική κατανομή και άρα ισχύει η υπόθεση της κανονικότητας

91

συνέχεια

Όταν διαπιστώνεται παραβίαση της κανονικότητας μπορούμε, σε αρκετές περιπτώσεις, να αντιμετωπίσουμε το πρόβλημα με κατάλληλους μετασχηματισμούς στις μεταβλητές.

92

Ανεξαρτησία (3η υπόθεση)

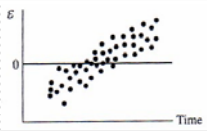
Εξαρτημένα ε_i (& Y_i) εμφανίζονται συνήθως σε περιπτώσεις που τα δεδομένα προκύπτουν:

1. από την ίδια πειραματική μονάδα σε διαφορετικές χρονικές στιγμές (π.χ. μετράμε την πίεση ή το βάρος του ιδίου ατόμου ανά εβδομάδα)
2. από μηχανές/όργανα που επηρεάζεται συν τω χρόνω η απόδοσή τους.

93

συνέχεια

Για τέτοιου είδους δεδομένα, σκόπιμο είναι να κάνουμε ένα διάγραμμα καταλοίπων ως προς το χρόνο (ακόμη και αν ο χρόνος δεν χρησιμοποιείται στο υπόδειγμα). Αν το διάγραμμα καταλοίπων (e_i) είναι:



94

συνέχεια

τότε είναι πιθανόν να υπάρχει στοχαστική εξάρτηση μεταξύ των σφαλμάτων.

Η υπόνοια αυτή ελέγχεται με τον έλεγχο ***Durbin-Watson***.

Αν διαπιστωθεί εξάρτηση των τιμών της Y τότε για την προσαρμογή κατάλληλου υποδείγματος και την εξαγωγή στατιστικών συμπερασμάτων πρέπει να χρησιμοποιηθούν ***ειδικές μέθοδοι***.

95

Έλεγχος αυτοσυσχέτισης

H_0: δεν υπάρχει αυτοσυσχέτιση ***ή $\rho = 0$***

H_1: υπάρχει αυτοσυσχέτιση ***ή $\rho \neq 0$***

$$DW = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

αν $r = 0 \rightarrow DW \cong 2$ (*όχι...αυτοσυχέτιση*)

αν $r = +1 \rightarrow DW \cong 0$ (*τέλεια...θετική..αυτ/ση*)

αν $r = -1 \rightarrow DW \cong 4$ (*τέλεια...αρνητική..αυτ/ση*)

r = *συντελεστής συσχέτισης*

96

συνέχεια

Μπορούμε επίσης να γράψουμε $DW \approx 2(1-r)$
Όπου r είναι ο εκτιμημένος συντελεστής συσχέτισης.
Αυτό σημαίνει ότι $-1 \leq r \leq 1$.
• Από τα παραπάνω έχουμε: $0 \leq DW \leq 4$.
• Εάν $r = 0$, $DW = 2$. Έτσι, σε γενικές γραμμές, μην απορρίπτετε την μηδενική υπόθεση εάν το DW είναι κοντά στο $2 \rightarrow$ δηλαδή δεν υπάρχουν στοιχεία για αυτοσυσχέτιση.
•Δυστυχώς, η DW έχει 2 κρίσιμες τιμές, μια ανώτερη κρίσιμη τιμή (d_u) και μια χαμηλότερη κρίσιμη τιμή (d_l), και υπάρχει επίσης μια ενδιάμεση περιοχή όπου δεν μπορούμε ούτε να απορρίψουμε ούτε να να μην απορρίψουμε την H_0.

97

συνέχεια

Γενικά:

$0 < DW < d_l \Rightarrow$ θετική αυτοσυσχέτιση

$4 - d_l < DW < 4 \Rightarrow$ αρνητική αυτοσυσχέτιση

$d_l < DW < d_u$ ή $4 - d_u < DW < 4 - d_l$

98

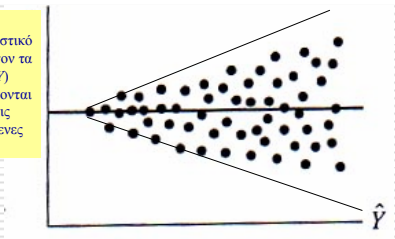
Ομοσκεδαστικότητα

Ένας πρώτος έλεγχος της σταθερότητας ή μη της διασποράς της Y (ή της e) για τα διάφορα επίπεδα της X μπορεί να γίνει με το *διάγραμμα διασποράς* και τα *διαγράμματα καταλοίπων*. Αν για παράδειγμα, το διάγραμμα καταλοίπων έχει μορφή όπως το παρακάτω, η πιο πιθανή αιτία είναι η μη σταθερότητα της διασποράς των τυχαίων σφαλμάτων ε.

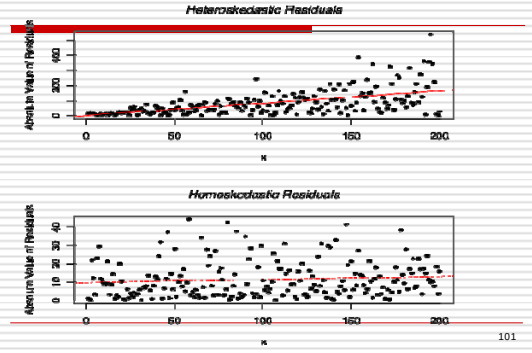
99

Ετεροσκεδαστικότητα

Ύπαρξη ετεροσκεδαστικότητας, εφόσον τα σημεία (e, Y) περιλαμβάνονται ανάμεσα στις δύο τεμνόμενες ευθείες.



100

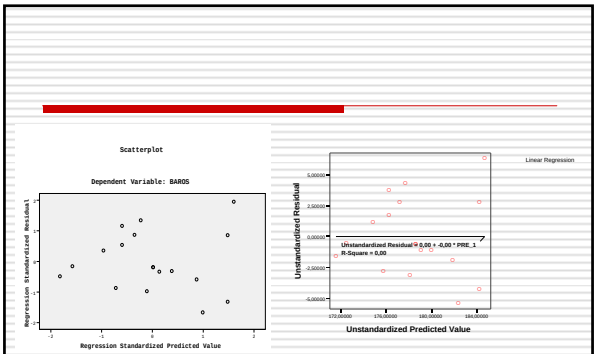


101

Γενικά, ερευνούμε διαγραμματικά για την ύπαρξη της ετεροσκεδαστικότητας, με:

- Ιστόγραμμα με την κατανομή των καταλοίπων: αυτό που αναζητούμε, είναι η κατανομή να ακολουθεί (όσο είναι εφικτό) κανονική κατανομή. Η μη κανονική κατανομή των καταλοίπων μπορεί να αντανακλά πρόβλημα κακής εξειδίκευσης της παλινδρόμησης.
- Διάγραμμα πιθανής κανονικότητας των καταλοίπων Normal probability plot of residuals όσο η κατανομή των καταλοίπων ακολουθεί την ευθεία γραμμή τόσο πιο κανονική είναι η κατανομή.
- Διάγραμμα διασποράς (Scatterplot) μεταξύ των τυποποιημένων καταλοίπων **ZRESID** και των τυποποιημένων προβλεπόμενων τιμών της εξαρτημένης μεταβλητής **ZPRED**.
- Διάγραμμα διασποράς (Scatterplot) μεταξύ των καταλοίπων **RESID** και των ερμηνευτικών μεταβλητών

102



103