30-3-2023

Διαγνωστικοί Έλεγχοι – Έλεγχοι Καταλοίπων

$Y = b_0 + b_1 X_1 + \varepsilon$

$Y = b_0 + b_1 X_1 + \dots + b_k X_k$ | αν $n-1$

$df_{er} = n-k-1$ | $df = n-1-k-1$

δείγμα $(x_1, y_1), \dots, (x_n, y_n)$

$Y_j = b_0 + b_1 x_j + \varepsilon_j$

$\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ $iid \sim N(0, \sigma^2)$

(τυχαίες αποκλίσεις)

1) $\varepsilon_j \sim N$ κανονικότητα
2) $Var(\varepsilon_j) = \sigma^2$ ομοσκεδαστικότητα
3) Ανεξαρτησία

(πώς ελέγχω αν ισχύουν οι υποθέσεις;)

Αν έχω ένα δείγμα $(z_1, z_2, \dots, z_n)$ από μια κατανομή $\Rightarrow$ 3 έλεγχοι {κανονικότητα, ανεξαρτησία

Όμως εδώ $\varepsilon_j = y_j - b_0 - b_1 x_j$

(ε_j: άγνωστα!!; b_0, b_1: άγνωστα!!)

Αν αντί των αγνώστων b_0, b_1 πάρουμε $\hat{b}_0, \hat{b}_1$

$$\hat{\varepsilon}_j = e_j = y_j - (\hat{b}_0 + \hat{b}_1 x_j), \quad j = 1, \dots, n$$

$$= y_j - \hat{y}_j = \text{κατάλοιπα}$$

$$\Sigma(y_i - \hat{y}_i)^2 = SSE = \Sigma e_i^2$$

Δυσκολίες

(1) Εστω $s_i^2 = Var(e_j)$ — s_i^2 δεν είναι ίσα

(2) Πάντα $\Sigma e_j = 0$ ⇐ δεν είναι ανεξάρτητα

$$\bar{Y} = \hat{b}_0 + \hat{b}_1 \bar{x}$$

? ∃ μετασχηματισμός των e_j που κανονικοποιούν τις διαταραχές

Μετρικοί Καταλοίπων

1) Τυποποιημένα κατάλοιπα (standardized)

$$\tilde{e}_j = \frac{e_j}{s_j} \Rightarrow Var(\tilde{e}_j) = 1$$

πρέπει να υπάρχει πρόβλημα με ανεξαρτησία.

2) Studentized residuals

$$r_j = \frac{\tilde{e}_j}{\sqrt{1-h_j}} = \frac{e_j}{s_j\sqrt{1-h_j}}$$

h_j = leverage (μόχλευση) της παρατήρησης

(H: hat matrix $H = X^T(X^TX)^{-1}X$)

h_j διαγώνια στοιχεία του H.

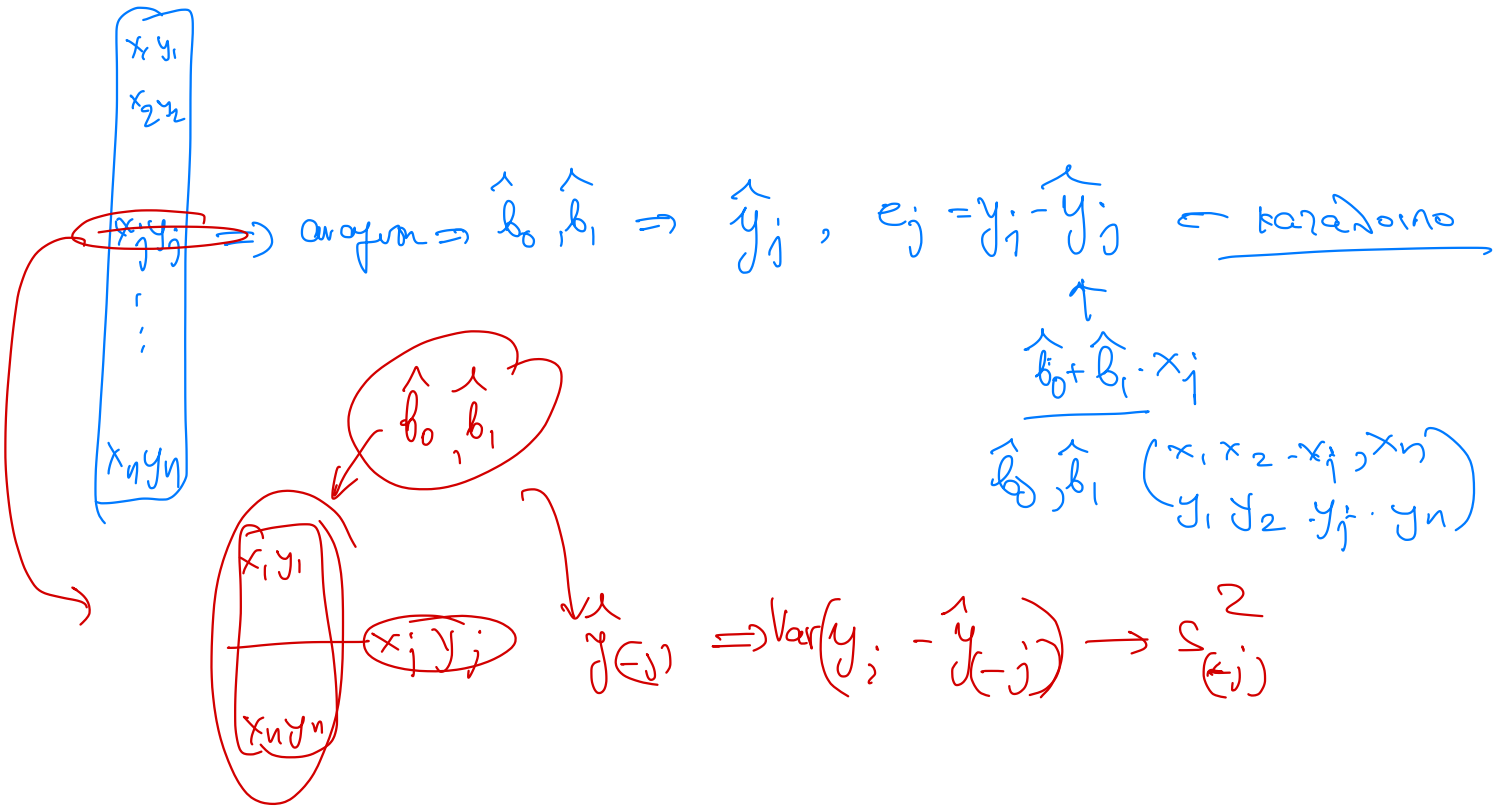
Αποδεικνύεται r_j, $j=1,\dots,n$ ασυσχετισμένα ανεξάρτητες τ.μ. $\sim t$

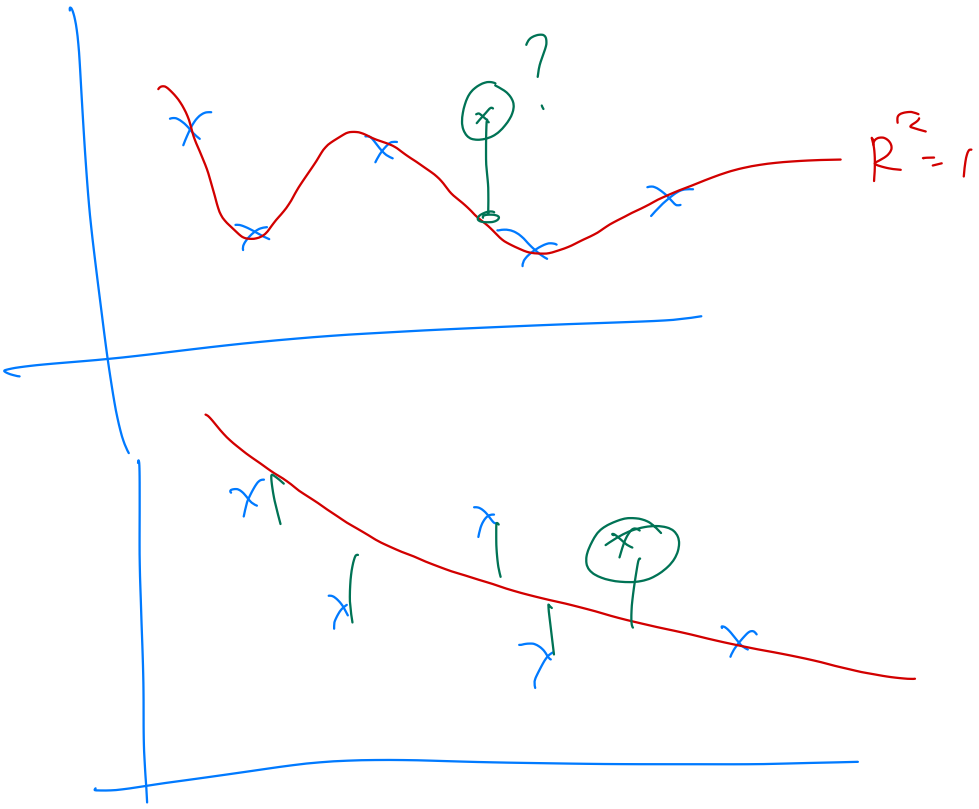
3) Jacknife residuals

$$\tilde{r}_j = \frac{e_j}{s_{(-j)}\sqrt{1-h_j}} \left.\begin{array}{l} \text{ασυσχετισμένα} \\ \sim t_{n-k-2} \left(\begin{array}{l} \text{όταν } n-k-2 \gg \\ t_{n-k-2} \approx N \end{array}\right) \end{array}\right\}$$

$s_{(-j)}$ = τυπ. απόκλιση των καταλοίπων παρατήρησης j $= y_j - \hat{y}_{(-j)}$

όπου όμως όλη η ανάλυση δε έχει συμπεριληφθεί η παρατήρηση j.





Στους διαγνωστικούς ελέγχους χρησιμοποιούμε είτε τα studentized residuals
" " jacknife

Στο Stata } → studentized → "standard"
κ' στο R } jacknife → "student"

Διαγνωστικοί Ελεγχοι

Ελεγχοι των υποθέσεων του γραμμικού μοντέλου με βάση τα κατάλοιπα (student ή jacknife)

1) Γραφήματα
2) Ελεγχοι υποθέσεων.

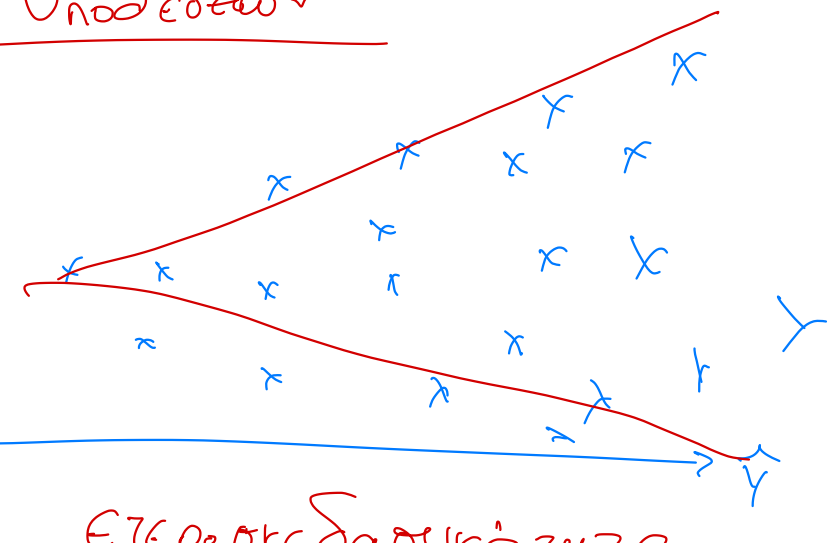
Συνήθως το καταλληλο σχήμα καλύτερα προβλ. τιμές

Θεωρία $Cov(e_j, \hat{Y}_j) = 0$

Παραβιάσεις Υποθέσεων

① Έλεγχος ετεροσκεδαστ.

έλεγχος Bartlett



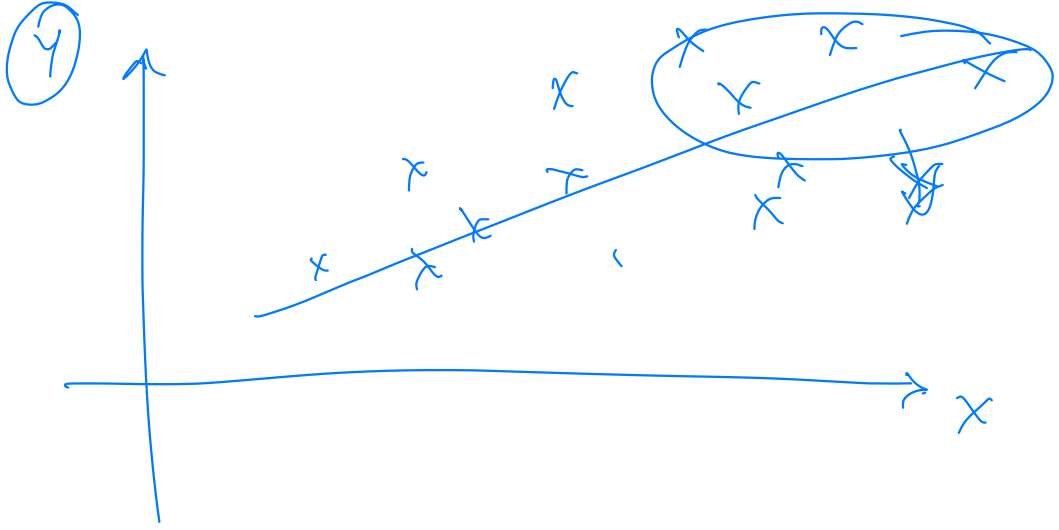
χοάνη

Ετεροσκεδαστικότητα

Διορθωτικά μέτρα

Μετασχηματισμοί της Y.

π.χ. $\tilde{Y} = \sqrt{Y}$ μειώνει τις τιμές της Y



Προσοχή $Y = b_0 + b_1 X + \varepsilon$: b_1 : μεταβολή της EY όταν X αυξηθεί κατά 1

αν $\sqrt{Y} = \gamma_0 + \gamma_1 X + \varepsilon$ $\quad \tilde{Y} = \gamma_0 + \gamma_1 X + \varepsilon$

$\tilde{Y}$

γ_1 : μεταβολή της $E(\tilde{Y}) = E(\sqrt{Y})$

$E(\sqrt{Y}) \neq \sqrt{EY}$

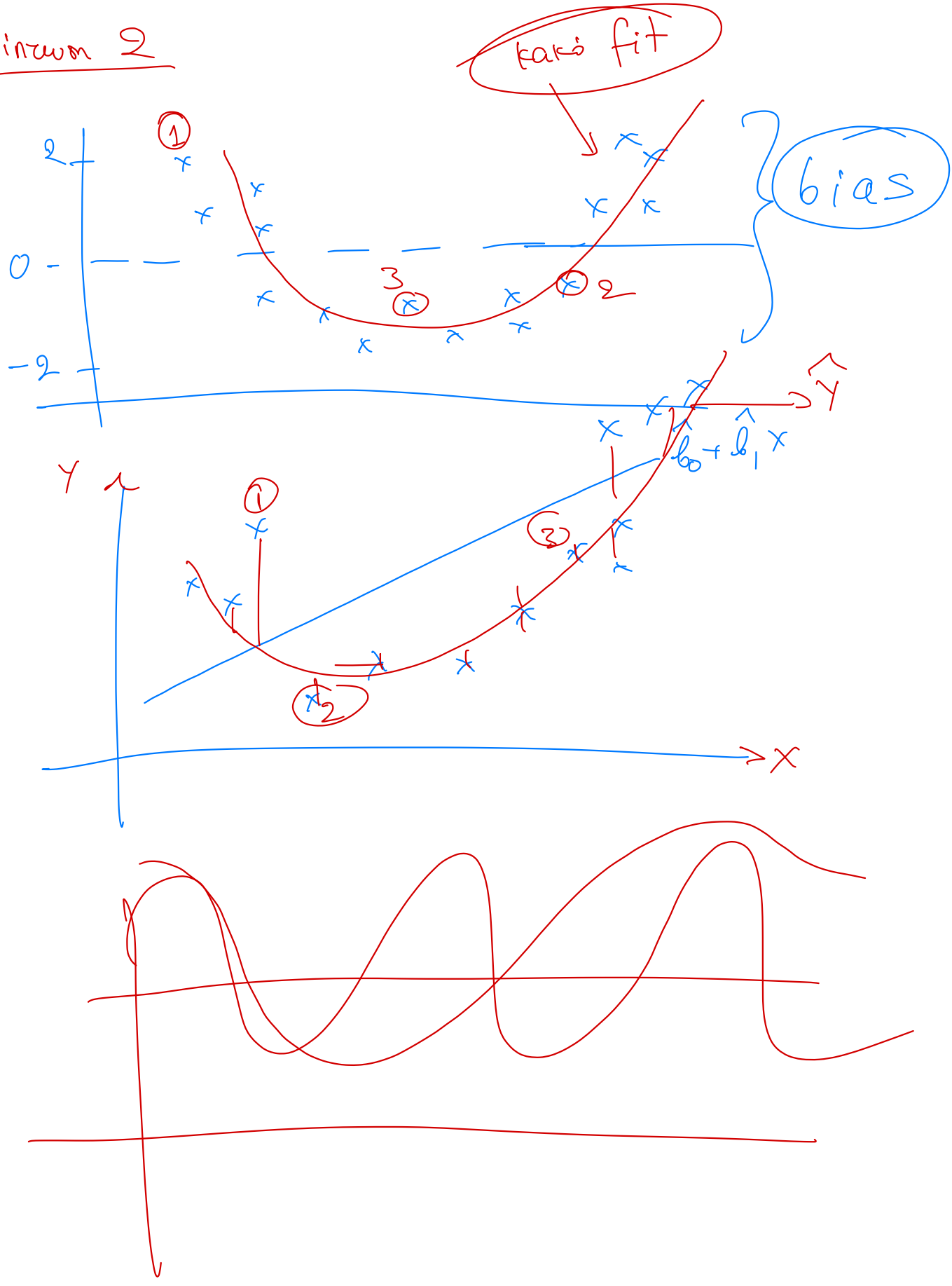
Οι κρίσεις δεν έχουν ερμηνεία ως μεταβολές της $E(Y)$

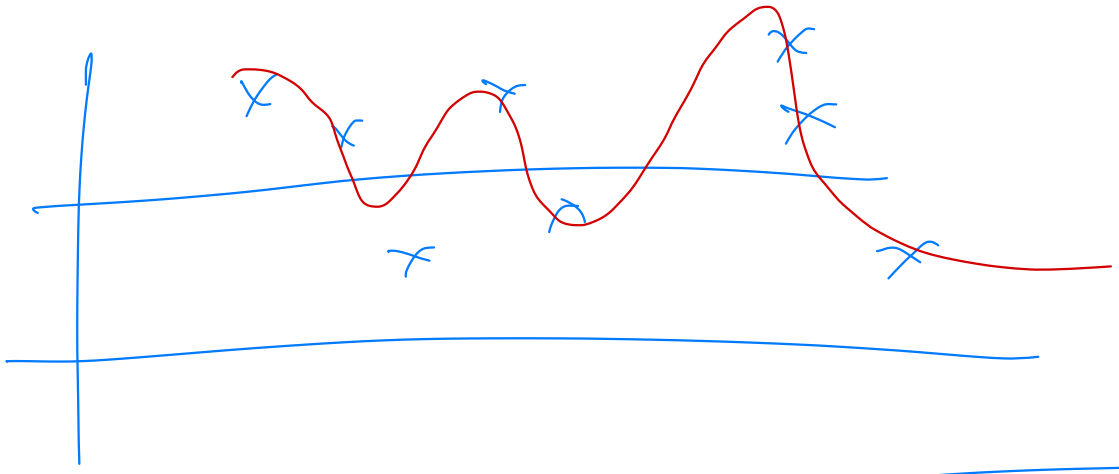
Άλλοι μετασχ. $\sqrt[3]{Y}$, $\log(Y)$

Τι γίνεται αν στο δείγμα υπάρχουν $Y \leq 0$?

$\tilde{Y}_j = Y_j - \min(Y) > 0$ ← μετασχ.

Περίπτωση 2





(3) Κανονικότητα

qplot γραφικά

έλεγχοι { KS
Shapiro Wilk.

(4) Αυτοσυσχέτιση Καταλοίπων.

$\tilde{r}_1, \tilde{r}_2, \tilde{r}_3, \ldots, \tilde{r}_n$ ← ανεξάρτητα

Η στερά των καταλοίπων μπορεί να παίζει ρόλο στην το δείγμα προέρχεται από χρονοσειρά !!

π.χ. χρονοσειρά θερμοκρασία x_i + αντ. όρος y_i } $i = 1, 2, \ldots$ ανά δεκαπ

Χρειάζονται μέθοδοι χρονοσειρών

Έλεγχος αυτοσυσχέτισης

Durbin – Watson test.

d : durbin watson statistic

$d \in [0, 4]$

$d \approx 2$ = Δ αυτοσυσχέτιση

$d \approx 0$ θετική αυτοσυσχ

$d \approx 4$ αρνητική "

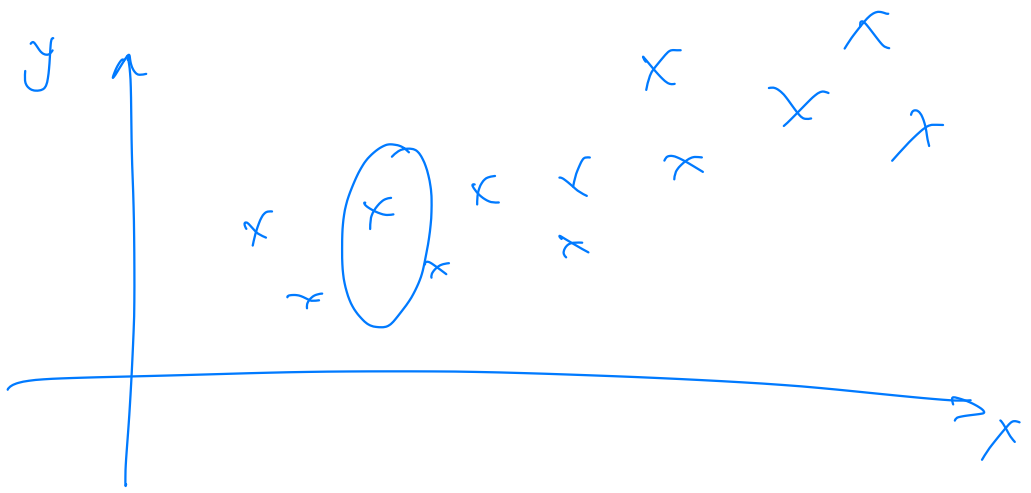
προσεγγιστικός έλεγχος

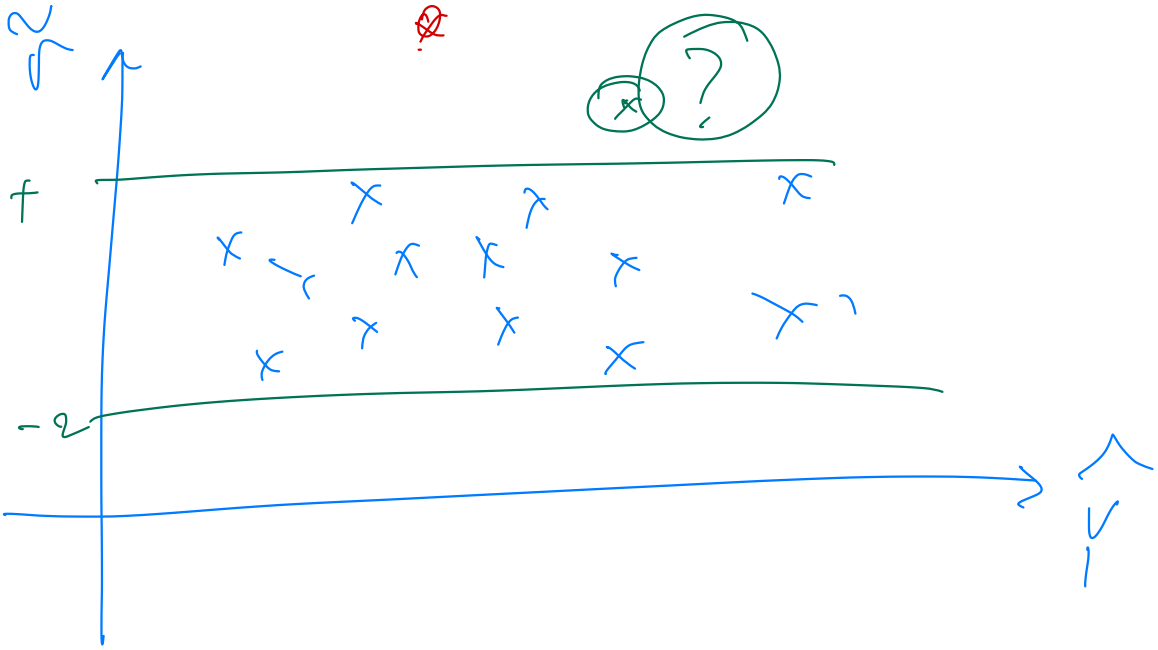
p-value προσέγγιση

Έκτοπες – Επιδραστικές παρατηρήσεις

(1) (Outliers – Influential Observation)

outlier ⊗ ← 27





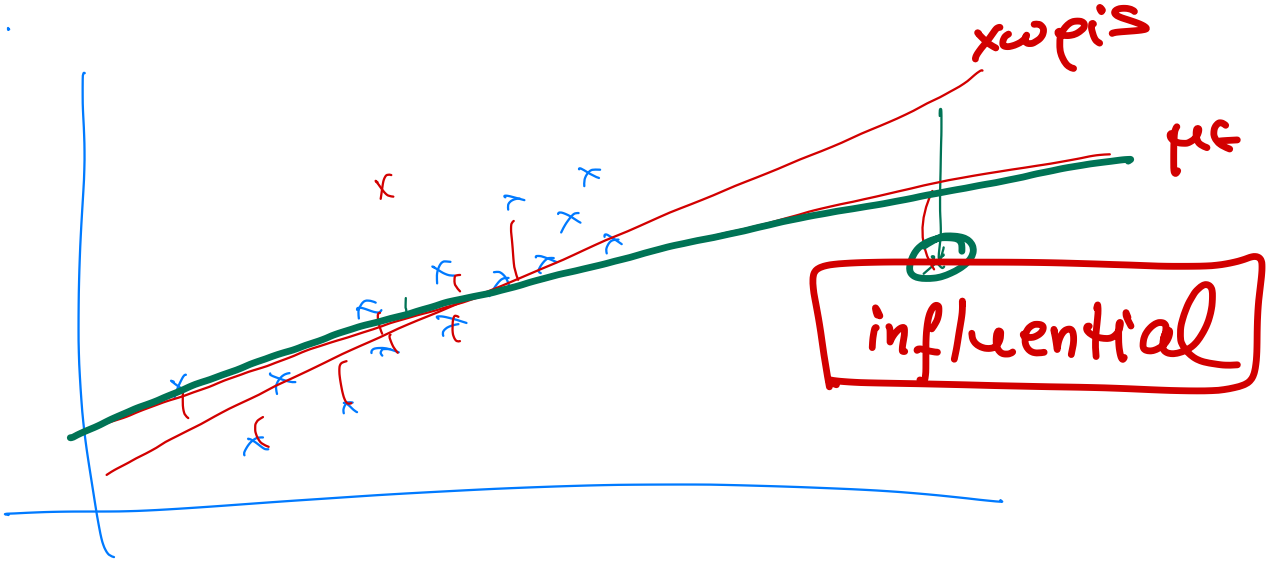
(1) λάθος καταγραφή

(2) " μέτρηση

??

Στατιστικοί έλεγχοι για outliers (*)

(2)



Outlier

$\tilde{r}_j$ πολύ μεγάλη ή πολύ μικρή τιμή → κριτήρια

Επιρροστικές { leverage h_j $0 \le h_j \le 1$
Cook's distance d_j

Πίνακες κρίσιμων τιμών για $\tilde{r}, h, d$.

Πολλαπλοί έλεγχοι

$P(\text{ωωτό}) = 1-\alpha$

$P(\text{η έλεγχοι ωωτοί}) \approx (1-\alpha)^n$ $\quad \left[\alpha \approx 0,\ (1-\alpha)^n \simeq 1-n\alpha\right]$

$\approx 1-n\alpha$

$= 1-\beta$ ← επ. σημ.

$\beta = n\alpha \Rightarrow \alpha = \frac{\beta}{n}$ → προσέγγιση διόρθωση Bonferroni

Πίνακες κρίσιμων τιμών $\tilde{r}$ έχουν
n διόρθωση
d Bonferroni

Επιλογή Μοντέλου

Βήματα Επιλογής Μοντέλου

1. Μέγιστο Μοντέλο (full model)
2. Κριτήρια σύγκρισης μοντέλων
3. Στρατηγικές επιλογής μεταβλητών
4. Αναζήτηση τελικού Μοντέλου
5. Εφαρμογές – Προβλέψεις

---

1. Μέγιστο Μοντέλο (full model)

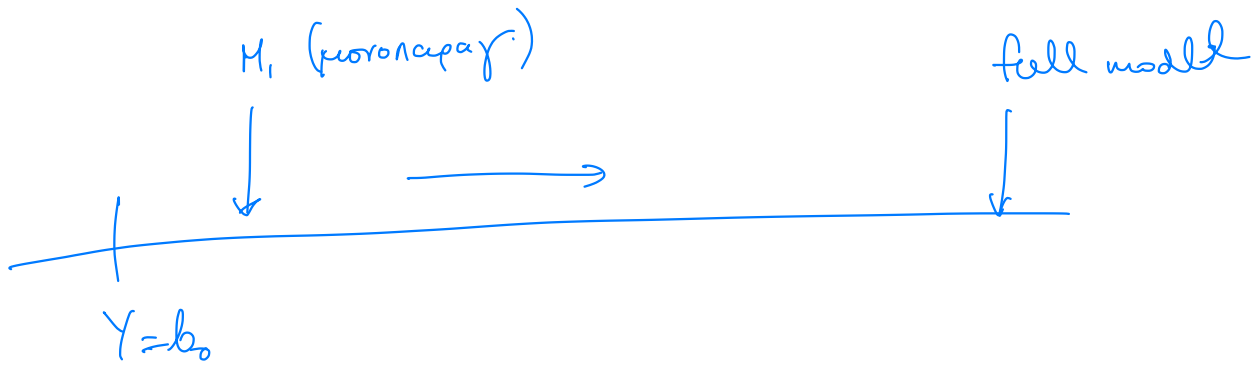
Περιέχει όλες τις υποψήφιες μεταβλητές
ή όλους τους επιτρεπτούς όρους
(πχ. μεγαλύτερης τάξης, αλληλεπιδράσεις κτλ)

$$Y = b_0 + b_1 X_1 + \cdots + b_k X_k + \varepsilon$$

df_{er} : "μεγάλο"

$df_{er} \geq 20$ ή 30

Πιο αυστηρός κανόνας $n \geq 5k$ ή και $n \geq 10k$.



---

(2) Κριτήρια Σύγκρισης Μοντέλων

(1) Για nested models

(full) $Y = b_0 + b_1 X_1 + \cdots + b_k X_k$ (full)

(part) $Y = b_0 + b_1 X_1 + \cdots + b_p X_p$ (υποσύνολο)

↖ nested στο full

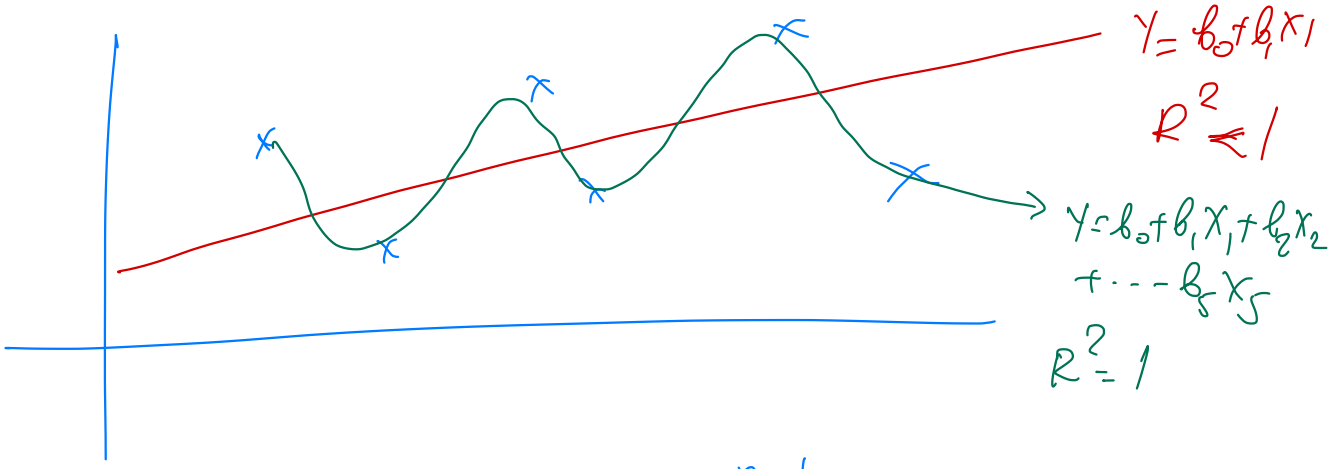
(a) Partial F-test

full $H_0 : b_{p+1} = b_{p+2} = \cdots = b_k = 0$ H_1: τουλάχιστον ένα $\neq 0$

(b) Adjusted R^2

$R^2 = \frac{SSR}{SST}$ = % μεταβ. του Y που εξηγείται από το μοντέλο

Για nested models $R^2_{full} \geq R^2_{part}$



$$adj\text{-}R^2 = 1 - (1-R^2)\cdot \underbrace{\frac{n-1}{n-p-1}}_{df_{er}} \qquad p = \text{αρ. μεταβλητών}$$

Όταν $n-1 \approx df_{er}$ $(p \ll n)$ $\frac{n-1}{n-p-1} \approx 1$

$$adj\ R^2 \approx 1-(1-R^2) = R^2$$

Όταν $p \to n$ $\frac{n-1}{n-p-1}$ μεγάλο $\Rightarrow$ $adj R^2 < R^2$

Πιο συντηρητική εκτίμηση για το % περιεκτικότητας των εξηγείται, λαμβάνοντας υπόψη τ' το μέγεθος του μοντέλου

(γ) Στατιστικό C_p - Mallows

Full model (k)
Partial " (p)

$$C_p = \frac{SSE(p)}{MSE(k)} - \left[n - 2(p+1)\right]$$

Αν $\text{Μοντέλο}(p) \approx \text{Μοντέλο}(k)$

$MSE(p) \approx MSE(k)$

Τότε $\underbrace{\frac{SSE(p)}{n-(p+1)}}_{df_{er,(p)}} \approx MSE(k) \Rightarrow$

$$\frac{SSE(p)}{MSE(k)} \approx n-(p+1) \Rightarrow$$

$$\Rightarrow C_p \approx n-(p+1) - \{n-2(p+1)\} = p+1$$

Όταν $\text{Model}(p) \approx \text{Model}(k) \Rightarrow$ $\boxed{C_p \approx p+1}$

Αν $\text{Μοντέλο}(p)$ χειρότερο από $\text{Μοντέλο}(k)$

$$\Rightarrow C_p > p+1$$

---

② Κριτήρια Πληροφορίας (όχι απαραίτητα για nested models)

AIC = Akaike Information Criterion

BIC = Bayesian " "

$$AIC = -2\log(L) + 2k$$

L : πιθανοφάνεια δείγματος τιμών από τις εκτιμήτριες των β.

k : μέγεθος μοντέλου

μικρές τιμές AIC ⇒ καλό μοντέλο

Στρατηγικές Επιλογής Μεταβλητών

Έστω ένα full model με k πολύ μεγάλο

Πόσα διαφορά υποσύνολα μπορώ να σχηματίσω 2^k !!

4 $\begin{cases} Y = b_0 \\ Y = b_0 + b_1 X_1 \\ Y = b_0 + b_1 X_2 \\ Y = b_0 + b_1 X_1 + b_2 X_2 \end{cases}$

Μέθοδοι stepwise regression

① Forward $(X_1, \dots, X_k)$ υποψήφιες μεταβλητές

R state

ⓐ Όλα τα μονοπαραγοντικά $\begin{cases} Y = b_0 + b_1 X_1 & \text{AIC} & \text{p-value} \\ Y = b_0 + b_1 X_2 & \text{AIC} & \text{p-value} \\ \vdots & \vdots & \\ Y = b_0 + b_1 X_k & \text{AIC} & \end{cases}$

Επιλέγουμε αυτό με το min p έστω X_j

$b_0 + b_1 X_j$ $\begin{cases} b_0 + b_1 X_j + b_2 X_1 \\ b_0 + b_1 X_j + b_2 X_2 \\ \vdots \\ b_0 + b_1 X_j + b_2 X_k \end{cases}$ min(p) εκτός των X_j

pvalue της νέας στο μοντέλο

Αν min(p) ⇒ p_{enter} ⇒ συμπεριλαμβάνεται

Για να μπει μια νέα μεταβλητή πρέπει να έχει το μικρότερο p από όλες που είναι εκτός κ' το $p < p_{enter}$.

---

② Backward.

Ξεκινάμε με full model

$$Y = b_0 + b_1 X_1 + \dots + b_j X_j + b_k X_p$$

p p p → max (p)

$p_{max} > p_{remove}$

↓ $b_j X_j$ φεύγει

⋮

συνέχεια μέχρι όλα $p < p_{remove}$

③ Stepwise

Κάθε φορά που προστίθεται μια μεταβλητή με $p < p_{enter}$ εξετάζονται όλες οι μεταβλητές στο τρέχον μοντέλο κ' αφαιρείται αυτή με max p αν αυτό είναι $p > p_{remove}$.

Προσοχή: Αν $P_{enter} = 0,07$

$P_{remove} = 0,05$

μπορεί να βρούμε σε κάποιο βήμα μια X_j εκτός που αν προσθέτει έχει $p = 0.06$

$P_{enter} < P_{remove}$.