

21-1-2026

Διαγνωστικοί Έλεγχοι

Έλεγχος καταλληλότητας μοντέλου :

κατά πόσον ικανοποιούνται οι υποθέσεις
του παμνδρόμιου

Υποθέσεις γραμμικού μοντέλου

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

Δείγμα n παρατηρήσεις : $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$

- 1) $\varepsilon_j \sim N(0, \sigma^2)$
- 2) ε_j ανεξάρτητα

- 1) Κανονικότητα
- 2) Ανεξαρτησία
- 3) Ομοσκεδαστικότητα

Διαγνωστικοί έλεγχοι
(έλεγχοι καταλοίπων)

- ① Αν έχω ένα δείγμα $(\varepsilon_1, \dots, \varepsilon_n)$ από παρατηρήσεις
κ' θέλω να ελέγξω π.χ. αν N
π.χ. έλεγχος Καλοστροφού-Smirnov
Shapiro-Wilk ...

- ② Εδώ $\varepsilon_1, \varepsilon_2, \dots ?$

$$\rightarrow \varepsilon_j = Y_j - E(Y_j) = Y_j - \underbrace{b_0}_{\text{αγνωστο}} - \underbrace{b_1}_{\text{αγνωστο}} X_{1j} - \underbrace{b_2}_{\text{αγνωστο}} X_{2j} - \dots - \underbrace{b_k}_{\text{αγνωστο}} X_{kj}$$

$\varepsilon_1, \dots, \varepsilon_n$: άγνωστα !!

Θυμίζετε όταν εκτιμούμε διασπορά $X \sim N(\mu, \sigma^2)$ με άγνωστο μ :

$$\hat{\sigma}^2 = \frac{(X_1 - \mu)^2 + \dots + (X_n - \mu)^2}{n}$$

τι κάνω όταν μ άγνωστο?

$$\hat{\sigma}^2 = \frac{(X_1 - \bar{X})^2 + \dots + (X_n - \bar{X})^2}{n-1}$$

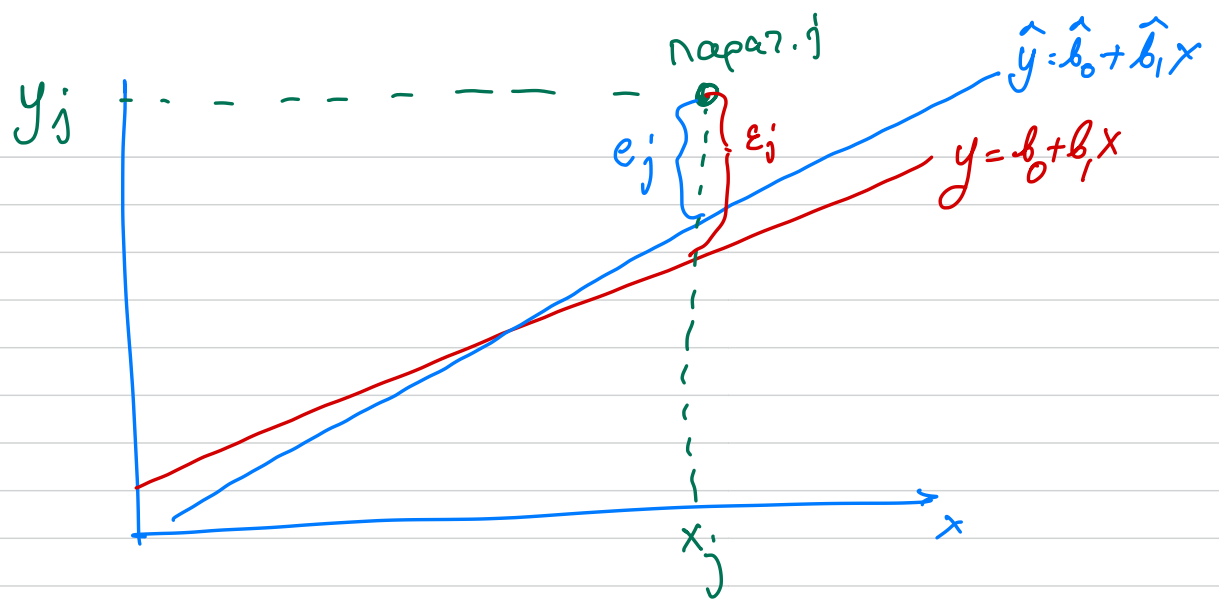
Αφού δε γνωρίζουμε τα $\varepsilon_1, \dots, \varepsilon_n$ θα θελήσει να τις "εκτιμήσουμε"

1) Ο προφανής τρόπος $b_0, b_1, \dots, b_k \leftarrow \hat{b}_0, \hat{b}_1, \dots, \hat{b}_k$

$$\hat{\varepsilon}_j = e_j = Y_j - (\hat{b}_0 + \hat{b}_1 X_{1j} + \dots + \hat{b}_k X_{kj}) = Y_j - \underbrace{\hat{Y}_j}_{\text{προβλεπ. τιμή}}$$

e_j : κατάλοιπο (residual) της παρατ. j .

$$\left(\text{π.χ. } SSE = \sum_j (Y_j - \hat{Y}_j)^2 = \sum_j e_j^2, \text{ MSE} = \frac{\sum e_j^2}{df_{\text{ερ}}} \right)$$



- ΟΜΩΣ :
- ① $e_j \sim N(0, \sigma_{e_j}^2)$ $\sigma_{e_j}^2$ διαφορετικά
 e_j : όχι ομοσκεδαστικά
 - ② Γενικά $\sum e_j = 0$ ($\bar{e}_j = 0$) οχι ανεξάρτητα

Μετασχηματισμοί Καταλοίπων

- ① Τυποποιημένα (standardized residuals)

$$\tilde{e}_j = \frac{e_j}{\hat{\sigma}_{e_j}} = \frac{e_j}{s_j} \quad \left(\sigma_{e_j}^2 \approx \left(\right) \sigma_{\text{αγνωστο}}^2 \right)$$

$$\text{Var}(\tilde{e}_j) = 1 \quad \forall j \quad \leftarrow \text{ομοσκεδαστικά}$$

όχι ανεξάρτητα

- ② Studentized residuals

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

"οχεδόν" ανεξάρτητα
για μεγάλο $n \sim N$
 $\sim t$
 $\text{var}(e_j') = 1$

$h_j = \text{leverage}$ (μόχλευση) της παρατήρησης j
 (σχετίζεται με το πόσο κοντά είναι η
 παρατήρηση στο κέντρο βάρους των δεδομένων)
 (υπολογίζεται)

$$\left[\begin{array}{l} \text{Hat matrix } H = X^T(X^T X)^{-1}X \quad H^2 = H \\ h_j : \text{διαγώνια στοιχεία του } H \end{array} \right]$$

3) Jackknife κατάλοιπα

$$r_j = \frac{e_j}{S_j \sqrt{1-h_j}}$$

S_j : τυπική απόκλιση του $y_j - \hat{y}_j$ σε
 ένα δείγμα από το οποίο έχει
 αφαιρεθεί η παρατήρηση j

(τεχνική jackknife)

Τρέχουμε το ίδιο μοντέλο στο δείγμα
 με παρατ. $(1, 2, \dots, j-1, j+1, \dots, n)$ [εκτός της j]

Τώρα η j είναι νέα παρατήρηση

(δεν έχει συμμετάσχει στην εκτίμηση των $\hat{\beta}$ κτλ)

$$y_j - \hat{y}_j \leftarrow$$

r_j ανεξάρτητα, $\sim t$, $\text{Var}(r_j) = 1$.

τα r_j για τα r_j με διαγνώσεις r_j

$(R, Stata) \rightarrow \text{studentized} \rightarrow \text{"standard"}$
 $\rightarrow \text{jackknife} \rightarrow \text{"student"}$

$$\hat{e}_j = r_j$$

Διαγνωτικοί Έλεγχοι

$\left\{ \begin{array}{l} \text{κανονικότητα} \\ \text{Ομοσκεδαστικότητα} \\ \text{Ανεξαρτησία} \end{array} \right\}$ καταλογογραφία

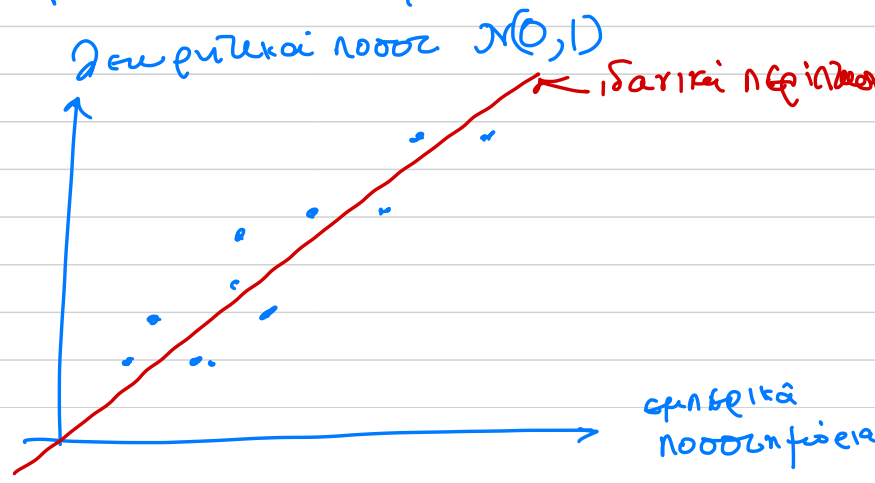
Έλεγχοι για εκτροπές « $\wedge$ » επιδραστικές
Outliers - influential

Διαγνωτικοί Έλεγχοι $\rightarrow$ διαγνωστικοί (διαγράμματα)
 $\rightarrow$ έλεγχοι υποθέσεων

① Κανονικότητα

Kolm-Smirnov
 or Shapiro-Wilk $\left\{ \begin{array}{l} \text{στα } r_j \end{array} \right\}$

Γραφικά: qqplots



Παράδειγμα

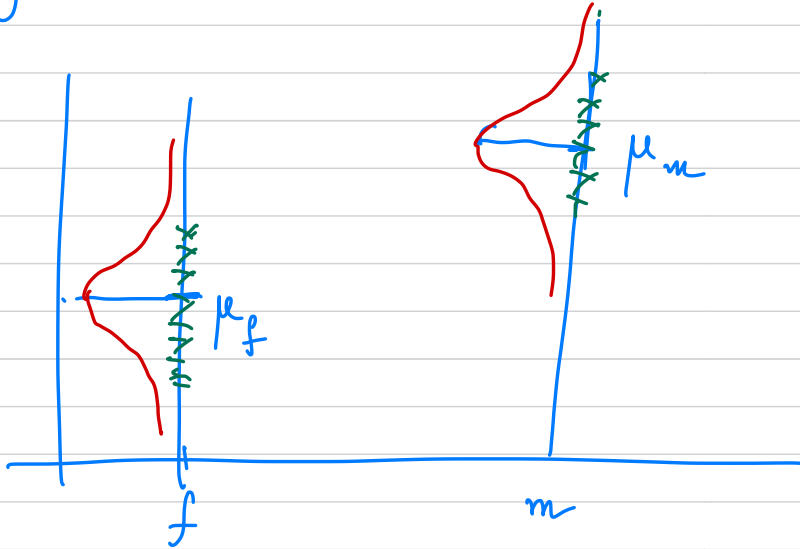
Στο μοντέλο μας

$$y_i \sim \mathcal{N}(b_0 + b_1 x_{i1} + \dots + b_k x_{ik}, \sigma^2)$$

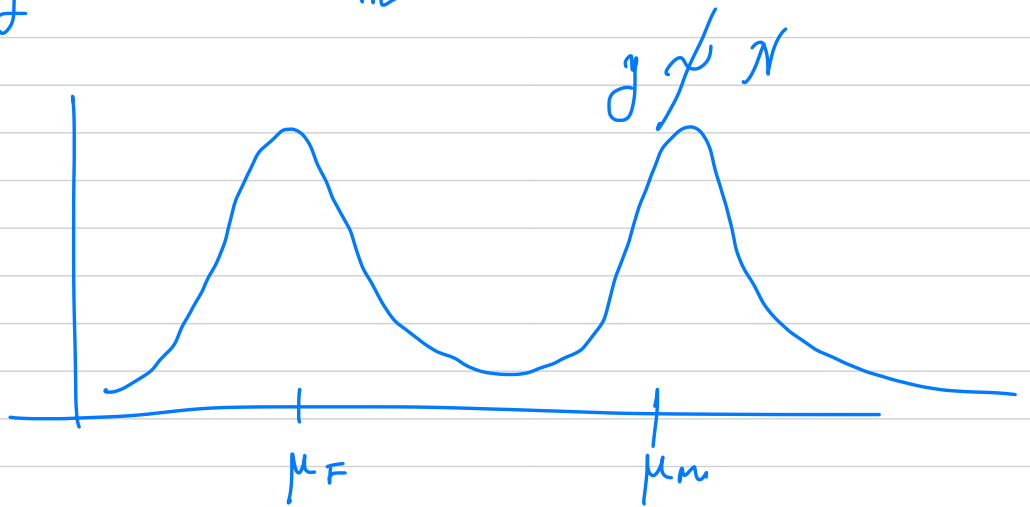
όχι με την ίδια μέση τιμή!!

π.χ. $y = b_0 + b_1 \cdot X_{\text{male}}$

$$X_{\text{male}} = \begin{cases} 1 & \text{sex} = \text{male} \\ 0 & \text{sex} = \text{female} \end{cases}$$



Γράφημα y

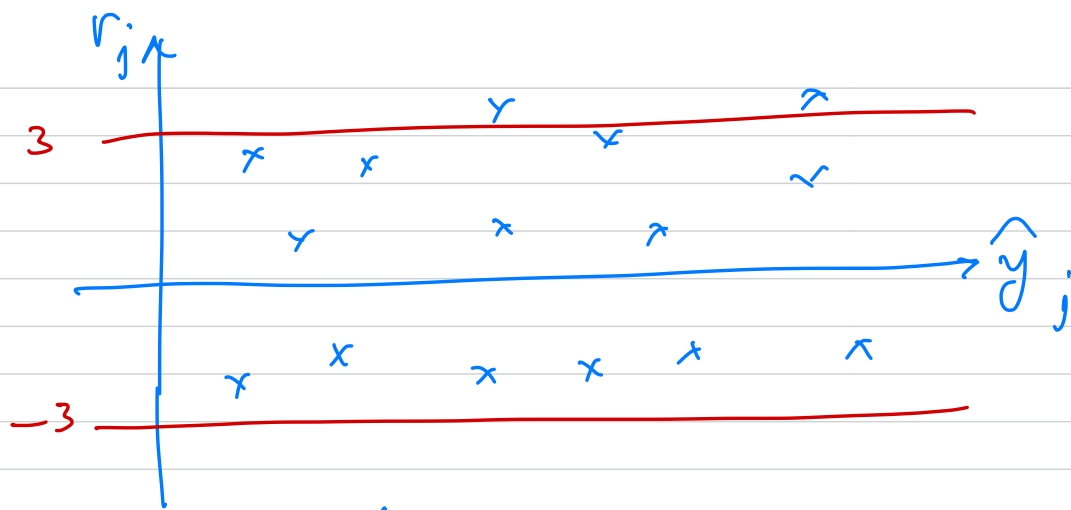


2) Ομοσκεδαστικότητα

Διάγραμμα κατανομών

Διάγραμμα (scatter)

r_j ως προς $\hat{y}_j$

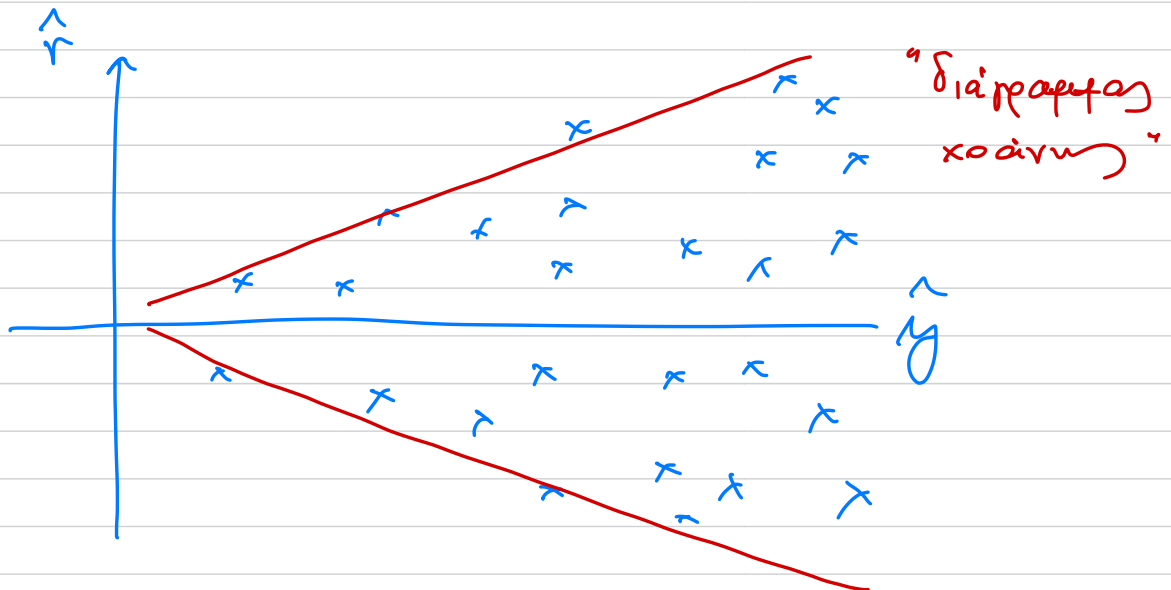


$$\text{Cov}(r_j, \hat{y}_j) = 0$$

$$\text{Ενώ } \text{Cov}(r_j, y_j) \neq 0$$

$$\text{Cov}(e_j, \hat{y}_j) = 0$$

"Τυπική εικόνα ετεροσκεδαστικότητας"



Bartlett test : κριτήριο σε δύο ομάδες

Αναμενόμενα

① Παραλλαγή μεθόδου LSE : weighted LSE

② Μετασχηματισμοί της Y $\begin{cases} Y_i \sim \sqrt{Y} \\ \sqrt{Y} = b_0 + b_1 X_1 + \dots + \varepsilon \end{cases}$

Προσοχή!! χάνονται οι ερμηνείες των b !!

$$E(\sqrt{Y}) = b_0 + b_1 X_1 + \dots$$

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

Πιο δραστικοί μετασχηματισμοί

$$Y_2 = \sqrt[3]{Y}$$

$$Y_3 = \log Y$$

3) Ανεξαρτησία — Έλεγχος αυτοσυσχέτισης καταλοίπων

$\varepsilon_1, \varepsilon_2, \dots$ ανεξάρτητα $\Leftrightarrow$

$y_1, y_2, \dots$ ανεξάρτητα $\leftarrow$ βασική υπόθεση

Παραβιάζεται σε περιπτώσεις $\left\{ \begin{array}{l} \text{χρονοσειρών} \\ \text{ή} \\ \text{πολ. με εναρμονισμένη} \\ \text{μεταβολή} \end{array} \right.$

Χρονοσειρά ακολουθία από παρατηρήσεις μέσα στο χρόνο

η χρονική σειρά συνήθως συνεπάγεται συσχέτιση μεταξύ διαδ. τιμών (πρώτης σειράς)



συσχέτιση μεταξύ διαδοχ. παρατηρήσεων : αυτοσυσχέτιση

δεν χρησιμοποιείται το κλασικό γραμμικό μοντέλο
(μέθοδοι κρουστικών)
ARMA, ARIMA, GARCH. -

Ελεγχος αυτοσυσχέτισης?

d = Durbin-Watson statistic

$$0 \leq d \leq 4$$

$$d \approx 2$$

δεν υπάρχει αυτοσυσχέτιση

$$d < 2$$

" αρνητική αυτοσυσχ.

συνήθως $\rightarrow$

$$d > 4$$

" θετική "

p-value : προσεγγιστικό

Outliers - Influential Observations

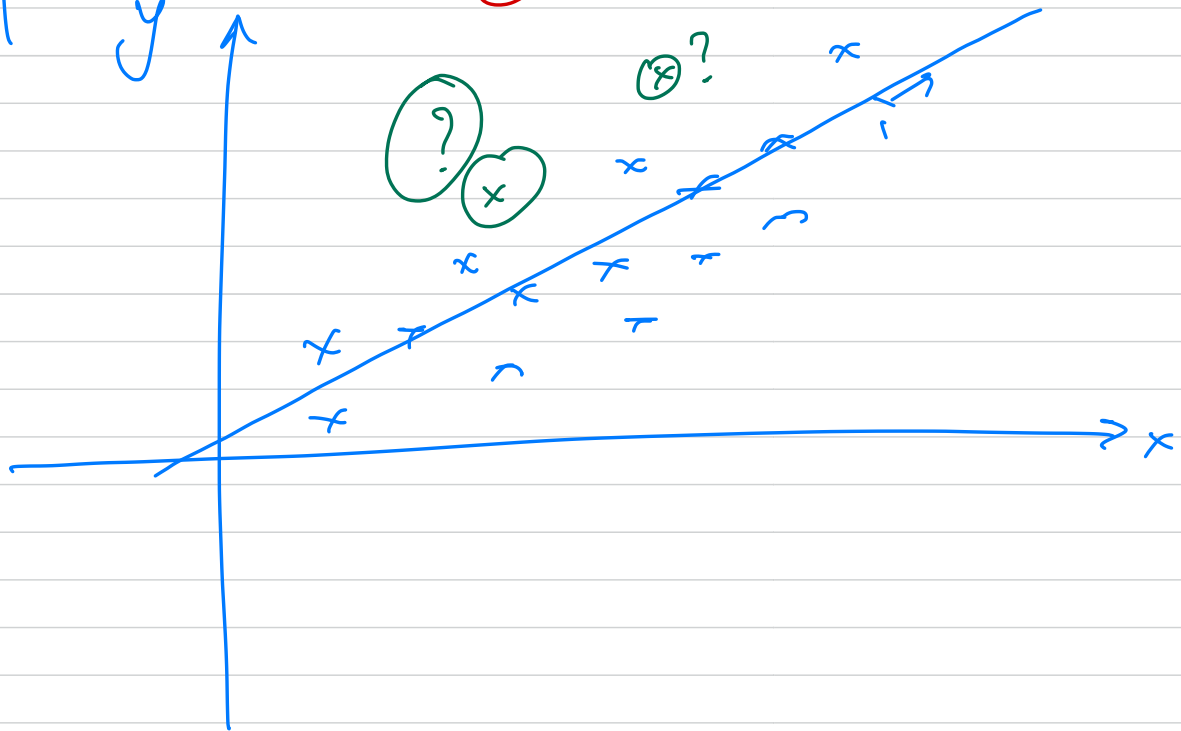
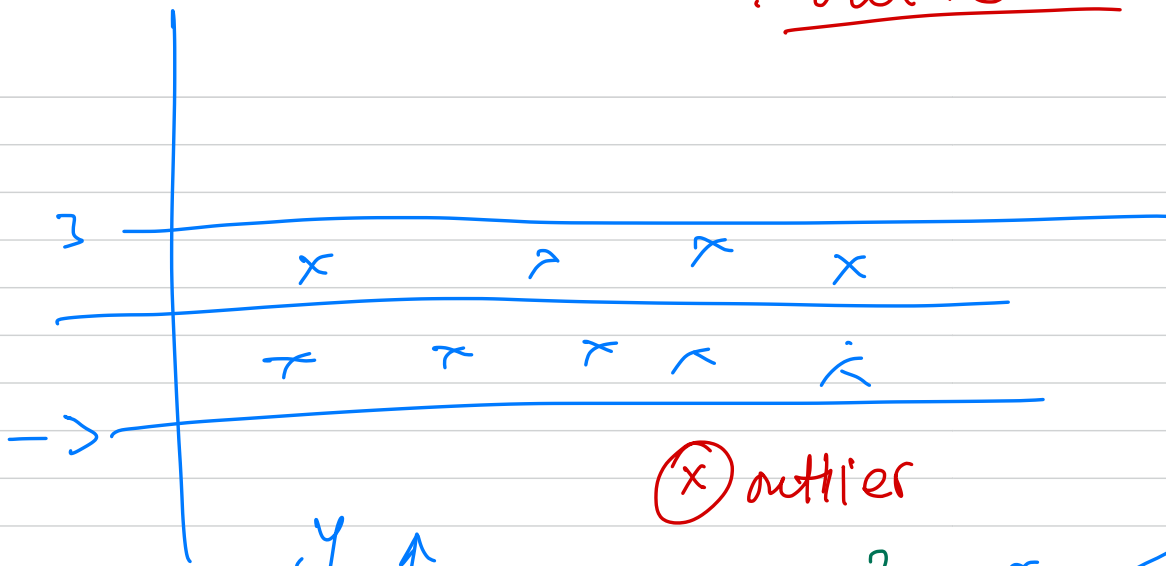
Παρατηρήσεις με διαφορετική συμπεριφορά
από τις υπόλοιπες

① Outliers - έκτροπες παρατηρήσεις

Παρατήρηση με $|r_j|$ πολύ μεγάλη

π. $r_j = 10$ " $r_j = -10$

(x) ? outlier



- 1) Λάθος καταγραφή (π.χ. πίεση: 31 ενώ 13)
- 2) Παρουσίαση που θα έπρεπε να έχει αποκλειστεί
- 3) Αν δεν μπορεί να αποκλειστεί/διορθωθεί ??

Συνήθως τρέχουμε το μοντέλο χωρίς αυτή κ' παρουσιάζουμε και τα δύο αποτελέσματα (με και χωρίς)

Ελέγχοι για outliers

Πινάκες με κριτικές τρέψ row r_i

εξαρτώνται από $n, k, \alpha = \epsilon_1$.
σημειώστε.

Διόρθωση Bonferroni

Όταν κάνουμε πολλούς ελέγχους στο ίδιο δείγμα
έ'σο ίδιο ποσοστό

έ'σοντας με $\alpha = 0.05$

το α δε ισχύει για ταυτόχρονο συμπέρασμα

Εσω ότι όλες οι ανακρίσεις είναι ανεξάρτητες
έ' κάθε μια έχει πιθαν. $(1-\alpha)$ να είναι σωστή

Α γίνουν m έλεγχοι

$$P(\text{όλοι σωστοί}) \approx (1-\alpha)^m$$

≈ 1 σημαίνει
πλήρως σωστό

$$\text{Για } \alpha \text{ μικρό } \Rightarrow (1-\alpha)^m \approx 1 - \alpha m$$

Επομένως αν θέσουμε συνολικό $\alpha m \approx 5\%$

$$\Rightarrow \text{για κάθε ένα πρέπει } \alpha \approx \frac{5\%}{m}$$

$$\alpha = \frac{0.05}{m}$$

Πχ. αν $m=10$ $\alpha = 5\% = 0.005$ $\leftarrow$ Διόρθωση Bonferroni

παρα ποτ' ανσμπι

Επιδραστικ' Παρεκτιρισθης →