Lezione n. 13 (a cura di Damiano Matteucci)

Analisi dei residui e previsione

Y $X_1, X_2, .., X_k$ k=50

Da una situazione di partenza con un numero di variabili pari a k si passa, tramite una delle procedure stepwise, a una situazione in cui troviamo solo variabili significative:

Y $X_1, X_2, .., X_k$ k=6

Queste variabili significative presenteranno tutte p-value < 0,05

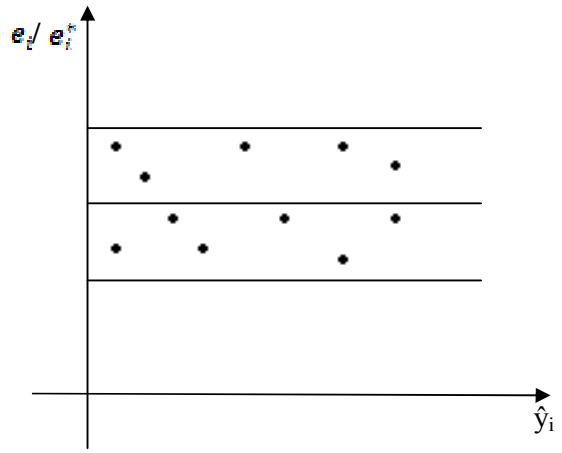
Ognuna di queste variabili avrà come coefficiente , il quale determina come ogni singola X influenza Y; pertanto è importante osservarne il segno.

Per svolgere delle previsioni corrette abbiamo bisogno di assunzioni del modello corrette.
Il modello si compone di 2 parti:

- ⇨ Deterministica $E(Y_i) = \beta_0+\beta_1 X_{i1}+...+\beta_k X_{ik}$
- ⇨ Stocastica ε_i $E(\varepsilon_i)= 0$, i=1, ..., n $\varepsilon_i \perp\!\!\!\perp \varepsilon_j$, $i \neq j$
  $V(\varepsilon_i)=\sigma^2$, i=1, ..., n $\varepsilon_i \sim N(0,\sigma^2)$

Diagnostica

Grafici dei residui



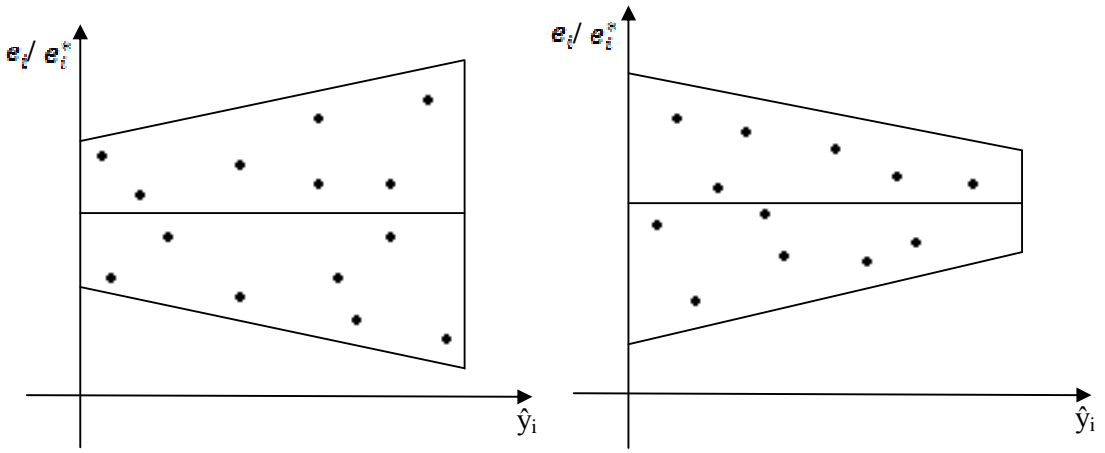
$$e_i = y_i - \hat{y}_i = y_i - x_i'\hat{\beta}$$

e_i è espresso nella stessa unità di misura della y; ad esempio con y uguale al reddito ed $e_i = 12$ il reddito di i è sottostimato di 12000 €.

e_i^* = residui standardizzati

Sull'asse delle ascisse non scelgo le variabili perché o le rappresento tutte insieme oppure ne scelgo una alla volta.

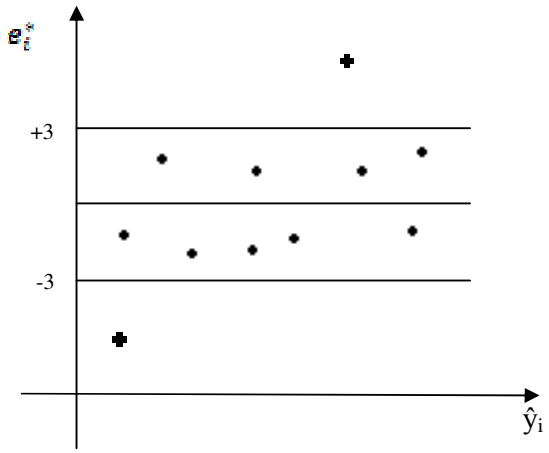
Nel caso in cui la varianza dell'errore non sia costante avrò eteroschedasticità.



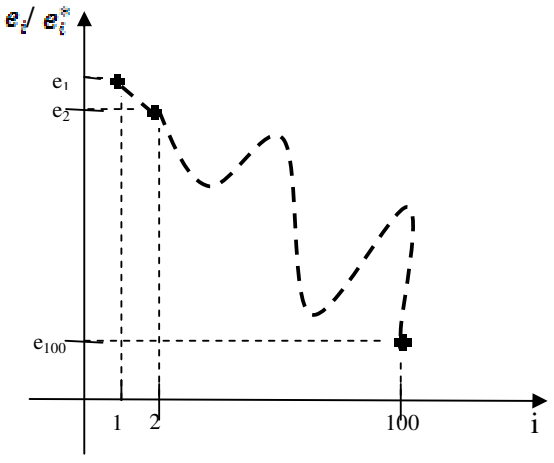
Nel primo grafico la varianza tende a crescere con la Y, formando cosi un megafono; nel secondo grafico, invece, la varianza tende a diminuire man mano che cresce la Y (questo secondo caso, tuttavia, è raro).

$$e_i^* = \frac{e_i}{\sqrt{S^2}} = \frac{e_i}{\sqrt{\frac{DR}{(n-(k+1))}}}$$

Delle volte i residui standardizzati permettono di trovare delle osservazioni anomale, le quali si discostano da tutte le altre osservazioni e sono chiamate outlier.



Forme di dipendenza tra i termini di errore



i= numero d'ordine del residuo nel data set

Se esiste una forma di dipendenza, questa spesso si manifesta tra unità vicine, ed è visivamente individuabile a seconda della struttura del grafico (ad esempio una sovrastima nei primi 10 soggetti).

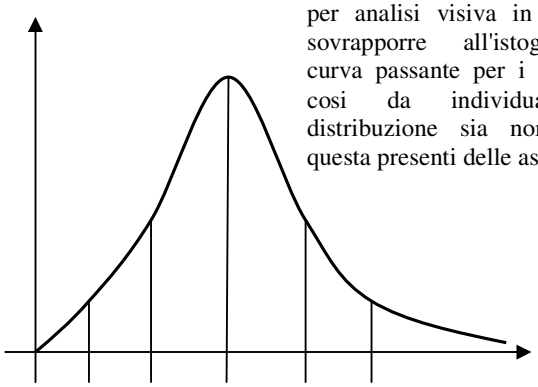
$$\varepsilon_i \sim N(0, \sigma^2), i = 1, \dots, n$$

{

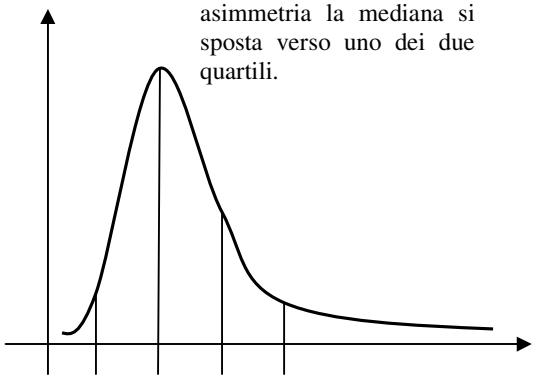
istogramma

boxplot

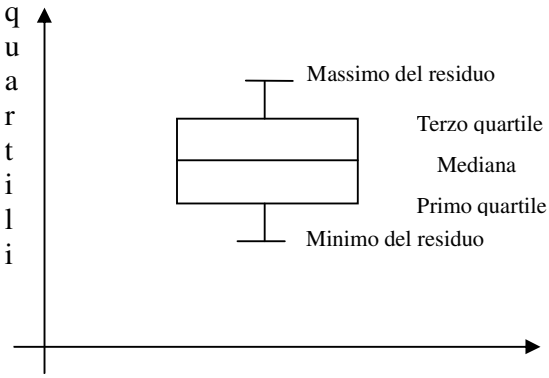
qq-Plot

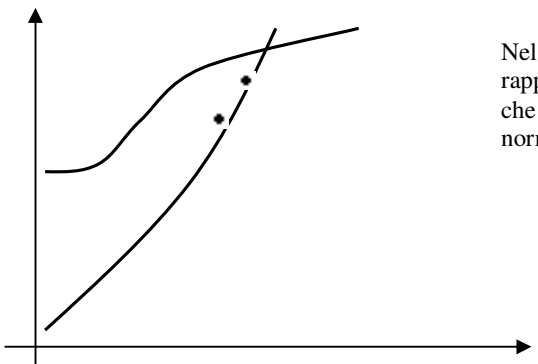


Nell'istogramma la verifica avviene per analisi visiva in quanto basta sovrapporre all'istogramma una curva passante per i diversi livelli, cosi da individuare se la distribuzione sia normale oppure questa presenti delle asimmetrie.



In questo caso quando vi è asimmetria la mediana si sposta verso uno dei due quartili.





Nel qq-plot viene effettuato un confronto tra il grafico che rappresenta i quantili della distribuzione dei residui e il grafico che rappresenta i quantili della distribuzione teorica. Se la normalità è rispettata i punti sono allineati.

Per verificare che la distribuzione della parte stocastica sia normale con media pari a 0 e varianza pari a sigma^2 si possono utilizzare 3 diversi grafici: istogramma, boxplot e qq-plot.

Al termine della diagnostica possono presentarsi due possibili situazioni:
1. Se le assunzioni sono rispettate siamo quasi pronti ad effettuare delle previsioni.
2. Se le assunzioni non sono rispettate necessitiamo di correttivi.

1. Come fare una previsione?

x_0 trasposto $\leftarrow$ $x_0' = (1 \cdot x_{01} . x_{02} \cdot \dots \cdot x_{0k})$

$$\hat{y}_o = \underline{x}_0' \hat{\underline{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \cdots + \hat{\beta}_k x_{0k}$$

$V(\hat{y}_0) = x_0'$

$V(\hat{\beta}) x_0 = \sigma^2 \underline{x}_0' (x'x)^{-1} \underline{x}_0$

non la conosco quindi stimo la varianza

$\hat{V}(\hat{y}_0) = S^2 x_0' (x'x)^{-1} x_0$ MISURA DELLA PRECISIONE DELLA MIA PREVISIONE

più è grande e meno è precisa la mia previsione

Intervallo di confidenza

$$\begin{cases} \hat{y}_0 - t_{\alpha/2} \sqrt{\hat{V}(\hat{y}_0)} \\ \hat{y}_0 + t_{\alpha/2} \sqrt{\hat{V}(\hat{y}_0)} \end{cases}$$

$t_{\frac{\alpha}{2}} = percentile\, t\big(n - (k+1)\big)$

$2 t_{\alpha/2} \sqrt{\hat{V}(\hat{y}_0)}$ AMPIEZZA INTERVALLO → risente direttamente della varianza

$y_0 = x_0' \beta + \varepsilon_0$

$E(y_0) - x_0'\beta$

reddito medio per una certa categoria di persone

Nel primo caso c'è una derivazione del singolo soggetto dal valore medio,mentre nel secondo non è presente il termine di errore.

La stima che faccio è del valore atteso, stessa cosa per l'intervallo di confidenza.
Se invece voglio la stima per un singolo soggetto trovo l'*intervallo predittivo* $y_0 = x_0'\beta + \varepsilon_0$

$$\begin{cases} \hat{y}_0 - t_{\alpha/2}\sqrt{\tilde{V}(\hat{y}_0)} & \tilde{V}\ tiene\ in\ considerazione\ l'errore \\ \hat{y}_0 + t_{\alpha/2}\sqrt{\tilde{V}(\hat{y}_0)} & t_{\frac{\alpha}{2}} = percentile\ t(n-(k+1)) \end{cases}$$

$$\tilde{V}(\hat{y}_0) = \sigma^2(1 + x_0'(x'x)^{-1}x_0 = \sigma^2\hat{V}(\hat{y}_0)$$

AMPIEZZA = $2t_{\alpha/2}\sqrt{\tilde{V}(\hat{y}_0)} > 2t_{\alpha/2}\sqrt{\hat{V}(\hat{y}_0)}$