LEZIONE N.8 (a cura di Teresa Fanelli)

- Forma matriciale del Modello di Regressione Semplice

L' assunzione di base del modello è: $Y_i=\beta_0+x_i\beta_1+\varepsilon_i$ i=1,2,…..n. Lo stesso modello può essere scritto attraverso vettori e matrici,si ha: $\mathbf{Y}=x\boldsymbol{\beta}+\boldsymbol{\varepsilon}$, dove:

- Y è un vettore di dimensione n×1 → $\mathbf{Y}_{n\times1}=\begin{pmatrix} y1 \\ y2 \\ . \\ . \\ yn \end{pmatrix}$

- X è una matrice di dimensione n×2 → $X_{n\times2}=\begin{pmatrix} 1 & x1 \\ 1 & x2 \\ . & . \\ . & . \\ 1 & xn \end{pmatrix}$

- $\boldsymbol{\beta}$ è il vettore dei PARAMENTRI di dimensione 2×1 → $\boldsymbol{\beta}_{2\times1}=\begin{pmatrix} \beta0 \\ \beta1 \end{pmatrix}$

- $\boldsymbol{\varepsilon}$ è un vettore di dimensioni n×1 → $\boldsymbol{\varepsilon}_{n\times1}=\begin{pmatrix} \varepsilon1 \\ \varepsilon2 \\ . \\ . \\ \varepsilon n \end{pmatrix}$

Questa forma risulta importante nel modello di regressione con più variabili.

Risolvendo tutti i passaggi ,prima il prodotto:

$X_{n\times2}\times\boldsymbol{\beta}_{2\times1}=\begin{pmatrix} \beta0+x1\,\beta1 \\ \beta0+x2\,\beta1 \\ . \\ . \\ \beta0+xn\,\beta1 \end{pmatrix}$ } il vettore della COMPONENTE DETERMINISTICA,di dimensione n×1

E aggiungendo il termine di errore al vettore ottenuto ,si ha:

$Y=x\,\boldsymbol{\beta}+\boldsymbol{\varepsilon}=\begin{pmatrix} \beta0+x1\beta1+\varepsilon1 \\ \beta0+x2\beta1+\varepsilon2 \\ . \\ . \\ \beta0+xn\beta1+\varepsilon n \end{pmatrix}$

Riprendendo le IPOTESI sul termine di errore,quali:

1. $E(\varepsilon i) = 0 \ \forall i$;
2. $V(\varepsilon i) = \sigma^2 \quad \forall i$;

3. $\varepsilon i \perp\!\!\!\perp \varepsilon j \quad \forall i \neq j$ dove $\perp\!\!\!\perp$ è il simbolo di <u>indipendenza;</u>

- in alcune considerazioni si ha che $\varepsilon_i \sim N(0,\sigma^2) \ \forall i$;

risulta importante vedere come queste ipotesi vengono applicate al vettore ε_i. Si ha:

1. $E(\boldsymbol{\varepsilon})=0$,ciò vuol dire che ε è un vettore con tutti elementi nulli,di dimensione n×1;
2. $V(\boldsymbol{\varepsilon})= \sigma^2 I$ dove σ^2è la varianza comune e I è la matrice identità $=\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$,da cui si ottiene

$$\rightarrow \begin{pmatrix} \sigma^2 & 0 & . & 0 \\ 0 & \sigma^2 & . & . \\ . & . & \sigma^2 & 0 \\ 0 & . & . & \sigma^2 \end{pmatrix}$$

Questa matrice corrisponde alla matrice varianza-covarianza,dove la diagonale principale è data dalle varianze e gli 0 corrispondono alle covarianze.

3. $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I})$ questo implica che gli errori sono INDIPENDENTI. Tale ipotesi sintetizza le altre ipotesi viste per **ε.** Nel modello di regressione con più variabili basta utilizzare solo quest'ipotesi.

- Stimatori

Se si conosce β,vuol dire che si dispone del vettore $\boldsymbol{\beta}=\begin{pmatrix}\beta 0\\ \beta 1\end{pmatrix}$,si ha:

ŷ=x **β** è un vettore di dimensione n×1,dove **ŷ** è il vettore delle previsioni che si possono calcolare per ogni soggetto. La differenza tra il vettore dei dati osservati e il vettore delle previsioni,da il vettore degli errori previsti di dimensioni n×1 → $\mathbf{y}-\hat{\mathbf{y}}_{n\times 1}=\begin{pmatrix}y1-\hat{y}\,1\\ y2-\hat{y}\,2\\ \cdot\\ \cdot\\ yn-\hat{y}\,n\end{pmatrix}$

sulla base di questo vettore si può definire la somma dei quadrati: $S(\beta)=(\mathbf{y}-\hat{\mathbf{y}})'_{1\times n}(\mathbf{y}-\hat{\mathbf{y}})_{n\times 1}$.

Il risultato di questa operazione è uno scalare,si ha:

$(y_1-\hat{y}_1, y_2-\hat{y}_2, \ldots\ldots, y-\hat{y}_n)\begin{pmatrix}y1-\hat{y}1\\ y2-\hat{y}2\\ \cdot\\ \cdot\\ yn-\hat{y}n\end{pmatrix}=(y_1-\hat{y}_1)^2+(y_2-\hat{y}_2)^2+\ldots\ldots+(y-\hat{y}_n)^2=\sum l(y1-\hat{y}1)_2$

In forma esplicita: $S(\boldsymbol{\beta})=(\mathbf{y}-\boldsymbol{\beta})'(\mathbf{y}-x\boldsymbol{\beta})$,da cui si ottiene $\bar{\boldsymbol{\beta}}_{2\times 1}=(x'x)^{-1}{}_{(2\times 2)}\,x'y_{(2\times 1)}$ →è lo stimatore dei minimi quadrati,ed è equivalente alle formule viste precedentemente $=\begin{pmatrix}\bar{\beta}0\\ \bar{\beta}1\end{pmatrix}=\begin{pmatrix}\bar{y}-\bar{\beta}_1\,\bar{x}\\ \dfrac{Cod(x,y)}{D(x)}\end{pmatrix}$

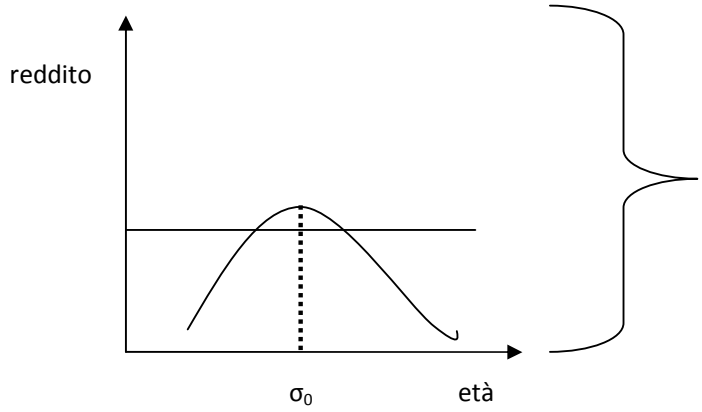
→quindi,concludendo, ci sono due modi per poter stimare $\boldsymbol{\beta_0}$ **e** $\boldsymbol{\beta_1}$**:**

1. Con le formule più lunghe;
2. Attraverso $(x'x)^{-1}(x'\mathbf{y})$

- Diagnostica

il modello di regressione semplice si basa su due assunzioni:

1. Parte deterministica->$E(y_i)=\beta_0+\beta_1 x_i$, ciò vuol dire che il valore atteso di y è una funzione lineare di x.
   In alcuni casi risulta essere più complesso,del tipo: $E(y_i)=\beta_0+\beta_1 x_i^2$
   o anche: $E(y_i)=\beta_0+\beta_1 x_i^2 + \beta_i x_i^2$.
   Queste situazioni sono presenti soprattutto nell'ambito economico, ad esempio nel la relazione tra reddito-età,graficamente:

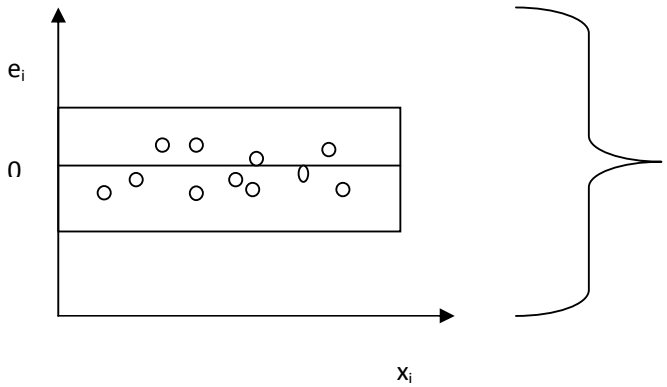


In questo caso il modello di regressione non è adeguato,perché la relazione è di tipo quadratica:y↑e poi↓;ciò porta a dire che l'età non influenza il reddito il che è del tutto sbagliato.

Se le assunzioni non sono vere,attraverso l'inferenza,le conclusioni possono essere sbagliate,questo perché il modello è del tipo lineare. Per capire se il modello è adeguato vengono utilizzati due indici:

1. $0 \leq r^2 \leq 1$;
2. Statistica F($H0:\beta_i=0$)

Per avere un'analisi più completa si utilizzano i residui di regressione: cioè quantità calcolate in corrispondenza di ogni stima→$e_i=y_i-\hat{y}_i$ (differenza tra valore osservato e valore stimato). Il metodo più semplice è rappresentato dal grafico dei residui:per ogni soggetto si ha (x_i,e_i):

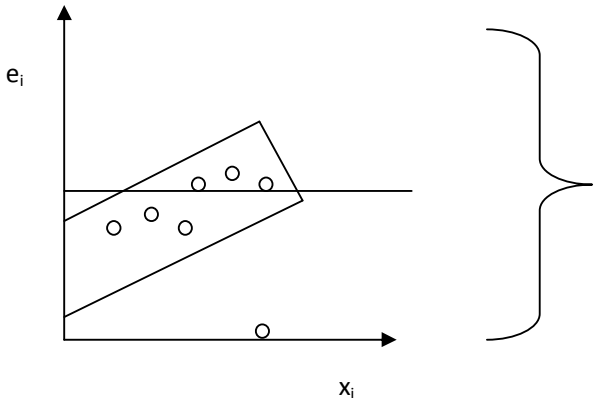


-Se le assunzioni sono vere i punti si concentrano intorno allo zero formando così un rettangolo;questo perché $\frac{1}{n}\varepsilon_i e_i=0$,ciò vuol dire che ci sono valori positivi negativi che si compensano tra di loro.

-Se ci sono problemi sulle assunzioni i punti portano

Risulta,così importante riconoscere alcuni grafici:

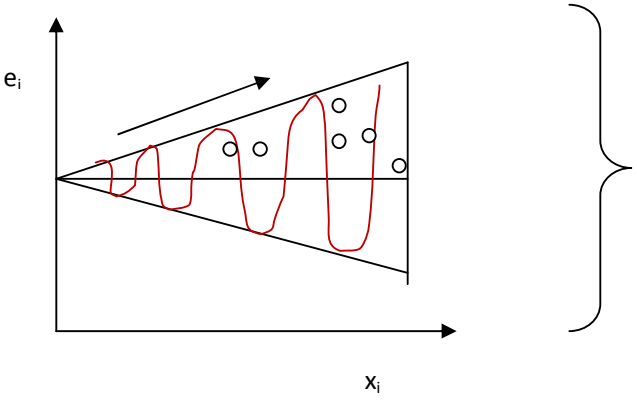
1. Problema:componente deterministica non specificata:



-I residui sono intorno allo zero,ma primo sotto e poi diventano positivi. Ciò vuol dire che il modello non è stato specificato nella forma lineare.

-Esiste,quindi una relazione sistematica tra x e e_i .

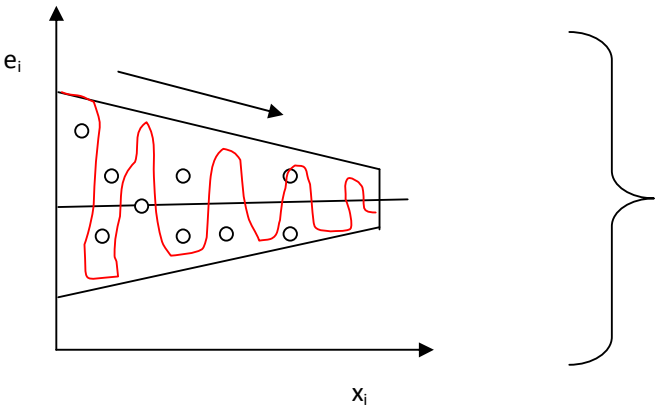
2. Problema:varianza costante(eteroschedasticità):



La varianza degli errori aumenta con x. L'andamento della varianza si nota dalle oscillazioni:

-oscillazione piccola,varianza bassa,viceversa nel caso contrario.

In questo caso non sono valide le conclusioni relative alla verifica di ipotesi e all'intervallo di confidenza.



Caso contrario:σ^2 diminuisce con la x.

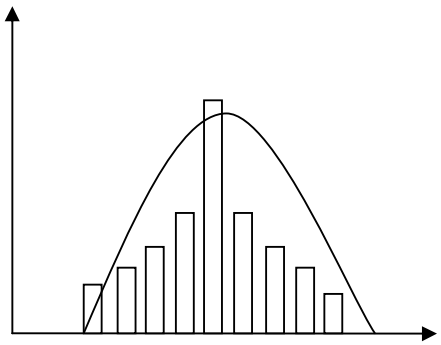
-assunzioni non rispettate;

-è valida sempre l'eteroschedasticità.

A volte può accadere di avere grafici in cui al posto di x_i vi è y_i;ciò succede perché x_i è una trasformazione lineare di y_i e viceversa. Lo stesso accade per il residuo standardizzato $e^*=\frac{e_i}{\sqrt{s^2}}$.

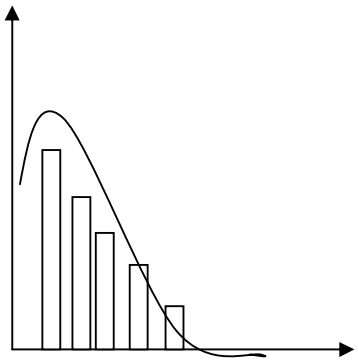
Un'altra possibilità è quella di verificare la normalità;questo è possibile attraverso diversi modi:

1. Modo:utilizzo dell'istogramma:



se le assunzioni sono rispettate l'istogramma avrà la forma di una normale;questa può essere verificata sovrapponendo la curva della normale.

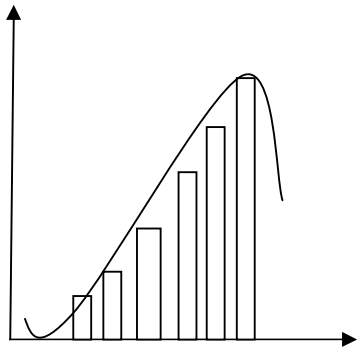
Tuttavia può accadere che queste assunzioni non vengano rispettate,si avranno così grafici differenti:



-Distribuzione asimmetrica verso destra;

-si hanno delle conseguenze sulla verifica d'ipotesi e sugli intervalli di confidenza;

- gli errori sono pochi,ma molto forti

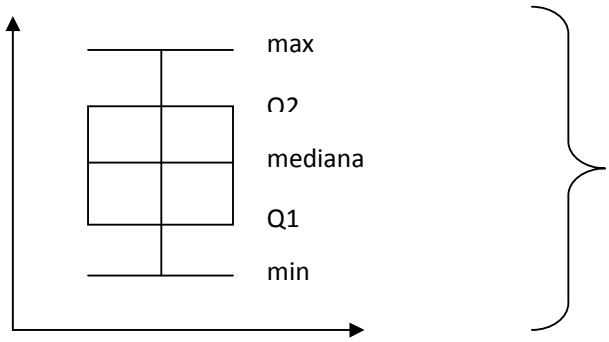


Caso opposto a quello precedente:

-asimmetria negativa;

-pochi errori e molo piccoli.

2. Modo:box plot(ε_i):

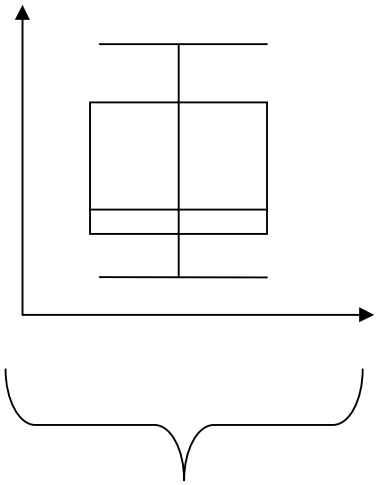


Se è normale,si ha:

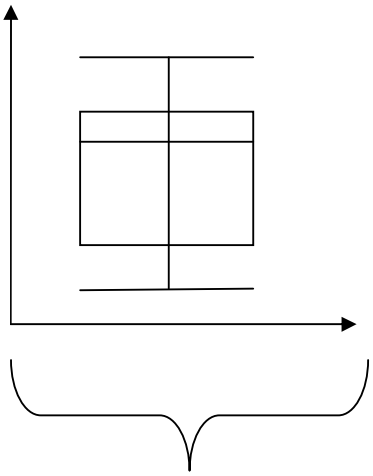
-mediana al centro;

-scatola tra il max e il min.

Possono verificarsi delle situazioni in cui non si ha la normale:

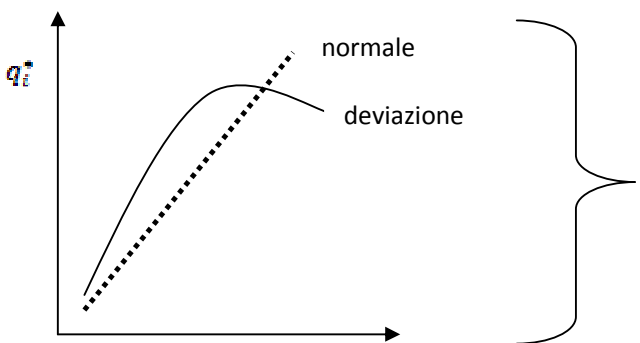


Mediana vicino al 1° quantile



Mediana vicino al 2° quantile

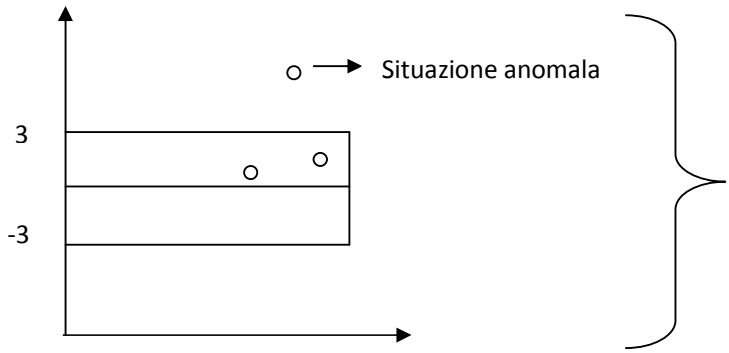
3. Modo:Q.Qplot:rappresentazione dei quantili della distribuzione di ε_i,con una percentuale tra 1% e 99%,e dei quantili teorici della normale(q_i,q_i^*):



Se c'è normalità,i punti si collocano su una retta.

Ogni volta che i punti deviano dalla retta,la condizione di normalità non è rispettata.

È possibile verificare che nel grafico ci siano delle osservazioni anomale,dette outelier;non sono altro che osservazioni che hanno una relazione diversa dagli altri punti. Questo è possibile attraverso un grafico:



Si ha un outlier quando un punto si trova fuori dalla fascia (-3,+3).

È impostante standardizzare. Residui standardizzati perdono l'unità di misura;il vantaggio è che per ogni data set si ha la stessa soglia.

È importante individuare outelier ,poiché possono esserci delle omissioni.