Lezione 1 - Stat model

5-03-2018

prof. Vittadini

Intro: Il corso comprende 3 argomenti: + modello lineare classico e estensione + modello lineare multivariato + applicazione di questo modello su dati gerarchici

Lezioni teoriche (con approccio pratico) e laboratori (SAS e R) + tutor (daniele.riggi@gmail.com).

Esame: 2 domande teoriche tra 15 preordinate (bastano le slide, possibilità di approfondire sui libri indicati). Esercizio da svolgere in classe o con SAS o con R.

Lezio: 1. Descrizione delle variabili e metriche. 2. Pulizia dei dati. 3. Statistiche descrittive: media, mediana, primo e terzo quantile, min e max, matrice di correlazione (eliminare variabili fortemente correlate, in quanto possono avere un'influenza negativa sul modello), scatter plot. 4. Costruzione di un modello statistico.

Quali sono gli indici che mi dicono se ho sviluppato un buon modello? + adattamento dei dati (fit dei dati) + semplicità del modello (numero dei parametri). Un modello serve per capire di più riguardo ad un fenomeno e non per complicarlo (parsimonia).

Qual è la differenza tra modello matematico e statistico? → Incertezza. Il modello statistico è caratterizzato dall'errore. L'errore considera la variazione individuale. → Variabili esplicative che possono essere considerate per migliorare il modello (quindi diminuire l'errore). → Nel caso stocastico errori nel rapporto campione-popolazione.

Il lavoro nel costruire un modello statistico sta sia nel diminuire l'errore ma anche nel capire da dove nasce.

Come nasce l'elaborazione di un modello statistico?

- Teoria, Ovvero formulazione di ipotesi, scoperta di relazion empiriche o rapporti di causa effetto tra variabili. Individuazione delle variabili esplicative.
- Dati. Capire quale metodo di raccolta utilizzare in base anche alla disponibilità economica che si ha per sviluppare il modello. Trattamenti preliminari (pulizia ecc.) e poi tornare al modello. Tenere conto dell'eterogeneità dei dati (es. considerando per esempio il livello di pericolosità delle acque di un lago, se valutiamo tutte le particelle nella loro totalità potremmo non concludere che le acque sono pericolose, questo potrebbe infatti risultare valido nella sua totalità ma magari identifichiamo delle zone in cui avvengono più morti rispetto alla normalità. Questo perché ci potrebbero essere delle zone maggiormente inquinate che non emergono da un'analisi totale

delle acque. Quindi considerare anche campionamenti di questo tipo , utilizzare tutti i dati potrebbe non dirci nulla). In questa fase rientra anche una prima analisi preliminare dei dati.
- Specificazione del modello (Probabilistico o descrittivo)
- Stima dei parametri e verifica dell'adattabilità ai dati
- Utilizzo

Ripetere più volte (se necessario).

Oss. Oggi un problema nella costruzione di un modello è anche la privacy. Ci sono modelli che potrebbero essere molto interessanti ma non si possono elaborare per problemi di privacy. Quindi devo usare il modello che ho per correggere i dati in questo senso (Teoria $\rightarrow$ Dati). Vale però anche il contrario, ovvero i Dati aiutano nell costruzione di un modello ($\Rightarrow$ Dati $\rightarrow$ Teoria)

Il modello di base è il *modello di regressione*. La regressione può essere *semplice*, *multipla* o *multivariata*. *Semplice*, se si ha una sola variabile dipendente ed una sola variabile esplicativa. *Multipla*, se si hanno più variabili esplicative e una sola dipendente. *Multivariata*, se si ha più di una variabile esplicativa e più di una variabile dipendente.

Stima, ovvero trovare i parametri per il modello. Uno dei metodi di stima è quello di *regressione lineare*.

Verifica dei risultati sia in termini descrittivi (adattamento ai dati), poi test statistici sulla significatività. Se la verifica non conduce ad un rifiuto del modello stimato allora lo si utilizza altrimenti si torna alla fase di specificazione.

Regressione multipla

Può essere espressa in termini matriciali. È costruita a partire dal vettore delle x e dal vettore delle y. La prima colonna è composta da 1, in quanto è quella che va a moltiplicarsi ai parametri da stimare e all'intercetta ignota b_{10}. La situazione è quindi la seguente:

RIVEDERE E COMPLETARE

dove appunto b è il vettore dei parametri ignoti da stimare, mentre ε il vettore degli errori casuali non osservabili. b_0 è l'intercetta, mentre gli altri termini del vettore b sono detti *coefficienti di regressione*. Gli $x_1 \dots x_n$ sono invece detti *regressori*.

L'errore ε è uno scalare che rappresenta tutti i fattori *rilevabili* e *non rilevabili* e può essere positivo o negativo. Non dipende dai valori dei regressori. Inoltre valgono le seguenti ipotesi:

1. Ipotesi di **omoschedasticità**, ovvero la variabilità dell'effetto di tutti i fattori non rilevati e/o non rilevabili non dipende dai valori dei regressori.

Quindi:

$$V(e|x_1, \ldots, x_n) = \sigma^2 \quad \Rightarrow V(Y|x_1, \ldots, x_n) = \sigma^2 \tag{1}$$

2. Ipotesi di **incorrelazione**, ovvero gli effetti su y dei fattori non rilevati ε per l'osservazione i non dipendono da quelli relativi all'osservazione j. Quindi:

$$Cov(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i, j \tag{2}$$

dove ε_i e ε_j sono appunto il valore delle variabili aleatorie per le osservazioni i e j.

3. Gli errori devono essere distribuiti in modo normale con media zero e varianza σ^2:

$$\varepsilon \sim N(0, \sigma^2) \tag{3}$$

A questo punto se abbiamo tante osservazioni la situazione diventa la seguente:

$$\underline{\underline{Y}} = \underline{\underline{X}} \cdot \underline{b} + \underline{\varepsilon} \tag{4}$$

con

$$Y = \begin{pmatrix} y_1 \\ . \\ . \\ . \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,k} \\ x_{2,1} & \ldots & \ldots & \\ \vdots & & \ddots & \vdots \\ x_{n,1} & \ldots & & x_{n,k} \end{pmatrix} \quad b = \begin{pmatrix} b_0 \\ . \\ . \\ . \\ b_n \end{pmatrix} \quad \epsilon = \begin{pmatrix} \varepsilon_0 \\ . \\ . \\ . \\ \varepsilon_n \end{pmatrix} \tag{5}$$

dove $\underline{b}$ rappresenta il vettore dei regressori, mentre $\underline{\varepsilon}$ quello degli errori per ciascuna osservazione.
Le condizioni ipotizzate prima sulle singole osservazioni del modello continuano a valere e possono essere riscritte in maniera più compatta come segue:

1. $\mathrm{E}(\underline{\varepsilon}) = 0$
2. $\mathrm{V}(\underline{\varepsilon}) = \mathrm{E}(\underline{\varepsilon}, \underline{\varepsilon}') = \underline{\underline{\Sigma}} = \sigma^2 \mathbb{1}_n$

La seconda condizione in particolare prende il nome di **ipotesi di sfericità degli errori** che include il fatto che gli errori sono omoschedastici e incorrelati tra loro (E indica il valore atteso).

Come metodo di stima dei parametri b si può utilizzare il *metodo dei minimi quadrati*.

Nell'ipotesi di perfetta dipendenza lineare tra Y e gli n regressori è possibile, facendo un campione di osservazioni, stimare i valori teorici previsti per la variabile dipendente y per tutte le unità del campione. Facendo la differenza tra questi valori teorici previsti e quelli empirici che risultano dall'osservazione si definiscono i residui come:

$$\underline{e} = \underline{y} - \underline{y'} = \underline{y} - \underline{\underline{X}} \cdot \underline{b} \tag{6}$$

dove con $\underline{y'}$ si è appunto indicato il vettore dei valori teorici previsti facendo una stima di $\underline{b}$ sul campione.

$$y'_i = b_0 + b_1 x_{1i} + b_2 \cdots + b_n x_{ni} \quad \text{per } i = 1, \ldots, k \tag{7}$$

Quindi il singolo residuo nella (6) si può anche riscrivere come:

$$e_i = y_i - y'_i = y_i - b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_n x_{ni} \quad \text{per } i = 1, \ldots, k \tag{8}$$

I valori di e_i sono k determinazioni campionarie (per i k campioni presi) del termine d'errore ε del modello.
Oss. È necessario fare campioni anche nel caso in cui abbiamo molti dati, in quanto se ho molti dati la regione di accettazione diventa piccolissima. Il metodo dei minimi quadrati ricerca il vettore di coefficienti $\underline{b}$ in modo da rendere minima la somma dei quadrati degli scarti tra ordinate empiriche e ordinate teoriche, o equivalentemente, la somma dei residui al quadrato:

$$\Phi(\underline{b}) = \sum_{i=1}^{k} (y_i - y'_i)^2 = \sum_{i=1}^{k} e_i^2 = \underline{e}^t \cdot \underline{e} = (\underline{y} - \underline{\underline{X}} \cdot \underline{b})^t \cdot (\underline{y} - \underline{\underline{X}} \cdot \underline{b}) \tag{9}$$

Sviluppando il calcolo si minimizza la funzione ponendo $= 0$ la derivata rispetto a b, ovvero:

$$\frac{\partial \Phi(\underline{b})}{\partial \underline{b}} = 0 \tag{10}$$

che porta alla seguente equazione, detta *equazione normale*:

$$\underline{\underline{X}}^t \underline{\underline{X}} \cdot \underline{b} = \underline{\underline{X}}^t \underline{y} \tag{11}$$

che corrisponde ad un sistema di $n + 1$ equazioni in $n + 1$ incognite. Nel caso in cui $k = 1$ si ha il modello di regressione lineare semplice. L'espressione tramite cui è stimato $\underline{b}$:

$$\underline{b} = (\underline{\underline{X}}^t X)^{-1} \underline{\underline{X}} \cdot \underline{y} \tag{12}$$

prende il nome di *stimatore dei minimi quadrati.*
Se le variabili sono standardizzate, ovvero divisi per lo scarto quadratico medio, lo stimatore dei minimi quadrati diventa:

$$\underline{b} = (\underline{\underline{X}}^{*t} \underline{\underline{X}}^*)^{-1} \underline{\underline{X}}^{*t} \cdot \underline{y}^* \tag{13}$$

dove $\underline{\underline{X}}^* = \underline{\underline{X}} \cdot \underline{\underline{D_x}}^{-1/2}$ con $\underline{\underline{D_x}}$ matrice i cui elementi diagonali sono le varianze delle variabili x (ottenendo in questo modo le x originali divise per il loro scarto quadratico medio), mentre y^* è y/σ_y.

Il problema si può anche vedere dal punto di vista geometrico. Possiamo infatti identificare il sottospazio lineare di $\mathbb{R}^N$ delle colonne di $\underline{\underline{X}}$, e in questo spazio la somma:

$$\sum_{i=1}^{k} (y_i - y'_i)^2 = \sum_{i=1}^{k} (y_i - \underline{\underline{X}} \cdot \underline{b})^2 \tag{14}$$

è il quadrato della distanza euclidea tra $\underline{y}$ e $\underline{\underline{X}} \cdot \underline{b}.$, ovvero:

$$\sum_{i=1}^{k}(y_i - \underline{\underline{X}} \cdot \underline{b})^2 = \|y_i - \underline{\underline{X}} \cdot \underline{b}\|^2 \tag{15}$$

Chiamiamo $\mu = \underline{\underline{X}} \cdot b$ il nostro vettore dei valori fittati (i valori teorici previsti) che corrisponde ad un vettore nello spazio delle colonne di $\underline{\underline{X}}$. Questo vettore μ rappresenta anche l'unica proiezione di $\underline{y}$ su $\underline{\underline{X}}$. La distanza tra $\underline{y}$ e μ è un vettore ortogonale (vettore dei residui) allo spazio $\underline{\underline{X}}$ e il metodo dei minimi quadrati ha lo scopo di minimizzare questo vettore. Se la dimensione dello spazio delle colonne è esattamente uguale al numero di variabili esplicative allora la soluzione è unica.

Figure 1: Rappresentazione dello spazio delle colonne e dei vettori di interesse.

La bontà di adattamento si stima in base a *devianza totale* e *devianza spiegata*.
La *devianza spiegata* è la somma delle differenze al quadrato tra i valori teorici della retta interpolante e la media dei valori empirici.
La *devianza residua* è la somma degli scarti al quadrato tra i valori osservati e teorici della y.
La *devianza totale* è la somma degli scarti dei valori di y empirici dalla loro media.

L'indice di adattamento è definito come:

$$R^2 = \frac{DevSpieg(Y)}{DevTot(Y)} \tag{16}$$

Nel caso di un modello di regressione lineare semplice si ha che:

$$DevSpieg(Y) = b_1 Codev(X, Y) \tag{17}$$

Dividendo per $n-1$ e con opportuni passaggi (si veda http://www2.stat.unibo.it/montanari/Didattica/lab3.pdf) si arriva a:

$$R^2 = \frac{Cov(X, Y)}{var(X)var(Y)} \tag{18}$$

R^2 è un numero che varia tra 0 e 1, è $= 0$ se non c'è correlazione lineare, $= 1$ se c'è perfetta correlazione.

Testare i risultati: t-test e F-measure (test di ipotesi)

Per testare i risultati ottenuti dei parametri si possono effettuare due misure di test sull'ipotesi nulla che il coefficiente stimato sia o meno uguale a zero, ovvero:

$$H_0 = \beta_i = 0 \tag{19}$$

Test di ipotesi e *p-value*

Un test di ipotesi è un procedimento tramite il quale si verifica la validità di una certa ipotesi. Solitamente si parte definendo un'*ipotesi nulla* tramite la quale si afferma che per la popolazione, o comunque più in generale per il fenomeno che si sta studiando, vale una determinata condizione, ovvero:

$$H_0 = \mu \tag{20}$$

dove μ sta a indicare una qualsiasi condizione che si vuole testare.
L'*ipotesi alternativa* invece specifica che cosa è vero nel caso in cui l'ipotesi nulla sia falsa. La più generale ipotesi alternativa è il contrario dell'ipotesi nulla ovvero:

$$H_1 \neq \mu \tag{21}$$

La valutazione dei risultati di un test di ipotesi avviene considerando il cosiddetto *p-value*. Per calcolare il p-value si calcola prima la distribuzione di probabilità per l'ipotesi nulla, assumendo quindi che essa sia vera si costruisce la distribuzione di probabilità centrata su questo valore, ed in seguito bisogna calcolare la probabilità di ottenere un valore che sia più grande del valore medio osservato per la quantità che si sta studiando. La distribuzione di probabilità che si costruisce infatti è una distribuzione delle medie, e il valore osservato è una media ottenuta da un campione.
Il p-value è quindi l'area sottesa alla parte di curva a destra del valore osservato. Vogliamo che la probabilità evidenziata sia la più piccola possibile, ovvero che il valore osservato sia il più possibile discostato dal centro della curva in quanto essa è centrata sull'ipotesi nulla. Se stabiliamo un livello di confidenza α ciò significa che la *probabilità* di ottenere un valore uguale o più grande del valore osservato deve essere minore o uguale a α. Quindi la parte di curva (probabilità) dentro le regioni delle code esterne individuate da α è uguale a $1 - \alpha$.

$$P(-Z_{\frac{\alpha}{2}} < \bar{a} < +Z_{\frac{\alpha}{2}}) = 1 - \alpha \tag{22}$$

Dove $-Z_{\frac{\alpha}{2}}$ e $+Z_{\frac{\alpha}{2}}$ rappresentano i valori di $\bar{a}$ tali per cui la probabilità racchiusa all'interno di questi valori è uguale a $1 - \alpha$. Se standardizziamo la distribuzione della quantità $\bar{a}$ in modo da renderla a media nulla e varianza 1 (quindi sottraiamo la media e dividiamo per la deviazione standard), otteniamo:

$$P\left(-Z'_{\frac{\alpha}{2}} < \frac{\bar{a} - \mu_a(H_0)}{\sigma} < +Z'_{\frac{\alpha}{2}}\right) = 1 - \alpha \tag{23}$$

che implica che $\bar{a}$ deve trovarsi entro i seguenti valori in termini di deviazione standard rispetto alla media:

$$P\left(\mu_a(H_0) - Z'_{\frac{\alpha}{2}} \cdot \sigma < \bar{a} < \mu_a(H_0) + Z'_{\frac{\alpha}{2}} \cdot \sigma\right) = 1 - \alpha \tag{24}$$

I valori di Z' sono tabulati dalla funzione degli errori.
La probabilità (p-value) per un valore osservato di $\bar{a}$ è:

$$p = P\left(\left|\frac{\bar{a} - \mu_a(H_0)}{\sigma}\right| > \left|\frac{\bar{a_{oss}} - \mu_a(H_0)}{\sigma}\right|\right) \tag{25}$$

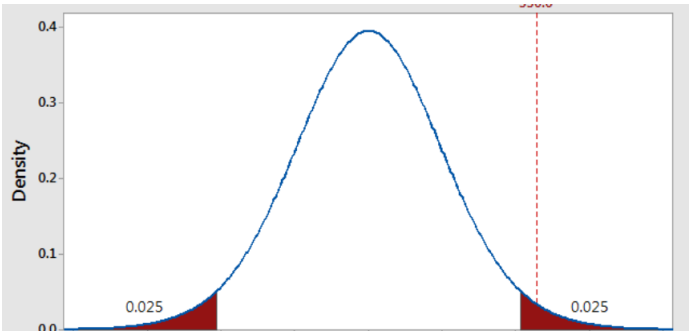

Figure 2: La curva rappresenta la distribuzione centrata sull'ipotesi nulla, le parti colorate di rosso rappresenta invece il livello α in questo caso uguale al 5%, quindi, essendo la curva simmetrica sarà $0,025$ a destra, e $0,025$ a sinistra. La linea tratteggiata rappresenta invece il valore osservato che in questo caso cade all'interno dell'area accettata per rifiutare l'ipotesi nulla. credits: http://blog.minitab.com/blog/adventures-in-statistics-2/understanding-hypothesis-tests:-significance-levels-alpha-and-p-values-in-statistics

che rappresenta quindi la probabilità di osservare un valore di $\bar{a}$ maggiore o uguale a quello che si è osservato (supponendo valida l'ipotesi nulla). Se questo valore di probabilità è inferiore al livello α fissato (quindi il valore $\bar{a_{oss}}$ è al di fuori dell'intervallo individuato in termini di Z' intervalli di confidenza) si può rigettare l'ipotesi nulla.

Il livello di confidenza α

A livello interpretativo il livello di confidenza α che si specifica (solitamente è fissato a $0,05$, ovvero al 5% totale della distribuzione) indica la probabilità di rigettare l'ipotesi nulla nel caso in cui poi essa sia effettivamente vera, ovvero le probabilità di errore. Nel valutare l'ipotesi nulla infatti possiamo incappare in due tipi di errori: ritenerla vera quando invece in realtà è falsa, e ritenerla falsa quando invece in realtà è vera. Specificando il livello α stiamo specificando la probabilità di commettere il secondo tipo di errore, infatti se diciamo che rifiutiamo l'ipotesi nulla se il valore che osserviamo cade nelle code, ovvero nella regione α della curva, stiamo comunque rifiutando tutti quei valori della distribuzione per cui l'ipotesi nulla è comunque vera, anche se poco probabile. Quindi nel caso in cui poi l'ipotesi nulla sia effettivamente vera, al massimo in α ($e.g 5\%$) dei casi concluderemo che è falsa, in quanto troveremo i valori nelle code per cui abbiamo deciso si rifiutare l'ipotesi, mentre nel rimanente 95% concluderemo correttamente che è vera. Se invece è effettivamente falsa queste probabilità non valgono più.

t-test

La *t-statistic*, detta anche *t-measure* o *t-test*, rappresenta un modo per valutare se la stima di una quantità risulta accettabile o meno rispetto ad un'ipotesi nulla. È definita come segue:

$$t = \frac{a - a_0}{SE(a)} \tag{26}$$

dove a_0 è il valore dell'ipotesi nulla che si sta testando per la quantità a e SE è lo *standard error* di questa variabile ovvero: $\frac{\sigma}{\sqrt{n}}$.
Quindi nel caso della stima dei parametri β della regressione lineare semplice, in cui per l'ipotesi nulla si suppone che β abbia valore zero, si ha:

$$t = \frac{\beta_i}{SE(\beta_i)} = \frac{\beta_i}{\frac{\sigma}{\sqrt{n\sigma_{jj}}}} \tag{27}$$

se la σ è nota, altrimenti si usa il suo stimatore s, quindi:

$$t = \frac{\beta_i}{\frac{s}{\sqrt{n\sigma_{jj}}}} \tag{28}$$

Se si fissa quindi un livello di confidenza α per il p-value, per esempio $\alpha = 0,05$, si ottiene che l'ipotesi nulla deve essere rigettata se $|t| > 1,96$.
Si può quindi riscrivere il p-value in termini della statistica t:

$$p = P(|t| > t_{oss}) \tag{29}$$

Se la quantità a è distribuita normalmente allora t è distribuito come un χ^2 a $n - 1$ gradi di libertà che tende ad una distribuzione normale per grandi n (si veda Stock p.87).

Intervallo di confidenza

Un altro metodo per valutare la validità dell'ipotesi nulla è calcolare l'*intervallo di confidenza* per il valore che si osserva.
Per costruire l'intervallo di confidenza ci si centra sul valore medio osservato e si costruisce su questo una distribuzione campionaria. Secondo l'espressione del t-value, l'ipotesi nulla è rifiutata, secondo un determinato livello di confidenza α, se è distante dal valore medio osservato più di t deviazioni standard. Nel caso di $\alpha = 5\%$, si avrà che l'ipotesi nulla non è rifiutata se $-1,96 \cdot SE(\bar{a}) < \bar{a} - \bar{a}_0 < +1,96 \cdot SE(\bar{a})$, ovvero se il valore dell'ipotesi nulla è contenuta all'interno dell'intervallo $[\bar{a} - 1,96 \cdot SE(\bar{a}), \bar{a} + 1,96 \cdot SE(\bar{a})]$ che rappresenta il 95% dei valori della distribuzione campionaria centrata su $\bar{a}$, in quanto è appunto la regione di curva compresa entro 1.96 deviazioni standard.

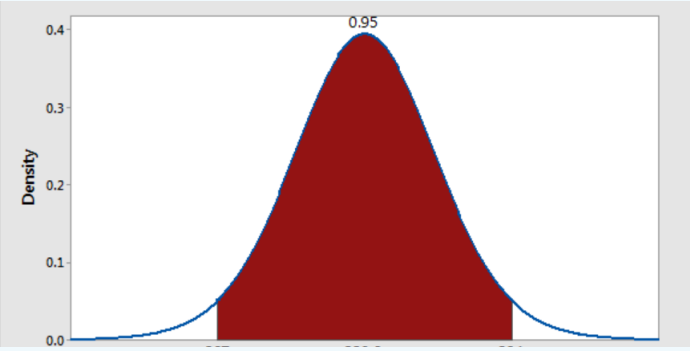

Figure 3: credits: http://blog.minitab.com/blog/adventures-in-statistics-2/understanding-hypothesis-tests%3A-confidence-intervals-and-confidence-levels
A questo punto se l'ipotesi nulla è effettivamente vera nel 95% dei casi sarà accettata, mentre solamente nel 5% dei casi sarà rifiutata, cioè per il 95% degli intervalli calcolati per diversi valori osservati sarà accettata, mentre per il rimanente 5% degli intervalli non sarà accettata . Questi intervalli di confidenza sono quelli centrati su valori osservati che sono nella regione del 5% delle code per la distribuzione centrata sull'ipotesi nulla, ovvero quei valori nelle regioni rosse in figura 2.

F measure

Nel caso sia condotta una regressione con più regressori si può effettuare un test di ipotesi congiunto per i vari parametri che vengono stimati, ovvero vedere se un determinato set di parametri è efficiente o meno nella stima del modello.

Quindi è definita la seguente ipotesi nulla:

$$\begin{aligned} H_0 : \ & \beta_0 = 0, \ \beta_1 = 0, \ \cdot, \ \beta_k = 0 \\ : \ & \beta_0 = \beta_1 = \cdot = \beta_k = 0 \end{aligned} \tag{30}$$

che implica l'ipotesi alternativa:

$$H_1 : \ \beta_0 \neq 0, \ \beta_1 \neq 0, \ \cdot, \ \beta_k \neq 0 \tag{31}$$

ovvero si testa se uno tra i parametri stimati sia nullo, se si rifiuta l'ipotesi nulla significa che almeno uno di essi non è nullo, e quindi è significativo.

Il test di ipotesi congiunta si effettua calcolando la F-measure per il modello lineare preso in esame, definita come segue:

$$F = \frac{SSR_r - SSR_{ur}/q}{SSR_{ur}/(n-(k+1))} \tag{32}$$

dove SSR rappresenta la somma dei residui al quadrato del modello, cioè la devianza residua. In particolare SSR_r rappresenta la devianza residua per il modello ristretto all'ipotesi nulla, ovvero, supponendo vera l'ipotesi nulla, SSR_r rappresenta la devianza residua del modello in cui i parametri sono posti uguale ai valori specificati dall'ipotesi, in questo caso sono posti uguale a zero. SSR_{ur} rappresenta invece la devianza residua per il modello non ristretto, ovvero quello stimato con tutti i parametri. La variabile q rappresenta invece il numero di restrizioni, ovvero il numero di parametri che sono testati congiuntamente, n rappresenta il numero di osservazioni e k il numero di variabili indipendenti nel modello non ristretto. Le due quantità rapportate nell'equazione (32) sono distribuite come un χ^2 che implica che la F sia distribuita come una F di Fisher-Snedecor con q e $n-(k+1)$ gradi di libertà. Possiamo quindi impostare un livello di significatività α con il quale rifiutare l'ipotesi nulla. L'ipotesi nulla è accettata se:

$$P(F_0 < F_\alpha) = 1 - \alpha \tag{33}$$

ovvero se si ottiene un valore per il test F_0 minore del valore F_α, valore per cui la probabilità è uguale a $1-\alpha$. Questo valore si può trovare tabulato. Se invece si trova un valore maggiore, tale per cui la probabilità di ottenere un valore maggiore o uguale è uguale o minore di α, allora si può rigettare l'ipotesi nulla.

Stima dei parametri tramite massima verosimiglianza

Nel modello lineare classico gli errori si distribuiscono, come si è detto, in modo gaussiano, ovvero:

$$\epsilon_i \sim N(0, \sigma^2) \tag{34}$$

così come anche i parametri e le variabili dipendenti, cioè:

$$\begin{aligned} b_{OLS} &\sim N(\beta, \sigma^2 (X^t X)^{-1}) \\ Y &\sim N(X\beta, \sigma^2 I_n) \end{aligned} \tag{35}$$

Si può quindi provare a utilizzare stime di massima verosimiglianza che chiedono l'esistenza di n variabili casuali $y_i - X_i\beta$ con $i = 1 \cdots n$ che siano identicamente e indipendentemente distribuite condizionatamente al valore X e dipendenti dai parametri β.
Si cerca quindi il valore dei parametri che rendono massima la probabilità di ottenere la verosimiglianza massima:

$$max_\beta L(y_i - X_i\beta) \quad i = 1 \cdots n \tag{36}$$

Da questo si arriva a dimostrare che lo stimatore dei β di massima verosimiglianza (ML) coincide con quello trovato in precedenza, e possiede le stesse proprietà di consistenza, correttezza ed efficienza, comprese le proprietà asintotiche richieste nel caso in cui le misure siano statisticamente indipendenti.

Variabili esplicative qualitative

Le variabili esplicative qualitative o categoriche sono determinate da attributi:

- nominali
- ordinali

e possono essere:

1. dicotomiche (*dummy variables*) se assumono solamente due valori (e.g. sesso);
2. Politomiche, se assumono più di due valori.

Ora se abbiamo un modello lineare del tipo:

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i \quad i = 1 \cdots n \tag{37}$$

in cui la D_i è una variabili dummy, si ha che essa va ad esercitare il proprio effetto sull'intercetta. Solitamente si riconducono le due possibilità per la variabile dummy ai due valori 0 e 1, per cui quando si ha $D = 0$ si ottiene:

$$Y_i = \beta_0 + \epsilon_i \tag{38}$$

mentre quando $D = 1$, si ottiene:

$$Y_i = \beta_0 + \beta_1 + \epsilon \tag{39}$$

ovvero l'intercetta rappresenta il *valore stimato* di Y quando la variabile esplicativa dummy è uguale a 0. Il coefficiente angolare è dato invece dalla differenza in Y per i due diversi valori della variabile esplicativa dummy, ovvero facendo la differenza tra la (38) e (39). La statistica inferenziale si fa anche in questo caso come prima, utilizzando stime e test.

Regressione logistica

[ref: https://codesachin.wordpress.com/2015/08/16/logistic-regression-for-dummies/ , https://www.youtube.com/watch?v=gNhogKJ_q7U]
La regressione logistica a differenza di altri tipi di regressione non si pone l'obiettivo di predire il valore di una variabile dato un certo numero di input. La regressione logistica si applica infatti a variabili categoriche. La regressione logistica restituisce in output la *probabilità* che un certo punto in input appartenga ad una certa classe. Assumiamo per semplicità di avere solamente due classi (classe binaria) in cui la variabile dipendente y può assumere valori. A differenza della regressione lineare classica quindi la variabile dipendente può assumere solamente un numero finito di valori e non uno spettro continuo. Possiamo comunque provare ad applicare una regressione lineare ad un problema che coinvolge variabili questo tipo, infatti a parte essere una variabile binaria, non c'è nulla di speciale nella variabile y che vogliamo prevedere, i problemi sorgono quando si devono poi interpretare i risultati restituiti da questo modello. Applichiamo quindi un modello lineare utilizzando tutte le variabili indipendenti $x_1, \cdots, x_n$ che si ritengono significative per la previsione di un'uscita della variabile y. Sia questa uscita $y = 1$, più il risultato restituito dal modello, per determinati valori delle variabili indipendenti, è alto, più sarà probabile che il valore per la variabile dipendente sia effettivamente uguale a 1. Per esempio possiamo considerare il caso di superare o meno un esame, la possibilità di successo la codifichiamo con il valore $y = 1$, mentre quella di insuccesso con il valore $y = 0$. Per convenzione solitamente si predice la variabile che viene indicata con 1, quindi in questo caso la possibilità di successo, di conseguenza inseriamo nel modello tutte quelle variabili che riteniamo significative per questa predizione. Supponendo di avere tra le variabili solamente il tempo di studio il modello lineare avrà la seguente forma:

$$y = \beta_0 + \beta_1 \cdot \textit{tempo di studio} \tag{40}$$

dai risultati di questa predizione otterremo dei valori per i coefficienti che vengono stimati tramite maximum likelihood (per approfondire si veda https://onlinecourses.science.psu.edu/stat504/node/150) ma cosa significa il risultato di questa regressione. Supponiamo per esempio di ottenere come risultati $\beta_0 = -1$ e $\beta_1 = 2$, se consideriamo un tempo di studio pari a 2 ore e lo inseriamo all'interno dell'equazione otteniamo per la variabile y, che in questo caso rappresenta la possibilità di superare l'esame, un valore pari a 1, ma cosa indica? Così com'è per ora non indica nulla, infatti se volessimo interpretarlo come la probabilità di superare l'esame sbaglieremmo in quanto quest'ultima deve sempre risiedere tra 0 e 1, mentre in questo caso se per esempio consideriamo un tempo di studio di 5 ore per esempio otteniamo un valore negativo, e se scegliamo il caso in cui il tempo di studio sia nullo si ottiene per y un risultato negativo. Per poter passare dai risultati della regressione lineare a delle probabilità P è necessario applicare una trasformazione non lineare che:

1. Assuma sempre valori positivi
2. Sia compresa tra 0 e 1

Per soddisfare il primo punto si applica una funzione esponenziale all'espressione della regressione lineare quindi dalla (40), che più in generale può essere riscritta come $y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots$, si passa a:

$$P = \exp(\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots) \tag{41}$$

Ora è necessario che P sia compresa risulti minore di 1, in quanto con il passaggio precedente ci assicuriamo già che assuma solo valori positivi. Per fare ciò ci basta divider l'espressione precedentemente ricavata per un numero che sia leggermente più grande, di modo da ottenere così una funzione che tenda asintoticamente a 1 e a 0. Per fare questo dividiamo per la stessa quantità a cui sommiamo 1, ottenendo così l'espressione per la probabilità:

$$P = \frac{\exp(y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots)}{\exp(y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots) + 1} \tag{42}$$

L'interpretazione del modello lineare si può però ancora mantenere in quanto:

$$\log(\frac{P}{1-P}) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots \tag{43}$$

dove $1 - P$ oltre a permettere il passaggio per ottenere l'espressione lineare, dal punto di vista interpretativo rappresenta la probabilità di ottenere $y = 0$. A volta ci si riferisce al rapporto $\frac{P}{1-P} = \frac{P(y=1)}{P(y=0)}$ con il termine di *Odds ratio*. Quando viene applicata la regressione logistica viene applicata la formula (43), detta anche *logit*, per cui per ottenere i valori di probabilità associati a determinati valori delle variabili indipendenti che si sono utilizzate è necessario applicare la formula inversa. nei software è spesso anche possibile impostare una soglia per la probabilità con la qual stabilire e quindi predire se una certa osservazione con una certa probabilità appartenga ad una determinata classe. In questo modo si possono poi calcolare delle contingency table dalle quali vedere quanti valori sono stati correttamente predetti e quanti invece no.

Violazioni del modello lineare classico

Eteroschedasticità

L'ipotesi di omoschedasticità suppone che il termine di errore sia uguale per tutte le variabili indipendenti, quindi dato un certo valore di X lo spread nella distribuzione delle Y è sempre lo stesso, come si può vedere dalla figura **??**. ovvero $Var(\varepsilon|x_i) = \sigma^2$ e $E(y|x) = 0$. Se effettuiamo un plot dei residui in funzione di x si ottiene un grafico di questo tipo:

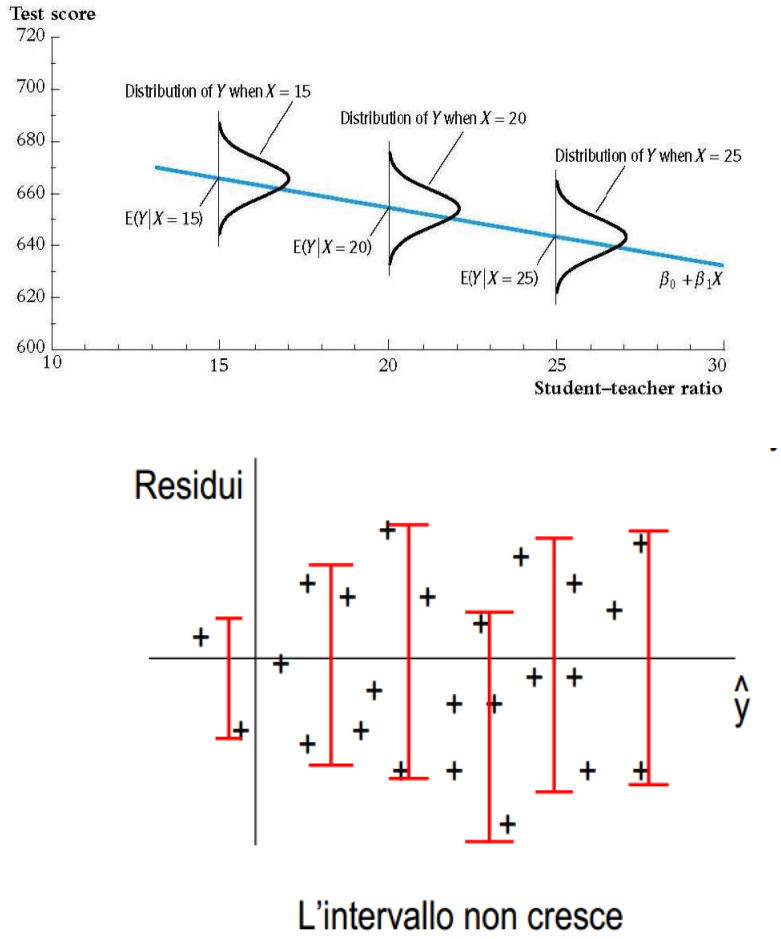

Figure 4: L'ampiezza intervallare dei residui non aumenta all'aumentare di x

Al contrario ci possono essere situazioni in cui il termine di errore varia tra le diverse variabili indipendenti, ottenendo così una situazione del genere: Ovvero lo scatter plot dei residui nel caso di errori eteroschedastici, in funzione dei valori predetti per la variabile dipendente è caratterizzato da un tipico andamento a cono. Questo andamento è conseguenza del fatto che la varianza del termine di errore, stimato dai residui del modello, dato un certo alore di X non è più costane e in particolare ora dipende da X, come in figura 6.

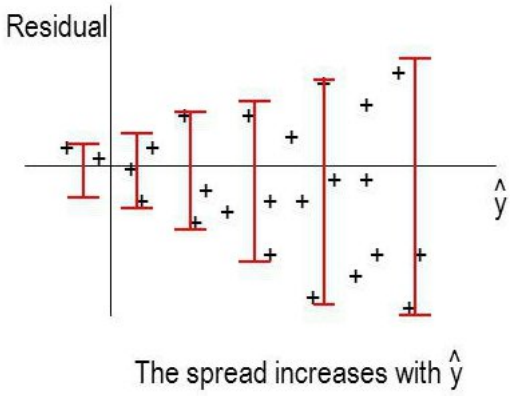

Figure 5: Da questo grafico si vede come l'ampiezza dell'intervallo residuale aumenta all'aumentare di y.

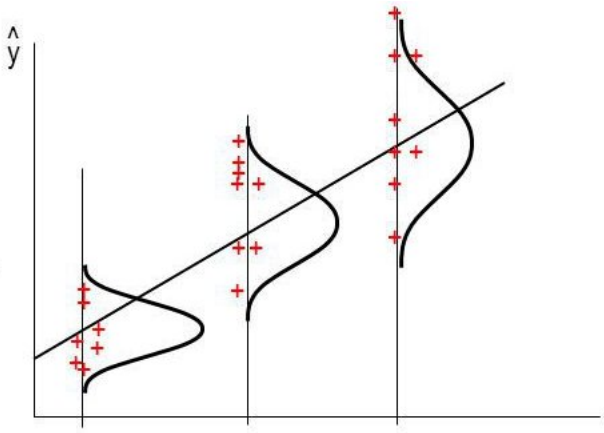

Figure 6: In figura si può osservare come la varianza della distribuzione della Y abbia una varianza sempre maggiore all'aumentare del valore di x, quindi l'errore è eteroschedastico. Il concetto chiave comunque per l'errore eteroschedastico è che la sua varianza cambia in base al valore delle variabile indipendente.

Un altro andamento anomalo del tipo in figura 5 lo si può osservare in un semplice grafico scatter plot della variabile dipendente in funzione della variabile indipendente: nel caso di errori omoschedastici i punti sono collocati in modo

equidistante dalla retta interpolante, mentre nel caso di errori eteroschedastici i punti sono distanziati in maniera diversa da questa. Il problema nel caso eteroschedastico si pone in quanto il metodo OLS dei minimi quadrati ordinari mira a minimizzare i residui ottenendo lo standard error minimo. Il metodo OLS pesa però tutte le osservazioni allo stesso modo, mentre quando si trattano errori eteroschedastici è necessario pesare meno i valori con più errore e pesare di più quelli che invece sono più rilevanti.
Se utilizziamo ancora stimatori OLS cosa succede?
→ Gli stimatori OLS dei parametri sono ancora unbiased, consistenti e distribuiti in modo asintoticamente normale. *Non* sono più stimatori *efficienti* tra tutti gli stimatori possibili dei parametri che sono lineari e unbiased in Y, dato un certo valore di X. In generale quindi possiamo dire che non sono più BLUE (Best Linear Unbiased Estimator).

La statistica t di Student calcolata in base al valore della deviazione standard utilizzato nel metodo OLS, non risulta più distribuita in modo normale nemmeno per grandi campioni se l'errore è eteroschedastico. Questo avviene principalmente per due ragioni:

1. Le stime campionare tendono a sottostimare il valore della varianza.
2. Non c'è più da calcolare una sola varianza ma diverse varianze.

Come conseguenza della sottostima della varianza si ha che la statistica t di Student ha valori erroneamente elevati, si possono considerare come significativi paramentri che in realtà non lo sono. Per lo stesso motivo la regione di accettazione diventa molto più piccola di quanto non lo sia in realtà e di conseguenza la regione di rifiuto molto grande.

L'ipotesi di omoschedasticità (uguali varianze) è infatti alla base di test come l'analisi della varianza ANOVA (Analysis Of Variance) e il t-test di Student.

ANOVA

L'analisi della varianza ANOVA è utilizzata per testare differenze tra medie, utilizzando appunto le varianze. Quando le medie sono solamente due è indifferente utilizzare questo test oppure il t-test, mentre si deve usare necessariamente il test ANOVA quado le medie da testare sono più di due.
Dati quindi un insieme di campioni di cui sono stati calcolati media e varianza per ciascuno si può procedere a costruire il test ANOVA, come segue:

$$F = \frac{\sigma^2_{between}}{\sigma^2_{within}} \tag{44}$$

dove $\sigma^2_{between}$ rappresenta la varianza tra gruppi, mentre σ^2_{within} quella in gruppi.
La varianza *within* in gruppi è la media delle varianze di ciascun campione pesata sul numero di gradi di libertà del campione, overo:

$$\sigma^2_{within} = \frac{1}{a(n-1)} \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_i)^2 \tag{45}$$

in cui appunto la $i = 1 \cdots a$ identifica il campione e $j = 1 \cdots n$ identifica invece il numero di osservazione che si sta prendendo in considerazione. Quindi si somma prima su j per calcolare le varianze del campione che si sta considerando, e poi si somma su j per fare la media delle varianze dei campioni, dividendo poi per il peso del campione pari a $n-1$, dove n è il numero delle osservazioni fatte per campione.
La varianza *between* tra gruppi invece è calcolata a partire dalla devianza totale che è la varianza stimata con tutte le osservazioni di tutti i campioni ovvero come se le varie osservazioni dei vari campioni appartenessero tutte ad un unico campione. A questo punto la varianza tra gruppi si ottiene moltiplicando questa per n, ovvero:

$$\sigma^2_{between} = \frac{n}{a-1} \sum_{i=1}^{a} (\bar{y}_i - \bar{\bar{y}}) \tag{46}$$

dove $\bar{\bar{y}}$ rappresenta appunto la media ottenuta considerando le osservazioni di tutti i campioni, mentre le $\bar{y}_i$ sono le medie dei singoli campioni.
Si può quindi a questo punto effettuare il test in (44). Poichè le due varianze rapportate sono stime di una stessa varianza parametrica (quella della distribuzione vera se si conoscesse tutta la popolazione), allora questo rapporto deve essere uguale a 1 in teoria. Se però i campioni provengono da popolazioni diverse si ottiene un valore al numeratore più grande rispetto al denominatore, risultando così in un numero maggiore di 1.
Per ogni combinazione di gradi di libertà di numeratore e denominatore, si confronta il valore ottenuto con la distribuzione di una variabile casuale distribuita come una F di Sendecor con pari gradi di libertà. Stabilendo il livello di confidenza, si confronta con il valore tabulato della F e si decide se confermare l'ipotesi nulla, cioè che le medie provengono tutte da una stessa distribuzione, oppure se rigettarla, affermando quindi che almeno una non appartiene alla distribuzione delle altre.
[ref:http://docenti.unimc.it/monica.raiteri/teaching/2013/12316/files/slides-i-parte-per-studenti-frequentanti/interpretazione-del-test-f-distribuzione-f-di]

Oltre alla visualizzazione grafiche, l'eteroschedasticità si può individuare anche tramite metodi analitici, ovvero eseguendo dei test come il *test di White* o il *test di Breuch-Pagan*.

Test di White

Il test di White testa l'ipotesi nulla di omoschedasticità degli errori:

$$H_0 = Var(\epsilon|X) = \sigma^2 \cdot I_n \tag{47}$$

e ovviamente ha come ipotesi alternativa la stessa espressione sopra in cui vale però una disuguaglianza. Il test di White si basa su na regressione OLS dei residui, ovvero:

1. si stima il modello lineare con il metodo OLS ottenendo:

$$\hat{Y_i} = \hat{\beta_0} + \hat{\beta_1} \cdot X_{i1} + \cdots + \beta_n \cdot X_{in} + \hat{\epsilon_i} \tag{48}$$

2. si fa una regressione OLS sugli errori assumendo che l'eteroschedasticità possa essere una funzione lineare dei regressori, del loro quadrato o della loro interazione $x_{ij} \cdot x_{ij}$. Quindi si ottiene:

$$\hat{\epsilon}_i^2 = \delta_0 + \delta_1 \hat{Y}_i + \delta_2 \hat{Y}_i^2 \tag{49}$$

   dove appunto si utilizza il risultato della regressione del modello lineare effettuato all'inizio. Considerando i quadrati si considerano tramite i doppi prodotti anche i termini di interazione.

3. si calcola $R^2_{\hat{\epsilon}_i^2}$

4. si effettua il test LM, definito come segue:

$$LM = nR^2_{\hat{\epsilon}_i^2} \tag{50}$$

   che si distribuisce come un χ^2 con un numero di gradi di libertà pari al numero di regressori inseriti nel modello.

5. scelto un livello di significatività α l'ipotesi nulla sarà rigettata se il test LM risulta superiore al valore soglia di χ^2 (tabulato), che è associato al livello di significatività scelto.

Test di Breusch-Pagan

Il test di Breuch Pagan a differenza del test di White ipotizza che l'eteroschedasticità sia solamente una funzione lineare delle variabili indipendenti, trascurando quindi i termini quadratici e di interazione. Si procede

quindi nello stesso modo definito precedentemente in cui però si assume che la forma funzionale per ϵ sia:

$$\hat{\epsilon}_i^2 = \delta_0 + \delta_1 \hat{Y}_i \tag{51}$$

CHIEDERE..

Autocorrelazione

A volte è possibile che gli errori (e quindi i residui) siano correlati tra loro, soprattutto in serie storiche o territoriali è ragionevole ipotizzare che ci sia correlazione tra gli errori che vengono stimati in momenti successivi o territori vicini.
Nel modello lineare classico si suppone che:

$$Cov(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i \neq j \tag{52}$$

Quando ciò non si verifica si dice che gli erroi sono correlati o che si è in presenza di una correlazione seriale. Se c'è omoschedasticità ma c'è correlazione, la matrice degli errori è:

$$\Sigma = \begin{pmatrix} \sigma^2 & \rho_{1,2} & \cdots & \rho_{1,n} \\ \rho_{1,2} & \sigma^2 & \cdots & \vdots \\ \vdots & & \ddots & \vdots \\ \rho_{1,n} & \cdots & \cdots & \sigma^2 \end{pmatrix} \tag{53}$$

dove i vari ρ rappresentano i termini di correlazione dei vari termini di errore tra loro. La matrice ovviamente risulta simmetrica.

L'errore può essere correlato con quello dell'osservazione immediatamente precedente e quindi avere una correlazione di primo livello, oppure ci può essere una correlazione di secondo livello o livello maggiore se gli errori correlati sono distanti due o più osservazioni. Ciò è espresso con la seguente formula:

$$\varepsilon'_i = \varepsilon'_{i-1} + \eta_j \tag{54}$$

dove si è utilizzata la notazione primata per identificare il fatto che gli errori sono correlati tra loro e non sono più sferici. Gli η_j sono invece identicamente e indipendentemente distribuiti in modo normale con $N(0, \sigma_i)$ per rappresentare quindi la parte di correlazione dell'errore con sè stesso (la diagonale della matrice).
Si ha autocorrelazione *positiva* quando residui consecutivi tendono ad essere dello stesso segno e simili in valore, *negativa* quando invece residui consecutivi sono di segno differente.

IMMAGINI AUTOCORRELAIZONE POSITIVA E NEGATIVA

In caso di autocorrelazione gli stimatori OLS dei parametri per la regressione lineare sono ancora lineari e corretti (unbiased) ma non sono più i migliori stimatori possibili, quindi non sono più BLUE, ovvero esistono altri stimatori

che risultano più efficienti. come conseguenza di ciò non si potrà più usare la varianza campionaria nella statistica t perchè non potrà più approssimare la varianza vera in quanto la t è costruita supponendo l'incorrelazione degli errori nella popolazione e il valore atteso della varianza campionaria non stima correttamente la varianza vera. Di conseguenza la statistica t assume valori erroneamente elevati, considerando significativi parametri quando in realtà non lo sono, ovvero si amplia la regione di rifiuto per l'ipotesi nulla con il conseguente restringimento della regione di accettazione.
Ragionamenti analoghi si possono fare per il test F, che corrisponde semplicemente al quadrato del t test.

Individuazione grafica

Si può notare un autocorrelazione nei reisui osservando:

1. scatter plot della vraibile dipendente in funzione in funzione del regressore x. Nel caso in cui ci fossero più regressori bisogna fare uno scatter plot in funzione di goni regressore. Se si nota una certa regolarità nell'andamento allora si è in presenza di correlazione.
2. scatter plot dei residui in funzione del regressore: anche qui valgono le stesse considerazioni fatte sopra. Se i residui oscillano intorno allo zero non c'è correlazione.
3. Residui in funzione dei residui ritardati
4. Correlogramma, permette di identificare chiaramente quali sono i gradi di correlazione che influiscono di più. Solitamente nei correlogramma è mostrata anche una banda di confidenza che indica il limite entro il quale non si ritiene che vi sia autocorrelazione. Se il coefficiente di autocorrelazione che si sta valutando esce da questa banda allora si è in presenza di autocorrelazione.

Test di Durbin-Watson

Il test di Durbin-Watson verifica l'ipotesi nulla:

$$H_0 : \ \rho = Corr(\varepsilon'_i, \varepsilon'_{i-1}) = 0 \tag{55}$$

dove ε'_i e ε'_{i-1} sono i residui relativi all'osservazione i-esima e i-1-esima.
La statistica di DW con cui si effettua il test è la seguente:

$$DW = \frac{\sum_i (\varepsilon'_i - \varepsilon'_{i-1})^2}{\sum_i \varepsilon'^2_i} \quad \text{per } i = 1 \dots n \tag{56}$$

La statistica di DW è centrata su 2 ed è sempre compresa tra 0 e 4. Nel caso in cui i residui siano correlati positivamente tende a 0, mentre nel caso in cui siano

correlati negativamente tende a 4.
Non si conosce la distribuzione teorica di questa statistica, comunque esistono dei valori tabulati in base al numero di regressori, il numero di osservazioni e livello di significatività con cui si vuole valutare l'ipotesi nulla, con in quali è possibile individuare dei valori critici d_l e d_u che delimitino le regioni di rifiuto e di accettazione. Se il valore che si trova dalla statistica in (56) è $d < d_l$ allora si può concludere che c'è autocorrelazione positiva, se $d > d_u$ allora si può concludere che c'è autocorrelazione negativa, mentre se d è compreso tra questi due valori allora non c'è sufficiente evidenza per concludere che ci sia autocorrelazione tra i residui. Convenzionalmente, nel caso in cui i valori di d_l e d_u non vengano specificati si fissa $d_l = 1$ e $d_u = 3$.

WLS e GLS