Chapter 1

Introduzione

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.

Chapter 2

Modello Lineare Classico

Un generico modello di regressione esprime una relazione del tipo:

$$y = f(x_1, \ldots, x_k) + \varepsilon \tag{2.1}$$

dove la y è la variabile dipendente che si cerca di spiegare in termini delle variabili indipendenti $x_1, \ldots, x_k$ dette anche variabili *esplicative*, *regressori* o *covariate*.ε rappresenta invece la componente di errore circa la relazione che lega la y alle covariate x, che può essere dovuta ad errori individuali, errori di misurazione o omissione di variabili esplicative che potrebbero ulteriormente cogliere la relazione che sussiste tra y e x.

Esistono diversi modelli di regressione, in questa dispensa in particolare viene indagato il metodo di regressione lineare che può essere:

- **semplice**, se si utilizza una sola variabile esplicativa, quindi:

  $$y_i = \beta_0 + \beta_1 x_1 + \varepsilon \tag{2.2}$$

  dove si è utilizzato il pedice i per la y per specificare la specifica relazione tra l'osservazione i del campione o della popolazione che si sta studiando e la variabile x_1 che in questo caso rappresenta il valore specifico che la covariata x_1 assume per l'osservazione i.

- **multiplo**, se si inserisce più di una variabile esplicativa, quindi:

  $$y_i = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k + \varepsilon \tag{2.3}$$

- **multivariato**, se si ha più di una variabile dipendente e più di una variabile esplicativa, quindi:

  $$\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_n} = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k + \varepsilon \tag{2.4}$$

Senza considerare per ora il caso multivariato che sarà analizzato in un prossimo capitolo, il modo più generale per esprimere la relazione tra una variabile dipendente e le variabili di regressione è la seguente:

$$Y = X \cdot \underline{\beta} + \underline{\varepsilon} \tag{2.5}$$

dove con Y si è indicato il vettore per tutte le osservazioni i in studio, e con X la matrice che contiene i valori delle covariate per ogni osservazione i del modello (più il termine noto).

$$Y = \begin{pmatrix} y_1 \\ . \\ . \\ . \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_{1,2} & \dots & x_{1,k} \\ 1 & x_{2,2} & \dots & \vdots \\ \vdots & \dots & \ddots & \vdots \\ 1 & \dots & & x_{n,k} \end{pmatrix} \quad \underline{\beta} = \begin{pmatrix} b_0 \\ . \\ . \\ . \\ b_n \end{pmatrix} \quad \underline{\varepsilon} = \begin{pmatrix} \varepsilon_0 \\ . \\ . \\ . \\ \varepsilon_n \end{pmatrix} \tag{2.6}$$

La matrice X, detta anche *matrice disegno*, ha la prima colonna composta da tutti 1 in quanto questa colonna è quella che va a moltiplicarsi con il termine noto, che è il primo valore nel successivo vettore $\underline{\beta}$. Ogni colonna della matrice rappresenta una variabile esplicativa quindi la prima colonna per esempio rappresenta la variabile esplicativa x_1 e ogni elemento in questa colonna è il valreo che tale variabile assume per l'osservazione i, così che se si considera per esempio la prima osservazione il valore di y per questa osservazione sarebbe la prima riga della moltiplicazione tra matrici a destra dell'equazione, ovvero:

$$y_1 = \beta_0 + \beta_1 \cdot x_{11} + \dots + \beta_n \cdot x_{1n} + \varepsilon_1 \tag{2.7}$$

$\underline{\beta}$ rappresenta come si è detto il vettore dei parametri ignoti per i regressori, mentre $\underline{\varepsilon}$ quello degli errori per ciascuna osservazione.

Ipotesi sul modello

Il modello di regressione lineare fin qui presentato poggia la sua validità su una serie di ipotesi che sono qui riassunte.

Ipotesi di Linearità

Secondo questa ipotesi, sia le variabili esplicative del modello (i valori della matrice $\underline{\underline{X}}$) sia i parametri $\underline{b}$ sono lineari.

Ipotesi di Non Sistematicità degli Errori

Quest'ipotesi sostiene che il valore atteso di ogni errore ε_i dato il corrispondente set di variabili esplicative corrispondente alla riga i-esima della matrice X è zero,

ovvero:

$$E\left(\varepsilon_i | X_{in}\right) = 0 \tag{2.8}$$

Di conseguenza il valore atteso della variabile dipendente è:

$$E\left(\underline{y}|X\right) = X\underline{\beta} \tag{2.9}$$

Ipotesi di Sfericità degli Errori

Per quest'ipotesi valgono le seguenti:

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:
$$V(e|x_1, \ldots, x_n) = \sigma^2 \quad \Rightarrow V(Y|x_1, \ldots, x_n) = \sigma^2 \tag{2.10}$$

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:
$$Cov(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i, j \tag{2.11}$$
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:
$$\varepsilon \sim N(0, \sigma^2) \tag{2.12}$$

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

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

Ipotesi di Non Stocasticità delle Variabili Esplicative

Si suppone che i valori delle variabili esplicative $\underline{x}_j$ siano fissi e non soggetti a fluttuazioni stocastiche. Di conseguenza vale:

$$E(\underline{\underline{X}}) = \underline{\underline{X}} \tag{2.13}$$

Ossia il valore di aspettazione della matrice delle covariate è la matrice stessa. Inoltre la correlazione tra questa e il vettore degli errori è nulla, cioè:

$$cov(\underline{\underline{X}}, \underline{\varepsilon}) = 0 \tag{2.14}$$

Ipotesi di Non Collinearità delle Variabili Esplicative

I vettori delle variabili esplicative contenuti nella matrice $\underline{\underline{X}}$ sono linearmente indipendenti. La matrice $\underline{\underline{X}}$ ha rango pieno, pari al numero delle covariate più uno dovuto al vettore costante, cioè:

$$Rank(\underline{\underline{X}}) = p + 1 \tag{2.15}$$

Di conseguenza, la matrice ha determinante non nullo, è invertibile e il prodotto $\underline{\underline{X}} \cdot \underline{\underline{X}}^T$ è definito.

Ipotesi sulla Numerosità della popolazione

Si suppone che il numero di eventi nel campione sia maggiore della numero di covariate (più uno a causa del vettore costante):

$$n > p + 1 \tag{2.16}$$

Questo garantisce che la matrice $\underline{\underline{X}} \cdot \underline{\underline{X}}^T$ abbia un'unica inversa, dunque la soluzione al problema di regressione sia unica.

Estimatori

Un'estimatore è una funzione da applicare su un certo campione di dati preso in maniera casuale dalla popolazione. La stima è il valore numerico che l'estimatore assume quando viene calcolato sul set. Ogni estimatore può essere classificato sulla base di tre criteri.

Correttezza (Unbiasedness)

Lo stimatore può essere applicato su diversi campioni casuali. Perchè sia un buon estimatore, ci si aspetta che la distribuzione dei suoi valori sia centrata sulla stima esatta che avrebbe sulla popolazione, cioè:

$$E(\hat{\mu}_Y) = \mu_Y \tag{2.17}$$

dove Y è la popolazione considerata, $\hat{\mu}_Y$ è l'estimatore applicato su un campione e μ_Y è l'estimatore applicato alla popolazione. Se un'estimatore possiede questa proprietà è detto corretto o unbiased, altrimento è detto scorretto o biased.

Consistenza (Consistency)

Un' estimatore è detto consistente se all'aumentare della dimensione del campione su cui viene applicato, questo tende ad assumere il valore vero. Ciò si può scrivere come:

$$\hat{\mu}_Y \xrightarrow{p} \mu_Y \tag{2.18}$$

Efficienza (Efficiency)

Siano $\tilde{\mu}_Y$ e $\hat{\mu}_Y$ due estimatori della stessa quantità μ_Y, entrambi corretti. Il criterio dell'efficienza dice che per scegliere quale estimatore tra i due sia migliore, si guarda alla larghezza della loro distribuzione (cioè alla loro varianza): l'estimatore con varianza minore è quello più efficiente, cioè:

$$var(\hat{\mu}_Y) < var(\tilde{\mu}_Y) \tag{2.19}$$

Il metodo dei minimi quadrati ordinari (OLS)

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

Nell'ipotesi di perfetta dipendenza lineare tra Ye gli n regressori è possibile, facendo un campione di osservazioni, stimare i valori <u>teorici</u> 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:

$$\underline{e} = \underline{y} - \underline{y'} = \underline{y} - \underline{\underline{X}} \cdot \underline{b} \tag{2.20}$$

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

$$y'_i = b_0 + b_1 x_{1i} + b_2 \cdots + b_n x_{ni} \quad \text{per } i = 1, \ldots, k \tag{2.21}$$

Quindi il singolo residuo nella (2.20) si può anche riscrivere come:

$$e_i = y_i - y'_i = y_i - b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_n x_{ni} \quad \text{per } i = 1, \ldots, k \tag{2.22}$$

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:

$$\Phi(\underline{b}) = \sum_{i=1}^{k}(y_i - y_i')^2 = \sum_{i=1}^{k} e_i^2 = \underline{e}^t \cdot \underline{e} = (\underline{y} - \underline{\underline{X}} \cdot \underline{b})^t \cdot (\underline{y} - \underline{\underline{X}} \cdot \underline{b}) \tag{2.23}$$

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

$$\frac{\partial \Phi(\underline{b})}{\partial \underline{b}} = 0 \tag{2.24}$$

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

$$\underline{\underline{X}}^t \underline{\underline{X}} \cdot \underline{b} = \underline{\underline{X}}^t \underline{y} \tag{2.25}$$

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 $\underline{b}$:

$$\underline{b} = (\underline{\underline{X}}^t X)^{-1} \underline{\underline{X}} \cdot \underline{y} \tag{2.26}$$

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:

$$\underline{b} = (\underline{\underline{X}}^{*t} \underline{\underline{X}}^*)^{-1} \underline{\underline{X}}^{*t} \cdot \underline{y}^* \tag{2.27}$$

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:

$$\sum_{i=1}^{k}(y_i - y_i')^2 = \sum_{i=1}^{k}(y_i - \underline{\underline{X}} \cdot \underline{b})^2 \tag{2.28}$$

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

$$\sum_{i=1}^{k}(y_i - \underline{\underline{X}} \cdot \underline{b})^2 = \|y_i - \underline{\underline{X}} \cdot \underline{b}\|^2 \tag{2.29}$$

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.

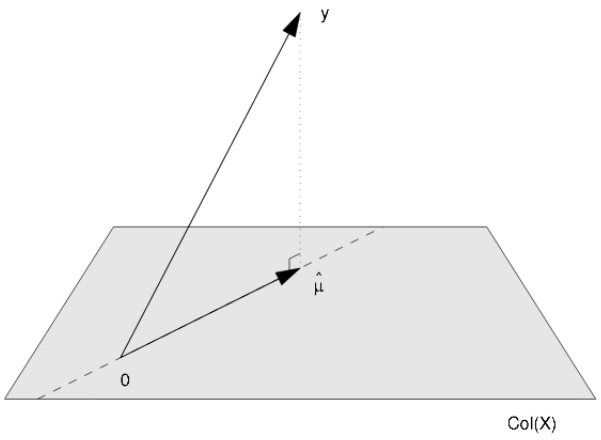

Figure 2.1: Rappresentazione dello spazio delle colonne e dei vettori di interesse.

Interpretazione

I coefficienti β indicano la variazione che si ha sulla variabile dipendente y per una variazione unitaria della variabile esplicativa x. Nessuna interpretazione addizionale è richiesta oltre alla stima dei coefficienti $\hat{\beta}$ in sè.

Bontà di adattamento

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:

$$R^2 = \frac{DevSpieg(Y)}{DevTot(Y)} \tag{2.30}$$

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

$$DevSpieg(Y) = b_1 Codev(X, Y) \tag{2.31}$$

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

$$R^2 = \frac{Cov(X,Y)}{var(X)var(Y)} \tag{2.32}$$

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:

$$H_0 = \beta_i = 0 \tag{2.33}$$

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:

$$H_0 = \mu \tag{2.34}$$

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:

$$H_1 \neq \mu \tag{2.35}$$

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 $1 - \alpha$.

$$P(-Z_{\frac{\alpha}{2}} < \bar{a} < +Z_{\frac{\alpha}{2}}) = 1 - \alpha \tag{2.36}$$

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

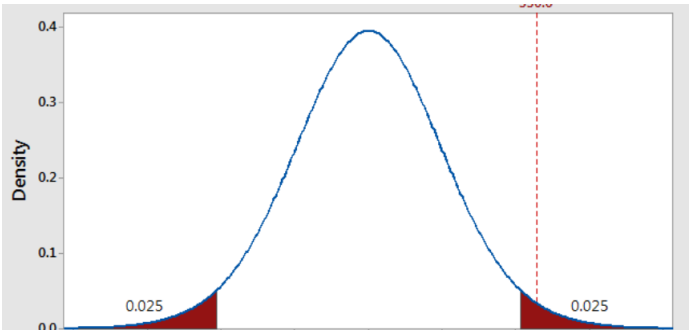

Figure 2.2: La curva rappresenta la distribuzione centrata sull'ipotesi nulla, le parti colorate di rosso rappresenta invece il livello α in questo caso uguale al 5%, quindi, essendo la curva simmetrica sarà $0,025$ a destra, e $0,025$ a sinistra. La linea tratteggiata rappresenta invece il valore osservato 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

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:

$$P\left(-Z'_{\frac{\alpha}{2}} < \frac{\bar{a} - \mu_a(H_0)}{\sigma} < +Z'_{\frac{\alpha}{2}}\right) = 1 - \alpha \tag{2.37}$$

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

$$P\left(\mu_a(H_0) - Z'_{\frac{\alpha}{2}} \cdot \sigma < \bar{a} < \mu_a(H_0) + Z'_{\frac{\alpha}{2}} \cdot \sigma\right) = 1 - \alpha \tag{2.38}$$

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

$$p = P\left(\left|\frac{\bar{a} - \mu_a(H_0)}{\sigma}\right| > \left|\frac{\bar{a_{oss}} - \mu_a(H_0)}{\sigma}\right|\right) \tag{2.39}$$

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:

$$t = \frac{a - a_0}{SE(a)} \tag{2.40}$$

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:

$$t = \frac{\beta_i}{SE(\beta_i)} = \frac{\beta_i}{\frac{\sigma}{\sqrt{n\sigma_{jj}}}} \tag{2.41}$$

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

$$t = \frac{\beta_i}{\frac{s}{\sqrt{n\sigma_{jj}}}} \tag{2.42}$$

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:

$$p = P(|t| > t_{oss}) \tag{2.43}$$

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.

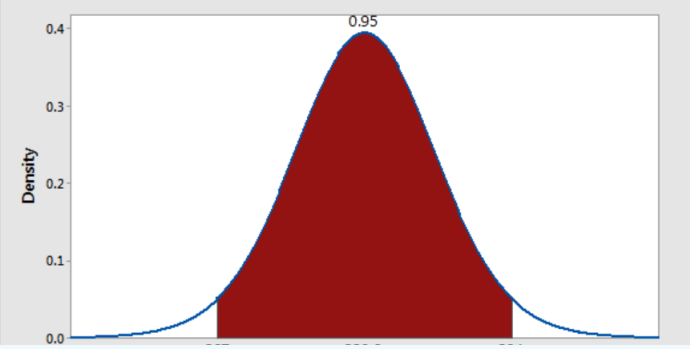

Figure 2.3: credits: `http://blog.minitab.com/blog/adventures-in-statistics-2/understanding-hypothesis-tests%3A-confidence-intervals-and-confidence-levels`
A questo punto se l'ipotesi nulla è effettivamente vera nel 95% dei casi sarà accettata, mentre solamente nel 5% dei casi sarà rifiutata, cioè per 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 2.2.

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:

$$\begin{aligned} H_0 : \ & \beta_0 = 0, \ \beta_1 = 0, \ \cdot, \ \beta_k = 0 \\ : \ & \beta_0 = \beta_1 = \cdot = \beta_k = 0 \end{aligned} \tag{2.44}$$

che implica l'ipotesi alternativa:

$$H_1 : \ \beta_0 \neq 0, \ \beta_1 \neq 0, \ \cdot, \ \beta_k \neq 0 \tag{2.45}$$

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:

$$F = \frac{SSR_r - SSR_{ur}/q}{SSR_{ur}/(n-(k+1))} \tag{2.46}$$

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 (2.46) 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:

$$P(F_0 < F_\alpha) = 1 - \alpha \tag{2.47}$$

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:

$$\epsilon_i \sim N(0, \sigma^2) \tag{2.48}$$

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

$$\begin{aligned} b_{OLS} &\sim N(\beta, \sigma^2(X^tX)^{-1}) \\ Y &\sim N(X\beta, \sigma^2 I_n) \end{aligned} \tag{2.49}$$

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:

$$max_\beta L(y_i - X_i\beta) \quad i = 1 \cdots n \tag{2.50}$$

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:

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i \quad i = 1 \cdots n \tag{2.51}$$

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:

$$Y_i = \beta_0 + \epsilon_i \tag{2.52}$$

mentre quando $D = 1$, si ottiene:

$$Y_i = \beta_0 + \beta_1 + \epsilon \tag{2.53}$$

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 (2.52) e (2.53). La statistica inferenziale si fa anche in questo caso come prima, utilizzando stime e test.

Si può avere anche il caso in cui sia presente sia una variabile qualitativa dummy che una variabile quantitativa. In questo caso si ha un modello del tipo:

$$y_i = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot D_i + \varepsilon_i \qquad \text{con} \quad i = 1 \cdots n \tag{2.54}$$

dove la variabile x_i è ovviamente una variabile quantitativa, mentre D_i è la dummy qualitativa che può assumere i due valori 0 e 1. In questo caso quindi il parametro β_2 rappresenta l'effetto della variabile dummy sull'intercetta e osservare quindi se tale effetto risulta significativo oppure no.
Il modello per una dummy dicotomica si può quindi scrivere anche come:

$$y_i = \begin{cases} \beta_0 + \beta_1 \cdot x_i + \varepsilon_i & \text{se} \quad D_i = 0 \\ (\beta_0 + \beta_2) + \beta_1 \cdot x_i + \varepsilon_i & \text{se} \quad D_i = 1 \end{cases} \tag{2.55}$$

da cui si può osservare meglio l'effetto della dummy sull'intercetta del modello. Graficamente questo caso si può visualizzare nel modo seguente:

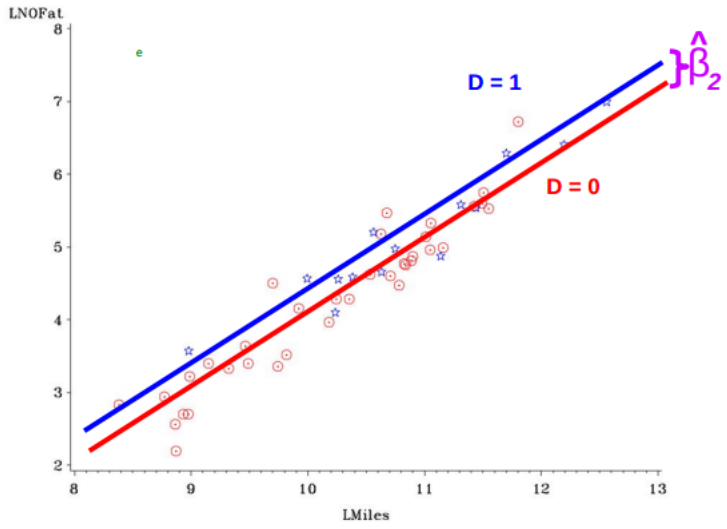

Figure 2.4: In figura è riportato un esempio di regressione con dummy variabile dicotomica. In particolare l'esempio rappresenta la regressione $\log(\hat{Y}) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \log(Miles) + \hat{\beta}_2 \cdot Seatbelt$, dove Y rappresenta il numero di incidenti fatali, e la variabile dicotomica D è $SeatBelt$.

Chapter 3

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 3.1. ovvero $Var(\varepsilon|x_i) = \sigma^2$ e $E(y|x) = 0$.

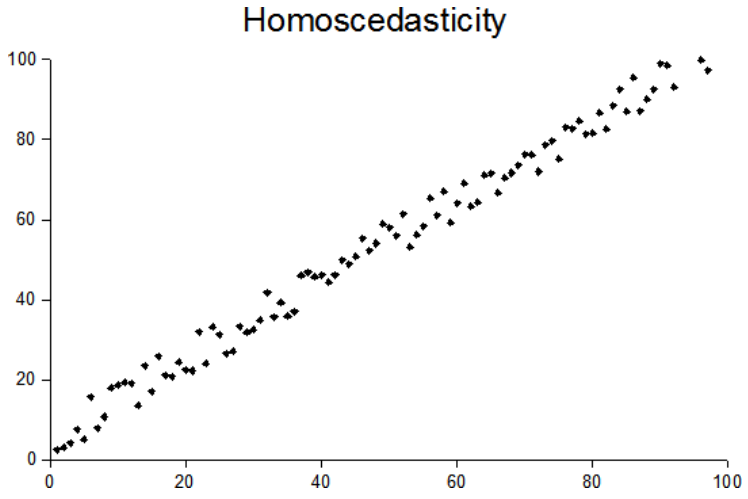

Figure 3.1: Regressione lineare con distribuzione omoschedastica degli errori.

Al contrario ci possono essere situazioni in cui il termine di errore varia tra le diverse variabili indipendenti, ottenendo così la situazione mostrata in figura **??** Questo andamento è conseguenza del fatto che la varianza del termine di errore,

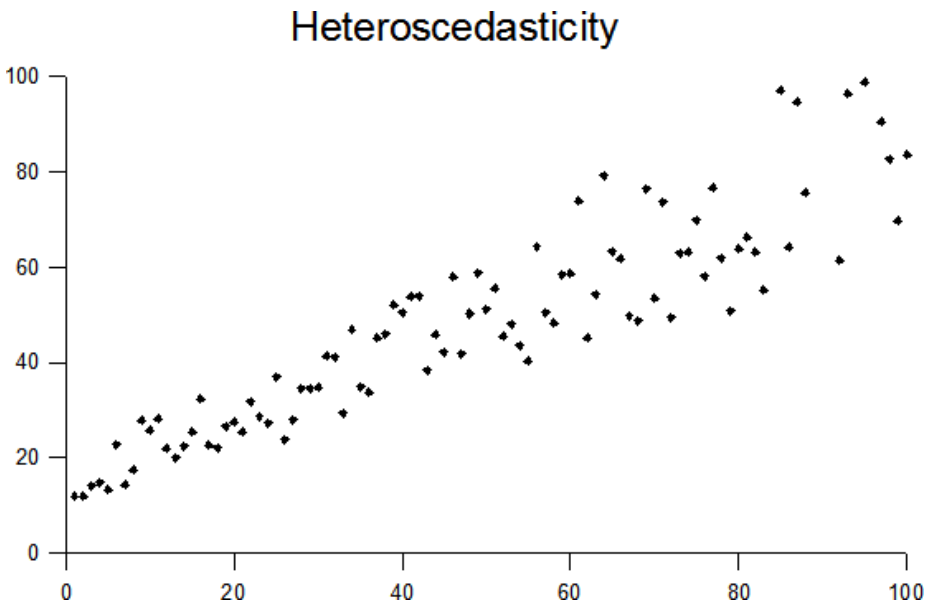

Figure 3.2: Regressione lineare con distribuzione eteroschedastica degli errori.

stimato dai residui del modello, dato un certo valore di X non è più costante e in particolare ora dipende da X.

Un altro andamento anomalo del tipo in figura **??** 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:

$$F = \frac{\sigma^2_{between}}{\sigma^2_{within}} \tag{3.1}$$

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:

$$\sigma^2_{within} = \frac{1}{a(n-1)} \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_i)^2 \tag{3.2}$$

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:

$$\sigma^2_{between} = \frac{n}{a-1} \sum_{i=1}^{a} (\bar{y}_i - \bar{\bar{y}}) \tag{3.3}$$

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 (3.1). 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:

$$H_0 = Var(\epsilon|X) = \sigma^2 \cdot I_n \tag{3.4}$$

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:
$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot X_{i1} + \cdots + \beta_n \cdot X_{in} + \hat{\epsilon}_i \tag{3.5}$$
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:
$$\hat{\epsilon}_i^2 = \delta_0 + \delta_1 \hat{Y}_i + \delta_2 \hat{Y}_i^2 \tag{3.6}$$
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:
$$LM = nR^2_{\hat{\epsilon}_i^2} \tag{3.7}$$
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:

$$\hat{\epsilon}_i^2 = \delta_0 + \delta_1 \hat{Y}_i \tag{3.8}$$

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:

$$Cov(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i \neq j \tag{3.9}$$

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

$$\Sigma = \begin{pmatrix} \sigma^2 & \rho_{1,2} & \cdots & \rho_{1,n} \\ \rho_{1,2} & \sigma^2 & \cdots & \vdots \\ \vdots & & \ddots & \vdots \\ \rho_{1,n} & \cdots & \cdots & \sigma^2 \end{pmatrix} \tag{3.10}$$

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:

$$\varepsilon'_i = \varepsilon'_{i-1} + \eta_j \tag{3.11}$$

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, in particolare se ρ è positivo la stima campionaria $\hat{\beta}$ sottostima il valore vero β. 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 residui osservando:

1. scatter plot della variabile dipendente in funzione del regressore x. Nel caso in cui ci fossero più regressori bisogna fare uno scatter plot in funzione di ogni 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:

$$H_0 : \ \rho = Corr(\varepsilon'_i, \varepsilon'_{i-1}) = 0 \tag{3.12}$$

dove ε'_i e ε'_{i-1} sono i residui relativi all'osservazione i-esima e i-1-esima. L'ipotesi nulla verifica quindi l'assenza di correlazione seriale al primo ordine.
La statistica di DW con cui si effettua il test è la seguente:

$$DW = \frac{\sum_i (\varepsilon'_i - \varepsilon'_{i-1})^2}{\sum_i \varepsilon'^2_i} \quad \text{per } i = 1 \dots n \tag{3.13}$$

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 (3.13) 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

Per tenere conto di eteroschedasticità e correlazione nei residui si possono applicare due metodi: il metodo dei minimi quadrati pesati WLS e il metodo dei minimi quadrati generalizzati GLS.

Il metodo WLS

Per tenere conto dell'eteroschedasticità, e fare in modo che i valori stimati dei parametri $\hat{\beta}$ siano ancora gli stimatori migliori, è stato dimostrato che tali stimatori risultano ancora i migliori se si minimizza una somma pesata dei residui dove i pesi sono il reciproco dello scarto quadratico medio per i valori previsti, ovvero:

$$S = \sum_{i=1}^{n} W_{ii} r_i^2 \quad \text{dove} \quad W_{ii} = \frac{1}{\sigma_i^2} \tag{3.14}$$

per poi procedere con la minimizzazione come nel caso OLS.
Alternativamente al posto di definire la (3.14) si può effettuare un cambiamento di variabili per il modello lineare classico, ovvero passare da:

$$Y = X\beta + \varepsilon^\star \tag{3.15}$$

dove $\varepsilon^\star$ indica l'errore eteroschedastico, a:

$$Y^\star = X^\star \beta + \varepsilon \tag{3.16}$$

dove:

$$\begin{aligned} Y^\star &= \frac{Y}{\sqrt{W_{ii}}} \\ X^\star &= \frac{X}{\sqrt{W_{ii}}} \end{aligned} \tag{3.17}$$

ovvero si dividono entrambi i membri dell'equazione di regressione lineare classica in (**??**) per W_{ii} e poi si procede con il metodo dei minimi quadrati ordinari OLS. Questa volta l'errore risulta omoschedastico, in quanto abbiamo diviso per lo scarto quadratico medio $\sqrt{\sigma_i^2}$ anche la parte di errore, quindi la diagonale della matrice delle varianze dei residui, che si calcola valutando $E(\varepsilon\varepsilon^t)$, REFF A

UNA MATRICE PRECEDENTE DI CORRELAZIONE DEI RESIDUI NEL CASO DI ERRORI ETEROSCHED (FARLA NEL CASO MANCHI) non è più composta da termini tutti diversi tra loro ma saranno tutti uguali e costanti.

Il metodo GLS

Anche il metodo GLS, di cui il metodo WLS è un caso particolare, si basa su una trasformazione di variabili. Data la matrice di varianza e covarianza per gli errori del modello lineare classico, caratterizzata, nel caso di errori omoschedastici ma correlati, da una diagonale tutta uguale e con termini diversi da zero fuori dalla diagonale, si supponga che esista una matrice V tale che:

$$\Sigma_{\varepsilon^\circ} = V\sigma^2 V^t \tag{3.18}$$

o analogamente:

$$\Sigma_{\varepsilon^\circ}^{-1} = (V)^{-1}\frac{1}{\sigma^2}(V^t)^{-1} \tag{3.19}$$

dove si è usata la notazione ε° per indicare che gli errori sono correlati.
Si possono definire gli errori trasformati, fatti nel seguente modo:

$$\varepsilon = V^{-1}\varepsilon^\circ \tag{3.20}$$

poichè vale la (3.18), allora si ottiene:

$$\Sigma_\varepsilon = VV^{-1}\sigma^2(V^t)^{-1}V^t = \sigma^2 \cdot I_n \tag{3.21}$$

A questo punto se passiamo sin da subito a delle variabili trasformate nel modello, moltiplicando a entrambi i membri l'equazione del modello lineare classico in (3.15), considerando di avere questa volta errori autocorrelati omoschedastici, per V^{-1}, si può applicare il metodo OLS classico per ottenere i parametri. La matrice V si può ottenere tramite decomposizione spettrale della matrice Σ_ε di correlazione degli errori. Lo stimatore GLS che si calcola in questo modo assegna un peso maggiore alle variabili caratterizzate da una minore varianza e quindi da considerarsi più attendibili. Inoltre lo stimatore è consistente, corretto (unbiased) e per il teorema di Aitken (analogo al Gauss-Markov per gli OLS) si dimostra che è anche efficiente.

Procedimento

Prima di procedere alla costruzione delle stime GLS, si deve stimare l'errore autocorrelato e i vari livelli di autocorrelazione. A questo scopo si costruisce un modello di regressione che spiega l'errore ε_i per l'osservazione i-esima in termini delle variabili esplicative e del residuo ritardato ε_{i-1}, ovvero:

$$\varepsilon_i^\circ = a_0 + a_1x_1 + \cdots + a_nx_n + a_\varepsilon\varepsilon_{i-1}^\circ \tag{3.22}$$

Il coefficiente di regressione a_ε che si ottiene rappresenta una stima per il coefficiente ρ di autocorrelazione al primo ordine.
A questo punto partendo dall'equazione di regressione lineare con gli errori autocorrelati

$$y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i^\circ \tag{3.23}$$

si considera l'equazione ritardata per il coefficiente ρ stimato dalla (3.22), quindi:

$$\rho y_{i-1} = \rho\beta_0 + \rho\beta_1 x_{i-1} + \rho\varepsilon_{i-1}^\circ \tag{3.24}$$

Si sottrae quindi questa alla (3.23) ottenendo:

$$y_i - \rho y_{i-1} = \beta_0(1-\rho) + \beta_1(x_i - \rho x_{i-1}) + w_i \tag{3.25}$$

da cui:

$$y_i' = \beta_0' + \beta_1' x_i' + w_i \tag{3.26}$$

con $y\grave{\imath}_i = y_i - \rho y_{i-1}$, $\beta_0' = \beta_0(1-\rho)$, $x_i' = (x_i - \rho x_{i-1})$ e $w_i = \varepsilon_i^\circ - \rho\varepsilon_{i-1}^\circ$. Ovvero si ottiene un'equazione di regressione lineare classica che si può risolvere tramite OLS. Il valore atteso per i w_i infatti ora è uguale a 0, come richiesto dal modello di regressione lineare:

$$E(w_i) = E(\varepsilon_i^\circ - \rho\varepsilon_{i-1}^\circ) = 0 \tag{3.27}$$

Alternativamente a questo metodo si può utilizzare un **metodo autoregressivo** che introduce tre le variabili esplicative anche un termine di errore ritardato che tenga conto dell'autocorrelazione al primo ordine. Se ci sono correlazioni anche a ordini successivi si introducono tante variabili esplicative che ne tengano conto in un numero pari agli ordini che si considerano.

Il metodo FGLS

Il problema del metodo GLS è che si basa su una conoscenza perfetta della matrice Σ_ε che però non sempre è nota, per questo si usa il metodo FGLS che si basa sulla matrice di varianza e covarianza campionaria. Gli stimatori che si calcolano con il FGLS sono consistenti nel senso che per $N \to +\infty \Rightarrow S_\varepsilon \to \Sigma_\varepsilon$, dove S_ε indica la matrice di varianza e covarianza per gli errori campionaria.
Oss.
Poichè il WLS è un caso particolare del GLS, tutte le procedure che si usano nel GLS si possono adottare anche nel WLS.

Multicollinearità

Se la matrice (X^tX) , con cui si stimano i coefficienti del modello lineare, non è invertibile (det = 0), le stime dei parametri non esistono o comunque non

sono stabili (coefficienti sottoidentificati). Ciò si verifica quando almeno una variabile è correlata linearmente con le altre e quindi si ha multicollinearità. Se in particolare si ha che una variabile esplicativa è una combinazione lineare *perfetta* delle altre ($\Rightarrow$ det $= 0$), allora diventa impossibile fare stime per i coefficienti in quanto ci sono infinite soluzioni al sistema. Se invece non c'è perfetta collinearità, ma si hanno comunque due variabili fortemente correlate, si ha che il determinante della matrice (X^tX) tende a 0. Questo implica un aumento della varianza per le stime dei coefficienti $\hat{\beta}$ in quanto ricordiamo che la loro varianza è uguale a:

$$Var(\hat{\beta}) = \sigma^2(X^tX)^{-1} \tag{3.28}$$

e $(X^tX)^{-1}$ dipende in modo inversamente proporzionale dal determinante di (X^tX) (per la formula dell'inversa), che se c'è collinearità si è detto che tende a zero. Di conseguenza $(X^tX)^{-1}$ diventa molto grande comportando un conseguente aumento della varianza (3.28). Il test t quindi sarà caratterizzato da una regione di accettazione dell'ipotesi nulla più stretta di quanto sia in realtà se le variabili non fossero collineari, e quindi aumenta la probabilità di ritenere un parametro come significativo quando in realtà non lo è. Lo stesso ragionamento può essere fatto per il test F e per gli intervalli di confidenza che risultano più ampi. Inoltre se ho due variabili fortemente correlate, se la prima spiega l'80% della varianza, se aggiungo la seconda variabile correlata con essa l'aumento di devianza spiegata da R^2 è di gran lunga inferiore all'aumento che avrei inserendo una variabile non correlata. Stessa cosa avviene se inserisco nel modello la seconda variabile suddetta e poi la prima, quindi non posso dire quale delle due variabili è più influente e utile ai fini del modello.

Indice di tolleranza, VIF e condition index

L'indice di tolleranza Tol misura quanto una variabile è correlata rispetto alle altre, è definito come segue:

$$Tol(x_j) = 1 - R_j^2(x_j|x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n) \tag{3.29}$$

dove il coefficiente R_j^2 risulta dalla regressione di x_j sulle altre variabili indipendenti. Se tutte le variabili sono incorrelate si ha $Tol(x_j) = 1$, mentre se una variabile è perfettamente correlata alle altre si ottiene $Tol(x_j) = 0$.

Associato all'indice di tolleranza vi è il VIF (Variance Inflation Factor), definito come:

$$VIF = \frac{1}{Tol} = \frac{1}{1 - R_j^2} \tag{3.30}$$

In caso di multicollinearità perfetta si ha $VIF = \infty$, e in generale un valore di questo indice superiore a 20, indica una forte collinearità tra la variabile j che si sta testando e le altre variabili.

Un altro indice di collinearità è il *condition index*, che si calcola a partire dagli autovalori per la matrice $(X^tX)^{-1}$. Dopo aver calcolato questi, si costruisce il condition index per ciascun autovalore prendendo la radice del rapporto tra l'autovalore di valore massimo e l'autovalore che si sta considerando, ovvero:

$$CI_i = \sqrt{\frac{\lambda_{max}}{\lambda_i}} \tag{3.31}$$

tipicamente un $CI \geq 30$ indica una probabile esistenza di collinearità tra le variabili.

Linearità

Tra le possibili violazioni del modello lineare classico vi potrebbe essere il fatto che l'approssimazione lineare non è sempre la migliore. Potremmo infatti avere una relazione non lineare sia nelle variabili che nei parametri del modello. Quando interpoliamo linearmente una relazione che non è lineare introduciamo un eccesso di residui sia positivi che negativi, che possono risultare anche in un'autocorrelazione.

Possiamo verificare la presenza di un andamento non lineare nelle variabili in diversi modi:

- **Scatter plot** della variabile dipendente in funzione della variabile esplicativa (o delle variabli esplicative), e vedere com'è l'adattamento del fit lineare ai dati.
- **Scatter plot** dei residui in funzione dei valori osservati o dei valori previsti e vedere se si ditribuiscono in un modo più o meno regolare in una banda, o invece se ci sono regioni in cui si addensano.
- **Risultati del fit**. Dai risultati del fit possiamo ottenere un R^2 elevato anche se c'è un andamento non prettamente lineare nei dati. Questo perchè magari la funzione con cui si interpolano correttamente i dati è una polinomiale con anche una componente lineare, quindi l'R^2 per questa componente risulta comunque elevato. La non linearità però può essere comunque individuata dai test di ipotesi, infatti spesso in caso di non linearità alcune variabili esplicative possono risultare non significative nonostante si abbia un R^2 elevato, così come il test F per l'analisi della varianza.

Nel caso di una non-linearità nei **parametri**, il parametro non lineare viene sostituito con un altro parametro in modo da ottenere una relazione lineare. In seguito si risolve il modello con questo nuovo parametro con il metodo OLS classico ed infine si ricava il parametro originario mediante la relazione che si è stabilita inserendo il nuovo parametro per la stima del modello.

Nel caso di una non-linearità nelle **variabili** si sostituisce la variabile non lineare eguagliandola ad una nuova, in modo che il nuovo modello risulti lineare. Si calcolano i parametri utilizzando il modello OLS classico ed infine si ritorna alle variabili originarie.

Ci possono essere però alcune relazioni *intrinsecamente* non lineari, per le quali non è possibile effettuare una linearizzazione e di cui non è possibile studiarne i parametri tramite OLS. In questi casi i parametri possono essere stimati tramite minimi quadrati non lineari (NLS). Un esempio di questo tipo è la curva di crescita esponenziale negativa:

$$y = \beta_0 - \alpha \exp(-\beta_1 x) \tag{3.32}$$

con β_0, β_1 e α parametri sconosciuti. I parametri di regressione in questo caso si trovano appunto con metodi di calcolo numerico attraverso NLS.

In tutti questi casi però la funzione non lineare è fornita (si conosce la forma funzionale non lineare) sin dall'inizio e da questa si procede poi alla linearizzazione. Nel caso la funzione non venga fornita (come accade nel caso si voglia studiare con un modello la relazione che sussiste tra una variabile dipendente di output e i regressori) è possibile studiare diversi casi di andamento non lineare nelle variabili tramite i metodi regressione multipla OLS. I principali effetti non lineari che si vogliono studiare sono:

- grado superiore in y (e.g. y^2)
- grado superiore in x (e.g. x^2)
- grado inferiore in x (e.g. $\log(x)$)
- grado inferiore in y (e.g. $\log(y)$)

Nel caso si proceda a stimare modelli non lineari di questo tipo nelle variabili, i coefficienti come si è detto possono essere stimati normalmente tramite il metodo OLS, e l'interpretazione dei test rimane la stessa che nel caso del modello lineare classico. La scelta della specificazione della forma funzionale deve essere guidata dall'analisi grafica (scatter plot) dei valori osservati e dei valori predetti.
Le trasformazioni logaritmiche in particolare possono aiutare nel caso il campo di variazione per la y o per la x sia molto ampio. Inoltre la trasformazione logaritmica permette un'interpretazione dei coefficienti di regressione in percentuale che può risultare utile in diverse applicazioni. L'interpretazione letterale dei coefficienti come variazione della variabile dipendente per variazioni unitarie della variabile (o delle variabili) indipendente vale ancora, anche nel caso in cui si utilizza un modello trasformato con un logaritmo, ma solitamente è più ragionevole interpretarlo non come cambi unitari per il logaritmo ma come variazioni percentuali. Consideriamo quindi i diversi casi che si possono presentare:

- **Linear-log model (Lin-log):** $y_i = \beta_0 + \beta_1 \log(x_i) + \varepsilon_i$.
  Se effettuiamo un cambio unitario su x in questo modello otteniamo:

$$\log(x) \to \log(x+1) = \log(x) + \log(e) = log(ex)$$

ovvero aumentare x di 1 equivale a moltiplicarla la per $e \simeq 2,72$, ovvero incrementarla del 172%.
$\hat{\beta}_1$ rappresenta quindi in questo caso il cambio atteso in y quando la x aumenta del 172%. Per altri cambi percentuali si può usare la formula $\hat{\beta}_1 \cdot \log(1+\frac{p}{100})$, quindi se per esempio si vuole considerare il cambio in y per un aumento del 10% nella x si ottiene che $\hat{\beta}_1 \cdot \log(1+0,1) = \hat{\beta}_1 \cdot \log(1,1) = 0,095\hat{\beta}_1$ è il cambiamento atteso in y, quando x è moltiplicata per 1.1, ovvero aumenta del 10%.

- **Log-linear model:** $\log(y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i$.
  In questo modello variazioni unitarie nella x implicano che il logaritmo di y sia aumentato di $\hat{\beta}_1$, ovvero che y sia moltiplicata per $e^{\hat{\beta}_1}$. Per piccoli valori di $\hat{\beta}_1$, $e^{\hat{\beta}_1}$ si può approssimare come $e^{\hat{\beta}_1} \approx 1 + \hat{\beta}_1$. Moltiplicando a questo punto $\hat{\beta}_1$ per 100, si può identificare il cambio percentuale nella y per una variazione unitaria nella x. Per esempio con $\hat{\beta}_1 = 0,06$, si ha che $e^{0,06}$, con cui si moltiplica la y, si può approssimare a $1,06$, ovvero si ottiene un aumento della y del 6%.

- **Log-Log model:** $\log(y_i) = \beta_0 + \beta_1 \log(x_i) + \varepsilon_i$.
  In quest'ultimo modello l'interpretazione dei coefficienti è una combinazione delle interpretazioni percentuali precedenti, ovvero quando x aumenta di una certa percentuale ci si aspetta un certo cambio percentuale sulle y. Qui si ottiene che moltiplicare la x per e comporta una moltiplicazione della y per $e^{\hat{\beta}_1}$. Per avere il cambio percentuale specifico per una variazione percentuale specifica sulla x bisogna calcolare prima $a = \log(1 + \frac{p}{100})$ coefficiente di moltiplicazione che starà davanti al $\hat{\beta}_1$ e poi calcolare $e^{a\hat{\beta}_1}$.

Normalità

Se gli errori non sono distribuiti in modo normale avvengono alcune conseguenze:

1. I parametri stimati $\hat{\beta}$ non sono normali. Essi infatti possono essere espressi come combinazione lineare degli errori, quindi se questi non sono normali anche i parametri non sono più normali.

2. Le stime che si ottengono tramite OLS non coincidono più con le stime di massima verosimiglianza. Dalle proprietà delle stime di massima verosimiglianza e dall'ipotesi di normalità tramite il teorema di Cramer-Rao si poteva concludere che gli stimatori OLS erano anche gli stimatori a minima varianza MVUE (Minimum Variance Unbiased Estimators). Tuttavia il teorema di Gauss-Markov assicura che sono ancora stimatori BLUE. Questo teorema infatti non fa alcuna richiesta sulla distribuzione che devono assumere gli errori, ma è sufficiente che siano:

   - a media nulla;

- incorrelati;
- varianza costante.

3. Non è più possibile applicare test e intervalli di confidenza.

Oss.
Distribuzioni non normali per gli errori si possono verificare nel caso in cui i campioni che si analizzano sono piccoli. Nel caso di campioni di grandi dimensioni la loro distribuzione tende asintoticamente alla distribuzione normale per il teorema del limite centrale. Quindi nel caso in cui si abbia una distribuzione non normale si può aumentare il numero di osservazioni per renderla normale. Tuttavia esistono dei casi in cui non è possibile fare questo, per esempio nel caso in cui si stiano facendo delle analisi statistiche riguardo una malattia non è possibile aumentare il numero di incidenza per avere più osservazioni. In questi casi è quindi necessario effettuare dei test di normalità.

Come si individua

Come nelle altre violazioni del modello lineare la non normalità si può individuare tramite via grafica oppure effettuando dei test statistici.

Metodo grafico

Per via grafica si può procedere in diversi modi:

Boxplot. Si rappresenta tramite boxplot la distribuzione dei residui, se la distribuzione non è normale la media non coinciderà con la mediana e il box sarà diviso in due rettangoli di diversa area, mentre se la dsitribuzione è normale la media deve coincidere con la mediana e i due rettangoli devono essere il più possibile uguali.

Rappresentazione grafica. Si visualizza il grafico della distribuzione dei residui. Non c'è normalità se c'è asimmetria a destra o *positiva*, che risulta con il picco della curva spostato a sinistra, oppure a sinistra o *negativa*, che risulta con il picco della curva spostato a destra.

PP-plot. Si costruisce un grafico in cui si rappresentano le probabilità cumulate teoriche in funzione di quelle empiriche. Se gli errori ε_i sono normali , allora i valori delle probabilità cumulate dei quantili delle distribuzione teoriche e dei residui coincidono e quindi si ottiene un grafico lineare in cui i punti sono distribuiti sulla bisettrice del quadrante.

IMMAGINE ESEMPIO PP PLOT.

QQ-plot. Sono l'inverso dei PP-plot, ovvero rappresentano sulle ordinate la distribuzione empirica dei residui e sulle ascisse la distribuzione teorica normale.

Anche qui nel caso in cui gli errori siano distribuiti in modo normale le probabilità coincidono e si ottiene un grafico lineare sulla bisettrice del quadrante.

IMMAGINI VARI CASI QQ PLOT.

Test non parametrici

I test non parametrici sono test che sono indipendenti dalla distribuzione e non necessitano di particolari distribuzioni per essere applicati. Per questo sono adatti per confrontare campioni molto piccoli e che non seguono alcuna distribuzione nota.

Test di Shapiro-Wilk

Il test di Shapiro-Wilk è definito come segue:

$$W = \frac{(\sum_{i=1}^{n} a_i \varepsilon_i)^2}{\sum_{i=1}^{n} (\varepsilon_i)^2} \tag{3.33}$$

ovvero come rapporto tra una devianza calcolata tramite dei pesi particolari a_i definiti in apposite tabelle di Shapiro in base alla dimensione del campione (si veda `http://www.real-statistics.com/statistics-tables/shapiro-wilk-table/`) e la devianza campionaria stimata dai residui.
W è compreso tra 0 e 1, più è vicino a 1 più la distribuzione è normale. In generale per concludere la normalità o la non normalità dal test di Shapiro si può guardare il p-value associato al risultato (anche in questocaso il p-vlaue associato si può ricavare da apposite tabelle). Se il p-value è superiore al livello di significatività α scelto, allora si può concludere che la distribuzione non sia normale.
Il test di Shapiro è un test fortemente asimmetrico, quindi anche se assume valori elevati approssimabili a 1 si potrebbe essere comunque in presenza di una distribuzione non normale. Per esempio con un valore di W pari a 0.82 si può già concludere che non ci sia normalità.

Test di Kolmogorov Smirnov

Il test di Kolmogorov misura le discrepanze sul QQ plot dalla diagonale e le valuta al quadrato. Se è uguale 0 la distribuzione osservata è normale, non normale altrimenti. Anche qui ci sono regioni di accettazione e di rifiuto a cui fare riferimento in apposite tabelle. Se il valore che si trova dal test supera il valore critico tabulato a livello di significatività scelto allora la distribuzione non è normale.

Test di skewness

Il test di skewness è un test direzionale. Si basa sul fatto che la distribuzione normale è simmetrica e quindi viene elaborato un indice per questa simmetria. Per effettuare questo tipo di test ènecessario avere un campione di grandi dimensioni, per questo spesso a partire dai dati che si hanno (che solitamente sono pochi quando si vuole effettuare un test di normalità), il valore dell'indice si ottiene dopo diverse iterazioni.
Esso è definito come:

$$S = \frac{(E(X-\mu)^3)^2}{(E(X-\mu)^2)^3} \tag{3.34}$$

Il valore del test è dato dalla distribuzione di S che in condizioni di normalità ha valore atteso uguale a zero, ovvero $E(S) = 0$. Rigettando l'ipotesi nulla si rigetta la normalità, ma non rigettandola si può concludere solamente che la distribuzione è simmetrica, non è detto che sia normale.

SPECCHIETTO SU TEST DIREZIONALI E NON DIREZIONALI

Test della kurtosi

Con il test della kurtosi si testa appunto la kurtosi della distribuzione che viene sottoposta al test. Ricordiamo che la kurtosi per una distribuzione normale è uguale a 3. Il test della kurtosi, come quello per la skewness è un test direzionale ed è definito come segue:

$$K = \frac{E(X-\mu)^4}{(E(X-\mu)^2)^2} \tag{3.35}$$

Il test è dato dalla distribuzione di $K - 3$, che quindi in condizioni di normalità avrà valore uguale a zero.
Rigettando l'ipotesi nulla si rifiuta la normalità, ma non rigettandola si conclude solo che la distribuzione che si osserva ha kurtosi uguale a 3.

Come si risolve

Come si è detto per portarsi in una condizione di normalità per la distribuzione degli errori bisogna innanzitutto pensare di allargare il campione, se è possibile farlo. Nel caso in cui non sia possibile si può provare ad applicare alcune trasformazioni in modo che la distribuzione risulti normale. Nel caso di eteroschedasticità classico degli errori la deviazione è data da variazioni individuali, mentre nel caso di non normalità o non linearità l'errore è sistematico, quindi se si riesce ad individuare la struttura sistematica sottesa alle deviazioni non lineari/non normali ci si può riportare alla distribuzione corretta. Le principali trasformazioni sono:

- $\log(y)$, quando lo scarto quadratico (errori) cresce con y o ha un'asimmetria positiva.

- y^2, quando lo scarto quadratico (errori) è proporzionale al valore atteso di y o asimmetria negativa.
- $y^{1/2}$ se lo scarto quadratico cresce proporzionalmente al valore atteso di y
- $\frac{1}{y}$ se lo scarto quadratico cresce significativamente al crescere di y.

Oss.
Ci potrebbero essere sia errori eteroschedastici che non lineari/non normali. Ovvero dopo aver applicato anche un'opportuna trasformazione potrei ancora avere delle variazioni individuali che non ho individuato. Quindi devo applicare anche WLS.

Outlier

I valori outlier sono quelle osservazioni che presentano valori estremamente elevati o estremamente bassi rispetto alla distribuzione del resto dei valori osservati. È comunque necessario osservare la distribuzione nella sua interezza per vedere se tali osservazioni corrispondono a casi isolati oppure no. Valori outlier possono influenzare molti indicatori come media, deviazione standard e asimmetria della curtosi, oltre agli indici di associazione tra le variabili come il coefficiente di correlazione di Pearson. Le osservazioni che influenzano di più le stime sono dette *punti influenti*. Non sempre un valore outlier è un valore influente, ma nel caso in cui gli outlier siano anche influenti è più indicato utilizzare la distanza di Cook.

Come si individuano

Rappresentazione grafica

Per via grafica gli outlier si possono individuare tramite:

1. **Box plot**, osservo i valori che sono fuori dagli estremi del box.
2. **Scatter plot** della variabile dipendente in funzione della variabile di regressione (o delle variabili). Valori isolati troppo distanti dalla retta di regressione sono da considerarsi outlier.
3. **Dispersione** non elevata dalla retta di regressione e elevato R^2. Ovvero tutti i punti devono disporsi più o meno attorno alla retta di regressione con qualche variabilità ma senza una dispersione elevata da essa.
4. **Direzione relazionale** positiva.

Indicatori

Alternativamente la metodo grafico si può far ricorso a diversi indicatori che ci possono permettere di capire se si è in presenza di valori outlier o meno.

Leverage values

Data la matrice $H = X(X^tX)^{-1}X^t$ detta *matrice di proiezione* (è la matrice che stima i minimi quadrati, ma in generale è una matrice $n \times n$ detta di proiezione). Gli elementi sulla diagonale principale prendono il nome di *leverage values*. Valori piccoli per questi elementi sulla diagonale principale indicano che lo stimatore di y è basato su molte osservazioni e che quindi la singola osservazione non è dominante, al contrario se è elevato vuol dire che l'osservazione corrispondente è influente nella stima di y. Analogamente quando si hanno valori vicini a 1 del rapporto tra h_{ii} e la somma di tutti i valori sulla diagonale principale, indica che la stima del modelle è determinata in modo predominante dalla singola osservazione.
Il valore medio del *leverage index* è calcolato come:

$$\frac{k+1}{n} \tag{3.36}$$

dove k è il numero delle variabili esplicative e n il numero di osservazioni. Generalmente se un valore leverage h_{ii} sulla diagonale è maggiore di 2 volte questo valore medio allora si è in presenza di possibili outlier.

Residui standardizzati

I residui standardizzati sono definiti come segue:

$$\varepsilon^\star = \frac{\varepsilon_i}{\sigma\sqrt{1-h_{ii}}} \tag{3.37}$$

Rappresentando in uno scatter-plot i residui standardizzati in funzione dei valori previsti dobbiamo cercare per residui molto elevati. In un campione distribuito normalmente il 95% dei valori osservati è compreso tra -2 e $+2$, mentre al 99% sono compresi tra $-2,5$ e $+2,5$. Un'osservazione con un residuo standardizzato con un valore di 3 è probabilmente un outlier.
Se più dell'1% dei casi osservati ha valori assoluti di residui standardizzati maggiori di $2,5$ il risultato del fit è basso e se più del 5% hanno valori assoluti di residui standardizzati maggiori di 2, allora il risultato del fit è molto basso. Non si possono però individuare delle soglie per i residui standardizzati in base alle quali stabilire con certezza se un'osservazione è un outlier o meno.

Residui studentizzati

Questo tipo si residui sono utilizzati per verificare la presenza di osservazioni anomale in campioni di elevata numerosità. Sono i residui divisi per una stima della deviazione standard che varia da punto a punto, ovvero:

$$\varepsilon_i^\star = \frac{\varepsilon_i}{s_{\varepsilon_i}}\sqrt{1-h_{ii}} \tag{3.38}$$

Allo stesso modo si possono definire dei residui studentizzati jacknife che sono i residui divisi per una stima della deviazione standard dei residui ottenuta eliminando dal dataset l'i-esima osservazione.
Osservazioni con valori di residui studentizzati maggiori di 3 sono considerate outliers.

COVRATI

La statistica COVRATI misura l'impatto di ciascuna osservazione sulle varianze dei coefficienti di regressione e sulle loro covarianze. Sono definiti come:

$$COVRATI = \frac{\det(\sigma_i X_i^t X_i)^{-1}}{\det(\sigma^2 (X_i^t X_i)^{-1})} \tag{3.39}$$

Dove i valori con il pedice i indicano che le matrici sono considerate eliminando l'i-esima osservazione.
Con questa statisica osservazioni con valori al di fuori dell'intervallo $1 \pm 3(\frac{k+1}{n})^{1/2}$ sono considerate outlier.

DFITTS

La statistica DFITTS di un'osservazione misura l'influenza di quell'osservazione sulla stima dei coefficienti di regressione e sulla loro varianza quando viene rimossa dal processo di stima.
Valori al di fuori dell'intervallo:

$$\pm\, 2\left(\frac{k+1}{n}\right)^{1/2} \tag{3.40}$$

sono da considerarsi outlier.

DFBETAS

La statistica DFBETAS misura l'influenza di un'osservazione, quando viene rimossa dal processo di stima, sulle stime di ogni coefficiente di regressione separatamente.
Come valore soglia oltre il qual un'ossrevazione viene considerata come outlier si prende il valore 2 o $2 \cdot n^{1/2}$ che tiene conto anche del numero di osservazioni.

Distanza di Cook

La distanza di Cook misura l'influenza di una singola osservazione sulla stima dei coefficienti di regressione, in termini di capacità del modello di predire tutti i casi quando la singola osservazione viene rimossa dal processo di stima. Un valore della distanza di Cook maggiore di 1 indica che il punto è influente.

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:

$$y = \beta_0 + \beta_1 \cdot \textit{tempo di studio} \tag{3.41}$$

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 (3.41), 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:

$$P = \exp(\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots) \tag{3.42}$$

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:

$$P = \frac{\exp(y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots)}{\exp(y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots) + 1} \tag{3.43}$$

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

$$\log(\frac{P}{1-P}) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots \tag{3.44}$$

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 (3.44), 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.