22/10/19

Associazione tra variabili categoriche (qualitative):

Tra due variabili (una risposta e l'altra esplicativa) esiste associazione (cioè dipendenza) se nella popolazione la distribuzione condizionata della variabile risposta (chiamata anche distribuzione parziale relativa) è diversa in corrispondenza delle diverse modalità di variabile esplicativa.

Esempio: tavola di contingenza su genere e orientamento politico

Analisi congiunta di 2 variabili (genere M,F → variabile esplicativa; orientamento politico, 3 tipologie → variabile risposta). Queste due variabili sono distribuzioni marginali in una tabella a doppia entrata; se fossero prese singolarmente sarebbero distribuzioni di frequenza assoluta.

Nella tabella abbiamo una serie di distribuzioni: quando ho una sola variabile ho la distribuzione di frequenza.

La distribuzione condizionata dell'orientamento politico esiste sia per maschi che per femmine; ci sono anche le distribuzioni parziali.

Distribuzione congiunta relativa → si ottiene dividendo ciascuna cella (ovvero il suo valore assoluto) per il totale generale (es. donne che preferiscono il partito democratico = 573/2771).

Dipendenza e indipendenza:

Indipendenza → se nella popolazione la distribuzione condizionata di ciascuna variabile è costante per tutte le modalità dell'altra → è una situazione estrema, ma serve per capire se sono indipendenti tramite un test in cui ipotizzo una situazione d'indipendenza (test del Chi-quadro).

Dipendenza (associazione) → quando nella popolazione le distribuzioni condizionate sono diverse tra loro.

Esempio di indipendenza statistica:

	Felicità		
Reddito	Molto	Abbastanza	Poco
Sopra	32%	55%	13%
In media	32%	55%	13%
Sotto	32%	55%	13%

Test statistico di indipendenza Chi – Quadro (χ^2):

Ha lo scopo di sintetizzare le differenze tra frequenze osservate e frequenze teoriche (ovvero quelle attese nel caso in cui H_0 sia vera, cioè se tutte le distribuzioni condizionate fossero state uguali) nelle varie celle della tavola di contingenza (in pratica faccio un confronto tra questi due valori). Per ***sintetizzare*** s'intende trovare un indice che abbia una distribuzione campionaria nota se H_0 è nulla.

Nel Chi – quadro le differenze tra frequenze osservate e teoriche vengono sintetizzate mediante la statistica:

$$\chi^2 = \frac{\sum(f_o - f_e)^2}{f_e}$$

Con la somma estesa a tutte le celle della tabella di contingenza.

L'uso della distribuzione Chi-quadro per approssimare l'effettiva distribuzione della statistica test è adeguata per grandi campioni, ovvero per tutte o quasi tutte le celle dove $fe \geq 5$; per piccoli campioni si applica il test esatto di Fischer. Il Chi-Quadro tratta le variabili come qualitative nominali, per variabili di tipo ordinale o quantitative sono utilizzabili altre tecniche di analisi.

IP:
H0: le variabili sono statisticamente indipendenti;
Ha: esiste un'associazione statistica tra le due variabili.
c → n° di colonne della tabella di contingenza
r → n° di righe della tabella di contingenza
fo → frequenze osservate
fe → frequenze attese → hanno distribuzioni condizionate costanti e uguali alla corrispondente distribuzione marginale; inoltre hanno la stessa distribuzione marginale sia di riga che di colonna delle frequenze osservate. Si possono calcolare così: **fe = (totale di riga) (totale di colonna)/n**
I totali di riga e colonna sono marginali; divido per n totale per relativizzare il tutto.
I valori delle fe posso lasciarli decimali.
Quando H0 è vera, la distribuzione campionaria di questa statistica si approssima (per n sufficientemente grande) alla distribuzione di probabilità Chi-Quadro.
Gdl = (r −1)(c −1) significa che dati i marginali di una tabella si possono fissare liberamente solo (r −1)(c−1) valori nella celle; gli altri sono determinati di conseguenza. Se z è una statistica con distribuzione normale standardizzata, allora z^2 ha una distribuzione Chi-quadro con gdl = 1.
La somma di *d* variabili indipendenti z con distribuzione normale standardizzata (prese al quadrato), ha una distribuzione χ^2 con gdl = d.
N.B: ***Elevati valori del χ^2*** mostrano una forte evidenza che vi sia ***associazione*** tra le variabili, ma **non** forniscono alcun elemento né sulla ***struttura*** dell'associazione né sulla sua ***forza***: occorre indagare sulla struttura dell'associazione.

Proprietà della distribuzione Chi – Quadro:
1) E' definita sull'asse positivo reale (0; + infinito);
2) E' asimmetrica positiva → tende a diventare simmetrica al crescere della dimensione campionaria;
3) Media e varianza dipendono entrambe dalla dimensione della tavola di contingenza attraverso i gradi di libertà (gdl = (r -1) (c – 1))→ è la meda della distribuzione.
2gdl = varianza della distribuzione.
4) Valori elevati della statistica χ^2 sono improbabili sotto l'ipotesi H0, pertanto il p-valore è rappresentato dai valori nella coda destra della distribuzione, maggiori della statistica test osservata.

Calcolo del Chi – Quadro sulla tavola felicità – reddito:
$\chi^2 = 172,3$ gdl = 4 p – valore < 0,001
-C'è un'evidenza molto forte contro l'ipotesi nulla H0:
Indipendenza tra le variabili (se H0 fosse vera, avremmo una probabilità < 0,001 di osservare un valore della statistica χ^2 maggiore o uguale a quello osservato, cioè 172,3);
-O è capitato un evento estremamente improbabile o l'ipotesi H0 non è corretta;
-Di conseguenza, al livello di significatività α = 0,05 (o α = 0,01 o α = 0,001), si respinge l'ipotesi H0 e si conclude che, nella popolazione, c'è associazione tra felicità e reddito.

24/10/19

Individuare la struttura dell'associazione: residui
-Possiamo indagare sulla struttura dell'associazione utilizzando i **residui** (cioè le contingenze) in ciascuna cella della tavola:
Residuo = **fo – fe**. I residui sono positivi/negativi se le frequenze osservate sono maggiori/minori di quelle attese sotto l'ipotesi di indipendenza;
Residui standardizzati: **z = fo – fe/es** (con es = errore standard del residuo) → misurano per quanti

errori standard la differenza fo – fe si allontana da zero sotto l'ipotesi Ho.
es = errore standard residuo = $\sqrt{fe\ (1 - proporzione\ di\ riga)(1 - proporzione\ di\ colonna)}$

Esempio: calcolo dei residui standardizzati sulla tavola felicità – reddito
Rapportando ciascun valore assoluto al totale della riga è possibile fare il confronto tra colonne (percentuale di riga es. 272 con 615), non il confronto tra righe perché relativizzerei gli stessi valori; così come per avere il confronto tra righe dobbiamo fare la percentuale di colonna.
Tavola di contingenza 2x2 → ha un carattere dicotomico come variabile esplicativa e uno dicotomico come variabile risposta. Le tavole di contingenza hanno un unico gdl e il test χ^2 è equivalente al test del confronto tra due proporzioni.
Statistica χ^2 → è il quadrato della statistica z.
Ci interessano in questo caso le percentuali calcolate in ciascuna riga.
Quando l'ipotesi d'indipendenza è vera, i residui standardizzati hanno distribuzione normale standardizzata con media 0 e varianza 1.
Lo z ci dice quante deviazioni standard dista dalla sua media; prescinde dall'entità dei valori e dall'unità di misura. Li abbiamo calcolati perché hanno una distribuzione normale standardizzata: se è maggiore di 2 o minore di -2 indica un significativo allontanamento dal valore che ci aspetteremmo per la relativa cella se l'ipotesi nulla fosse vera; questo capiterà per effetto del caso, solo circa 5 volte su 100. Se invece è maggiore di 3 o minore di -3 per quella cella (combinazione di modalità) vi è una evidenza molto forte di associazione.

Misura dell'associazione:
-Il test Chi-quadro risponde alla domanda: "C'è associazione tra due variabili?".
-I residui standardizzati ci aiutano a comprendere la struttura dell'associazione e rispondono alla domanda: "Quanto i dati osservati si allontanano dalla situazione di indipendenza?"
-Se ci chiediamo: "Quanto forte è l'associazione tra due variabili?" possiamo rispondere utilizzando la differenza tra proporzioni.

29/10/19

Confronto mediante rapporti:
Due proporzioni possono essere confrontate mediante un rapporto (rischio relativo) anziché una differenza.
Nell'esempio precedente i due confronti effettuati con differenze danno luogo a:
0,44/0,20 = 2,2
0,23/0,08 = 2,875
Un procedimento alternativo per confrontare proporzioni, tipico delle tavole 2×2 è rappresentato dall'**odds ratio**.

Odds:
Rapporto tra probabilità di successo e di insuccesso (che sono i possibili risultati di una variabile) o probabilità di successo diviso 1 – P di successo.
Es: se P(successo) = 0,8 e P(insuccesso) = 0,2 odds = 0,8/0,2 = 4,0;
se P(successo) = 0,2 e P(insuccesso) = 0,8 odds = 0,2/0,8 = 1/4 = 0,25
E' un'operazione che si fa tra le probabilità; da un odds possiamo risalire alla probabilità di successo rapportando l'odds all'odds stesso aumentato di un unità →Probabilità = (odds/(odds + 1))
Es: odds = 4,0 ⇒ P(successo) = 4/(4 + 1) = 4/5 = 0,8.
N.B: Non esistono odds negativi perché lavoriamo con le probabilità.
Un odds molto piccolo si approssima a zero, se è molto grande vuol dire che la probabilità di

successo è molto superiore rispetto a quella di insuccesso. (es. odds = 4 → la probabilità di successo è 4 volte quella di insuccesso). Gli odds vengono confrontati tra loro perché li calcoliamo per le varie modalità del carattere. Quando nella tavola il numero di soggetti è elevato ottenere un valore di Chi-quadro che ci dica che non c'è indipendenza tra i caratteri è facile, ma bisogna capire se questo allontanamento all'atto pratico può avere una rilevanza oppure no.

Odds ratio (θ):

In una tavola 2x2 → $\boldsymbol{\theta = \frac{(odds\ riga\ 1)}{(odds\ riga\ 2)}}$

Proprietà dell'odds ratio:

1) Il suo valore non dipende dalla variabile scelta come esplicativa o risposta → Es: l'odds ratio stimato che chi beve alcolici sia un fumatore è: (1449/500)/(46/281) = 2,90/0,163 = 17,7. Ottengo 17,7 indipendentemente che scelga il fumo o l'alcol come variabile esplicativa;
2) Assume valori positivi ed è pari a 1,0 in assenza di associazione, mentre l'associazione tra le due variabili che stiamo considerando è tanto maggiore quanto più il suo valore si allontana da 1,0;
3) Può essere calcolato come rapporto dei prodotti in croce nella tavola → nell'esempio su alcol e fumo: θ = (1449)(281)/(46)(500) = 17,7;
4) E' un rapporto tra odds e non tra proporzioni come lo è il rischio relativo.

Limiti del test Chi-quadro:

-Misura esclusivamente l'evidenza a favore della presenza di associazione tra le variabili;
-Non ci dice nulla riguardo alla struttura dell'associazione (per cui usiamo i residui standardizzati);
-Non ci dice niente riguardo alla forza dell'associazione (che può essere evidenziata dalle differenze tra proporzioni, dai rapporti tra proporzioni o dagli odds ratio).

N.B: Elevati valori del Chi-quadro e piccoli P-valori indicano forte evidenza di associazione, ma non necessariamente una forte associazione.

Effetto di n sul valore del Chi-quadro per un dato grado di associazione:

Il χ^2 raddoppia se le unità raddoppiano; quando aumentiamo la dimensione campionaria riusciamo ad avere risultati significativi per qualsiasi cosa, ma non conta che sia significativa, quello che conta è la dimensione dell'effetto, perché potrebbe venire significativa anche se le differenze sono irrisorie. Con n sufficientemente grande è possibile ottenere un alto valore di χ^2 (e quindi un piccolo P-valore) anche in corrispondenza di debole associazione.
Quando il p-valore è 0,0000 vuol dire che è talmente piccolo che possiamo non scriverlo.
Rischi relativi → rapporti → es. Quanto il fumo aumenta i rischi di assunzione di alcol.
Rapporto tra due probabilità che mi dà come risultato un'altra probabilità:
0,97/0,64 → rapporto tra probabilità tra fumatori e non → ho probabilità di bere alcol se fumo = 1,5 superiore rispetto a quelli che non fumano.

Associazioni tra variabili ordinali (si possono ordinare dal più basso al più alto o viceversa):

-Il χ^2 si basa su frequenze osservate e frequenze attese, sotto H0. Non utilizza le modalità delle variabili in tabella;
-Altre misure di analisi sono possibili se le variabili sono di natura ordinale o quantitativa;
-Per le variabili quantitative si usa l'analisi della regressione e correlazione, mentre per quelle ordinali esistono misure dell'associazione che si basano sui concetti di:
1) **Concordanza** → c'è tra due caratteri per tutti i soggetti che occupano una posizione elevata (o non elevata) per ambedue i caratteri. Prefigura un'associazione positiva.
2) **Discordanza** → per tutti i soggetti che occupano una posizione elevata per uno dei due caratteri e non elevata per l'altro. Prefigura un'associazione negativa.

Due caratteri ordinali in una tavola di contingenza hanno un'associazione positiva se nella tavola le coppie di soggetti concordanti sono in prevalenza, associazione negativa se prevalgono le coppie discordanti, nessuna associazione se le coppie di concordanti e di discordanti si compensano.

Conteggio delle coppie concordanti (C) e discordanti (D):

Per le coppie concordanti parto dal primo valore in alto a sx (perché è quello che presenta la massima concordanza) della tavola (poco felici e reddito basso) → quanti altri formano coppie concordanti con loro? Quelli che hanno felicità maggiore e reddito maggiore: elimino la riga e la colonna relativa al valore di partenza, perché hanno lo stesso basso reddito e bassa felicità. Quindi: 21 x (84 + 45 + 29 +27) → queste sono tutte coppie concordanti. Faccio lo stesso procedimento anche per quelli abbastanza felici con un reddito sotto la media; con loro concordano quelli che hanno un reddito superiore e un livello di felicità inferiore: tolgo i risultati di riga e colonna (53 concorda con 45 e 27), quindi al risultato precedente aggiungo 53 x (45 + 27). Con 19 non posso mettere nessuno che concorda con loro, trovo solo quelli che hanno reddito superiore, ma per essere coppie concordanti devono crescere entrambi i caratteri e siamo arrivati al grado di massima felicità. Passo al valore 13 → 13 x (29 + 27). Il valore 84 concorda solo con 27.
Per le coppie discordanti si parte dal primo valore in alto a dx (perché è quello che presenta la massima discordanza), e con lo stesso procedimento elimino la riga e la colonna corrispondente: 19 x (13 + 84 +5 +29) e ripeto il procedimento con 53, 45 e 84.
Nella tabella c'è una prevalenza di coppie concordanti → l'associazione è di tipo positivo.
Un indice statistico può scaturire dal confronto (differenza o rapporto) tra C e D.

31/10/19

Indice gamma ($\hat{\gamma}$):

E' dato dal rapporto tra la differenza tra C e D e la loro somma:

$$\hat{\gamma} = \frac{C - D}{C + D}$$

Ha dei limiti prefissati (come il coefficiente di correlazione), il ***risultato*** che viene fuori è **compreso tra –1 e +1**; se le coppie concordanti sono meno di quelle discordanti il numeratore è negativo, mentre è uguale a 0 se C = D → in questo caso c'è assenza di associazione ma non necessariamente indipendenza statistica tra le variabili
I due massimi 1 (completa concordanza, corrisponde a una retta che passa per tutti i punti che rappresentano le coppie di valori osservati) **e –1** (completa discordanza), **si raggiungono quando nella tabella tutte le coppie sono concordanti oppure quando tutte sono discordanti**.
Il segno + o – mostra se l'associazione è positiva o negativa.
Più è alto il valore assoluto di $\hat{\gamma}$ e più forte è l'associazione.

Associazione e causalità:

Per interpretare l'associazione tra due variabili è normale assegnare ad una il significato di variabile risposta e all'altra quello di variabile esplicativa. Questo porta inevitabilmente a pensare che l'esplicativa abbia un'influenza sulla variabile risposta e talvolta che ne sia la causa, ma la presenza di associazione tra due variabili non implica un rapporto di causa-effetto tra le stesse.
Quando si ritiene che una variabile eserciti un'influenza su un'altra si utilizza la seguente notazione:
X Y
Per poter stabilire che esista un rapporto di causa-effetto, devono essere soddisfatte tre condizioni:
1) Associazione tra variabili;
2) Un appropriato ordine cronologico;
3) Assenza di spiegazioni alternative.

Anche se non è mai possibile provare in modo definitivo un nesso di causa-effetto, è possibile confutare un nesso di causalità mostrando che l'evidenza empirica contraddice una delle tre condizioni necessarie.
Per stabilire un nesso causale è necessario poter eliminare spiegazioni alternative a quella ipotizzata: per far questo valutiamo se l'associazione tra due x e y permane anche quando rimuoviamo l'influenza di altre variabili. Nell'associazione tra variabili diciamo che una variabile è controllata quando la sua influenza è rimossa; nell'ambito della ricerca sociale, controllare una variabile significa normalmente raggruppare le osservazioni che per essa presentano la stessa modalità, quindi verificare se all'interno dei gruppi così ottenuti è ancora presente l'associazione tra le variabili che stiamo studiando.
Una variabile la cui influenza è provata ma non misurata, viene definita ***variabile in agguato*** o ***variabile confondente***; se studio l'influenza di X Y, ma so che Z ha molta influenza su Y, ma che ha anche una qualche associazione con X, appare evidente che l'influenza di X su Y può essere in parte (o anche del tutto) spiegata da Z. Questa è una situazione molto frequente nella ricerca medica, dove è doveroso tenere conto dell'età nello studio dei fattori di rischio per determinate patologie.

Tipi di relazioni multivariate:
Nella ricerca sociale è normale spiegare una variabile risposta in funzione di diverse esplicative; le relazioni multivariate che possono presentarsi appartengono alle seguenti tipologie:
1) Associazioni (correlazioni) spurie;
2) Relazioni concatenate;
3) Cause multiple;
4) Variabili sopprimenti;
5) Interazioni.

Associazione spuria tra X_1 e Y:

L'associazione è dovuta all'effetto che una terza variabile X_2 ha sia sulla variabile esplicativa X_1, sia sulla variabile risposta Y.
N.B: Controllando per X_2 l'associazione scompare.
Es. In una serie numerosa di incendi si è rilevata una correlazione positiva tra il numero delle vittime (X_1) e il numero di pompieri (Y) impegnati nello spegnimento dell'incendio (X_2): si potrebbe paradossalmente pensare che per ridurre il numero delle vittime fosse sufficiente ridurre il numero dei pompieri; in realtà vittime e pompieri sono correlati con la dimensione dell'incendio.

Relazioni concatenate:
X_1 è solo causa indiretta di Y attraverso la sua influenza su X_2, detta variabile interveniente o mediatrice che, a sua volta, influenza Y.

X_1 X_2 Y

Es. Studi sulla longevità umana hanno rilevato un'associazione positiva tra lunghezza della vita e livello d'istruzione. Si è propensi a pensare che nella maggior parte dei casi la variabile reddito intervenga a determinare questa associazione:
Istruzione → Reddito → Durata vita.
Controllando per "Reddito", l'associazione scompare.

Cause multiple
Una variabile Y può avere più di una causa

Se le variabili X_1 e X_2 sono statisticamente indipendenti, il controllo per X_2 lascia inalterata la relazione X_1,Y. Nella ricerca sociale, le variabili esplicative sono spesso associate: es. X_1 può avere sulla Y anche un'influenza indiretta:
Il controllo per X_2 può alterare la relazione X_1, Y.

Variabili sopprimenti:

In alcune situazioni due variabili possono risultare non associate fino a quando non si inserisce nella relazione una terza variabile.

Interazione:

Si dice che c'è interazione tra la variabile X_1 e la variabile X_2 se la relazione tra X_1 e Y varia al variare dei possibili livelli di X_2. Il grafo che segue mostra chiaramente che la variabile X_2 influisce direttamente sulla relazione tra X_1 e Y.

Un esempio è rappresentato dagli effetti che l'età può avere sulla relazione tra fattori di rischio e patologie: tra numero di sigarette fumate e insorgenza di tumore polmonare la relazione è nulla tra i giovani e molto forte tra gli anziani.

Sintesi delle relazioni multivariate:

Per le relazioni spurie (X_2 influenza sia X_1 che Y) e per le relazioni a catena (X_2 interviene tra X_1 e Y), l'associazione tra X_1 e Y scompare quando controlliamo per X_2. In presenza di cause multiple, un'associazione può cambiare quando vi è un controllo ma non scomparire. Quando c'è una variabile sopprimente un'associazione compare soltanto in presenza di un controllo; quando c'è un'interazione statistica, un'associazione ha differente forza e/o direzione in relazione a diversi livelli della variabile di controllo.

14/11/19

Regressione e correlazione multipla:

Regressione lineare semplice → mette in relazione lineare una variabile risposta con una esplicativa (predittore o regressore), può essere scritto come segue:

$$\mathbf{E(y) = \alpha + \beta x}$$

Dove:
y è il valore atteso;
α è un valore che assume y quando x = 0, non si può realizzare nella pratica;
β ci dice di quanto varia la variabile risposta ad una variazione unitaria della variabile esplicativa.
Il modello lineare è una retta che passa tra questi punti.
Nella pratica ci sono altre variabili che influenzano la variabile risposta; si può generalizzare il modello binario tramite il **modello di regressione multipla**:

$$\mathbf{E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k}$$

Sono **k** le variabili che si ritengono collegate linearmente con la variabile risposta (cioè sono le variabili esplicative), e i β_k sono i ***coefficienti di regressione parziali*** che indicano sempre quanto varia la y a seguito di un incremento unitario della x, per una qualsiasi combinazione di valori costanti delle altre variabili (ovvero controllando per le altre variabili).

Regressione multipla:

Occorre non inserire variabili troppo correlate tra loro nella regressione multipla; in essa gli effetti dei vari coefficienti possono essere definiti "netti" in quanto non risentono degli effetti delle altre variabili nel modello, diversamente dalla regressione binaria in cui gli effetti di altre possibili

variabili rispetto a quella del modello sono completamente ignorati. L'effetto parziale di x_1 (tenendo sotto controllo x_2) è uguale all'effetto della stessa variabile nella regressione binaria solo se la correlazione tra x_1 e x_2 è uguale a 0. Questo è auspicabile anche se difficilmente realizzabile nelle indagini osservazionali.
L'effetto parziale di ciascuna variabile esplicativa in questo modello di regressione è lo stesso per qualsiasi valore fisso dell'altra variabile.
Questo parallelismo delle due linee rette mostra l'assenza di effetto del valore di x_2 sulla relazione tra y e x_1: in questo caso parliamo di un modello con assenza di interazione tra le due variabili esplicative. Se vi fosse un'interazione, questo parallelismo non sarebbe presente e il modello andrebbe modificato:
$E(y) = \alpha+\beta_1x_1+\beta_2x_2 \rightarrow y = \alpha+\beta_1x_1+\beta_2x_2+\varepsilon$ dove $\varepsilon = y-E(y)$ è un errore che ha $E(\varepsilon) = 0$.

Equazione di previsione:
Si usa il metodo dei minimi quadrati per stimare i parametri del modello a partire dai dati campionari; i valori dei parametri minimizzano la **somma dei quadrati dei residui (SSE)**.
SSE = Σ(y osservati – y previsti)^2 = Σ(y – $\hat{y}$)^2
Stimati i parametri, possiamo scrivere l'equazione di previsione dei minimi quadrati:
$\hat{y} = a + b_1x_1 + b_2x_2 \ldots + b_kx_k$

19/11/19

Correlazione multipla (R) e coefficiente di determinazione multiplo (R^2):
Ci permettono di dire in che misura le variabili esplicative scelte per il modello riescano a prevedere i valori della variabile risposta.
R è la correlazione tra i valori y osservati e i corrispondenti valori ricavati dall'equazione lineare di previsione: $\hat{y} = \alpha+\beta_1x_1+\ldots+\beta_kx_k$ cioè la correlazione ordinaria tra le coppie di valori osservati della variabile risposta, y, e quelli teorici $\hat{y}$, provenienti dal modello lineare di previsione per ciascuna delle n unità del campione osservato.
R^2 ci dice quanta parte della variabilità complessiva viene spiegata dal modello lineare: se ne spiega il 100%, passa per tutti i punti che abbiamo disegnato.
Infatti consiste nella riduzione relativa (quota di riduzione) nell'errore totale che si ottiene stimando (cioè prevedendo) la y (variabile risposta) mediante il modello lineare anziché utilizzando soltanto la sua media campionaria.

$$R^2 = \frac{TSS - SSE}{TSS} = \frac{\Sigma(y - \bar{y})^2 - \Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2}$$

TSS = somma degli scarti di ciascun valore osservato dalla media (TSS = SSE + RSS);
SSE = somma dei residui;
TSS – SSE → riduco TSS di una quota. Corrisponde alla somma degli scarti di ciascun valore teorico dalla media → è chiamato ***somma dei quadrati di regressione*** e rappresenta la variabilità dei valori y previsti o spiegati dal modello.
La media dei valori previsti è uguale alla media dei valori osservati per il metodo dei minimi quadrati, quindi posso scrivere $\bar{y}$ o $\hat{y}$.
Se si aggiunge una variabile nel modello, si riduce l'errore che si commette, quindi aumenta R^2; più sono le variabili più cresce R^2. La media dei quadrati è il rapporto tra regressione e gdl, e residui e gdl; F è la F di Fischer, la distribuzione campionaria del rapporto delle medie dei quadrati nell'ipotesi che il coefficiente di regressione sia 0. La significatività sta per il p-valore e nel nostro caso mostra un'evidenza molto forte contro l'ipotesi nulla.

21/11/19

Proprietà di R e R^2:

1) $0 \leq R^2 \leq 1$

2) R = + $\sqrt{R^2}$ → $0 \leq R \leq 1$ → R non può essere negativo

3) Maggiore è il loro valore e migliore è la capacità predittiva delle variabili esplicative del modello.

4) R^2 =1 quando tutti i valori y osservati sono uguali a quelli previsti; in questo caso abbiamo ovviamente SSE = 0.

5) R^2 = 0 quando tutti i valori previsti $\hat{y}$ sono uguali alla media $\bar{y}$ e, quindi, TSS = SSE. Quando accade questo, $b_1 = b_2 = ... = b_k = 0$ (cioè sono uguali a zero anche tutti i coefficienti di regressione parziali) e, ovviamente, anche il coefficiente di correlazione r tra la y e ciascun regressore x è uguale a 0.

6) R^2 non può diminuire se aggiungiamo una variabile esplicativa al modello, tuttavia se ne aggiungiamo una che è fortemente correlata con una o più variabili già presenti nel modello, l'R^2 potrebbe rimanere invariato o quasi. Questa situazione è detta **multicollinearità** → aggiungo una variabile che però spiega una variabilità già spiegata da altri.

7) R^2 è additivo: è uguale alla somma degli r^2 bivariati quando le coppie di variabili esplicative sono tra loro incorrelate (questo sarebbe l'ideale ma in pratica non avviene mai). Ciò sarebbe sempre auspicabile, ma difficilmente si realizza nelle indagini osservazionali.

8) Lo stimatore campionario di R^2 è distorto e tende a sovrastimare il valore di R^2 nella popolazione. Il software riporta il calcolo dell'R^2 corretto: si tratta in realtà di un coefficiente che è affetto da una distorsione minore e che ha un valore di norma inferiore a quello dell'R^2 classico.

Inferenza sui parametri del modello di regressione multipla:

Si basa su una serie di assunzioni:

-Il modello deve essere approssimativamente adeguato → il modello che uso (es. lineare, logistico) è giusto per quello che sto studiando?;

-La distribuzione nella popolazione della y condizionatamente a ciascuno dei regressori $x_1,\ldots, x_k$ è normale;

-La deviazione standard della distribuzione condizionata della y è costante per ogni combinazione di valori $x_1,\ldots,x_k$;

-Il campione è selezionato in modo casuale.

Queste assunzioni raramente sono completamente soddisfatte, tuttavia se la prima è rispettata, l'inferenza di tipo bidirezionale è robusta rispetto a violazioni non pesanti di normalità e di variabilità costante della distribuzione condizionata della y.

Test sull'inferenza complessiva delle variabili esplicative:

-I coefficienti di regressione esprimono l'influenza delle variabili esplicative sulla variabile risposta, l'inferenza può riguardare la loro influenza complessiva o quella di ciascuna variabile;

-Per valutare l'influenza complessiva si formulano le seguenti ipotesi:

H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$

Ha: almeno un $\beta i \neq 0$; (i=1,...,k)

(si può scrivere anche come H_0: $R^2 = 0$; Ha: $R^2 >$ di 0) → entrambe indicano ipotesi nulla di indipendenza lineare tra le variabili esplicative e la variabile risposta → ovvero che y sia linearmente indipendente da ciascuna delle variabili esplicative contro l'alternativa che almeno una di queste abbia influenza sulla y.

Il test da usare per la verifica di queste ipotesi è la **statistica F** (vedi formula in cui k è il numero di

variabili esplicative del modello; k+1 = numero complessivo di parametri nel modello). Quando H_0 è vera, si conosce la distribuzione campionaria della statistica F. Maggiori valori di R (o di R^2) producono maggiori valori di F, cioè a maggior evidenza contro l'ipotesi nulla.

Proprietà della distribuzione F:

Formula test F:

$$F = \frac{R^2/k}{(1-R^2)/[n-(k+1)]}$$

-Assume solo valori positivi o nulli;
-La forma della distribuzione è asimmetrica positiva (graficamente assomiglia al Chi-quadro);
-La sua media è approssimativamente uguale a 1 (l'approssimazione migliora al crescere di n);
-La distribuzione dipende da una coppia di gdl:
$gdl_1 = k$ → numero di variabili esplicative;
$gdl_2 = [n - (k+1)]$ → dimensione del campione – numero dei parametri nel modello.

Inferenza per singoli coefficienti i regressione:

Le variabili esplicative del modello sono tutte necessarie? Occorre valutare l'effetto parziale di ciascuna variabile esplicativa (xi), controllando quindi per le altre variabili → si sottopone a test l'ipotesi nulla utilizzando il test t; se H_0 è vera, questa statistica è una t con n – (k + 1) gdl.
$H_0: \beta i = 0$
$t = (bi - 0)/se$ con $gdl = n - (k+1)$
Un altro modo per fare inferenza sul parametro → si può costruire un IC che è dato dalla stima del coefficiente di regressione parziale ± M.E; la tavola da usare è sempre la t di Student. Se il test è significativo, vuol dire che l'IC non comprende il valore.
IC = bi ± t(se) con il t critico in corrispondenza di gdl = n – (k+1).
I principali software per la regressione forniscono: stime dei coefficienti, errore standard, t-test e p-valore, usualmente riferito al test bidirezionale.

28/11/19

Attenzione alla multicollinearità:

Non possiamo passare direttamente al t-test sui singoli coefficienti perché è possibile che dal test F si ottenga un piccolo P−valore mentre nessuno dei test t produce un analogo risultato. E' ugualmente possibile che si ottenga un piccolo P−valore nella regressione binaria per una variabile esplicativa e non altrettanto se questa variabile viene controllata per altre variabili; ciò accade in presenza di multicollinearità. In questo caso la variabilità parziale spiegata da una singola variabile è piccola (il valore di una esplicativa può essere facilmente previsto dalle altre esplicative).
Ovviamente assurdo, ma chiarificatore è il modello: y= statura; x_1= lunghezza gamba destra; x_2= lunghezza gamba sinistra.
In presenza di multicollinearità:
-L'errore standard di ciascun coefficiente di regressione può essere abbastanza elevato e, conseguentemente, non significativo il relativo t-test;
-R^2 può mantenersi elevato anche se vengono eliminate una o più variabili esplicative (perché quella che abbiamo messo e poi tolto non modificava nulla);
-E' consigliabile eliminare le variabili esplicative che non apportano sostanziali benefici al modello (un modello è migliore tanto quanto ha meno variabili, perché è più facile farne un'interpretazione).
Esistono strumenti diagnostici (VIF - fattori di incremento della varianza) per valutare la presenza e l'entità della multicollinearità.

Interazione tra i regressori (cioè le variabili) del modello:

Il modello $\mathbf{E(y) = \alpha+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k}$ dove le x sono variabili esplicative assume che ciascun coefficiente di regressione parziale βi resti costante per qualunque combinazione dei rimanenti regressori x_j;($j \neq i$). Ciò equivale a assumere che nel modello non vi sia interazione tra le variabili esplicative (come nell'esempio sul disagio mentale, eventi e SES). Se c'è interazione tra le variabili x_1 e x_2 allora l'effetto di x_1 sulla variabile dipendente y, può variare al variare di x_2.

Il più semplice modello d'interazione → Es. k = 2 → $E(Y) = \alpha+\beta_1x_1+\beta_2x_2+\beta_3(x_1x_2)$ può essere considerato come un caso speciale del modello: $E(Y) = \alpha+\beta_1x_1+\beta_2x_2+\mathbf{\beta_3x_3}$ **con $x_3 = x_1x_2$.**

Per vedere l'interazione (es. tra eventi e stato socio – economico) inserisco un altro termine (x_3) chiamato termine d'interazione.

Se inseriamo il termine d'interazione nel modello sul disagio mentale R^2 aumenta, passa da 0,339 (senza interazione) a 0,347 e otteniamo la seguente espressione:

$\hat{y} = 26+0{,}156x_1-0{,}060x_2-0{,}00087x_1x_2$

Al crescere della variabile x_2 decresce l'effetto della variabile x_1 sulla variabile risposta y.

Devo inserire tutte le variabili che credo abbiano effetto sulla variabile risposta, sperando che ci sia poca multicollinearità.

Test sul termine d'interazione:

Per sottoporre a test l'ipotesi nulla → H_0: non c'è nessuna interazione nel modello

$E(y) = \alpha+\beta_1x_1+\beta_2x_2+\beta_3x_1x_2$, si sottopone a test H_0: $\beta_3 = 0$, per mezzo del test $t = b_3/se$.

Nell'esempio sul disagio mentale: $t = -0{,}00087/0{,}0013 = -0{,}67$ $gdl = n - 4 = 36$, $P-valore = 0{,}51$ per H_a: $\beta \neq 0$ evidenza insufficiente per concludere che ci sia interazione, ma con il data set completo, di oltre 1000 osservazioni, si ottiene un valore significativo del test t. In un modello con k > 2 l'interazione può riguardare tutte le possibili coppie di variabili:

$E(y) = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_1x_2 + \beta_5x_1x_3 + \beta_6x_2x_3$

Significatività statistica → la trovo sempre aumentando n;

Significatività effettiva → mi dice realmente qualcosa rispetto a quello che sto studiando; quando vengono moltiplicati per un regressore è difficile interpretare i coefficienti.

Confronto tra modelli:

Per valutare il miglior modello da usare confrontiamo per esempio i seguenti modelli:

$E(y) = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_1x_2 + \beta_5x_1x_3 + \beta_6x_2x_3$

$E(y) = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3$

E sottoponiamo a test l'ipotesi nulla di assenza di interazione formulandola come segue:

H_0: $\beta_4 = \beta_5 = \beta_6 = 0$

Il confronto tra modelli può essere effettuato con un test F confrontando le SSE dei due modelli o, equivalentemente, i loro R^2. Il modello più complesso (*modello completo*), sarà migliore se il valore di SSE sarà sufficientemente inferiore a quello del modello più semplice (*modello ridotto*) o, equivalentemente, se il suo R^2 sarà sufficientemente maggiore. Indichiamo rispettivamente con SSE_c e con SSE_r le devianze di errore dei modelli completo e ridotto e utilizziamo in modo analogo la notazione R^2_c e R^2_r per i loro R^2.

$$F = \frac{(SSE_r - SSE_c)/gdl_1}{SSE_c/gdl_2} = \frac{(R_c^2 - R_r^2)/gdl_1}{(1 - R_c^2)/gdl_2}$$

gdl_1 = **differenza** tra il numero dei parametri nei due modelli;

$gdl_2 = n - (k + 1)$ con k = numero di variabili esplicative del modello completo.