Lezione 1 - Stat model

5-03-2018

prof. Vittadini

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

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

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

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

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

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

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

Come nasce l'elaborazione di un modello statistico?

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

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

Ripetere più volte (se necessario).

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

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

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

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

Regressione multipla

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

RIVEDERE E COMPLETARE

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

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

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

Quindi:

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

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

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

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

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

$$\underline{\underline{Y}} = \underline{\underline{X}} \cdot \underline{b} + \underline{\varepsilon} \tag{3}$$

con

$$Y = \begin{pmatrix} y_1 \\ . \\ . \\ . \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,k} \\ x_{2,1} & \ldots & \ldots & \\ \vdots & & \ddots & \vdots \\ x_{n,1} & \ldots & & x_{n,k} \end{pmatrix} \quad b = \begin{pmatrix} b_0 \\ . \\ . \\ . \\ b_n \end{pmatrix} \quad \epsilon = \begin{pmatrix} \varepsilon_0 \\ . \\ . \\ . \\ \varepsilon_n \end{pmatrix} \tag{4}$$

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

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

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

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

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

$$\underline{e} = \underline{y} - \underline{y'} = \underline{y} - \underline{\underline{X}} \cdot \underline{b} \tag{5}$$

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

$$y'_i = b_0 + b_1 x_{1i} + b_2 \cdots + b_n x_{ni} \quad \text{per } i = 1, \ldots, k \tag{6}$$

Quindi il singolo residuo nella (5) si può anche riscrivere come:

$$e_i = y_i - y'_i = y_i - b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_n x_{ni} \quad \text{per } i = 1, \ldots, k \tag{7}$$

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:

$$\Phi(\underline{b}) = \sum_{i=1}^{k} (y_i - y'_i)^2 = \sum_{i=1}^{k} e_i^2 = \underline{e}^t \cdot \underline{e} = (\underline{y} - \underline{\underline{X}} \cdot \underline{b})^t \cdot (\underline{y} - \underline{\underline{X}} \cdot \underline{b}) \tag{8}$$

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

$$\frac{\partial \Phi(\underline{b})}{\partial \underline{b}} = 0 \tag{9}$$

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

$$\underline{\underline{X}}^t \underline{\underline{X}} \cdot \underline{b} = \underline{\underline{X}}^t \underline{y} \tag{10}$$

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 $\underline{b}$:

$$\underline{b} = (\underline{\underline{X}}^t X)^{-1} \underline{\underline{X}} \cdot \underline{y} \tag{11}$$

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:

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

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:

$$\sum_{i=1}^{k} (y_i - y'_i)^2 = \sum_{i=1}^{k} (y_i - \underline{\underline{X}} \cdot \underline{b})^2 \tag{13}$$

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

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

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

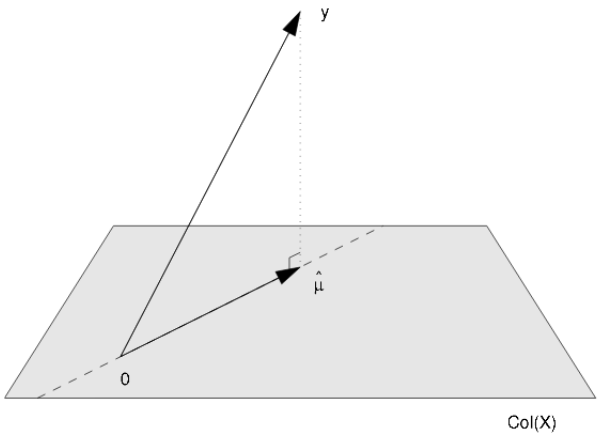

Figure 1: Rappresentazione dello spazio delle colonne e dei vettori di interesse.

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

L'indice di adattamento è definito come:

$$R^2 = \frac{DevSpieg(Y)}{DevTot(Y)} \tag{15}$$

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

$$DevSpieg(Y) = b_1 Codev(X, Y) \tag{16}$$

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

si arriva a:

$$R^2 = \frac{Cov(X,Y)}{var(X)var(Y)} \tag{17}$$

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.

Modello lineare classico

Estensione della regressione al caso stocastico. Regole **necessarie** per avere un modello (che valgono sempre):

- **Ipotesi di linearità**.
- **Ipotesi di non sistematicità degli errori**. Vale per tutti i modelli, l'errore è considerato casuale ed ha valore atteso 0 (si aggira lì intorno), altrimenti è sistematico.
- **Sfericità degli errori** + Non collinearità. Affinchè matrice dei minimi quadrati abbia un'inversa unica $x'x$ deve essere uguale al rango. Se c'è collinearità non ho soluzione unica del modello, questo perchè due variabili sono appunto colineari (si è inserita due volte la stessa variabile oppure una combinazione linearmente dipendente). + Numerosità. Il numero di elementi nel campione che si sta utilizzando deve essere maggiore del numero dei parametri.

+

+ **Sfericità degli errori**. Per quanto riguarda la Non collinearità, affinchè matrice dei minimi quadrati abbia un'inversa unica $x'x$ deve essere uguale al rango. Se c'è collinearità non ho soluzione unica del modello (non ho inversa unica), questo perchè due variabili sono appunto colineari (si è inserita due volte la stessa variabile oppure una combinazione linearmente dipendente). + Numerosità. Il numero di elementi nel campione che si sta utilizzando deve essere maggiore del numero dei parametri.

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:

$$y = \beta_0 + \beta_1 \cdot \textit{tempo di studio} \tag{18}$$

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 (18), 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:

$$P = \exp(\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots) \tag{19}$$

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:

$$P = \frac{\exp(y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots)}{\exp(y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots) + 1} \tag{20}$$

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

$$\log(\frac{P}{1-P}) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots \tag{21}$$

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 (21), detta anche *logit*, per cui per ottenere i valori di probabilità associati a determinati valori delle variabili indipendenti che si sono utilizzate è necessario applicare la formula inversa. nei software è spesso anche possibile impostare una soglia per la probabilità con la qual stabilire e quindi predire se una certa osservazione con una certa probabilità appartenga ad una determinata classe. In questo modo si possono poi calcolare delle contingency table dalle quali vedere quanti valori sono stati correttamente predetti e quanti invece no.

Violazioni del modello lineare classico

Eteroschedasticità

L'ipotesi di omoschedasticità suppone che il termine di errore sia uguale per tutte le variabili indipendenti. Però ci possono essere situazioni in cui ciò non è vero, ovvero situazioni in cui il termine di errore varia tra le diverse variabili indipendenti, in questo caso se osserviamo uno scatter plot dei residui in funzione dei valori predetti per la variabile dipendente osserviamo un tipico andamento a cono. Un altro andamento anomalo di qeusto tipo 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 (Ordinary Least Squares) cioè il metodo 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?

Valgono ancora le ipotesi di correttezza e linearità.
È ancora consistente.
Non è più uno stimatore
In questo caso lo stimatore non è più efficiente.

La statistica t di Student non può approssimare nel modo corretto la varianza σ^2 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 alla base di test come l'analisi della varianza ANOVA (Analysis Of Variance) e il t-test di Student.

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 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:

$$H_0 = \mu \tag{22}$$

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:

$$H_1 \neq \mu \tag{23}$$

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 di ogni osservazione assumendo che l'ipotesi nulla sia vera ed in seguito per calcolare il p-value bisogna calcolare la probabilità di ottenere un valore che sia più grande del valore osservato. Quindi l'area sottesa alla parte di curva a destra del valore osservato.
FIGURAAAA 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.