Linee guida per prova d'esame di statistica per il marketing, per argomenti

Regressione

Si tratta di prevedere una variabile quantitativa in funzione lineare di altre (*k*-1), anzi, problema di "model selection", di sceglierne s≤(k-1), le quali spieghino al meglio la variabile dipendente.

Questi i passi principali:

- Descrizione e presentazione del dataset, e sua preparazione – eliminazione dati mancanti, variabili quasi collineari, etc
- Selezione delle variabili esplicative, due approcci:

a) Top down
   1. Si considera il modello con tutte le k-1 variabili
   2. Si escludono quelle il cui coefficiente di regressione risulta non significativamente diverso da zero (test *t*)
   3. Con altre considerazioni, che dipendono anche dalla sensibilità del ricercatore, si perviene ad un modello con un numero s<k-1 variabili esplicative, che si ritiene candidato a prevedere la variabile dipendente
   4. Si effettua il test *F* che verifica l'ipotesi nulla che valga il modello con *s* variabili, contro l'alternativa che valga il modello con tutte le k-1 variabili. Si accetta il modello associato all'ipotesi ritenuta vera.

b) Bottom up
   1. Si considera il modello che prevede la variabile dipendente in funzione lineare di quella che ha coefficiente di correlazione più elevato con essa
   2. E via via si aggiungono ad una ad una le variabili con più alta correlazione (sarebbe meglio considerare le

correlazioni parziali, meno bene utilizzare le correlazioni grezze) sino a pervenire ad un modello tale che aggiunta di altre variabili parrebbe non migliorare la previsione

3. Si effettua il test F che verifica l'ipotesi nulla che valga il modello di cui al punto 2., contro l'alternativa che valga il modello con tutte le k-1 variabili. Si accetta il modello associato all'ipotesi ritenuta vera.

- Si deve verificare che le ipotesi del modello sui residui siano soddisfatte, in particolare:
  1. Normalità: si costruisce il grafico delle frequenze relative specifiche dei residui, e si vede se imita l'andamento di una normale standard N(0,1), quasi simmetrico unimodale campanulare. Esistono anche dei test di verifica dell'ipotesi, Kolmogorov-Smirnov e altri (si veda SPSS)
  2. Incorrelazione: test Di Durbin-Watson e altri
  3. Omoschedasticità: si vede se in intervalli di valori differenti i residui presentino variabilità differente, se ciò non avviene si può concludere per varianza costante, in linea con le ipotesi del modello.
- Previsione
- Conclusioni scientifiche finali
- Bibliografia, sitografia, appendici.

Warning: se la numerosità del campione è almeno 100, le verifiche d'ipotesi del modello sui residui possono essere omesse.

Cluster analysis

Si consideri una matrice *X* di *n* righe e *p* colonne, composta da *p* variabili quantitative rilevate su *n* unità statistiche.

- Descrizione e presentazione del dataset, e sua preparazione – eliminazione dati mancanti, variabili quasi collineari, etc
- Eventuale riduzione delle variabili:

a) Se *p* ed *n* sono ragionevolmente piccoli, si procede con l'analisi dei cluster
b) Se *p* e/o *n* sono grandi, al fine di far girare gli algoritmi di cluster in tempi ragionevoli sarà necessario ridurre le variabili attraverso l'analisi delle componenti principali, scegliendo le componenti principali ottenute dalla matri ce di correlazione, associate ad autovalori maggiori di uno, valutando la percentuale cumulata di varianza spiegata. L'analisi dei cluster sarà effettuata sulle nuove variabili, le componenti principali (le quali vanno interpretate nel loro legame con le variabili iniziali)

- Determinazione del numero ottimale dei cluster: si applica un metodo di clustering gerarchico, pervenendo ad un dendrogramma, che verrà tagliato in corrispondenza della soluzione "a gomito" (si vedano tecniche opportune), il numero di cluster potrebbe essere anche aumentato o diminuito di 1, e fare confronti di performance.
- Identificazione dei cluster e loro interpretazione: si considerano le sintesi delle distribuzioni delle variabili sulle unità statistiche di tutta la popolazione e i loro principali quantili. Sintesi: min, max, media, scarto quadratico, varianza, asimmetria, curtosi. Quantili principali di ordine 0,05 0,10 0,25 0,5 (mediana) 0,75 0,90 0,95. Queste sintesi e quantili di popolazione vanno confrontati con i corrispondenti delle distribuzioni delle variabili per le unità statistiche di ogni cluster, al fine di avere una descrizione delle caratteristiche delle unità statistiche di cui è composto ogni cluster rispetto

alle variabili. [WARNING: se la clusterizzazione è avvenuta sulle componenti principali, l'interpretazione dei cluster avverrà sulle sintesi delle distribuzioni delle variabili di partenza (più informative) e NON sulle sintesi sulle componenti principali (meno informative)

- Applicazione di una tecnica di clustering non gerarchico:
  1. si considera un numero di cluster determinato attraverso il dendrogramma ottenuto precedentemente dall'applicazione di una tecnica di clustering gerarchico.
  2. Interpretazione dei cluster con le sintesi
- Confronto fra i cluster formati con il metodo gerarchico con quelli formati col metodo non gerarchico
- Conclusioni scientifiche finali
- Bibliografia, sitografia, appendici.

Analisi discriminante

k gruppi, l'*i*-esimo di numerosità n_i, su ognuna delle *n* unità statistiche si rilevano p variabili quantitative.

- Descrizione e presentazione del dataset, e sua preparazione – eliminazione dati mancanti, variabili quasi collineari, etc
- Diagonalizzazione della matrice $S^{-1}S_b$, determinazione degli autovalori (componenti varianza) e autovettori (funzioni discriminati).
- Scelta del numero di funzioni discriminati, in base alla percentuale di varianza cumulata
- Interpretazione logico-statistica delle funzioni discriminanti, in base al loro legame con le *p* variabili
- Tasso di corretta classificazione

- Se vi sono delle osservazioni aggiuntive, imputazione di ugni unità statistiche a un gruppo con confronto con i centroidi di gruppo
- Conclusioni scientifiche finali
- Bibliografia, sitografia, appendici.

Analisi delle corrispondenze

- Descrizione e presentazione del dataset, e sua preparazione – eliminazione dati mancanti, variabili quasi collineari, etc
- Due casi
  1. Bivariata

     Si trovano le soluzioni per i fattori e gli assi fattoriali, e se ne sceglie il numero in base alla percentuale di varianza cumulata. S'interpretano in senso logico-statistico assi e fattori in relazioni alle variabili. Si svolgono commenti appropriati. Qualora si scelgano due componenti, si può considerare il b-plot, e la sua interpretazione. Infine si svolge l'analisi quantitativa dei contributi puntuali e globali di fattori agli assi e di assi ai fattori (si vedano teoria, applicazioni, e manuale SPSS)
  2. Multivariata. Si considera uno dei tre approcci, matrice disgiuntiva, soluzione di Burt, e la terza, e si procede all'interpretazione secondo quando prescritto dalla teoria (si vedano teoria, applicazioni, e manuale SPSS)
- Qualora esistano osservazioni supplementari, si può pensare di imputare ogni singola unità statistica ad un asse o ad un fattore
- Conclusioni scientifiche finali
- Bibliografia, sitografia, appendici.

Si consiglia di esaminare con attenzione i paper e le applicazioni proposti nelle due sezioni dello spazio dedicato al modulo di analisi quantitative di mercato, in modo da trarre ispirazione e impratichirsi, al fine di raggiungere il miglio risultato possibile nella stesura della relazione.