Statistica multivariata

saverio.vicario@ba.itb.cnr.it

Statistica multivariata

- Con i modelli lineari andiamo esplicitamente a lavorare con l'interazioni di diverse variabili aleatorie
- Il test del chi quadro utilizzato nel caso speciale di indipendenza fra variabili discrete era un primo approccio alla statistica multivariata
- Il t-test può essere considerato un caso speciale di test per verificare l'interazione tra una variabile continua e una variabile discreta con due valori possibili (gruppo1 e gruppo 2). Se non c'e' differenza tra i due gruppi non c'e interazione tra le due variabili (es. il diverso insetticida non cambia i valori degli insetti trovati)

Impostazione modelli lineari

- Dato un campione di unita statistiche tratte da una popolazione di interesse misuriamo due attributi. L'insieme delle due misurazioni crea due v.a. X e Y continue.
- Le variabili X e Y provengono da un campionamento accoppiato
- Diversamente dal t test le X e Y non devono essere misurare simili attributi ma possono descrivere attributi arbitrariamente diversi, ma deve essere possibile applicare una trasformazione lineare per passare da una variabile all'altra. L'esatto significato dipende dalla domanda posta

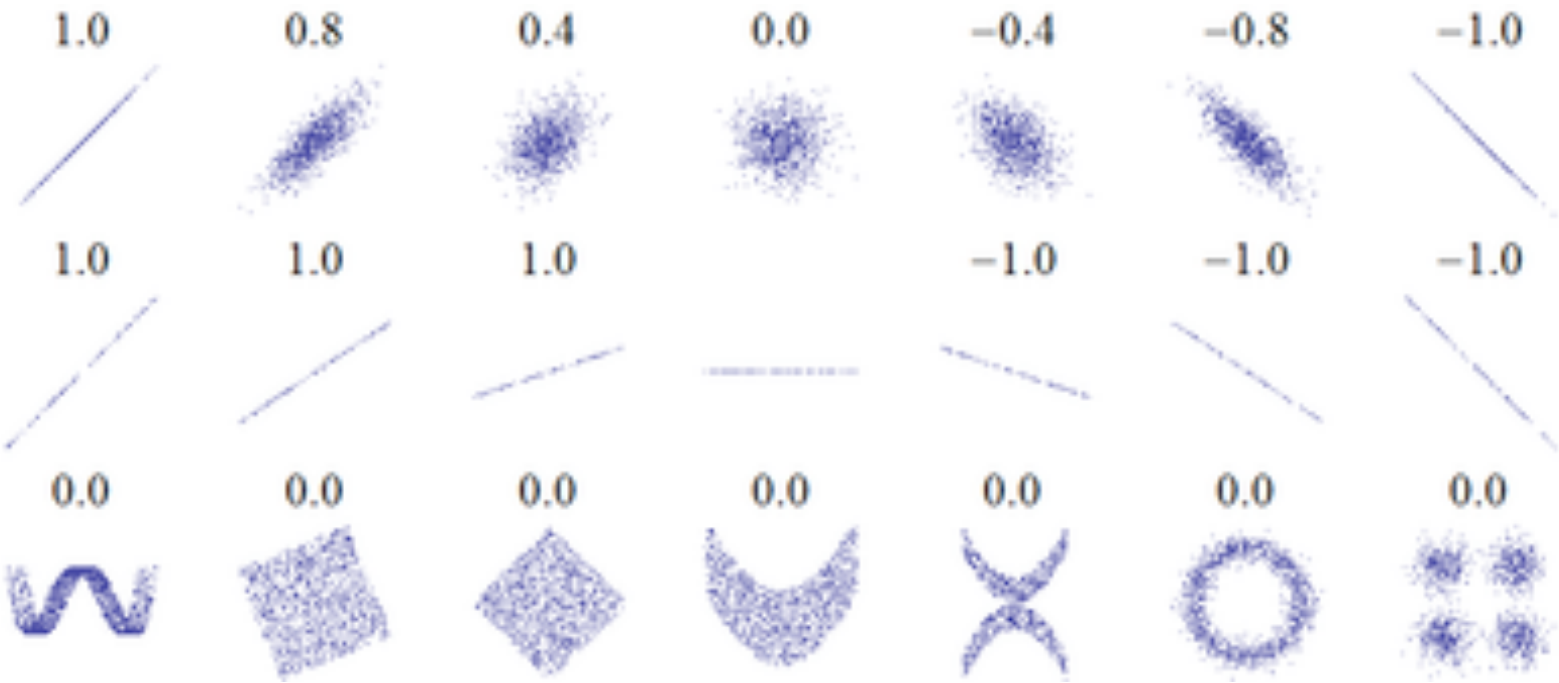
Assunti dei modelli lineari

- L'errore , definito come Y-(aX+c), deve essere distribuito con legge $N(0, \sigma^2)$
- L'errore e' distribuito in maniera omogenea tra tutte le osservazioni
- Non ci sono limitazioni sulla distribuzioni di X e Y

Test di correlazione

- Data l'impostazione precedente ci si può domandare se esiste una correlazione lineare tra due variabili, ovverossia se provengono dalla stessa distribuzione dopo aver applicato una trasformazione lineare arbitraria del tipo
- Y=aX+c
- La statistica del test e' ρ il coefficiente di correlazione di person, espresso come la covarianza diviso il prodotto delle deviazione standard delle due variabili
- $$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n \sqrt{\frac{n^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}}$$

Esempi di valori del coefficiente correlazione



Impostazione test di correlazione

- H0 : (ρ =0) X e Y sono non correlate;
- H1 : (ρ ≠0) X e Y sono correlate.
- Per scalare la statistica ρ con una distribuzione t di student

$$t = \rho\sqrt{\frac{n-1}{1-\rho^2}}$$

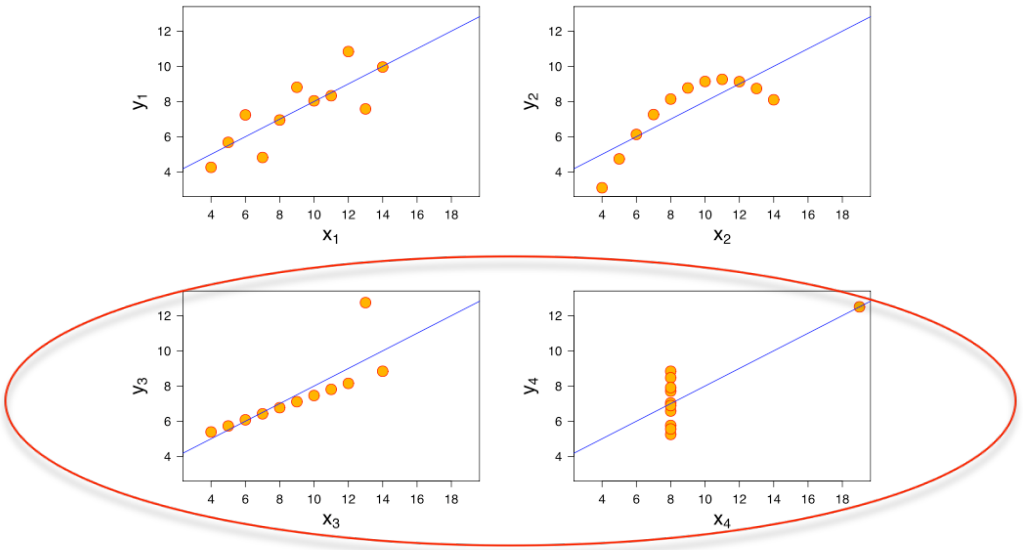
- I gradi liberta come d'uso sono n-2

Esempi

```
> x=rexp(100)
>hist(x)
> y=3*x+0.4 +rnorm(100,sd=3)
> cor.test(x,y)
>plot(x,y)
>y=3*x+0.4 +rnorm(100,sd=12)
> cor.test(x,y)
>y=rexp(100)
> cor.test(x,y)
```

Modello lineare: assunti e limitazioni

- Concentrandosi su trasformazioni lineari il test il coefficiente di correlazione non riassume sempre al meglio la correlazione tra due variabili



Relazione non lineare tra le variabili

Errore distribuito in maniera non omogenea

4 distribuzioni di due variabili con 11 osservazioni accoppiate che hanno lo stesso coefficiente di correlazione (0.86) e dunque danno gli stessi risultati al test di correlazione

Modelli lineari per predizioni

- L'enfasi viene messa sulla possibilità predire i valori di Y conoscendo i valori di X.
- I risultati sono diversi se X e Y sono scambiati
- Con questo quadro interpretativo e' possibili avere diverse variabili X o predittori di natura continua o categorica non ordinata
- E' possibile avere una sola Y o variabile dipendente di natura continua
- $Y=a_1X_1+a_{...}X_{...}+a_pX_p+c$
- Queste limitazioni possono essere superate utilizzando il quadro interpretativo di un modello lineare generalizato

Stima dei coefficienti

- I coefficienti sono definiti in maniera da minimizzare lo scarto quadratico medio
- $E=Y-(a_1X_1+a_{...}X_{...}+a_pX_p+c)$
- Se l'assunto di normalità dell'errore e' mantenuto, minimizzare lo scarto quadratico e' la procedura più efficiente per estrarre l'informazione dai dati.

Impostare una regressione lineare in R

- >summary(iris)
- >iris.lm<-lm(Sepal.Length~Sepal.Width+Species, data=iris)
- >plot(Sepal.Length~Sepal.Width, data=iris, col=Species)
- names(Iris)
- >summary(iris.lm)
- >anova(iris.lm)
- ?lm

Esempio di problemi

- >iris.lm<-lm(Sepal.Length~Sepal.Width, data=iris)
- >summary(iris.lm)
- Perche al crescere della largezza dei sepali cresce la lunghezza (cedi coefficiente negativo di sepal.width)?
- Questo e' difficile da riconciliare con quello che conosciamo dei fiori. Uno stesso tipo di fiori tende a mantenere le stesse proporzioni tra le sue parti, mentre qui le proporzioni cambiano

Controllo degli assunti

- Usando la funzione plot su l'oggetto modello lineare R restituisce una serie di grafici diagnostici per controllare che l'errore sia distribuito con legge normale e in maniera omogenea al crescere della variabile dipendente
- >plot(iris.lm)
- Il primo grafico mostra i residui contro i valori predetti (fitted). Nel grafico vedete che la linea rossa non e' piatta
- Il secondo grafico mostra il QQplot dei residui contro la distribuzione normale dove si identifica una lieve mancanza di scarti negativi rispetto ad una equivalente distribuzione normale
- Il terzo grafico e' simile al primo ma usando scarti standardizzati (radice dello scarto quadratico diviso per la deviazione standard)
- Il quarto grafico identifica quelle osservazioni che influenzano molto il modello (high leverage). Devono essere distribuiti in maniera uniformi tra scarti positivi e negativi.
- Vedi cosa succede in una situazione non aberrante
- >plot(lm(Sepal.Length~Sepal.Width+Species, data=iris))

Uso di variabili categoriche nei modelli lineari

- I modelli lineari sono impostati per v.a. quantitative, meglio se continue. Per poter includere variabili categoriche tra i predittori bisogna effettuare l'artificio di trasformare le v.a. in cosiddetti contrasti.
- Esistono varie procedure riconosciute, ne esploreremo solo una: i contrasti per trattamento.
- La procedura di ricodifica e' molto semplice:
- Si definiscono per ogni livello di una variabile categorica un v.a. discreta con solo due valori 1 e 0 per identificare quali osservazioni posseggono quel livello. Per evitare ridondanza di informazioni, per convenzione, per il primo livello non e' definita una v.a. discreta: ove tutte le altre sono zero vuol dire che quel livello e' stato osservato

Esempi di contrasti per variabili categoriche

Ricodifica di una variabile categorica con tre livelli (‘setosa’, ‘versicolor’, ‘virginica’).

```
>contrasts(iris$Species)
           versicolor virginica
setosa              0         0
versicolor          1         0
virginica           0         1
```

Significato grafico dei modelli lineare e dei contrasti di tipo trattamento

```
>irisS.lm<-lm(Sepal.Length~Sepal.Width+Species, data=iris)
>irisS.lm$coef
>plot(iris$Sepal.Width, iris$Sepal.Length, col=iris$Species)
>abline(a=irisS.lm$coef[1], b=irisS.lm$coef[2])
>abline(a=sum(irisS.lm$coef[c(1,3)]), b=irisS.lm$coef[2],
   col=2)
>abline(a=sum(irisS.lm$coef[c(1,4)]), b=irisS.lm$coef[2],
   col=3)
```

I coefficienti calcolati per ogni contrasto servono a modificare l'intercetta

Significato grafico dei modelli lineare e dei contrasti di tipo trattamento

- E' possibile modificare al il coefficiente moltiplicativo usando i contrasti usando la seguente formula

```
>irisSper.lm<-lm(Sepal.Length~Sepal.Width*Species,
  data=iris)
>summary(irisSper.lm)
>irisSper.lm$coef
>abline(a=irisSper.lm$coef[1],b=irisSper.lm$coef[2],lty=2)
>abline(a=sum(irisSper.lm$coef[c(1,3)]),b=sum(irisSper.lm
  $coef[c(2,5)]),lty=2, col=2)
>abline(a=sum(irisSper.lm$coef[c(1,4)]),b=sum(irisSper.lm
  $coef[c(2,6)]),lty=2, col=3)
```

Metodi di diagnosi

- In una regressione lineare e' possibile oltre che ottenere la migliore trasformazione lineare da X a Y anche testare l'adeguatezza del modello e la probabilità che singoli elementi del modello siano ininfluenti
- Il test dell'adeguatezza generale e' basato sulla identità della varianza totale con la varianza delle predizione del modello tenendo conto che le predizioni sono limitate dai parametri del modello.
- Il paragone delle due varianze scarti quadratici medi viene fatto con la distribuzione F di Fisher che descrive il rapporto fra scarti quadratici ovverossia il rapporto tra due distribuzioni del chi quadro

Considerazioni sull'adeguatezza

- Il valore della statistica ci informa se il modello riesce a descrivere i dati meglio che un modello perfettamente casuale
- Dunque il fatto che il modello sia significativo non vuol dire che le predizioni siano buone o che predittori descrivano bene la variabile dipendente
- Per valutare il potere esplicativo del modello si usa R^2 che non e' altri che il coefficiente di correlazione al quadrato
- Il R^2 rappresenta la varianza spiegata dal modello.
- Se il numero e' prossimo a 1 e' probabile che non ci siano altri predittori da cercare, se il numero e' piccolo altre variabile non prese in considerazione sono importanti

Singoli t test per ogni predittore

- Nel quadro dei modelli lineari la significatività della contribuzione di un predittore X_i viene valutata con un t test
- E' un calcolo esplorativo che stabilisce se il coefficiente a_i e' significativamente diverso da zero.
- Dunque il test non stabilisce l'importanza della contribuzione ma solo se e' significativa
- Va inteso come test esplorativo perché non tiene conto dell'insieme del modello ma solo di quell'unico predittore

Ulteriori test di controllo

- La funzione "anova" permette di paragonare due modelli diversi sulla base del numero di parametri e sulla entità dei scarti quadrati tra i valori del modello e i valori osservati
- RSS o residual sum of square = $\sum^{n}_{i=1}(y_i-(ax_i+c))^2$
- Il confronto e' fatto con un test di Fisher di paragone tra due RSS

Buona prassi nella definizione di un modello

- Un buon modello deve essere significativo e spiegare la variazione della variabile dipendente Y con il minor numero di predittori possibili
- Si può cominciare da un modello che includa tutti i predittori e poi andare a scendere oppure il contrario
- Sceglieremo la prima via.
- Dal modello completo si sceglie la variabile meno significativa con il test t.
- Si costruisce un nuovo modello senza quella variabile
- Si paragona i risultati tra il modello completo e il modello senza la variabile ritenuta inutile

Esempio

- >iriscompl.lm<-lm(Sepal.Length~Sepal.Width+Petal.Width +Petal.Length+Species, data=iris)
- >summary(iriscompl.lm) #noto che il contributo della larghezza petali non e' altamente significativo e il contributo delle specie e' basso. Provo vari modelli che includono parzialmente l'informazione dei petali e della specie
- >irisnoPW.lm<-lm(Sepal.Length~Sepal.Width+Petal.Length+Species, data=iris)
- >irisnoS.lm<-lm(Sepal.Length~Sepal.Width+Petal.Width+Species, data=iris)
- >irisnoSPW.lm<-lm(Sepal.Length~Sepal.Width+Species, data=iris)
- >anova(iriscompl.lm,irisnoPW.lm) #verifico che esiste una differenza marginalmente significativa
- >anova(iriscompl.lm,irisnoSPW.lm)
- >anova(iriscompl.lm,irisnoS.lm)

Esempio (cont.)

- Risulta che tutti modelli differiscono con una significatività di soglia maggiore di 0.05. Il p-value del modello senza larghezza dei petali e' marginalmente significativo essendo 0.03. Vista la disponibilità di dati (150 osservazioni) la possibilità di essere di fronte ad una mancanza di potenza del test e' ridotta e probabilmente sarebbe meglio utilizzare soglie più basse come 0.01 o meno.
- >plot(iriscompl.lm)
- >plot(irisnoPW.lm)
- Si puo vedere che i residui sembrano meno omogenii nel modello piu semplice, forse e' preferibile quello piu complesso

Esempio II

- Il caso del conteggio di insetti e dell'insetticida
- >summary(InsectSprays)
- La domanda che sottende la raccolta di dati e' se c'e' differenza tra l'uso di questi vari insetticidi.
- Usando il t test e' possibile fare raffronti coppie a coppie ma non c'e' la possibilità di dare un giudizio complessivo che non soffra di problemi di test multipli e parzialmente dipendenti.
- Usiamo l'approccio della regressione lineare per esplorare questi dati

Esempio II

- Applichiamo un modello lineare scegliendo come variabile dipendente il conteggio degli insetti sulle piante e come predittore il trattamento di insetticida utilizzato. Questo segue anche il consiglio di scegliere la variabile dipendente quella variabile che probabilmente e' un effetto dei predittori.
- >insect.lm<-lm(count~spray, data=InsectSprays)
- >summary(insect.lm)
- Notiamo che gli insetticidi B e F non sono significativamente diversi per ogni soglia ragionevole dall'intercetta che rappresenta l'insetticida A
- Inoltre CDE sono molto simili con D leggermente a parte (paragona le differenze fra coefficienti con il valore della deviazione standard
- Per formalizzare questa impressione applichiamo il Tukey Honest Signficant Difference test (TukeyHSD) applicabile ai fattori

Svolgimento esempio II

>insect.lm<-lm(count~spray, data=InsectSprays)

>summary(insect.lm)

>TukeyHSD(aov(insect.lm))

Identifico quali livelli sono fra loro non molti diversi (p-value grandi)

In questo caso la procedura indica A-B, F-B, E-C, E-D, F-A come molto simili, e DC leggermente divergenti

Questo propone una ricodifica dei 6 insetticidi in due gruppi (ABF e CDE)

Svolgimento esempio II (cont.)

Costruisco un vettore alternativo di insettidici usati

```
>InsectSprays<-data.frame(InsectSprays,
   spray2=InsectSprays$spray)
```

Modifico i livelli in modo da tenere conto dei raggruppamenti proposti dal Tukey test

```
>levels(InsectSprays$spray2)
>levels(InsectSprays$spray2)<-
   c('ABF','ABF','CDE','CDE','CDE','ABF')
>levels(InsectSprays$spray2)
```

Svolgimento esempio II (cont.)

Costruisco il nuovo modello con la lista degli insetticidi semplificata

```
>insect2.lm<-lm(count~spray2, data=InsectSprays)
```

Con la funzione anova paragono i gli scarti quadratici residui dei due modelli e controllo se la differenza puo essere spiegata dal numero di parametri diverso

```
>anova(insect2.lm,insect.lm)
Analysis of Variance Table
Model 1: count ~ spray2
Model 2: count ~ spray
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     70 1092.00
2     66 1015.17  4     76.83 1.2488 0.2991
```

P-value uguale a 0.2991 non permette di rigettare l'ipotesi nulla di nessuna differenza fra modello 1 e 2 per una qualunque soglia sensata alpha

Conclusioni esercizio II

- Il modello insect2.lm e' da preferire in quanto con meno parametri spiega i dati in maniera non differente dal modello più completo insect.lm.
- Si può contemporaneamente affermare che le differenze tra gli insetticidi A, B e F e quelli C,D e E sono dovute al caso con una probabilità del 0.29, mentre la probabilità che le singole differenze sono dovute al caso può essere vista dai risultati del test di Tukey HSD.

Nota Bene

- I risultati di un modello lineare dipendono anche dall'interazione fra i predittori, overossia i casi di predittori che riportano informazioni parzialmente sovrapposte
- Queste interazione non sono visibili ai vari test e ai sistema di diagnosi
- E' possibile vederle solo aggiungendo e togliendo variabili e osservando come le altre variabili possono diventare significative o non essere più significative

Esempio di interazione

Torniamo all'esempio sulle piante. Se usiamo il Tukey HSD test sul modello completo

>TukeyHSD(iriscompl.lm)

La differenza fra le varie specie e' altamente non significativa e la variabile specie sembra ridondante

Ma l'eliminazione delle specie da un risultato globalmente significativo

>anova(iriscompl.lm, irisnoS.lm)

Eliminando le informazioni sui petali invece si recupera significatività nella differenza delle varie specie

>irisnoP.lm<-lm(Sepal.Length~Sepal.Width+Species, data=iris)

>TukeyHSD(irisnoP.lm)

Le informazioni sui petali e quelle di appartenenza a specie interferiscono una sull'altra dimostrando che in parte portano lo stesso tipo di informazione.

Vedi pure che la R^2 dei modelli irisnoP.lm (0.7203), irisnoS.lm(0.8557) , iriscompl.lm (0.8627). Dunque l'informazione del modello senza specie riassume quasi quello del modello completo. Informazione che sembra essere presente nel modello senza petali anche se meno completa