Regressione

Esempio

Un'azienda manifatturiera vuole analizzare il legame che intercorre tra il volume produttivo X per uno dei propri stabilimenti e il corrispondente costo mensile Y di produzione.

Volume X (ton.)	Costo Y (K€)	Volume X (ton.)	Costo Y (K€)
10.11	*1.53*	42.87	*13.51*
50.56	*13.14*	61.53	*23.65*
90.28	*31.24*	24.60	*9.43*
15.50	*5.47*	46.85	*15.12*
69.52	*22.27*	50.63	*18.94*
98.40	*26.47*	89.68	*26.06*
86.66	*24.32*	27.91	*10.08*

Modelli di stima

- Lo scopo è di cogliere un legame semplice e tendenziale tra la variabile dipendente Y e le variabili indipendenti X.
- Si ipotizza l'esistenza di una funzione $f: IR^n \rightarrow IR$ che esprime il legame tra la variabile dipendente Y e le n variabili esplicative X_j

$$Y = f(X_1, X_2, \ldots, X_n).$$

- La funzione f può essere:
  - Lineare: $Y = b + \omega X$
  - Quadratica: $Y = b + \omega X + d X^2$
    - Posto $Z=X^2$, il modello è $Y = b + \omega X + d Z$
  - Esponenziale: $Y = e^{b+\omega X}$
    - Posto $Z = \log Y$, il modello è $Z = b + \omega X$

Modello probabilistico

- E' improbabile che le coppie (X, Y) si dispongano lungo una retta del piano.
- E' più realistico supporre un legame di natura approssimata tra X e Y, espresso dal modello

$$Y = \omega X + b + \varepsilon$$

con ε variabile casuale detta *scarto* o *errore*, che deve soddisfare alcune ipotesi di natura stocastica.

Calcolo della retta di regressione

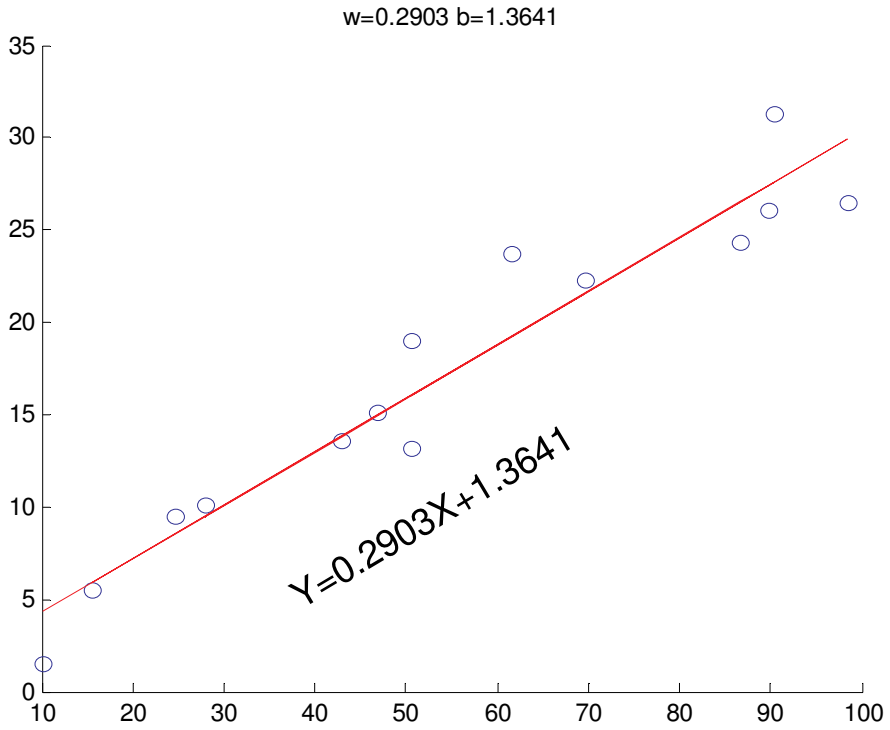
- L'identificazione della retta di regressione si riduce all'identificazione del coefficiente angolare ω e dell'intercetta b. della retta $Y = \omega X + b + \varepsilon$
- Minimizzazione della funzione SSE (*sum of squared errors*):

$$SSE = \sum_{i=1}^{m} e_i^2 = \sum_{i=1}^{m} [y_i - f(x_i)]^2 = \sum_{i=1}^{m} [y_i - \omega x_i - b]^2$$

  - In Matlab: polyfit(), polyval()

Esempio azienda manifatturiera

```
>> scatter(X,Y)
>> p = polyfit(X,Y,1);
>> FX= polyval(p,X);
>> hold on
>> plot(X,FX)
```



Regressione lineare multipla

- Se indichiamo con e il vettore dei residui, deve valere:

  $$y_i = \omega_1 x_1 + \omega_2 x_2 + \cdots + \omega_n x_n + b + e_i = \sum_{j=1}^{n} \omega_j x_j + b,$$

  che in notazione matriciale diventa:

  $$\mathbf{y}=\mathbf{X}\mathbf{w}+\mathbf{e}$$

- Nel caso di una regressione con $n+1$ parametri, ω_j e b possono essere determinati minimizzando la somma degli errori:

$$SSE = \sum_{i=1}^{m} e_i^2 = \|\mathbf{e}\|^2 = \sum_{i=1}^{m} (y_i - w' x_i)^2 = (\mathbf{y}-\mathbf{X}\mathbf{w})'(\mathbf{y}-\mathbf{X}\mathbf{w})$$

Calcolo dei coefficienti di regressione

- La soluzione del problema di minimo è:

$$\hat{\mathbf{w}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

- Possiamo ricavare il valore delle variabili di risposta Y come

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{w}} = (\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y}) = \mathbf{H}\mathbf{y}$$

dove

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

- è detta matrice di proiezione (*hat matrix*)

Assunzioni relative ai residui

- Minimizzando SSE, la variabile aleatoria ε deve seguire una distribuzione normale di media 0 e deviazione standard σ.

- Si richiede inoltre che i residui ε_i e ε_k, corrispondenti a due distinte osservazioni x_i e x_k siano indipendenti per ogni scelta di i e k.

- Un modello è tanto più accurato quanto più la deviazione σ risulta prossima a zero.

Esercizi

- Determinare un modello di regressione lineare per il dataset http://statmaster.sdu.dk/courses/st111/data/data/tvads.txt
- Cosa accade se si usa una scala logaritmica?
- Cosa si può dire per il dataset http://statmaster.sdu.dk/courses/st111/data/data/velocity.txt

Trattamento di attributi predittivi categorici

- Ad un attributo categorico che può assumere H valori v_h distinti è possibile associare H-1 variabili binarie fittizie $D_{j1}, D_{j2}, \ldots, D_{jH-1}$.

- Per il campione i il cui attributo categorico j vale v_h, solo la $D_{ih} = 1$ e tutte le altre 0.

- Il livello della variabile omessa è arbitrario.

Valutazione dei modelli di regressione

- Normalità e indipendenza dei residui
- Significatività dei coefficienti
- Coefficiente di correlazione lineare
- Multi-collinearità delle variabili indipendenti
- Limiti di confidenza e predizione
  - In Matlab, regstats()

Normalità e indipendenza dei residui

- Diagramma di dispersione dei residui rispetto ai valori predetti.
  - Un andamento regolare dei residui indica l'esistenza di fattori esplicativi non considerati nel modello.

- Diagramma di dispersione della radice dei residui
  - I valori sono tutti positivi ed attenuati rispetto ai precedenti

Significatività dei coefficienti

- Lo z-indice del valore stimato di ω può essere utilizzato per stimare la bontà della previsione:
  - z-indice< 0.05 || z-indice> 2 $\rightarrow$ con confidenza del 95% un intervallo attorno a ω non contiene lo 0

- Lo stesso si può dire per b.

- Nell'esempio dell'azienda manifatturiera gli z-indici sono rispettivamente 11.980 e 0.916
  - La mancanza di significatività dell'intercetta non pregiudica la bontà del modello.

Covarianza

- La *covarianza* quantifica la forza della relazione tra due insiemi di valori, ovvero misura quanto lineare è la dipendenza tra i due attributi;
- La covarianza è la media del prodotto delle deviazioni dei valori dalla media degli insiemi dei dati

$$v_{jk} = cov(a_j, a_k) = \frac{1}{m-2} \sum_{i=1}^{m} (x_{ij} - \bar{\mu}_j)(x_{ik} - \bar{\mu}_k)$$

  - In Matlab cov()
- un valore positivo indica una variazione di X e Y nella stessa direzione, un valore negativo l'opposto

Correlazione

- Un limite della covarianza è la sua dipendenza dall'unità di misura.
- Per esempio possiamo aumentare il fattore covarianza di 1000, semplicemente usando come unità di misura € in luogo di K€
  - Nel caso le unità sono appropriate
- La misura di *correlazione* risolve il problema producendo un risultato indipendente dalle unità di misura e compreso tra –1 e 1

$$r_{jk} = corr(a_j, a_k) = \frac{v_{jk}}{\bar{\sigma}_j \bar{\sigma}_k}$$

Correlazione

- Un valore della correlazione vicino a –1 indica che i due insiemi di valori tendono a variare in senso opposto
- Un valore della correlazione vicino a +1 indica che i due insiemi di valori tendono a variare nello stesso senso
- Una indipendenza nelle variazioni dei due valori produce un indice di correlazione uguale a 0
- Ma, attenzione: l'indice di correlazione è rilevante solo per relazioni *lineari*
- L'indice può risultare vicino a 0 anche se esiste una relazione non lineare tra i due insiemi di valori.

Multi-collinearità

- Si parla di multi-collinearità quando sono presenti relazioni lineari tra le variabili indipendenti.
- Si parla di multi-collinearità esatta quando almeno una delle variabili esplicative è correlata con altre variabili indipendenti.
  - Esempio: la produzione settimanale è la somma delle produzioni giornaliere, e tutte le variabili sono incluse nel modello.
- In presenza di multi-collinearità esatta la matrice (X^TX) è singolare e non ammetta inversa.
- La multi-collinearità esatta è piuttosto rara e tipicamente causata da una errore nella definizione del modello.

Limiti di confidenza e di predizione

- Conseguenze della multi-collinearità nelle variabili indipendenti:
  - difficoltà di determinare i contributi individuali delle variabili, perché i loro effetti vengono mescolati o confusi;
  - alta variabilità delle stime con conseguente bassa significatività dei coefficienti di regressione;
  - forte instabilità delle stime dei coefficienti di regressione (piccole variazioni nei dati o l'aggiunta/eliminazione di una variabile dal modello possono portare a grandi variazioni nella stima).
- La multicollinearità non invalida il modello,ma l'interpretazione dei singoli coefficienti di regressione.

Selezione delle variabili predittive

- Per identificare la multi-collinearità si possono calcolare i coefficienti di correlazione tra tutte le coppie di variabili esplicative.
  - Valori elevati (> 0,90) indicano la forte collinearità
  - Valori bassi, non assicurano l'assenza di multi-collinearità
    - Effetto congiunto di due o più variabili esplicative.
- Rimedio:
  - Eliminare una o più variabili indipendenti altamente correlate, senza eliminare variabili significative;

Passi per la costruzione del modello

1. Individuazione dei valori anomali
2. Scelta del modello
3. Individuazione dei parametri
4. Significatività dei coefficienti
5. Previsione per diversi valori della variabile indipendente

Sommario

- La regressione lineare semplice e multipla permette di determinare semplici modelli.
- È possibile valutare la bontà di tali modelli valutando la normalità, l'indipendenza dei residui e la significatività dei coefficienti
- Tramite il coefficiente di correlazione è possibile stabilire se ci sono dipendenze lineari tra le variabili indipendenti.
- Le conseguenze della multi-linearità vanno affrontate alla luce delle soluzioni esistenti.