Regressione

Esempio

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

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

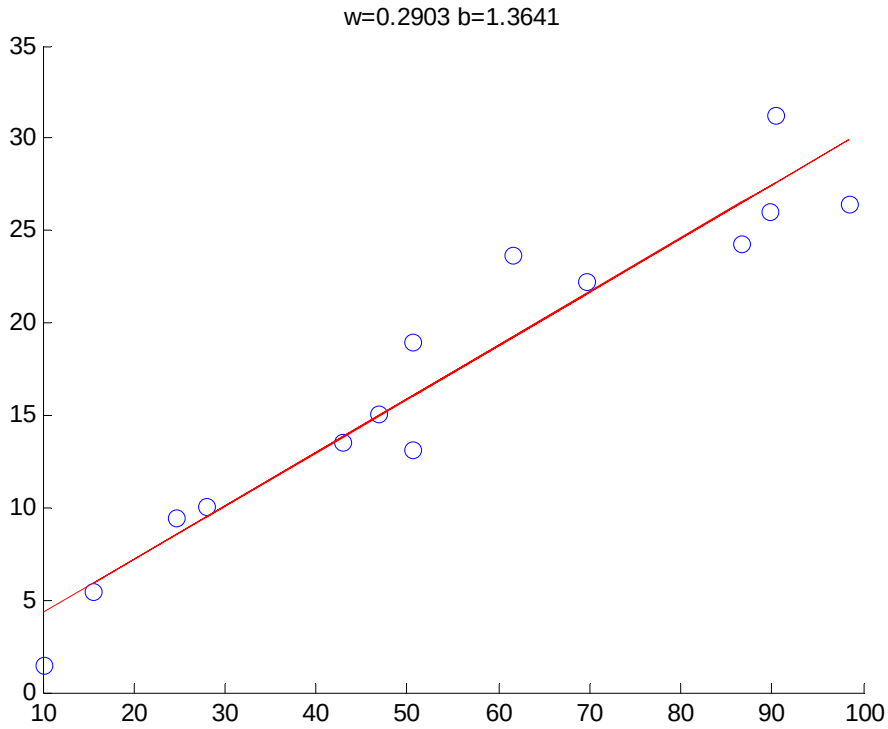
- L'identificazione della retta di regressione si riduce all'identificazione del coefficiente angolare ω e dell'intercetta b. della retta $Y = \omega X + b + \varepsilon$
- Minimizzazione della funzione SSE (*sum of squared errors*):

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

  - 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 $D_{j1}, D_{j2},\ldots, D_{jH\text{-}1}$.
- 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
- Analisi della varianza
- Coefficiente di determinazione
- 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 rispetto ai valori predetti.
  - Un andamento regolare dei residui indica làesistenza 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 per b.
- Nell'esempio dellà'azienda manifatturiera gli z-indici sono rispettivamente 11.980 e 0.916
  - La mancaza di significatività dell'intercetta non pregiudica la bontà del modello.

Nella prossima lezione

- Analisi delle serie storiche
  - Valutazione dei metodi
  - Modelli di previsione