Introduzione all'inferenza statistica
VI lezione
Analisi di regressione

Carla Rampichini

Dipartimento di Statistica "Giuseppe Parenti" - Firenze - Italia

rampichini@ds.unifi.it - www.ds.unifi.it/rampi/

Regressione lineare semplice

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

- Y_i *risposta* è una v.c., X_i *predittore* osservabile

Ipotesi alla base del modello:

- la relazione tra X e $E(Y \mid X)$ è lineare nei parametri. Sempre vero se $(X, Y) \sim BN$
- ε_i: $E(X\varepsilon) = 0, Var(\varepsilon \mid X) = \sigma^2$,$Cov(\varepsilon_i, \varepsilon_{i'} \mid X) = 0$
- $\varepsilon_i - iid \sim N(0, \sigma^2)$

- Due problemi:

✓ stimare i *parametri* incogniti α, β, σ^2

✓ valutare la bontà di *adattamento* del modello

Soluzione matematica: Minimi Quadrati (Least Squares)

- Non si fanno assunzioni sulle coppie di valori (x_i, y_i)
- minimizzo distanza verticale tra (x_i, y_i) e la retta: $\min_{\alpha,\beta} \sum_{i=1}^n (y_i - (\alpha + \beta x_i))^2$
- soluzione: $\hat\beta = \frac{S_{xy}}{S_{xx}}$, $\hat\alpha = \bar y - \hat\beta \bar x$

$\bar x = (1/n) \sum_{i=1}^n x_i$, $\bar y = (1/n) \sum_{i=1}^n y_i$
$S_{xy} = \sum_{i=1}^n (x_i - \bar x)(y_i - \bar y)$, $S_{xx} = \sum_{i=1}^n (x_i - \bar x)^2$

Proprietà statistiche dello stimatore dei MQ

Teorema di Gauss-Markov

Lo stimatore dei MQ $(\hat\alpha, \hat\beta)$ è lo stimatore lineare corretto a varianza minima di (α, β) (stimatore BLU). Lo stimatore BLU di qualsiasi trasformazione lineare di (α, β) è dato dalla corrispondente trasformazione lineare dello stimatore dei minimi quadrati.

Regressione lineare multipla:

$$Y_i = \mathbf{x}_i' \boldsymbol{\beta} + \boldsymbol{\varepsilon}_i$$

con $\mathbf{X}'$ a pieno rango.
Lo stimatore dei MQ di $\boldsymbol{\beta}$ è

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

(*esempi in regressione_dottorato08.ppt*)

Inferenza per il modello di Regressione lineare

Esempio

tratto da Fox, J. (1997), *Applied Regression Analysis, linear models and related methods*, Sage publications

duncan.dat

prestigio, reddito e livello di istruzione delle professioni in America nel 1950

var	descrizione	modalità
1. occup	professione	
2. occ_type	tipo profess.	prof =Profess. e dirigenti wc= impiegati bc= operai
3. Income	% maschi con stipendio $\geq$ \$3500	0-100
4. Education	% maschi diplomati (high-school)	0-100
5. Prestige	% di esperti (NORC study) che valutano il prestigio della professione eccellente o buono	0-100

(Fonte: Tabella VI-1 in O. D. Duncan (1961), "A socioeconomic index for all occupations," in A. J. Reiss, Jr., Occupations and Social Status, New York, Free Press.).

Lettura di dati in formato testo in SPSS

```
 GET DATA  /TYPE = TXT
 /FILE = 'percorso \duncan.dat'
/FIXCASE = 1
/ARRANGEMENT = FIXED
/FIRSTCASE = 1
 /IMPORTCASE = ALL
/VARIABLES =
/1 profess 0-35 A36
occ_type 36-39 A4
income 40-42 F3.2
education 43-46 F4.2
prestige 47-49 F3.2 .
CACHE.
EXECUTE.
```


Analisi di regressione con SPSS

- leggiamo i dati con SPSS e salviamoli con nome `duncan.sav`
- sul sito:

http://www.ats.ucla.edu/STAT/spss/examples/ara/default.htm

sono disponibili gli esempi in SPSS di *Applied Regression Analysis*, di Fox

- sul sito: http://www.ats.ucla.edu/stat/spss/topics/regression.htm

si può trovare materiale vario su analisi di regressione con SPSS

Stimatori M.Q. con SPSS

```
GET  FILE='D:\duncan.sav'.
REGRESSION
/STATISTICS COEFF OUTS R ANOVA
/DEPENDENT prestige
/METHOD=ENTER educ income.
```

Output fornisce molte informazioni (duncan.doc)!

Model Summary

- **R Square** $R^2 = SS_{reg}/SS_{TOT} = 0.828$, $SS_{TOT} = \sum(y_i - \bar{y})^2$, $SS_{reg} = \sum(\hat{y}_i - \bar{y})^2$
- **R** $R = \sqrt{R^2} = 0.910$, coeff. corr. multipla e anche $Corr(Y, \hat{Y})$
- **Std. Error of the Estimate** $S_E = \sqrt{\frac{\sum e_i^2}{n-k-1}} = 13.67$. Dimensione media dei residui e_i; k predittori, n obs

Coefficienti stimati $\hat{\beta}$

- **Std. Error** $V(\hat{\beta}) = \frac{1}{1-R_j^2} \times \frac{\hat{\sigma}_\varepsilon^2}{\sum (x_{ij}-\bar{x}_j)^2}$,
  R_j^2 calcolato per la regressione di X_j su tutte le altre X,
  $\hat{\sigma}_\varepsilon^2 = (\sum e_i^2)/(n-k1)$
- **Standardized Coefficients** $\hat{\beta}_j^* = (s_j/s_Y)\hat{\beta}_j$,
  s_j deviazione standard di X_j
- **Sig.** è il *p-value* per il test bidirezionale
- $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$, statistica test $t = \hat{\beta}_j / std(\hat{\beta}) \sim t_{\alpha/2}$
- intervallo di confidenza $1-\alpha$: $\hat{\beta}_j \pm t_{\alpha/2} \times std(\hat{\beta}_j)$.

⋆ Per *income* si ha: $\hat{\beta}_j = 0.599$, $std(\hat{\beta}_j) = 0.120$
✓ statistica test $t = 5.003$, $p - value = 0.000 \Rightarrow$ rifiuto H_0
✓ intervallo di confidenza $(1-\alpha) = 0.95$:
$0.599 \pm 2.018 \times 0.120 = [0.357; 0.841]$, $t_{0.025} = 2.018$

ANOVA

$H_0 : \beta_1 = \ldots = \beta_k = 0$ vs H_1 : almeno un $\beta_j \neq 0$

- **Sum of Squares** devianze
- **df** gradi di libertà
- **Mean Square** varianze
- **F** statistica test $F = \frac{SS_{reg}/k}{SS_{res}/(n-k-1)} = \frac{(n-k-1)}{k} \times \frac{R^2}{(1-R^2)}$
- **Sig.** p-value=$Pr(F > f \mid H_0)$, $F \mid H_0 \sim F_{k,(n-k-1)}$

⇝ Sotto H_0 la statistica test deve essere vicina a 1 perchè la varianza di regressione è una stima indipendente della varianza di errore. Sotto H_1 la varianza di regressione è più grande della varianza di errore. Rifiutiamo quindi H_0 quando F è significativamente più grande di 1.

▶ nel ns esempio: $p - value = Pr(F > 101.216 \mid H_0) = 0.000$.

Test su un sottoinsieme di parametri

$H_0 : \beta_1 = \ldots = \beta_q = 0$ vs H_1 : modello completo

$q < k$

si confronta il modello vincolato con quello completo utilizzando la statistica test:

$$F = \frac{(SS_{reg1} - SS_{reg0})/q}{SS_{res1}/(n-k-1)} = \frac{(n-k-1)}{q} \times \frac{R_1^2 - R_0^2}{(1-R_1^2)}$$

Esempio

```
REGRESSION
/STATISTICS COEFF OUTS R ANOVA
/DEPENDENT prestige
/METHOD=ENTER income /method=test (educ).
```

guardiamo a **Subset Tests** nella tabella **ANOVA**:
$F = \frac{(SS_{reg1} - SS_{reg0})/q}{SS_{res1}/(n-k-1)} = \frac{(36181-30665)/1}{7506.7/(45-2-1)} = 30.86$

Link analisi di regressione con SPSS

▶ sul sito:

http://www.ats.ucla.edu/STAT/spss/webbooks/reg/default.htm

è disponibile il testo WEB *Regression with SPSS*, di Xiao Chen, Phil Ender, Michael Mitchell & Christine Wells.

▶ sul sito:

http://www.ats.ucla.edu/stat/spss/seminars/SPSSRegression/default.htm

è disponibile il materiale del seminario *Regression using SPSS*

Questo materiale spiega come fare analisi di regressione con SPSS e presuppone che si conoscano l'analisi di regressione e gli elementi di base di SPSS.