Lezione 10 (a cura di Nicola Mincioni)

Il vettore $\hat{\beta}$ ha distribuzione normale tale che $\hat{\beta} \sim N(\underline{\beta}, \sigma^2 (X'X)^{-1})$: questa è la struttura del vettore $\underline{\beta}$ che serve per fare inferenza, ovvero stima dell'intervallo e verifica dell'ipotesi.

VARIANZA STIMATORE $\hat{\beta}$: $V(\hat{\beta}) = s^2 (X'X)^{-1}$, dove:

$$s^2 = \frac{DR}{n-(k+1)}$$

$\hat{V}(\hat{\beta}_j) = diag_j(\hat{V}(\hat{\beta}))$ è la stima della varianza del singolo β_j

s.e. $(\hat{\beta}_j)$ è la misura dell'errore standard, l'errore del singolo stimatore.

Ad esempio, se k=3, la matrice di covarianza ha dimensione 4×4.

$$\hat{V}(\hat{\beta}_J) = \begin{pmatrix} \hat{V}(\hat{\beta}_0) & & & \\ & \hat{V}(\hat{\beta}_1) & & \\ & & \hat{V}(\hat{\beta}_2) & \\ & & & \hat{V}(\hat{\beta}_3) \end{pmatrix}$$

INTERVALLO DI CONFIDENZA PER IL SINGOLO β_j

$$\beta_j \longrightarrow \left[\hat{\beta}_j - t_{\alpha/2} s.e.(\hat{\beta}_j), \hat{\beta}_j + t_{\alpha/2} s.e.(\hat{\beta}_j)\right]$$

$t_{\alpha/2}$ è il percentile della distribuzione t di Student con (n-(k+1)) gradi di libertà.

Se n abbastanza grande, si può assumere che $t_{\alpha/2} = z_{\alpha/2}$ e, considerando $\alpha = 0.05$ si ha $z_{\alpha/2}$=1.96.

VERIFICA IPOTESI

H_0: $\beta_j = \beta_{j0}$ con β_{j0} fissato; ad sempio si valuta l'effetto degli anni di istruzione sul reddito.

Si calcola la statistica test $t = \frac{\hat{\beta}_j - \beta_{jo}}{s.e.(\hat{\beta}_j)}$, ovvero la discrepanza tra la stima e il valore osservato.

La statistica test t ha distribuzione t-student sotto l'ipotesi nulla, ovvero:

t~t(n-(k+1)) sotto H_0.

Se $|t| > t_{\alpha/2}$, rifiuto l'ipotesi nulla in favore dell'ipotesi alternativa $H_1: \beta_j \neq \beta_{j0}$

Se β_{j0} appartiene all'intervallo di confidenza, allora è un valore ritenuto plausibile e quindi accetto l'ipotesi nulla (AH_0) .

L'intervallo di confidenza come logica contiene tutti i valori ritenuti plausibili del parametro.

p-value: quando lo p-value$\leq \alpha$, allora rifiuto l'ipotesi nulla H_0.

$H_0: \beta_j = 0$

$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon$

j=0 $\longrightarrow$ $H_0: \beta_0 = 0$ $\longrightarrow$ la y è proporzionale alla x, ovvero $E(y_i)=0$ se $x_{i1} = x_{i2} = \ldots = x_{ik} = 0$.

Nel caso di regressione lineare semplice, il modello sarebbe rappresentato da una retta passante per l'origine.

j>0 $\longrightarrow$ $H_0: \beta_j = 0$ $\longrightarrow$ la covariata x_j non ha influenza lineare sulla y.

$y_i = \beta_0 + \beta_{j-1} x_{i,j-1} + \beta_{j+1} x_{i,j+1} \beta_k x_{ik}$ $\longrightarrow$ il β_j non c'è più così come il termine β_j, la covariata x_j non ha più influenza sulla y.

La statistica test diventa $t = \frac{\hat{\beta}_j}{s.e(\hat{\beta}_j)}$ e la verifica viene fatta nel modo tradizionale.

Tavola verifica ipotesi nulla in R.

PARAMETRI	STIMATORE	s.e	t-stat	p-value
INTERCETTA	$\hat{\beta}_0$	s.e ($\hat{\beta}_0$)	$\hat{\beta}_0$ /s.e ($\hat{\beta}_0$)	p-value$_0$
β_1	$\hat{\beta}_1$	s.e.($\hat{\beta}_1$)	$\hat{\beta}_1$ /s.e.($\hat{\beta}_1$)	p-value$_1$
β_K	$\hat{\beta}_k$	s.e.($\hat{\beta}_k$)	$\hat{\beta}_k$ /s.e.($\hat{\beta}_k$)	p-value$_k$

Come si fa a verificare l'ipotesi nulla? Se t-stat è piccolo o lo p-value>0.05.

Anche il segno è importante per i parametri stimati, al fine di comprendere se l'influenza sulla y è positiva o negativa. Se lo p-value è basso, rifiuto l'ipotesi nulla

sul coefficiente e la covariata ha reale influenza sulla y.

Se p-value$_0$=0.01, rifiuto l'ipotesi nulla H_0: β_0=0, l'intercetta è significativamente diversa da 0 e va utilizzata nel modello.

Se p-value$_1$=0.3, accetto l'ipotesi nulla H_0= β_1=0 e concludo che la covariata non è importante e si può non considerare nel modello.

Accettare l'ipotesi nulla significa togliere qualcosa dal modello.

INDICE R^2

Questo indice R^2 è l'indice di determinazione multipla e serve a misurare la bontà dell'adattamento del modello di dati.

$R^2 = 1-\frac{DR}{Dev(Y)}$ $DR=\sum_i(y_i-\hat{y}_i)^2$; $Dev(Y)=\sum_i(y_i-\bar{y})^2$; $DS=\sum_i(\hat{y}_i-\bar{y})^2$

In virtù della decomposizione Dev(Y)=DS+DR, si può scrivere anche l'indice R^2 come

$R^2=\frac{DS}{Dev(Y)}$

A differenza della regressione lineare semplice, non vale la formula $\frac{Cod^2(x,y)}{Dev(x)Dev(y)}$; quella di $R^2=\frac{DS}{Dev(Y)}$ è la formula esclusiva che può essere applicata per la regressione lineare multipla.

INTERPRETAZIONE

$0\le R^2\le 1$ è l'intervallo di valori in cui può ricadere R^2 : nella pratica questo indice assume valori intermedi e per stabilire la bontà dell'adattamento si assume il valore soglia $R^2 \ge 0.75$

$R^2=0 \longrightarrow$ DR=Dev(Y) e DS=0 : l'adattamento del modello è pessimo.

$R^2=1 \longrightarrow$ DR=0 e DS=Dev(Y), l'adattamento del modello è ottimo, la regressione lineare è un'ottima interpolazione.

Se $R^2 \ge 0.75$, il modello regressione lineare è adeguato.

$R=+\sqrt{R^2}$ è l'indice di correlazione multipla, è considerato positivo perché non si conosce il segno della covariata associata.

$r=\pm\sqrt{r^2}$ è l'indice di correlazione semplice, il segno dipende dal tipo di correlazione tra le due variabili.

$R^2_{adj} = 1 - \frac{DR(n-(k+1))}{Dev(Y)/(n-1)}$

Qui viene tenuto in considerazione anche il numero dei parametri.
L'R^2 aumenta con k, ovvero all'aumentare delle covariate
Es.
R^2=0,7, k=1
R^2=0,8, k=2
R^2=0,9, k=3
Attraverso ciò ottengo un modello troppo complesso con tutte le covariate.
R^2_{adj} aumenta solo se la nuova covariata è realmente importante.

TABELLA ANOVA:

SS	**g.d.l.**	**Ms**	**F**
Ds	K	Ds/k	(Ds/k)/(Dr/(n-(k+1)))
Dr	(n-(k+1))	Dr/(n-(k+1))	
Dev(Y)	n-1		

Se il modello è buono Ds>Dr di molto.

R^2=Ds/Dev(Y)

Statistica F serve a verificare che H_0:
H_0: $\beta 1 = \beta 2 = ... = \beta k = 0 \Rightarrow$ Yi= $\beta_0 + \varepsilon_i$ (non c'è diff. tra i soggetti)
H_1: almeno una Xj è utile (non si specifica quale)

$F = \frac{Ds/k}{Dr/(n-(k+1))} \sim F(k, n-(k+1))$ sotto H_0

$F \geq F\alpha \Rightarrow RH_0 \Rightarrow$ se F molto grande almeno una Xj utile

p-value $\leq \alpha \Rightarrow RH_0$
p-value$> \alpha \Rightarrow AH_0$

es.
Y=reddito p-value=0,5$\Rightarrow AH_0$
x_1= età p-value=0,006$\Rightarrow RH_0$
x_2=anni istruzione

Situazione ottimale:

p-value basso, R^2 alto