

## Test di ipotesi e intervalli di confidenza sui parametri

Durante l'esame di un modello lineare classico, è necessario identificare le relazioni presenti tra la variabile dipendente e i regressori, ma è fondamentale anche comprendere se tali relazioni risultino essere significative dal punto di vista statistico, ed è necessario comprendere quale sia la bontà di adattamento del modello ai dati.

A tal scopo si ricavano test sui parametri e sull'adattamento del modello, si definisce un'ipotesi nulla ed in particolare tale ipotesi nulla divide in due parti l'area sottesa alla distribuzione di un parametro:

- Si definisce l'area di accettazione del test in cui si considerano la quasi totalità dei valori associati alle probabilità cumulate della distribuzione.
- Si definisce l'area di rifiuto che comprende i valori relativi alle code della distribuzione.

Se il valore del parametro campionario ricade nell'area di accettazione si accetta l'ipotesi nulla legata al valore del campione; si rifiuta l'ipotesi nulla altrimenti.

Si minimizza così l'errore di prima specie cioè la probabilità di rifiutare l'ipotesi nulla quando è vera.

### Test della normale per i parametri varianza nota

Si definisca  $\beta_j$  il valore parametro campionario.

Si consideri  $N(\beta_j, \frac{\sigma^2}{n\sigma_{jj}^{-1}})$ ,

con  $\sigma_{jj}^{-1}$  il generico valore diagonale della matrice  $(X'X)^{-1}$ .

Si consideri l'ipotesi nulla  $H_0: \beta_j=0$

Come precedentemente detto, si rifiuta  $H_0$  se l'intervallo centrale considerato non comprende il valore di  $\beta_j$  ricavato dal campione.

La regione di accettazione è:

$$P \left[ -z_{\frac{\alpha}{2}} < \frac{\beta_j - \theta}{\sqrt{m\theta_{jj}^{-1}}} < z_{\frac{\alpha}{2}} \right] = P \left[ -z_{\frac{\alpha}{2}} \frac{\theta}{\sqrt{m\theta_{jj}^{-1}}} < \beta_j < z_{\frac{\alpha}{2}} \frac{\theta}{\sqrt{m\theta_{jj}^{-1}}} \right] = 1 - \alpha$$

### Test t di Student per i parametri con varianza ignota

Si consideri  $H_0 : \beta_j = 0$

La regione di accettazione è:

$$P \left[ -t_{\frac{\alpha}{2}} < \frac{\beta_j - 0}{\sqrt{m\theta_{jj}^{-1}}} < t_{\frac{\alpha}{2}} \right]$$

Si osservi che raramente nella realtà si conosce la varianza della popolazione, dunque si costruisca la variabile *t di Student* come rapporto tra una normale standardizzata e una  $\chi^2_{(n-k-1)}$  divisa per i suoi gradi di libertà e indipendente da normale:

Se non conosco varianza popolazione:

$$\frac{\beta_j - 0}{\sqrt{m\theta_{jj}^{-1}}} \quad / \quad \frac{\beta_j - S}{\sqrt{m\theta_{jj}^{-1}}}$$

Tale rapporto si distribuisce come una *T di student*.

### Test F per il modello

Dati  $SST = \frac{\text{MODEL SS}}{\text{SSR}}$  varianza totale

$R^2 = \frac{\text{MODEL SS}}{\text{TSS}}$  coefficiente di determinazione, definito come il rapporto tra la varianza spiegata e la varianza totale.

Sia  $H_0: R^2=0$  ipotesi nulla.

Si consideri:

$$\frac{\text{Model SS} / k}{\text{SSR} / (n - k - 1)}$$

Il rapporto tra due  $\chi^2$  divise per i loro gradi di libertà indipendenti tra loro si distribuisce come una *F di Snedecor*(k, n-k-1).

Il testo F è basato sull'ipotesi nulla  $H_0$ : tutti i parametri nulli.

La regione di accettazione è data da:

$$P \left[ \frac{\text{Model SS} / k}{\text{SSR} / (n - k - 1)} < F_{(k, n - k - 1)} \right]$$

### Test F su uno o più parametri

Si ipotizzi che vi siano q parametri uguali a 0, con  $q < k$

Ossia  $H_0: b_1, \dots, b_j, \dots, b_q = 0$

Si consideri  $TSS = SSR1 + \text{MODEL SS1}$

$$\frac{\text{Model SS1} / (k - q)}{\text{SSR1} / (n - k - q)}$$

Il rapporto tra  $\chi^2$  indipendenti divise per i g.d.l. è distribuito come  $F(k-q, n-k-q)$

Si osservi in particolare che:

Se  $q=k-1$  il test F può essere effettuato sui singoli parametri, risultando dunque uguale al quadrato del test T di student.

$$F = \frac{\chi^2(1)}{\chi^2/(n-p-1)} \quad \text{con } H_0: b=0$$

### **Relazione fra test di ipotesi e intervalli di confidenza**

La costruzione di un intervallo di confidenza per il parametro della regressione al 95%, implica che se si considerassero 100 campioni e si calcolasse un intervallo di confidenza del 95% per ogni campione, il 95% degli intervalli conterrà il vero valore della popolazione.

Si osservi che l'intervallo di confidenza non riflette la variabilità del parametro sconosciuto, riflette la quantità di errori casuali presenti nel campione e fornisce un intervallo di valori che potrebbero includere il parametro sconosciuto.

### **Test di ipotesi**

Per quanto riguarda il test di ipotesi: si respinge  $H_0$  se il p-value è “piccolo”, dunque se c'è solo una piccola probabilità che il parametro nel campione ha un valore estremo anche nel caso in cui  $H_0$  sia vera.

### **Intervallo di confidenza**

Un intervallo di confidenza al 95% fornisce un intervallo di valori plausibili per il coefficiente angolare in cui il 95% degli intervalli comprende il vero valore del coefficiente angolare stesso.

## Modello SURE

Utilizzare gli stessi regressori per diverse variabili dipendenti può risultare essere una scelta eccessivamente rigida. Si consideri il caso di differenti variabili esplicative per ogni equazione e correlazione tra residui casuali di disturbo associati a diverse equazioni:

$$\begin{aligned} y_1 &= \beta_{10} + \beta_{11} z_{11} + \dots + \beta_{1m} z_{1m} + \varepsilon_1 \\ &\vdots \\ y_j &= \beta_{j0} + \beta_{j2} z_{j2} + \dots + \beta_{jr} z_{jr} + \varepsilon_j \\ &\vdots \\ y_m &= \beta_{m0} + \beta_{mm} z_{mm} + \dots + \beta_{mr} z_{mr} + \varepsilon_m \end{aligned}$$

Per il modello lineare multivariato si ha:

$$B^{\wedge} = Y^{\circ} Z^{\circ\prime} (Z^{\circ} Z^{\circ\prime})^{-1} = Y \Sigma_E^{-1} Z' (Z \Sigma_E^{-1} Z')^{-1}$$

e per ogni equazione j in particolare:

$$\beta_{j\wedge} = y_j \Sigma_E^{-1} Z' (Z \Sigma_E^{-1} Z')^{-1}$$

Si osservi che le variabili esplicative in questo caso variano da equazione a equazione, dunque al fine di ottenere soluzioni che considerino la differenza in numero di variabili esplicative, si esprimono le matrici Y, B, ed E in termini di supervettori e Z come matrice diagonale a blocchi:

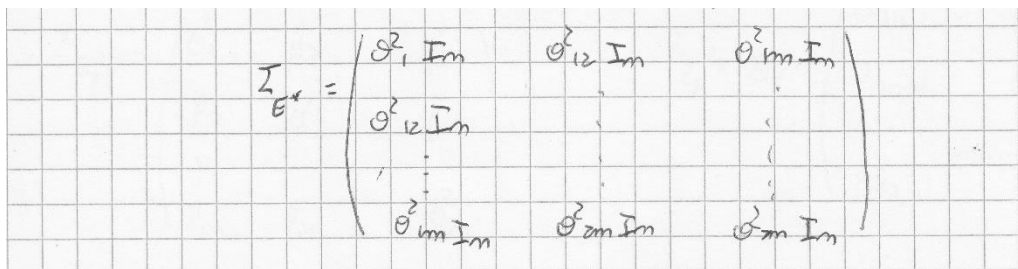
$$\begin{aligned} Y^* &= \begin{pmatrix} y_1' \\ y_2' \\ \vdots \\ y_m' \end{pmatrix} & B^* &= \begin{pmatrix} \beta_1' \\ \beta_2' \\ \vdots \\ \beta_m' \end{pmatrix} & E^* &= \begin{pmatrix} \varepsilon_1' \\ \varepsilon_2' \\ \vdots \\ \varepsilon_m' \end{pmatrix} \\ Z^* &= \begin{pmatrix} z_1' & & & \\ & z_2' & & \\ & & \ddots & \\ & & & z_3' \end{pmatrix} \end{aligned}$$

Si ottiene così il modello lineare multivariato SURE:

$$y^{*o} = b^{*o} Z^{*o} + e^{*o}$$

Vengono utilizzati solo i regressori effettivamente legati alle diverse variabili dipendenti, ciò permette di risolvere anche il problema in cui si ha un numero diverso di osservazioni nelle diverse equazioni.

Per la matrice di varianze covarianze degli errori individuali si considera:



$$\Sigma_{E^{*o}} = \begin{pmatrix} \sigma_{11}^2 I_m & \sigma_{12}^2 I_m & \sigma_{13}^2 I_m & \dots & \sigma_{1m}^2 I_m \\ \sigma_{12}^2 I_m & \sigma_{22}^2 I_m & \sigma_{23}^2 I_m & \dots & \sigma_{2m}^2 I_m \\ \sigma_{13}^2 I_m & \sigma_{23}^2 I_m & \sigma_{33}^2 I_m & \dots & \sigma_{3m}^2 I_m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{1m}^2 I_m & \sigma_{2m}^2 I_m & \sigma_{3m}^2 I_m & \dots & \sigma_{mm}^2 I_m \end{pmatrix}$$

Dunque si considerano:

1. Errori omoschedastici all'interno delle stesse equazioni.
2. Errori eteroschedastici tra diverse equazioni.
3. Errori incorrelati all'interno delle stesse equazioni.
4. Errori incorrelati fra equazioni diverse.
5. Errori correlati fra equazioni diverse per uguali osservazioni.

Si calcola la soluzione con il metodo dei minimi quadrati generalizzati, come per il modello lineare multivariato generalizzato si ha:

$$b^{*o} = y^{*o} Z^{*o'} (Z^{*o} Z^{*o'})^{-1} = y^{*o} \Sigma_{E^{*o}}^{-1} Z^{*o'} (Z^{*o} \Sigma_{E^{*o}}^{-1} Z^{*o'})^{-1}$$

con  $b^{*o}$  che ha dimensione  $(1, \Sigma_{jr})$ ,

$$y^{*o}(1, nm),$$

$$\Sigma_{E^{*o}}(nm, nm),$$

$$Z^{*o}(nm, \Sigma_{jrj})$$

Si osservano le principali differenze con la soluzione dei minimi quadrati generalizzati:

- Il modello è caratterizzato dalla presenza di un numero di variabili esplicative diverso da equazione ed equazione
- Gli errori sono omoschedastici e incorrelati nella stessa equazione, eteroschedastici, correlati per lo stesso individuo e incorrelati fra individui diversi fra diverse equazioni.