

ECONOMETRIA – Cod. 6089

Soluzioni della prima prova parziale (*Mock version 2*)

Di seguito, trovate (in forma sintetica) alcune possibili risposte alle domande della prima prova parziale (*Mock version 2*).

Domanda 1

a) Sotto le assunzioni di linearità (A1), assenza di multicollinearità perfetta (A2), campionamento casuale (A3) e media condizionata nulla (A4: $E(u|\mathbf{x}) = 0$), $\hat{\beta} = (X'X)^{-1}X'y$ è uno stimatore non distorto di β , cioè: $E(\hat{\beta}) = \beta$. Infatti:

$$E[\hat{\beta}] = E[(X'X)^{-1}X'y] = E[(X'X)^{-1}X'X\beta + (X'X)^{-1}X'u] = \beta + (X'X)^{-1}X'E[u|X] = \beta.$$

b) Sotto il nuovo modello per la popolazione, lo stimatore $\hat{\beta} = (X'X)^{-1}X'y$ non è più uno stimatore non distorto di β , visto che viene violata la condizione A4. In particolare:

$$E(\hat{\beta}) = \beta + (X'X)^{-1}X'P\lambda,$$

dove si vede che la distorsione scompare soltanto se i) le variabili in X e P sono ortogonali oppure ii) le variabili in P non hanno nessun impatto sulla y .

c) L'ipotesi nulla di perfetta specificazione funzionale può essere testata con la procedura RESET. La regressione ausiliaria del test è:

$$y = \beta_1 + \beta_2x_2 + \dots + \beta_kx_k + \delta_1\hat{y}^2 + \delta_2\hat{y}^3 + \delta_3\hat{y}^4 + \eta,$$

dove gli \hat{y} sono i valori fittati del modello originario. L'ipotesi nulla di linearità è quindi data da: $H_0 : \delta_1 = \delta_2 = \delta_3 = 0$. Possiamo usare il test F per decidere se rifiutare H_0 .

d) Il nuovo modello è: $\mathbf{y} = \tilde{X}\tilde{\beta}$, dove \tilde{X} è una matrice $[N \times (K + 2)]$ e $\tilde{\beta}$ un vettore $[(K + 2) \times 1]$. Dobbiamo testare l'ipotesi $H_0 : R\tilde{\beta} = \mathbf{q}$, dove R è una matrice $[2 \times (K + 2)]$ e \mathbf{q} un vettore $[2 \times 1]$. Più precisamente:

$$R = \begin{pmatrix} 0 & \dots & 0 & 1 & 0 \\ 0 & \dots & 0 & 0 & 1 \end{pmatrix},$$
$$\mathbf{q} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Il test F è dato da:

$$\frac{(R\hat{\beta} - \mathbf{q})'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - \mathbf{q})}{\mathbf{e}'\mathbf{e}} \cdot \frac{N - K - 2}{2} \sim F_{2, N - K - 2},$$

dove $\hat{\beta}$ è lo stimatore OLS di $\tilde{\beta}$ ed \mathbf{e} il vettore dei residui. Si noti che il test F può anche essere specificato come:

$$F = \frac{(SQR_r - SQR_c)/2}{SQR_c/(N - K - 2)} \sim F_{2, N-K-2}.$$

Domanda 2

a) Sotto le assunzioni di linearità (A1: $\mathbf{y} = X\beta + \varepsilon$), assenza di multicollinearità perfetta (A2: $\text{rang}(X) = K$), campionamento casuale (A3: i i.i.d.), media condizionata dell'errore nulla (A4: $E(\varepsilon|\mathbf{x}) = \mathbf{0}$) e omoschedasticità (A5: $\text{Var}(\varepsilon_i|X) = \sigma^2$), la varianza dello stimatore OLS è $\text{Var}(\hat{\beta}_{OLS}) = \sigma^2(X'X)^{-1}$. Il suo stimatore è $\widehat{\text{Var}}(\hat{\beta}_{OLS}) = s^2(X'X)^{-1}$, dove $s^2 = \frac{1}{N-K}\mathbf{e}'\mathbf{e}$ è uno stimatore non distorto di σ^2 .

b) Lo stimatore OLS è BLUE (*Best Linear Unbiased Estimator*), ovvero lo stimatore più efficiente (a varianza minima) tra tutti gli stimatori lineari non distorti, se valgono le ipotesi A1-A5 (Teorema di Gauss-Markov).

c) La varianza dell'errore del nuovo modello è:

$$\text{Var}(\bar{\varepsilon}_s) = \text{Var}\left(\frac{1}{N_s} \sum_{i=1}^{N_s} \varepsilon_i\right) = \frac{1}{N_s^2} \sum_{i=1}^{N_s} \text{Var}(\varepsilon_i) = \frac{1}{N_s^2} N_s \sigma^2 = \frac{1}{N_s} \sigma^2.$$

La varianza dell'errore è quindi diversa per ogni settore e inversamente correlata al numero di imprese nel settore stesso.

d) Per stimare i coefficienti con il metodo dei minimi quadrati ponderati (WLS) è sufficiente moltiplicare ogni osservazione per $\sqrt{N_s}$ e stimare con OLS il modello così trasformato (si noti che $h_s = 1/N_s$). Infatti:

$$\text{Var}\left(\sqrt{N_s}\bar{\varepsilon}_s\right) = N_s \frac{1}{N_s} \sigma^2 = \sigma^2.$$

Domanda 3

a) I coefficienti stimati di ECOPRC e di REGPRC hanno il segno atteso: la funzione di domanda è infatti inclinata negativamente ($\hat{\beta}_2 < 0$) e le mele biologiche e le mele normali risultano, come da attendersi, beni sostituti ($\hat{\beta}_3 > 0$). La stima dell'elasticità della domanda di mele biologiche al proprio prezzo (calcolata per valori medi delle due variabili) è:

$$\epsilon = -\frac{\partial ECOLBS}{\partial ECOPRC} \frac{\overline{ECOPRC}}{\overline{ECOLBS}} = -(-2.9) \frac{1.08}{1.47} = 2.13$$

Si noti che l'elasticità al proprio prezzo è convenzionalmente definita positiva.

b) La domanda suggerisce un cambiamento dell'unità di misura che equivale a dividere per 1000 le osservazioni su FAMINC. Il corrispondente coefficiente risulterà quindi moltiplicato per 1000 (quindi sarà pari a 3.08), ma continuerà ad essere non significativo.

c) E' possibile riscrivere l'ipotesi nulla come $\beta_2 + \beta_3 = 0$. Il test t è:

$$t = \frac{(\hat{\beta}_2 + \hat{\beta}_3) - 0}{S.E.(\hat{\beta}_2 + \hat{\beta}_3)} \sim t_{N-K}$$

dove $N - K = 656$. Dunque:

$$t = \frac{-2.9 + 3.03}{\sqrt{0.35 + 0.51 - 2 * 0.35}} \approx 0.33$$

Poichè $-1.96 < t = 0.18 < 1.96$ non è possibile rifiutare l'ipotesi nulla.

d) Il modello ristretto è: $ECOLBS_i = \beta_1 + \beta_3(REGPRC_i - ECOPRC_i) + \beta_4FAMINC_i + \varepsilon_i$. Utilizzando la somma dei residui al quadrato di questo modello (RSS_R) e la somma dei residui al quadrato del modello completo (RSS_U) è possibile costruire il test F :

$$F = \frac{(RSS_R - RSS_U)/1}{RSS_U/(N - K)} \sim F_{1, N-K}$$

I due test, ovviamente, devono dare lo stesso risultato in termini di significatività. In questo caso, poichè vi è un'unica restrizione, utilizzando l'equivalenza tra il test F ed il test t : $F = t^2 = (0.18)^2 = 0.0324$.

Domanda 4

a) La presenza di eteroschedasticità crea problemi alle proprietà di efficienza e alle procedure inferenziali standard degli stimatori OLS. Se $E(\mathbf{u}\mathbf{u}') = \sigma^2\Omega \neq \sigma^2I$ (eteroschedasticità), la formula per la matrice di varianza e covarianza degli stimatori OLS diventa:

$$Var(\hat{\beta}) = E[(X'X)^{-1}X'\mathbf{u}\mathbf{u}'X(X'X)^{-1}] = \sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1},$$

che è diversa da quella derivata sotto condizione di omoschedasticità ($\sigma^2(X'X)^{-1}$). La stima della varianza, la dimostrazione del teorema di Gauss-Markov e le procedure inferenziali derivate sotto questa assunzione cessano di essere valide.

b) Una possibile soluzione per il problema di eteroschedasticità è quella di usare il seguente stimatore per la matrice di varianza e covarianza definita sopra:

$$\hat{Var}(\hat{\beta}) = (X'X)^{-1} \sum e_i^2 \mathbf{x}_i' \mathbf{x}_i (X'X)^{-1},$$

dove \mathbf{x}_i è il vettore delle k variabili osservate per l'osservazione i . Una volta applicato questo metodo (detto appunto Huber-White) per stimare la matrice di varianza e covarianza,

anche in presenza di eteroschedasticità, si può procedere nel solito modo sia per individuare lo *standard error* di ogni coefficiente stimato e quindi il test t , sia per individuare il test F su ipotesi multiple. Il confronto tra le stime riportate nell'esercizio, tuttavia, mostra poche differenze significative. I coefficienti stimati, ovviamente, sono gli stessi (OLS), mentre cambiano gli standard error e le statistiche test, ma le variabili significativamente diverse da zero in precedenza rimangono tali anche con lo stimatore Huber-White.

c) Il coefficiente di determinazione R^2 è definito come:

$$R^2 = \frac{SQS}{SQT} = 1 - \frac{SQR}{SQT}.$$

L' R^2 è compreso tra zero e uno (se c'è una costante nel modello) e fornisce una misura della qualità dell'approssimazione lineare rispetto ai dati del nostro campione. In altre parole, se il nostro modello è giusto, l' R^2 quantifica la capacità del modello di predire nel campione. Il suo valore, infatti, cattura la frazione della variazione campionaria di y spiegata dalla variazione delle x_j . Il confronto tra l' R^2 dei due modelli stimati nell'esercizio non è del tutto agevole perché nel secondo modello abbiamo aggiunto un regressore, e sappiamo che l'aggiunta di un qualsiasi regressore aumenta sempre il valore dell' R^2 . Comunque, un salto da una frazione di varianza spiegata intorno al 4% (primo modello) a una frazione di varianza spiegata di quasi il 28% (secondo modello), ci fa azzardare l'ipotesi che la capacità predittiva del modello nel campione sia aumentata in maniera consistente.

d) Il coefficiente della variabile *train* cambia addirittura di segno a seguito dell'introduzione di un regressore aggiuntivo (*unem75*). Sulla base della significatività economica e statistica del nuovo regressore, possiamo concludere che la prima stima (che indicava addirittura un effetto negativo del corso di formazione) fosse distorta per un problema di variabile omessa. La distorsione della prima stima era negativa, come ci saremmo attesi per il fatto che la variabile *unem75* ha un effetto negativo sulla variabile dipendente (*em78*), ma è probabilmente correlata positivamente alla probabilità di seguire il corso di formazione (*unem75*). La formula per la distorsione da variabile omessa (si veda anche il punto b della domanda 1) può aiutare a capire questa conclusione.