ECONOMETRIA – Cod. 6089
PRIMA PROVA PARZIALE:
TRACCIA SOMMARIA DELLE SOLUZIONI

PRIMA PARTE (“TEORICA”)

Domanda 1

Considerate il seguente modello di regressione lineare multipla $y_i = x_i'\beta + \epsilon_i$ e assumete che le ipotesi classiche A1-A5 siano verificate.

(**a**) Scrivete la formula dell’R^2. Che cosa rappresenta? Calcolatelo con riferimento alla regressione il cui output è riportato nella tabella sottostante.

```
      Source |       SS       df       MS              Number of obs =     130
-------------+------------------------------           F(          ) =25317.94
       Model |  14735.7495     2  7367.87474           Prob > F      =  0.0000
    Residual |  36.9587758   127  .291013982           R-squared     =
-------------+------------------------------           Adj R-squared =
       Total |  14772.7083   129  114.517118           Root MSE      =  .53946

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          X2 |   .8009959   .0328644    24.37   0.000     .7359632    .8660286
          X3 |   1.604939   .0117895   136.13   0.000      1.58161    1.628268
       _cons |   .8910662   .2287712     3.90   0.000     .4383692    1.343763
------------------------------------------------------------------------------
```

$$R^2 = \frac{SQS}{SQT} = 1 - \frac{SQR}{SQT} = 1 - \frac{36.96}{14772.7} = 0.9975$$

Il coefficente di determinzaione R^2 rappresenta una misura della bontà del fit (ovvero della capacità del modello lineare di predire i dati nel campione): è il rapporto tra la varianza spiegata e quella totale.

(**b**) Spiegate come potete testare la significatività congiunta dei regressori utilizzando il valore dell’R^2 così trovato, avendo cura di scrivere la statistica test utilizzata, la sua distribuzione ed il criterio di decisione.

Stimare la significatività congiunta dei regressione significa testare l’ipotesi nulla: $H_0 : \beta_2 = \beta_3 = 0$, utilizzando un test F con la seguente statistica test e distribuzione:

$$F = \frac{(SQR_r - SQR_c)/(K-1)}{SQR_c/(N-K)} \sim F_{j,N-k}$$

Dato che $SQT = SQS + SQR$,

possiamo anche scrivere $SQR_c = (1 - R_c^2)SQT$ e $SQR_r = (1 - R_r^2)SQT$.

La statistica F può essere riscritta come:

$$F = \frac{(R_c^2 - R_r^2)/(K-1)}{(1-R_c^2)/(N-K)} \sim F_{j,N-k}$$

Per valori del p-value inferiori al 5%, rifiuto l'ipotesi nulla: i coefficienti non sono congiuntamente uguali a zero.

(**c**) Dimostrate che $R^2 = corr^2(y, \hat{y})$.

$$R^2 = \frac{ESS}{TSS} = \frac{\sum(\hat{y}-\bar{\hat{y}})^2}{\sum(\hat{y}-\bar{y})^2}$$

$$corr^2(y,\hat{y}) = \frac{cov^2(y,\hat{y})}{var(y)var(\hat{y})} = \frac{cov^2(\hat{y}+\hat{e},\hat{y})}{var(y)var(\hat{y})} = \frac{var^2(\hat{y})}{var(y)var(\hat{y})} = \frac{var(\hat{y})}{var(y)} = \frac{\sum(\hat{y}-\bar{y})^2}{\sum(\hat{y}-\bar{y})^2} = R^2$$

(**d**) Utilizzate la formula dell'R^2 per mostrare cosa accade all'R^2 stesso quando sia la y che tutte le variabili esplicative sono moltiplicate per 10.

Riscriviamo il modello come: $Y_i = Z_i'\beta + \epsilon_i$ *dove* $Y_i = ky_i$ *e* $Z_i = kx_i$

Calcolo il coefficiente di determinazione come:

$$R^2 = \frac{SQS}{SQT} = 1 - \frac{SQR}{SQT} = 1 - \frac{\sum(\hat{Y}_i-\bar{Y})^2}{\sum(Y_i-\bar{Y})^2}$$

posso riscriverlo come:

$$R^2 = 1 - \frac{\sum(\hat{Y}_i-\bar{Y})^2}{\sum(Y_i-\bar{Y})^2} = 1 - \frac{\sum(\widehat{ky}-k\bar{y})^2}{\sum(ky_i-k\bar{y})^2} = 1 - \frac{k^2}{k^2}\frac{\sum(\hat{y}-\bar{y})^2}{\sum(y_i-\bar{y})^2} = 1 - \frac{\sum(\hat{y}-\bar{y})^2}{\sum(y_i-\bar{y})^2}$$

Domanda 2

Considerate il seguente modello di regressione lineare multivariata $y = X\beta + \epsilon$ e assumete che le ipotesi classiche A1-A5 siano verificate.

(**a**) Ricavate la formula dello stimatore OLS. Tutte le ipotesi A1-A5 sono necessarie? Argomentate la vostra risposta.

Nel caso multivariato il problema della minimizzazione dei quadrati può essere scritto così:

$$\underset{\beta}{Min}\ [S(\beta) = (y - X\beta)'(y - X\beta)]$$

Dato che possiamo riscrivere

$$S(\beta) = y'y + \beta'X'X\beta - \beta'X'y - y'X\beta = y'y + \beta'X'X\beta - 2\beta'X'y,$$

dalle condizioni del primo ordine otteniamo che: $2X'X\beta - 2X'y = 0.$

Risolvendo si ottiene che: $\hat{\beta} = (X'X)^{-1}X'y.$

Abbiamo quindi utilizzato le ipotesi A1-A2.

(**b**) Dimostrate la correttezza dello stimatore OLS. Tutte le ipotesi A1-A5 sono necessarie? Argomentate la vostra risposta.

Sono necessarie le ipotesi A1-A4. In altre parole, deve essere garantita l'esistenza di una relazione lineare tra x e y; dobbiamo poter escludere l'esistenza di multicollinearità perfetta; e dobbiamo assumere la condizione di

esogeneità (A3-A4).

Uno stimatore è corretto se $E(\hat{\beta}) = \beta$

$E[\hat{\beta}] = E[(X'X)^{-1}X'y] = E_{x,\varepsilon}[(X'X)^{-1}X'X\beta + (X'X)^{-1}X'X\varepsilon] =$

$= E_x[\beta + (X'X)^{-1}X'E(\varepsilon|X)] \ = \beta$ *dato che* $E(\varepsilon|X) = 0$

(**c**) Utilizzate la formula ricavata al punto (a) per mostrare cosa accade allo stimatore quando sia la y che tutte le variabili esplicative sono moltiplicate per 10.

Supponiamo di voler calcolare il valore dello stimatore $\hat{\delta} = (Z'Z)^{-1}Z'Y$

dove $Z = kX$ *e* $Y = ky$

È possible riscrivere come:

$\hat{\delta} = (X'k'kX)^{-1}X'k'ky$

da cui si ottiene $\hat{\delta} = (X'X)^{-1}X'y = \hat{\beta}$ *dato che* $(k'k)^{-1'}k'k = I$

(**d**) Assumete ora che $Var(\epsilon) = \Omega \neq \sigma^2 I$. Quali sono le proprietà dello stimatore OLS in questo caso?

Se viene infranta l'ipotesi A5 sull'omoschedasticità, lo stimatore OLs è ancora corretto, ma non più BLUE.

SECONDA PARTE ("APPLICATA")

Domanda 3

Avete a disposizione un dataset contenente informazioni sul logaritmo dello stipendio (*lwage*) di 428 donne residenti negli Stati Uniti, e volete stimare le determinanti di questa variabile. Il dataset contiene anche le seguenti informazioni: *exper* (anni di esperienza), *expersq* (anni di esperienza al quadrato), *faminc* (il reddito familiare in dollari) e *motheduc* (il livello di istruzione della madre).

Supponete di stimare il seguente modello:

$$\text{lwage} = \beta_1 + \beta_2\text{exper} + \beta_3\text{expersq} + \beta_4\text{motheduc} + \beta_5\text{faminc} + \varepsilon$$

i cui risultati sono riportati qui sotto:

Source	SS	df	MS
Model	36.751461	4	9.18786525
Residual	186.57598	423	.441077967
Total	223.327441	427	.523015084

Number of obs =	428
F(4, 423) =	20.83
Prob > F =	0.0000
R-squared =	0.1646
Adj R-squared =	0.1567
Root MSE =	.66414

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
exper	.0426518	.0131223	3.25	0.001	.0168588	.0684449
expersq	-.000826	.0003922	-2.11	0.036	-.0015968	-.0000551
motheduc	.0047702	.0098531	0.48	0.629	-.0145969	.0241373
faminc	.0000215	2.78e-06	7.76	0.000	.0000161	.000027
_cons	.2626053	.1437752	1.83	0.068	-.0199975	.545208

(**a**) Valutate il livello di significatività del coefficente associato a *faminc.* Come potete interpretare il suo valore? Ricavate l'espressione dell'elasticità del salario al reddito familiare.

Il coefficiente associato a faminc è significativo all'1%.

In questo caso abbiamo un modello log-lineare e quindi il coefficiente riportato in tabella rappresenta la semi-elesticità. Possiamo dire che un incremento della variabile faminc ha un impatto positivo sulla variabile dipendente.

Per ricavare l'espressione dell'elasticità partiamo dal significato del coefficiente.Dato che la variabile dipendente è in logaritmi (lwage), mentre la variabile esplicativa è in livelli (faminc), il valore coefficente non è di immediata interpretazione (non è l'effetto marginale)

$$\beta_5 = \frac{\vartheta lwage}{\vartheta faminc}$$

da cui moltiplicando per entrambi i termini per faminc otteniamo

$$\beta_5 * faminc = \frac{\frac{1}{wage} dwage}{\frac{1}{faminc} \vartheta faminc} = \varepsilon_{wage,faminc}$$

(**b**) Come potete interpretare l'impatto dell'esperienza sul salario femminile? Si tratta di un impatto costante al variare degli anni di esperienza? Esiste un numero di anni a cui tale impatto cambia di segno?

Prendendo in considerazione entrambi i coefficienti delle variabili collegate all'esperienza, sia di exper che di expersq, possiamo dedurre che siamo in presenza di una funzione concava del salario all'esperienza.

Maggiore è l'esperienza maggiore sarà il salario (effetto del coefficiente positivo di exper), ma all' aumentare degli anni di esperienza, l'impatto dell' esperienza sul salario diminuirà (effetto del coefficente negativo di expersq).

Per calcolare in che anno l'impatto cambia segno, si risolve l'equazione:

$$0.04265 * x - 0.000826 * x^2 = 0 \Rightarrow 0.04265 - 2 * 0.000826 * x = 0 \quad \Rightarrow \quad x = 25.81$$

0Sui residui della regressione OLS viene effettuato il test di Breusch-Pagan, il cui risultato è riportato di seguito. Come è costruito il test? Specificate l'ipotesi

nulla, l'ipotesi alternativa, la statistica test e la sua distribuzione. Quali sono le implicazioni del risultato qui sotto per le proprietà dello stimatore OLS?

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of lwage

         chi2(1)      =    10.61
         Prob > chi2  =   0.0011
```

Il Test di Breusch-Pagan è utilizzato per verificare la presenza di eteroschedasticità. Ottenuti i residui della regressione di partenza, si utilizziano come variabile indipendente e come le variabili esplicative i regressori del nostro modello:

$e^2 = \delta_1 + \delta_2$exper+$\delta_3$expersq+$\delta_4$motheduc+$\delta_5$faminc+$\eta$

dove viene testata l'ipotesi nulla $H_0 : \delta_2 = \delta_3 = \delta_4 = \delta_5 = 0$

La statistica test e la sua distribuzione sono: $\frac{R_{\hat{e}}^2/(5-1)}{(1-R_{\hat{e}}^2)/(428-5)} \sim F_{4,423}$

i risultati riportati nell'output del test portano a rifiutiare l'ipotesi nulla: ci sono quindi problemi di eteroschedasticità

Data la presenza di eteroschedasticità, lo stimatore OLS è ancora corretto, ma non è più BLUE

(**d**) Supponete ora di voler ora testare l'ipotesi nulla: $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$. Che test utilizzereste? Qual'è la sua statistica test? Calcolate il valore della statistica test sulla base della stima del modello ristretto riportata di seguito.

Source	SS	df	MS
Model	27.4800213	1	27.4800213
Residual	195.84742	426	.459735727
Total	223.327441	427	.523015084

```
Number of obs =     428
F(  1,   426) =   59.77
Prob > F      =  0.0000
R-squared     =  0.1230
Adj R-squared =  0.1210
Root MSE      =  .67804
```

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
faminc	.0000217	2.81e-06	7.73	0.000	.0000162	.0000273
_cons	.6656776	.0753422	8.84	0.000	.5175888	.8137664

Si utilizza un test F, che testa la significatività congiunta di tutti i coefficienti delle variabili indipendenti, ad esclusione della costante:

l'ipotesi nulla è:

$H_0 : \beta_2 = \beta_3 = \beta_4 = 0$

$J = 3;\ N = 428$

mentre la statistica test ed il suo valore sono:

$$F = \frac{(SQR_r - SQR_c)/J}{SQR_c/N-K} = \frac{(SQR_r - SQR_c)/3}{SQR_c/423} =$$

$$= \frac{(R_e^2 - R_r^2)/3}{(1-R_e^2)/(N-5)} = \frac{(0.1646-0.123)/3}{(1-0.1646)/423} = 7,0213 \sim F_{(3,423)}$$

Domanda 4

Supponete di avere un campione composto da 3271 donne, contenente osservazioni sul numero di figli (*children*), sul livello di educazione della madre (*edu*), sull'età del primo figlio (*agefbrth),* sull'età della madre (*age)*. Sapete inoltre se la donna possiede una televisione (*tv*=1) e se vive in città (*urban*=1*).* Stimate con il seguente modello lineare le determinanti del numero di figli per donna:

$$\text{children} = \beta_0 + \beta_1\text{educ} + \beta_2\text{tv} + \beta_3\text{agefbrth} + \beta_4\text{age} + \beta_5\text{agesq} + \beta_6\text{urban} + \varepsilon$$

dove *agesq* è il quadrato dell'età della donna, e ottenete i seguenti risultati:

Source	SS	df	MS
Model	7901.99832	6	1316.99972
Residual	6156.50366	3264	1.88618372
Total	14058.502	3270	4.29923608

Number of obs =	3271
F(6, 3264) =	698.24
Prob > F =	0.0000
R-squared =	0.5621
Adj R-squared =	0.5613
Root MSE =	1.3734

children	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	-.044728	.006851	-6.53	0.000	-.0581608	-.0312953
tv	-.2359453	.0922769	-2.56	0.011	-.4168719	-.0550188
agefbrth	-.2100026	.0084514	-24.85	0.000	-.2265731	-.1934321
age	.3777368	.0222689	16.96	0.000	.3340743	.4213992
agesq	-.0029441	.0003426	-8.59	0.000	-.0036159	-.0022723
urban	-.3192239	.0506493	-6.30	0.000	-.4185315	-.2199163
_cons	-1.063848	.3619346	-2.94	0.003	-1.77349	-.3542063

(**a**) Commentate la significatività ed il valore del coefficente del livello di educazione della madre (*edu*) . Scrivete l'espressione dell'effetto marginale dell'educazione della madre sul numero di figli e la corrispondente elasticità.

Il coefficiente di edu è significativamente diverso da zero ad un livello di significatività dell'1%. Essendo un modello in cui le variabili sono espresse in livelli il coefficiente rappresenta l'effetto marginale: aumentando di un anno il livello di istruzione della madre, si ottiene una riduzione (segno negativo) del numero dei figli pari al valore del coefficente della variabili analizzata (0.044728).

dato che il coefficiente è uguale a:

$$\beta_1 = \frac{\partial children}{\partial educ}$$

possiamo scrivere l'elasticità come:

$\varepsilon_{children,educ} = \frac{\vartheta children}{\vartheta educ} \frac{educ}{children} = \beta_1 \frac{educ}{children}$

(**b**) Una volta osservato l'output della regressione decidete di testare l'ipotesi nulla $\beta_1 = -\beta_3$ al 5%. Specificate l'ipotesi nulla, l'ipotesi alternativa, la statistica test e la sua distribuzione. Effettuate il test e commentatene il risultato, utilizzando la matrice di varianza dello stimatore OLS sotto riportata, e sapendo che il valore critico della corrispondente distribuzione t al 5% è pari a 1.96.

	educ	tv	agefbrth	age	agesq	urban	_cons
educ	.00004694						
tv	-.00020284	.00851503					
agefbrth	-.00001159	-.00005304	.00007143				
age	.00001315	-.00010503	-.00001815	.0004959			
agesq	-8.377e-08	9.395e-07	1.263e-07	-7.542e-06	1.174e-07		
urban	-.00005994	-.00083178	.00001601	-.00006167	1.036e-06	.00256535	
_cons	-.00029848	.0039998	-.00087543	-.0072959	.00011057	-.00035693	.13099669

Si riscriva l'ipotesi nulla come $H_0 : \beta_1 + \beta_3 = 0$

dove l'ipotesi alternativa è : $H_1 : \beta_1 + \beta_3 \neq 0$

La statistica test t è uguale a:

$t = \frac{(\widehat{\beta_1}+\widehat{\beta_3})-0}{S.E.(\widehat{\beta_1}+\widehat{\beta_3})} \sim t_{N-K}$ *dove* $\widehat{\beta_1}, \widehat{\beta_3}$ *sono i valori stimati dei coefficenti*

in questo caso $N - K = 3271 - 7 = 3264$

la statistica test può essere riscritta come:

$t = \frac{(-0.045-0.21)-0}{\sqrt{0.000047+0.000071-2*0.000012}} = \frac{-0.254}{0.00976} \approx -26.107$

è possibile rifiutare l'ipotesi nulla dato che $t = -26.107 < -1.96$

(**c**) La variabile *urban* è una dummy. Quale interpretazione date al suo coefficente?

La variabili urban è una variabile dummy, che assume valore 1 se l'individuo vive in vittà, 0 altrimenti.

Dato che il coefficiente risulta statisticamente diverso da zero, è possibile affermare che, a parità di tutte le altre condizioni, gli individui che vivono in città hanno un numero di figli inferiore (coefficiente negativo) rispetto a chi vive fuori dalla città pari al valore del coefficiente (-0.319).

(**d**) Il test di Ramsey effettuato in seguito alla regressione OLS fornisce il seguente risultato:

```
Ramsey RESET test using powers of the fitted values of children
       Ho:  model has no omitted variables
                 F(3, 3261) =     26.42
                   Prob > F =      0.0000
```

Commentate il risultato, descrivendo l'ipotesi nulla ed alternativa e la statistica test. Che implicazioni ne traete per il modello stimato?

Il test di Ramsey è utilizzato per valutare la presenza di mispecificazione del

modello. Il Ramsey-Reset test svolge una regressione aumentata del modello originale che include i regressori originali, le potenze dei valori predetti della regressione originale (vengono inserite le prime 3 potenze di $\hat{y}$: la seconda , la terza, la quarta, sapendo che altre scelte sono comunque plausibili).

Data la regressione originale.

children= $\beta_0 + \beta_1$educ+β_2TV+$\beta_3$agefbrth+$\beta_4$age+$\beta_5$agesq+$\beta_6$urban+ε,

si stimia il modello ed otteniamo $\hat{y}$,

si costruisce la regressione aumentata:

$y = \beta_0 + \beta_1$educ+β_2TV+$\beta_3$agefbrth+$\beta_4$age+$\beta_5$agesq+$\beta_6$urban+$\delta_1\hat{y}^2 + \delta_2\hat{y}^3 + \delta_3\hat{y}^4 +$

dove l'ipotesi nulla è : $H_0 : \delta_1 = \delta_2 = \delta_3 = 0$

e l'ipotesi alternativa : $H_1 : \exists\delta_k \neq 0$.

Si utilizza un test F per decidere se rifiutare o accettare l'ipotesi nulla H_0

In questo caso il p-value è inferiore a qualsiasi livello di significatività, quindi bisogna rifiutare l'ipotesi nulla, e di conseguenza il modello presenta problemi di mispecificazione.