ECONOMETRIA – Cod. 6089
ESAME PARZIALE (*MOCK VERSION 1*)

Di seguito, trovate (in forma sintetica) alcune possibili risposte alle domande della prima prova parziale (*Mock version 1*).

PRIMA PARTE ("TEORICA")

1. Si consideri il seguente modello lineare: $y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$

   a) Si spieghi sinteticamente come si ottiene lo stimatore OLS di $\beta = (\beta_1, \beta_2, \beta_3)'$ specificando la funzione obiettivo e le tre condizioni del primo ordine. Si riporti quindi la formula in notazione matriciale di $\hat{\beta}_{OLS}$.

   $\hat{\beta}_{OLS}$ *si ottiene minimizzando la somma dei residui al quadrato. Quindi:*
   $$\min_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_1 - \beta_2 x_{2i} - \beta_3 x_{3i}\right)^2$$
   Le condizioni del primo ordine sono:
   1. $-2\sum_{i=1}^{N}\left(y_i - \beta_1 - \beta_2 x_{2i} - \beta_3 x_{3i}\right) = 0$
   2. $-2\sum_{i=1}^{N}\left(y_i - \beta_1 - \beta_2 x_{2i} - \beta_3 x_{3i}\right)x_{2i} = 0$
   3. $-2\sum_{i=1}^{N}\left(y_i - \beta_1 - \beta_2 x_{2i} - \beta_3 x_{3i}\right)x_{3i} = 0$

   La formula in notazione matriciale è: $\hat{\beta}_{OLS} = \left(X'X\right)^{-1} X'y$

   b) Supponendo che il modello sia correttamente specificato, quali sono le proprietà di cui gode lo stimatore OLS nei piccoli campioni e quali sono le ipotesi che garantiscono tali proprietà?

   Assunzioni:

   A1) Modello lineare nei parametri: $y = X\beta + \varepsilon$ *, dove X è una matrice (NxK)*
   A2) Rango(X) = K (assenza di multicollinearità perfetta)
   A3) Campionamento casuale (i=1…N).
   A4) Media condizionata nulla (esogeneità): $E(\varepsilon_i \mid X) = 0$.
   A5) Varianza condizionata costante (omoschedasticità): $Var(\varepsilon_i \mid X) = \sigma^2$.
   A6) Normalità: ε_i iid $N(0, \sigma^2)$.

   Proprietà dello stimatore OLS:

1. Non distorto ($E(\hat{\beta}) = \beta$ *)*	*[da A1 ad A4]*
2. BLUE, $Var(\hat{\beta}) = \sigma^2 (X'X)^{-1}$	*[da A1 ad A5]*
3. $\hat{\beta} \sim N(\beta, \sigma^2 (X'X)^{-1})$	*[da A1 ad A6]*

c) Ipotizzate che $x_{2i} = 3 + 2 * x_{3i}$. Cosa succederà se tentate di stimare il modello? Perché?

Se x_2 è una combinazione lineare della costante e di x_3, come nel caso ipotizzato, $\hat{\beta}_{OLS}$ non si può calcolare perché non si può invertire la matrice X'X. Si tratta del problema della multicollinearità perfetta. Viene dunque violata l'ipotesi A2.

d) Ipotizzate ora che il modello sia: $y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{2i} * x_{3i} + \varepsilon_i$. Qual è l'effetto marginale di x_3? Come lo interpretate?

L'effetto marginale è: $\partial y_i / \partial x_{3i} = \beta_3 + \beta_4 x_{2i}$. L'effetto marginale di x_3 non è costante, ma cresce al crescere di x_2 se $\beta_4 > 0$, mentre diminuisce al crescere di x_2 se $\beta_4 < 0$.

2. Considerate il modello di regressione: $y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \beta_4 x_i^3 + u_i$.

a) Sulla base delle stime OLS $\hat{\beta}_i$ (con *i=1,2,3,4*), si stimi l'effetto marginale di x su y. È costante?

L'effetto marginale non è costante per via dei due termini non-lineari. Infatti, è dato da: $dy/dx = \hat{\beta}_2 + 2\hat{\beta}_3 x + 3\hat{\beta}_4 x^2$.

b) Assumendo che $\hat{\beta}_4 < 0$, per quale valore di x l'effetto marginale trovato al punto a) è massimo?

L'effetto marginale è massimo se (condizione del primo ordine): $2\hat{\beta}_3 + 6\hat{\beta}_4 x = 0$. *Cioè, se:* $x = -\hat{\beta}_3 / 3\hat{\beta}_4$. *La condizione del secondo ordine è soddisfatta visto che* $\hat{\beta}_4 < 0$.

c) Si supponga di voler sottoporre a test l'ipotesi nulla che la forma funzionale del modello è correttamente specificata. Quale procedura potremmo usare?

Possiamo applicare il test di Ramsey (RESET). Prendiamo i valori fittati del modello specificato sopra (che diventa il nostro modello ristretto) e stimiamo il modello completo: $y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \beta_4 x_i^3 + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + \delta_3 \hat{y}^4 + v_i$. *Dopodiché, applichiamo un test F per l'ipotesi nulla* $H_0 : \delta_1 = \delta_2 = \delta_3 = 0$, *del tipo:* $\dfrac{(SQR_r - SQR_c)/3}{SQR_c/(N-7)}$. *Se l'ipotesi nulla è rifiutata al livello di significatività prescelto, abbiamo un problema di specificazione funzionale.*

d) Se invece volessimo confrontare direttamente il modello specificato con un modello alternativo puramente lineare della relazione tra y ed x (cioè, senza x^2 e x^3), che procedura inferenziale potremmo applicare?

In questo caso, stimeremmo $y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \beta_4 x_i^3 + u_i$ *come modello completo e* $y_i = \beta_1 + \beta_2 x_i + \eta_i$ *come modello ristretto. Per poi sottoporre a test l'ipotesi nulla* $H_0 : \beta_3 = \beta_4 = 0$ *con un test F:* $\dfrac{(SQR_r - SQR_c)/2}{SQR_c/(N-4)}$.

SECONDA PARTE ("APPLICATA")

3. Uno studio recente ha evidenziato il ruolo delle principali determinanti della remunerazione dei CEO, utilizzando un campione di 209 dirigenti di imprese attive in uno dei quattro fondamentali raggruppamenti settoriali: finanza, beni di consumo, produzione e distribuzione di energia e, infine, trasporti. La relazione stimata è la seguente:

$$\log(salary_i) = \beta_1 + \beta_2 \log(sales_i) + \beta_3 ROE_i + \beta_4 finance_i + \beta_5 consprod_i + \beta_6 utility_i + \varepsilon_i$$

dove *salary* è la remunerazione del manager, *sales* il fatturato dell'impresa, *ROE* è una misura di profittabilità, e *finance*, *consprod* e *utility* sono tre variabili dummy uguali ad 1 se l'impresa è attiva, rispettivamente, nel settore della finanza, dei beni di consumo o della produzione e distribuzione di energia.

a) Perché nel modello non è inclusa la variabile dummy *transport*, uguale ad 1 se l'impresa è attiva nel settore dei trasporti? Cosa rappresentano i coefficienti delle variabili *finance*, *consprod* e *utility*?

Se si inserisse anche la variabile transport si genererebbe multicollinearità perfetta, essendo la dummy transport una combinazione lineare delle altre 3:

transport = 1 – (finance + consprod + utility)

Il coefficiente della variabile dummy finance rappresenta la differenza di remunerazione tra il CEO di un'impresa attiva nel settore finance relativamente al CEO di un'impresa attiva nel settore transport (la dummy esclusa), a parità di altri fattori (sales e ROE). I coefficienti delle altre due dummies hanno interpretazione analoga.

b) In seguito alla stima OLS del modello sopra riportato viene effettuato il test di Ramsey che fornisce il seguente risultato:

```
Ramsey RESET test using powers of the fitted values of lsalary
      Ho:  model has no omitted variables
               F(3, 200) =      4.24
                Prob > F =      0.0063
```

Quali sono l'ipotesi nulla e l'ipotesi alternativa del test? Quale è la regressione ausiliaria su cui è costruito? Specificate, inoltre, la statistica test e la sua distribuzione. Sulla base dei valori riportati, rifiutate l'ipotesi nulla?

Sotto l'ipotesi nulla di linearità, il vero modello è dato da:

$$\log(salary_i) = \beta_1 + \beta_2 \log(sales_i) + \beta_3 ROE_i + \beta_4 finance_i + \beta_5 consprod_i + \beta_6 utility_i + \varepsilon_i .$$

L'ipotesi alternativa, invece, prevede rilevanti non-linearità:

$$\log(salary_i) = \beta_1 + \beta_2 \log(sales_i) + \beta_3 ROE_i + \beta_4 finance_i + \beta_5 consprod_i + \beta_6 utility_i + h(\alpha, x) + \varepsilon_i ,$$

dove $h(\alpha, x)$ *è una funzione non lineare delle variabili esplicative incluse nel modello. Il test di Ramsey serve dunque a testare l'ipotesi di linearità della regressione.*
La regressione ausiliaria del test è:

$$\log(salary_i) = \beta_1 + \beta_2 \log(sales_i) + \beta_3 ROE_i + \beta_4 finance_i + \beta_5 consprod_i + \beta_6 utility_i + \\ + \gamma_1 \hat{y}_i^2 + \gamma_2 \hat{y}_i^3 + \gamma_3 \hat{y}_i^4 + v_i$$

dove $\hat{y}$ *è il valore fittato ottenuto dalla regressione OLS sul vero modello sotto* H_0.

Dunque:

$$H_0 : \gamma_1 = 0, \gamma_2 = 0, \gamma_3 = 0 \quad vs. \quad H_a : \exists\, \gamma_j \neq 0 \quad j = 1,2,3$$

$$F = \frac{(SQR_r - SQR_c)/3}{SQR_c /(N-K-3))} \sim \mathrm{F(3, N - K - 3)}$$

Essendo il p-value inferiore anche all'1%, l'ipotesi nulla è rifiutata, ovvero il modello lineare non è correttamente specificato.

c) Nello studio è a questo punto effettuata la regressione OLS riportata nella tabella sotto riportata, dove al modello iniziale è stata aggiunta la variabile $lsales2 = \log(sales)^2$.

Source	SS	df	MS
Model	25.4928645	6	4.24881075
Residual	41.2292987	202	.204105439
Total	66.7221632	208	.320779631

Number of obs =	209
F(6, 202) =	20.82
Prob > F =	0.0000
R-squared =	0.3821
Adj R-squared =	0.3637
Root MSE =	.45178

lsalary	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lsales	1.177207	.3220406	3.66	0.000	.5422143	1.812199
lsales2	-.0552807	.0192577	-2.87	0.005	-.0932526	-.0173088
roe	.0102955	.0042354	2.43	0.016	.0019442	.0186468
finance	.1194572	.0884781	1.35	0.178	-.055002	.2939163
consprod	.2024786	.0836347	2.42	0.016	.0375695	.3673876
utility	-.3388436	.0994316	-3.41	0.001	-.5349005	-.1427867
_cons	.8433673	1.336347	0.63	0.529	-1.791611	3.478346

Come potete valutare la significatività complessiva della regressione? Spiegate quale test usate, specificando l'ipotesi nulla e quella alternativa, la statistica test e la sua distribuzione. Utilizzate quindi l'output della regressione riportato nella tabella per testare l'ipotesi da voi specificata.

Per valutare la significatività della regressione utilizzo un test F, che testa la significatività congiunta di tutti i coefficienti delle variabili indipendenti, ad esclusione della costante:

$$H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0 \quad vs. \quad H_a : \exists\, \beta_j \neq 0 \quad j = 1,...,7$$

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

Dove SQR_r è la somma dei residui al quadrato del modello ristretto (ovvero con la costante come unica variabile esplicativa), SQR_c è la somma dei residui al quadrato del modello completo, $K - 1 = 6$ e $N - K = 202$.

La statistica F è pari a 20.82 ed il corrispondente p-value è 0 $\Rightarrow$ si rifiuta l'ipotesi nulla: la regressione è significativa.

d) Qual è l'elasticità della remunerazione al fatturato? Come varia al crescere del fatturato?

$$elasticità = \frac{\partial salary}{\partial sales}\frac{sales}{salary} = \frac{\partial \log(salary)}{\partial \log(sales)} = 1.177 + 2 * (-0.055)\log(sales)$$

L'elasticità diminuisce al crescere del fatturato.

4. Si consideri un campione di 1260 lavoratori. Conosciamo il loro salario in forma logaritmica (*lwage*), la loro esperienza lavorativa in anni (*exper*), i loro anni di studio (*educ*), se sono donne (*female*), e se la loro bellezza – secondo valutazioni realizzate sulla base di fotografie da persone che non li conoscono – è superiore (*abvavg*) o inferiore (*belavg*) al valore medio. In altre parole, su una scala di bellezza da 1 a 5, gli individui con valori 4 o 5 hanno *abvavg=1*, mentre gli individui con 1 o 2 hanno *belavg=1*. Si stimi il seguente modello di regressione.

```
. reg  lwage exper educ female belavg abvavg

      Source |       SS       df       MS              Number of obs =    1260
-------------+------------------------------           F(  5,  1254) =  128.55
       Model |  150.793202     5  30.1586404           Prob > F      =  0.0000
    Residual |   294.18677  1254    .2345987           R-squared     =  0.3389
-------------+------------------------------           Adj R-squared =  0.3362
       Total |  444.979972  1259  .353439215           Root MSE      =  .48435

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0137182   .0012111    11.33   0.000     .0113423    .0160941
        educ |   .0712334   .0053365    13.35   0.000     .0607639    .0817029
      female |  -.4617922   .0296475   -15.58   0.000    -.5199564    -.403628
      belavg |  -.1484752   .0429883    -3.45   0.001    -.2328121   -.0641384
      abvavg |  -.0097954   .0311361    -0.31   0.753    -.0708801    .0512893
       _cons |   .6951378   .0778697     8.93   0.000     .5423685    .8479071
------------------------------------------------------------------------------
```

a) Come possiamo interpretare il segno e la significatività dei coefficienti stimati? In particolare, a parità di altre condizioni, le donne guadagnano di più o di meno degli uomini? Di quanto in termini percentuali?

Tutti i coefficienti stimati sono significativamente diversi da zero a un livello di significatività dell'1% (ad esclusione di abvavg) ed hanno il segno atteso: a parità di altre condizioni, si guadagna di più se si ha una maggiore esperienza, più anni di studio, se si è uomini e non si ha un aspetto fisico valutato negativamente dagli altri. In particolare, le donne guadagnano meno degli uomini approssimativamente di un 46% (più precisamente, di un 37%: $e^{-.4617922} - 1 \cong 0.36985$*).*

b) Dieci anni in più di esperienza lavorativa si traducono in un bonus salariale di quale consistenza (sempre in termini percentuali)? La bellezza influenza il salario? Come?

Un anno di esperienza in più aumenta il salario dell'1.4% ($e^{.0137182} - 1 \cong 0.0138$*). Quindi, 10 anni di esperienza in più aumentano il salario del 14%. Il coefficiente di belavg è negativo e significativo, ma quello di abvavg non è significativamente diverso da zero. In altre parole: essere belli non aiuta, ma essere brutti penalizza (all'incirca del 15%, rispetto alla categoria di riferimento con una valutazione di bellezza pari a 3).*

c) Chiediamoci, adesso, se l'impatto di istruzione, esperienza e bellezza sul salario è significativamente diverso tra uomini e donne. Quale procedura di test potremmo usare? Sulla base dell'output di Stata riportato qui sotto (dove *fembel= female*belavg*, *femabv=female*abvavg*, *femeduc=female*educ*, *femexp=female*exper*), si calcoli il valore della statistica del test e si commenti il risultato ottenuto.

Se il modello iniziale era:

$$\log(wage_i) = \beta_1 + \beta_2 \exp er_i + \beta_3 educ_i + \beta_4 female_i + \beta_5 belavg_i + \beta_6 abvavg_i + u_i .$$

Adesso dobbiamo stimare anche il modello completo:

$$\log(wage_i) = \beta_1 + \beta_2 \exp er_i + \beta_3 educ_i + \beta_4 female_i + \beta_5 belavg_i + \beta_6 abvavg_i + \\ + female_i * (\beta_7 belavg_i + \beta_8 abvavg_i + \beta_9 \exp er_i + \beta_{10} educ_i) + \eta_i$$

L'ipotesi che l'impatto degli altri repressori non è diverso tra uomini e donne equivale a un test di Chow soltanto sulle "pendenze" della regressione. Quindi, nel modello completo, testiamo l'ipotesi nulla: $H_0 : \beta_7 = \beta_8 = \beta_9 = \beta_{10} = 0$*. La statistica F è data da:*

$$F = \frac{(SQR_r - SQR_c)/4}{SQR_c/(N-10)} = \frac{(294.1868 - 292.5230)/4}{292.5230/(1260-10)} \cong 1.77 \sim \mathrm{F}(4,1250)$$

I valori critici della distribuzione F con gradi di libertà 4 e sopra 1000 sono 2.37 (5%) e 3.32 (1%), quindi non possiamo rifiutare l'ipotesi nulla che l'impatto di esperienza, istruzione e bellezza sul salario sia uguale per gli uomini e per le donne. [NB: solitamente, i valori critici saranno riportati direttamente nell'esercizio.]

```
. reg  lwage exper educ female belavg abvavg fembel femabv femeduc femexp

      Source |       SS       df       MS              Number of obs =    1260
-------------+------------------------------           F(  9,  1250) =   72.39
       Model |  152.456961     9  16.9396624           Prob > F      =  0.0000
    Residual |  292.523011  1250  .234018409           R-squared     =  0.3426
-------------+------------------------------           Adj R-squared =  0.3379
       Total |  444.979972  1259  .353439215           Root MSE      =  .48375

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0146529   .0014324    10.23   0.000     .0118427     .017463
        educ |   .0661737   .0065478    10.11   0.000     .0533279    .0790195
      female |  -.6459314   .1590362    -4.06   0.000    -.9579386   -.3339241
      belavg |  -.1732432   .0541665    -3.20   0.001    -.2795106   -.0669759
      abvavg |  -.0375949   .0387855    -0.97   0.333    -.1136869    .0384971
      fembel |   .0649208   .0889604     0.73   0.466    -.1096073    .2394489
      femabv |   .0760927   .0650056     1.17   0.242    -.0514394    .2036249
     femeduc |   .0166178   .0113061     1.47   0.142    -.0055632    .0387988
    femexper |  -.0036383   .0026866    -1.35   0.176    -.0089091    .0016325
       _cons |   .7505399   .0949302     7.91   0.000     .5642997      .93678
------------------------------------------------------------------------------
```

d) Per capire se il modello stimato al punto c) soffre di un problema di eteroschedasticità, applichiamo il test di Breusch-Pagan. Specificate l'ipotesi nulla e quella alternativa, la statistica test e la sua distribuzione. Sulla base dell'output di Stata riportato qui sotto, che cosa concludiamo?

```
. estat hettest,iid rhs

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: exper educ female belavg abvavg fembel femabv femeduc femexper

         chi2(9)      =     3.43
         Prob > chi2  =   0.9446
```

Il test di Breusch-Pagan è implementato come segue. Si stima il modello

$$\log(wage_i) = \beta_1 + \beta_2 \exp er_i + \beta_3 educ_i + \beta_4 female_i + \beta_5 belavg_i + \beta_6 abvavg_i + \\ + female_i * (\beta_7 belavg_i + \beta_8 abvavg_i + \beta_9 \exp er_i + \beta_{10} educ_i) + \eta_i$$

e si ricavano i residui (e_i) come stime dei termini di errore (u_i). Dopodiché, si stima

$$e_i^2 = \beta_1 + \beta_2 \exp er_i + \beta_3 educ_i + \beta_4 female_i + \beta_5 belavg_i + \beta_6 abvavg_i + \\ + female_i * (\beta_7 belavg_i + \beta_8 abvavg_i + \beta_9 \exp er_i + \beta_{10} educ_i) + v_i$$

e si stima la significatività congiunta della regressione con un test F (ipotesi nulla di omoschedasticità, $H_0 : \beta_2 = ... = \beta_{10} = 0$*). Stata implementa una versione leggermente diversa del test (versione LM con distribuzione* χ^2*), ma l'interpretazione è identica. Dato un p-value di 0.9446, non c'è evidenza empirica sufficiente a rifiutare l'ipotesi nulla di omoschedasticità. (NB: la statistica test sarà distribuita come una* $F_{9,1250}$*.)*