

**ECONOMETRIA**  
**QUARTO PROBLEM SET**  
**SOLUZIONI**  
**Esercizio 1**

Considerate un modello dove la variabile da spiegare è di tipo binario:

$$y_i = \begin{cases} 1 & \text{se l'individuo è laureato} \\ 0 & \text{se l'individuo non è laureato} \end{cases}$$

Il modello lineare che spiega  $y_i$  in funzione, per esempio, del livello del reddito e di altre variabili esplicative, è:

$$y_i = x_i' \beta + \varepsilon_i$$

a) Considerando che  $E(\varepsilon_i|x_i) = 0$ , derivate  $E(y_i|x_i)$ .  
*In questo modello:*

$$E(y_i|x_i) = x_i' \beta$$

*La variabile  $y_i$  ha distribuzione di Bernoulli e quindi:*

$$\begin{aligned} E(y_i|x_i) &= 1 \Pr(y_i = 1|x_i) + 0 \Pr(y_i = 0|x_i) = \\ &= \Pr(y_i = 1|x_i) \end{aligned}$$

*Il modello che stiamo stimando spiega la probabilità che un evento si realizzi:*

$$y_i = \Pr(y_i = 1|x_i) + \varepsilon_i$$

*quindi la probabilità che l'evento si realizzi è una funzione lineare:*

$$\Pr(y_i = 1|x_i) = x_i' \beta$$

b)  $\varepsilon_i$  è distribuito normalmente?

*No, in questo caso  $\varepsilon_i$  può assumere solo due forme ed è distribuito secondo una Bernoulli:*

$$y_i = \begin{cases} 1 & \text{allora } 1 - x_i' \beta \\ 0 & \text{allora } -x_i' \beta \end{cases}$$

*Quindi  $\hat{\beta}_{OLS}$  non è più distribuito secondo una normale e dobbiamo considerare una distribuzione asintotica.*

c)  $\varepsilon_i$  è omoschedastico?

No,  $\varepsilon_i$  è eteroschedastico, infatti la varianza di  $\varepsilon_i$  è data da:

$$\text{var}(\varepsilon_i|x_i) = x_i'\beta(1 - x_i'\beta)$$

La varianza dell'errore dipende dalle  $x$ .

Di conseguenza,

$$V(\hat{\beta}_{OLS}) \neq \sigma^2 (X'X)^{-1}$$

Quindi i test sulla significatività dei coefficienti non sono attendibili.

Inoltre,  $\hat{\beta}_{OLS}$  non è BLUE. Si può risolvere il problema dell'eteroschedasticità usando i minimi quadrati ponderati.

d)  $x_i'\hat{\beta}_{OLS}$  assume valori solo compresi nell'intervallo  $[0, 1]$ ?

No, può assumere valori maggiori oppure minori di zero.

## Esercizio 2

Supponiamo di avere dati su un campione di famiglie e di essere interessati a determinare le variabili rilevanti nella scelta di acquisto di una barca per le vacanze estive.

a) Che tipo di modello teorico può essere considerato? Che tipo di modello stimabile può essere considerato?

In questo caso, l'acquisto di una barca dipende dalla disponibilità di reddito della famiglia. Infatti, possiamo pensare che esista una soglia di costo, per metri, della barca che fa sì che la famiglia scelga di acquistare la barca, invece magari di prenderla in affitto oppure orientarsi in altro modo per le vacanze estive. Nel modello stimato oppure detto modello con variabile dipendente limitata, la variabile dipendente osservata è di tipo qualitativo, binario (acquisto della barca oppure non acquisto della barca). Invece il modello teorico sottostante è definito modello latente.

b) Scrivere un modello teorico di tipo lineare.

$$y_i^* = x_i'\beta + \varepsilon_i^*$$

dove  $y_i^*$  è la variabile latente.

c) Scrivete il modello stimabile.

Consideriamo che

$$\begin{aligned} y_i &= 1 \text{ se } y_i^* > 0 \\ &= 0 \text{ se } y_i^* \leq 0 \end{aligned}$$

Possiamo scrivere il modello:

$$y_i = E(y_i|x_i) + \varepsilon_i$$

dove

$$\begin{aligned}
E(y_i|x_i) &= \Pr(y_i = 1|x_i) = \\
&= \Pr(y_i^* > 0|x_i) = \\
&= \Pr(x_i'\beta + \varepsilon_i^* > 0|x_i) = \\
&= \Pr(\varepsilon_i^* > -x_i'\beta|x_i) = \\
&= \Pr(\varepsilon_i^* < x_i'\beta|x_i) = \\
&= F(x_i'\beta)
\end{aligned}$$

Quindi il modello da stimare:

$$y_i = F(x_i'\beta) + \varepsilon_i$$

Il modello così ottenuto è non lineare ed assicura che la probabilità stimata sia compresa tra zero ed uno, poiché  $0 \leq F(x_i'\beta) \leq 1$  per definizione della funzione di ripartizione.

d) Come sono gli effetti marginali?

Gli effetti marginali del modello teorico o latente non sono costanti, ma variano al variare di  $x_i$ :

$$\beta_j = \frac{\partial E(y_i^*|x_i)}{\partial x_{ij}}$$

Invece gli effetti del modello stimato indicano la variazione della probabilità che un evento si realizzi conseguente ad una variazione unitaria della variabile  $j$ :

$$\begin{aligned}
ME_j &= \frac{\partial E(y_i|x_i)}{\partial x_{ij}} = \\
&= \frac{\partial F(x_i'\beta)}{\partial x_{ij}} = \\
&= f(x_i'\beta) \beta_j
\end{aligned}$$

dove  $f$  indica la funzione di densità di  $\varepsilon_i^*$ , questi effetti marginali non sono costanti, ma sono funzione delle variabili esplicative.

### Esercizio 3

Considerate un modello in cui stimiamo un modello che vuole spiegare le scelte lavorative di alcune donne intervistate, considerando l'età, lo stato civile, il livello di istruzione. La variabile dipendente è:

$$\begin{aligned}
y_i &= 1 \text{ se la donna lavora} \\
&= 0 \text{ se la donna non lavora}
\end{aligned}$$

Il modello stimato è:

$$y_i = \beta_0 + \beta_1 age_i + \beta_2 married_i + \beta_3 educ_i + \beta_4 children_i + \varepsilon_i$$

Ecco la tabella delle statistiche descrittive e l'output di un logit.

```
. sum work age education married
```

Variable	Obs	Mean	Std. Dev.	Min	Max
work	2000	.6715	.4697852	0	1
age	2000	36.208	8.28656	20	59
education	2000	13.084	3.045912	10	20
married	2000	.6705	.4701492	0	1

```
. logit work age married children education
```

```
Iteration 0: log likelihood = -1266.2225
Iteration 1: log likelihood = -1046.086
Iteration 2: log likelihood = -1028.5151
Iteration 3: log likelihood = -1027.9154
Iteration 4: log likelihood = -1027.9144
```

```
Logistic regression
```

Number of obs	=	2000
LR chi2(4)	=	476.62
Prob > chi2	=	0.0000
Pseudo R2	=	0.1882

```
Log likelihood = -1027.9144
```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
work					
age	.0579303	.007221	8.02	0.000	.0437774 .0720833
married	.7417775	.1264704	5.87	0.000	.4939001 .9896549
children	.7644882	.0515287	14.84	0.000	.6634938 .8654827
education	.0982513	.0186522	5.27	0.000	.0616936 .1348089
_cons	-4.159247	.3320397	-12.53	0.000	-4.810033 -3.508462

a) Commentate la significatività dei coefficienti?

*In questa regressione tutti i coefficienti sono significativi e per valutarla si usa il test t, ma lo stimatore MLE è asintoticamente normale e quindi usiamo la statistica della Normale.*

b) Quale test è utilizzato per la significatività della regressione?

*In questo tipo di modello viene costruita una statistica, likelihood ratio, in cui  $L_0$  è il valore della verosimiglianza associato a  $\max \ln L$  ottenuta dal modello ristretto, in cui tutti i parametri sono uguali a zero, tranne la costante e  $L_1$  è il valore della verosimiglianza associato a  $\max \ln L$  ottenuta dal modello completo.*

*La statistica test è:*

$$2 [\ln L_1 - \ln L_0] \stackrel{a}{\sim} \chi_{K-1}^2$$

dove  $K-1$  è il numero delle variabili esplicative del modello completo meno la costante, ovvero il numero di restrizioni.

In questo caso rifiutiamo l'ipotesi nulla, quindi la regressione è significativa.

c) Quale test è utilizzato per valutare la bontà del modello?

In questo caso non è possibile usare  $R^2$ , ma misure alternative basate su  $L_0$  e  $L_1$ . I due indicatori maggiormente utilizzati sono pseudo  $R^2$  e McFadden  $R^2$ .

$$PseudoR^2 = 1 - \frac{1}{1 + 2(\ln L_1 - \ln L_0)/N}$$

### Esercizio 4

Consideriamo lo stesso modello dell'esercizio precedente. Questa volta usiamo una stima probit. Ecco l'output della stima e degli effetti marginali.

```

probit work age married children education

Iteration 0:  log likelihood = -1266.2225
Iteration 1:  log likelihood = -1040.0608
Iteration 2:  log likelihood = -1027.2398
Iteration 3:  log likelihood = -1027.0616
Iteration 4:  log likelihood = -1027.0616

Probit regression                               Number of obs   =       2000
                                                LR chi2(4)      =       478.32
                                                Prob > chi2     =       0.0000
Log likelihood = -1027.0616                    Pseudo R2      =       0.1889

-----+-----
      work |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
      age |   .0347211   .0042293     8.21   0.000   .0264318   .0430105
    married |   .4308575   .074208    5.81   0.000   .2854125   .5763025
  children |   .4473249   .0287417   15.56   0.000   .3909922   .5036576
  education |   .0583645   .0109742    5.32   0.000   .0368555   .0798735
      _cons |  -2.467365   .1925635   -12.81   0.000   -2.844782  -2.089948
-----+-----

. mfx compute

Marginal effects after probit
      y = Pr(work) (predict)
      = .71835948

-----+-----
variable |      dy/dx   Std. Err.      z    P>|z|     [ 95% C.I.   ]     X
-----+-----
      age |   .011721    .00142     8.25   0.000   .008935   .014507    36.208
    married* |   .150478    .02641     5.70   0.000   .098716   .20224     .6705
  children |   .1510059    .00922    16.38   0.000   .132939   .169073    1.6445
  educat~n |   .0197024    .0037     5.32   0.000   .012442   .026963    13.084
-----+-----

(*) dy/dx is for discrete change of dummy variable from 0 to 1

```

a) Che differenza c'è tra una stima logit e probit?

*Una stima probit si basa sul presupposto che gli errori del modello teorico sono distribuiti secondo una normale, invece la stima logit si basa sul presupposto che gli errori del modello teorico si distribuiscono secondo una logistica standard.*

b) Qual è l'effetto marginale dell'istruzione sulla variabile dipendente?

*L'effetto marginale, in questo caso di probit, dell'istruzione sulla variabile dipendente è pari 0.0197.*

*L'effetto marginale nel caso di probit è pari a :*

$$ME_j = \frac{\partial E(y_i|x_i)}{\partial x_{ij}} = \frac{\partial \Phi(x'_i \beta)}{\partial x_{ij}} = \phi(x'_i \beta) \beta_j$$

dove

$$\begin{aligned}F(x'_i\beta) &= \Phi(x'_i\beta) \\ y_i &= \Phi(x'_i\beta) + \varepsilon_i\end{aligned}$$

dove  $\phi$  indica la funzione di densità della  $N(0, 1)$

c) Secondo voi, il modello è significativo?

*Osservando sia la significatività dei coefficienti e sia il LR test, possiamo definire i coefficienti significativi, invece il pseudo  $R^2$  non è molto alto, spieghiamo solo al 18%.*

### **Esercizio 5**

Ecco lo stesso modello ristimato con il logit.

```
. logit work age married children education
```

```
Iteration 0: log likelihood = -1266.2225
Iteration 1: log likelihood = -1046.086
Iteration 2: log likelihood = -1028.5151
Iteration 3: log likelihood = -1027.9154
Iteration 4: log likelihood = -1027.9144
```

```
Logistic regression                Number of obs =      2000
LR chi2(4)                        =      476.62
Prob > chi2                       =      0.0000
Pseudo R2                         =      0.1882

Log likelihood = -1027.9144
```

work	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
age	.0579303	.007221	8.02	0.000	.0437774	.0720833
married	.7417775	.1264704	5.87	0.000	.4939001	.9896549
children	.7644882	.0515287	14.84	0.000	.6634938	.8654827
education	.0982513	.0186522	5.27	0.000	.0616936	.1348089
_cons	-4.159247	.3320397	-12.53	0.000	-4.810033	-3.508462

```
. logit work age married children education, or
```

```
Iteration 0: log likelihood = -1266.2225
Iteration 1: log likelihood = -1046.086
Iteration 2: log likelihood = -1028.5151
Iteration 3: log likelihood = -1027.9154
Iteration 4: log likelihood = -1027.9144
```

```
Logistic regression                Number of obs =      2000
LR chi2(4)                        =      476.62
Prob > chi2                       =      0.0000
Pseudo R2                         =      0.1882

Log likelihood = -1027.9144
```

work	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
age	1.059641	.0076517	8.02	0.000	1.04475	1.074745
married	2.099664	.2655454	5.87	0.000	1.638695	2.690306
children	2.147895	.1106783	14.84	0.000	1.941564	2.376153
education	1.10324	.0205779	5.27	0.000	1.063636	1.144318

```
. mfx
```

```
Marginal effects after logit
y = Pr(work) (predict)
= .72678588
```

variable	dy/dx	Std. Err.	z	P> z	[ 95% C.I. ]		X
age	.0115031	.00142	8.08	0.000	.008713	.014293	36.208
married*	.1545671	.02703	5.72	0.000	.101592	.207542	.6705
children	.151803	.00938	16.19	0.000	.133425	.170181	1.6445
educat~n	.0195096	.0037	5.27	0.000	.01226	.02676	13.084

```
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

a) Qual è l'effetto marginale dell'istruzione sulla variabile dipendente? E' diverso rispetto al caso dei stima probit?

L'effetto marginale dell'istruzione sulla variabile dipendente è pari a 0.0195 e non è molto diverso rispetto a quello nel caso di stima probit.

In questo caso di logit, l'effetto marginale è uguale a :

$$ME_j = \frac{\partial E(y_i|x_i)}{\partial x_{ij}} = f(x'_i\beta)\beta_j = \frac{e^{x'_i\beta}}{(1 + e^{x'_i\beta})^2}\beta_j$$

b) Che cosa rappresentano i odds-ratio?

Il modello logit è disponibile anche in modo alternativo al fine di descrivere l'effetto delle variabili esplicative sulla probabilità:

$$\Omega(y_i = 1|x_i) = \frac{\Pr(y_i = 1|x_i)}{\Pr(y_i = 0|x_i)} = \frac{\Lambda(x'_i\hat{\beta})}{1 - \Lambda(x'_i\hat{\beta})} = \exp(x'_i\beta)$$

Se consideriamo variazioni unitarie  $\Delta x_{ij} = 1$ , l'effetto sulla probabilità relativa è pari a  $e^{\beta_j}$

c) Che cosa rappresenta  $\beta_j$ ? Qual è il valore di  $\beta_j$  nel caso di variazioni dell'istruzione sulla variabile dipendente?

$\beta_j$  rappresenta l'effetto marginale di  $x_{ij}$  sui log odds ratio, infatti  $\beta_j$  per l'istruzione è pari a 0.098 e quindi ciò indica che la probabilità che l'individuo lavori aumenta con l'aumentare dell'istruzione. Inoltre questo valore è esattamente il logaritmo naturale della stima del coefficiente degli odds-ratio.