Esame Econometria I – 01-06-2021

Nome e cognome:

Domanda 1 (punti 9)

Si assuma di stimare il seguente modello lineare con il metodo OLS:

$$\log(wage_i) = \beta_0 + \beta_1 educ_i + \beta_2 experience_i + u_i \quad (1),$$

ottenendo i seguenti risultati:

Source	SS	df	MS	Number of obs	=	15,481
				F(2, 15478)	=	1396.90
Model	512.579786	2	256.289893	Prob > F	=	0.0000
Residual	2839.75133	15,478	.183470173	R-squared	=	0.1529
				Adj R-squared	=	0.1528
Total	3352.33112	15,480	.216558858	Root MSE	=	.42833

log_m_wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
years_educ	.0508986	.0010531	.	.	.	.
exp	.0130541	.0003286	39.73	0.000	.0124099	.0136982
_cons	6.402625	.0182681	350.48	0.000	6.366817	6.438432

dove log_m_wages è il logaritmo del salario mensile, years_educ sono gli anni di studio e exp cattura gli anni di esperienza lavorativa.

a) Si intepreti il coefficiente stimato di years_educ

b) Si calcoli, per il medesimo regressore (years_educ) la statistica t e l'intervallo di confidenza al 95%

Si supponga adesso che gli individui differiscano in termini di abilità e che la corretta specificazione del modello includa anche l'abilità individuale, tale che:

$$u_i = \beta_3 ability_i + \varepsilon_i \quad (2).$$

Assumendo che l'abilità sia osservabile e inclusa come ulteriore controllo si ha:

$$\log(wage_i) = \beta_0 + \beta_1 educ_i + \beta_2 experience_i + \beta_3 ability_i + \varepsilon_i \quad (3).$$

c) Il coefficiente β_1 stimato dall'equazione (3), cioè controllando per abilità, sarà maggiore o minore rispetto al coefficiente β_1 stimato dall'equazione (1), cioè senza controllare per abilità? Si motivi la risposta.

Domanda 2 -Punti (12)

Si consideri il seguente modello di popolazione:

$$y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 z_i + \varepsilon_i \quad (5),$$

con:

- $cov(x_i, \varepsilon_i) = 0$ per assunzione.
- $cov(x_i, x_i^2) > 0$ per costruzione.
- $cov(x_i, z_i) > 0$ per assunzione.
- $\beta_1 > 0$
- $\beta_2 > 0$

- $\beta_3 < 0$

Sfortunatamente, assumendo che il vero processo generatore dei dati descritto nell' equazione (5) è sconosciuto, z_i è non osservabile, e x_i è misurata con errore, è possibile stimare solo il seguente modello:

$$y_i = \alpha + \beta_1 \tilde{x}_i + u_i \qquad (6),$$

dove:

$$\tilde{x}_i = x_i + w_i \qquad (7),$$

$$u_i = \beta_2 x_i^2 + \beta_3 z_i + \varepsilon_i \qquad (8),$$

e $w_i \sim (0, \sigma_w^2)$ è una componente di errore di misurazione classico.

In questo contesto empirico:

a. *Si derivi il limite in probabilità di $\hat{\beta}_1$ ottenuto dall'equazione (6).*
b. *Senza conoscere I valori esatti di β_1, β_2, β_3 e di tutte le covarianze, è possibile dire se $\hat{\beta}_1$ è asintoticamente distorto verso il basso o verso l'alto?*
c. *Si assuma adesso che β_2=0 e $\sigma_w^2 = 0$. Qual è la nuova formulazione del limite in probabilità di $\hat{\beta}_1$.*
d. *In quest'ultimo caso, è possibile dire se $\hat{\beta}_1$ è asintoticamente distorto verso l'alto o verso il basso?*

Domanda 3 (punti 10.5)

In un lavoro di ricerca gli autori utilizzano informazioni contenute in un'indagine campionaria su 1000 imprese per predire il successo imprenditoriale sfruttando un vettore contenete 100 caratteristiche delle imprese.

Per valutare la performance predittiva di diversi stimatori, gli autori utilizzando un training sample contenente l'80% delle osservazioni e un test sample con il 20% delle osservazioni rimanenti. Successivamente, per calibrare i metodi di regolarizzazione e scegliere il parametro λ, gli autori utlizzano il metodo 5-fold cross-validation (cross-validazione incrociata a 5 fold). I risultati di questo esercizio empirico di statistical learning sono presentati nella Tabella 1.

Tabella 1: performance predittiva di diversi stimatori

Stimatore	Ordine	Numero di regressori con coefficiente diverso da zero	MSE in-sample	MSE-out-of-sample	Parametro di penalizzazione: λ
		100	0.152	6.321	0
		18	1.211	0.344	0.156
		100	0.877	0.999	1.434

Secondo le informazioni presentate nella Tabella 1:

a. Si compili la prima colonna della tabella 1 con il nome dello stimatore utilizzato.
b. Si compili la seconda colonna ordinando gli stimatori in base alla loro performance predittiva (il migliore avrà il valore 1)
c. Qual è lo stimatore con la peggiore performance in termini di overfitting (sovradattamento)? Si spieghi la risposta
d. Come si calcolano i MSE in-sample and out-of-sample?
e. Si descriva il processo "5-fold cross-validation" adottato per scegliere il parametro λ.
f. Cosa si intende per "shrinking" o penalizzazione di un coefficiente quando si usa un metodo di regolarizzazione?
g. Quale degli stimatori presentati nella Tabella 1 può essere utilizzato per selezionare i regressori (variable selection)?