

Egzamin z ekonometrii IiE17.06.2014

Pytania teoretyczne

1. Opisz dwustopniową procedurę znajdowania estymatora UMM z optymalną macierzą wag.
2. Jakie są podobieństwa i różnice między ocenianą próbą a przypadkiem, kiedy próba zawiera obserwacje będące wynikiem rozwiązań brzegowych? Jaki model powinno się stosować dla takich prób?
3. Co to są ilorazy szans i dlaczego w kontekście modelu logitowego lepiej jest używać ilorazów szans niż efektów krańcowych?
4. Czym różni się panel od próby przekrojowo-czasowej?

ZADANIE 1 W modelu

$$y_i = \beta_0 + \beta_0\beta_1x_{1i} + \beta_1x_{2i} + \varepsilon_i \quad (*)$$
$$\varepsilon \sim N(0, \sigma^2 \mathbf{I})$$

1. Jaką postać ma logarytm funkcji wiarygodności dla tego modelu i jakie są warunki pierwszego rzędu na maksymalizację tej funkcji?
2. Czy dla estymatorów MNW parametrów β_0, β_1
 - (a) suma reszt $e_i = y_i - \tilde{\beta}_0 - \tilde{\beta}_0\tilde{\beta}_1x_{1i} - \tilde{\beta}_1x_{2i}$ w modelu (*) jest równa zero?
 - (b) wektory obserwacji x_1 i x_2 są ortogonalne do wektora reszt?
3. Jaka jest postać statystyki LM dla hipotezy $\beta_1 = 0$ w modelu (*)?
4. Wyjaśnij w jaki sposób parametry modelu (*) można oszacować za pomocą MNK. Wyjaśnij dlaczego estymator ten jest w tym przypadku zgodny. Czy taki estymator będzie także estymatorem efektywnym?

Rozwiązanie:

1.

$$y_i \sim N(\beta_0 + \beta_0\beta_1x_{1i} + \beta_1x_{2i}, \sigma^2 I)$$

Więc funkcja gęstości jest dana przez

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{[y_i - \beta_0 - \beta_0\beta_1x_{1i} - \beta_1x_{2i}]^2}{2\sigma^2} \right\}$$

W tym przypadku MNW sprowadza się do nieliniowej metody najmniejszych kwadratów, w której minimalizujemy $S(b) = \frac{1}{2} \sum [y_i - h(x_i, \beta)]^2$. Warunki pierwszego rzędu

$$\ln f(y_1 \dots y_n) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \sum \frac{[y_i - \beta_0 - \beta_0\beta_1x_{1i} - \beta_1x_{2i}]^2}{2\sigma^2}$$
$$\frac{\partial \ln f}{\partial \beta_0} = \sum \frac{(y_i - \beta_0 - \beta_0\beta_1x_{1i} - \beta_1x_{2i})(1 + \beta_1x_{1i})}{\sigma^2}$$

w maksimum

$$\sum (y_i - \tilde{\beta}_0 - \tilde{\beta}_0\tilde{\beta}_1x_{1i} - \tilde{\beta}_1x_{2i})(1 + \tilde{\beta}_1x_{1i}) = \sum e_i(1 + \tilde{\beta}_1x_{1i}) = 0$$

a więc w maksimum

$$\sum e_i(1 + \tilde{\beta}_1x_{1i}) = 0 \Leftrightarrow \sum e_i = -\tilde{\beta}_1 \sum e_ix_{1i} \neq 0$$

Co więcej

$$\frac{\partial \ln f}{\partial \beta_1} = \sum \frac{(y_i - \tilde{\beta}_0 - \tilde{\beta}_0\tilde{\beta}_1x_{1i} - \tilde{\beta}_1x_{2i})(\beta_0x_{1i} + x_{2i})}{\sigma^2} = 0$$

w maksimum

$$\sum e_i(\tilde{\beta}_0x_{1i} + x_{2i}) = 0$$

IMIĘ NAZWISKO.....

$$\sum e_i x_{2i} = -\tilde{\beta}_0 \sum e_i x_{1i} \neq 0$$

$$\frac{\partial \ln f}{\partial \sigma^2} = -\frac{n}{2} \frac{1}{\sigma^2} + \sum_{i=1}^n \frac{[y_i - \tilde{\beta}_0 - \tilde{\beta}_0 \tilde{\beta}_1 x_{1i} - \tilde{\beta}_1 x_{2i}]^2}{2\sigma^4}$$

w maksimum

$$\widehat{\sigma^2} = \frac{1}{n} \sum_{i=1}^n [y_i - \tilde{\beta}_0 - \tilde{\beta}_0 \tilde{\beta}_1 x_{1i} - \tilde{\beta}_1 x_{2i}]^2 = \frac{1}{n} \sum_{i=1}^n e_i^2$$

- (a) Z pierwszego warunku pierwszego rzędu wynika, że suma reszt nie jest równa zero.
- (b) Z pierwszego i drugiego warunku wynika, że wektory obserwacji dla x_{1i} i x_{2i} nie są ortogonalne do reszt.

2. Dla $H_0 : \beta_2 = 0$ otrzymujemy model liniowy, który możemy oszacować MNK. Statystykę LM można policzyć w następujący sposób:

- (a) oszacować model, w którym zakładamy, że $H_0 : \beta_1 = 0$. Taki model ma postać $y_i = \beta_0 + \varepsilon_i$ jest więc zwykłym model liniowym. Znajdujemy oszacowania MNK parametru β_0 . Oznaczmy reszty z MNK jako $e_i = y_i - \beta_0$, użykujemy $\tilde{\beta}_0 = \bar{y}$.
- (b) Liczymy gradienty funkcji wiarygodności dla modelu bez ograniczeń dla wartości oszacowanych w modelu z ograniczeniami (wektory score):

$$\left. \frac{\partial \ln f_i}{\partial \beta_0} \right|_{\theta = \tilde{\theta}_R} = \frac{y_i - \tilde{\beta}_0}{s^2} = \frac{e_i}{s^2}$$

$$\left. \frac{\partial \ln f_i}{\partial \beta_1} \right|_{\theta = \tilde{\theta}_R} = \frac{(y_i - \tilde{\beta}_0)(\tilde{\beta}_0 x_{1i} + x_{2i})}{s^2} = \frac{e_i(\tilde{\beta}_0 x_{1i} + x_{2i})}{s^2}$$

$$\left. \frac{\partial \ln f_i}{\partial \sigma^2} \right|_{\theta = \tilde{\theta}_R} = -\frac{1}{2} \frac{1}{s^2} + \frac{[y_i - \tilde{\beta}_0]^2}{2s^4} = -\frac{1}{2s^2} \left[1 - \left(\frac{\hat{e}_i}{s} \right)^2 \right]$$

- (c) Znajdujemy statystykę LM jako sumę wartości dopasowanych w regresji 1 na scorach. Statystykę tą można znaleźć przy zastosowaniu MNK. Statystyki LR i W wymagałyby oszacowania modelu nieliniowego.

3. Estymator MNK, przy podanych założeniach, jest estymatorem zgodnym w modelu:

$$y_i = \beta_0 + \gamma x_{1i} + \beta_1 x_{2i} + \varepsilon_i$$

bez względu na to, czy prawdą jest, że $\beta_0 \beta_1 = \gamma$. Oczywiście jeśli model (*) jest prawdziwy, to $\gamma \xrightarrow{p} \beta_0 \beta_1$. Estymator MNK nie będzie jednak estymatorem efektywnym, ponieważ zgodnie z tw. Rao-Cramera asymptotycznie efektywnym estymatorem parametrów jest estymator MNW.

ZADANIE 2 Na podstawie wyników badań Polskiego Generalnego Sondażu Społecznego z lat 1999, 2002, 2005, 2008, 2010 starano się zidentyfikować zmienne, które wpływają na poziom wykształcenia ankietowanych. Zmienną zależną w modelu jest zmienna wyksz, która przyjmuje wartość 1 dla osób, które mają wykształcenie podstawowe, 2 dla osób z wykształceniem średnim i 3 dla osób z wykształceniem wyższym. Jako zmienne objaśniające znalazły się następujące zmienne: płeć (1 mężczyzna, 2 kobieta), wiek, wiek2, pgssyear (rok badania). Poniżej znajdują się oszacowania wielkości parametrów i oszacowania efektów cząstkowych dla alternatywy wykształcenie średnie dla zmiennej wyksz modelowanej za pomocą uporządkowanego probita. Testy przeprowadzamy na poziomie istotności $\alpha = 0.05$.

REGRESJA

Ordered probit regression

Number of obs = 8588

IMIĘ NAZWISKO.....

Log likelihood = -7877.3723

LR chi2(4) = 547.03
 Prob > chi2 = 0.0000
 Pseudo R2 = 0.0336

wyksz	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
plec						
kobieta	.2315973	.0256208	9.04	0.000	.1813814	.2818131
pgssyear	.0326579	.0031964	10.22	0.000	.026393	.0389228
wiek	.0273503	.0040806	6.70	0.000	.0193524	.0353481
wiek2	-.0004147	.0000419	-9.90	0.000	-.0004968	-.0003326
/cut1	65.88426	6.408716			53.32341	78.44511
/cut2	67.08136	6.409737			54.51851	79.64421

PSEUDO-R2

McKelvey and Zavoina's R2: 0.092
 Count R2: 0.577
 Adj Count R2: 0.024

EFEKTY CZĄSTKOWE

	dy/dx	Std. Err.	z	P> z	[95% Conf. Interval]	
plec						
kobieta	.0465577	.0052584	8.85	0.000	.0362515	.0568639
pgssyear	.006468	.0006307	10.26	0.000	.0052319	.0077041
wiek	.0054168	.0008037	6.74	0.000	.0038417	.0069919
wiek2	-.0000821	8.21e-06	-10.01	0.000	-.0000982	-.0000661

Note: dy/dx for factor levels is the discrete change from the base level

TEST TYPU ZWIĄZKU

wyksz	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
_hat	-79.61997	11.98614	-6.64	0.000	-103.1124	-56.12756
_hatsq	.6137051	.0912528	6.73	0.000	.4348529	.7925574
/cut1	-2581.731	393.5928			-3353.158	-1810.303
/cut2	-2580.528	393.5919			-3351.953	-1809.102

1. Wypisz założenia modelu uporządkowanego probita.
2. Wyjaśnij dlaczego w przypadku takiego badania stosuje się model uporządkowanego probita a nie zwykłą regresję liniową lub zwykły model probitowy.
3. Zinterpretuj wielkości pseudo- R^2 McKalveya i Zavoiny oraz pseudo- R^2 liczebnościowego i skorygowanego

IMIĘ NAZWISKO.....

ZADANIE 3 Przy użyciu danych dotyczących 133 krajów świata z lat 1970-2009 badacz zbudował panel niezbilansowany. Przy użyciu tego panelu badacz oszacował model, którego celem jest wyjaśnienie zróżnicowania średniej długości trwania życia w poszczególnych krajach. Zmienną zależną jest zmienna *life* oznaczająca oczekiwaną długość życia w latach a zmiennymi objaśniającymi poziom alfabetyzacji (*lit* - wyrażony w procentach udział ludności umiejącej czytać), PKB per capita (*GDP_pc* wyrażone w dolarach), trend czasowy (*year* - rok z którego pochodzą dane) umieszczony, by uwzględnić postęp medycyny. Dodatkowo do regresji włączono zmienną zerojedynkową *post_comm* oznaczającą, że dany kraj był w pewnym okresie lub dalej jest krajem komunistycznym po to, by zbadać, czy rzeczywiście w krajach komunistycznych opieka medyczna była na względnie dobrym poziomie. Model został przez badacza oszacowany za pomocą MNK, estymatora efektów losowych i estymatora efektów stałych. Wyniki estymacji znajdują się na następnej stronie.

Założony poziom istotności przy testowaniu hipotez statystycznych $\alpha = 0.05$. Uzyskane wyniki testów należy uzasadnić wielkościami odpowiednich statystyk bądź wartościami p.

1. Wyjaśnij, dlaczego badacz użył w regresji MNK odpornego warstwowego estymatora macierzy wariancji i kowariancji.
2. Weźmy pod uwagę jedynie wyniki dla MNK i estymatora efektów losowych. Który z nich jest estymatorem efektywnym w przypadku rozpatrywanego problemu? Odpowiedź uzasadnij odpowiednią statystyką testową.
3. Na podstawie znajdujących się na wydruku statystyk testowych wybierz spośród estymatorów POLS, RE i FE estymator, który powinno się użyć w kontekście analizowanego problemu. Wyjaśnij jakie hipotezy zerowe testujemy za pomocą użytych testów.
4. Dlaczego w przypadku estymatora efektów stałych nie udało się oszacować współczynnika dla zmiennej *post_comm*?
5. Zinterpretuj wielkości wszystkich trzech statystyk R^2 uzyskanych dla estymatora efektów stałych.
6. Zinterpretuj wielkość istotnych współczynników w modelu efektów stałych.
7. Czy w modelu efektów stałych poprawne byłoby pominięcie efektów indywidualnych dla krajów? Odpowiedź uzasadnij wielkością odpowiedniej statystyki.
8. Dlaczego zarówno w modelu efektów stałych jak i w modelu efektów losowych uzyskujemy dwa oszacowania błędów standardowych czynników losowych (*sigma_u*, *sigma_e*)?

Linear regression

	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
life						
lit	.3069304	.0160516	19.12	0.000	.2751788	.338682
GDP_pc	.0002563	.0000399	6.43	0.000	.0001775	.0003352
post_comm	-1.302394	.8825781	-1.48	0.142	-3.048221	.4434327
year	-.0098472	.0260935	-0.38	0.706	-.0614626	.0417683
_cons	60.10586	51.21889	1.17	0.243	-41.21018	161.4219

Random-effects GLS regression

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
life						
lit	.1054896	.007226	14.60	0.000	.0913269	.1196522
GDP_pc	.0000376	7.16e-06	5.24	0.000	.0000235	.0000516
post_comm	1.562903	1.079774	1.45	0.148	-.5534152	3.679221
year	.1900848	.0068268	27.84	0.000	.1767046	.2034649
_cons	-321.9032	13.15153	-24.48	0.000	-347.6797	-296.1266
sigma_u	4.4966449					
sigma_e	2.5231168					
rho	.76054561	(fraction of variance due to u_i)				

Breusch and Pagan Lagrangian multiplier test for random effects

chi2(1) = 43672.49
 Prob > chi2 = 0.0000

Fixed-effects (within) regression

R-sq: within = 0.5793
 between = 0.4524
 overall = 0.2970

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
life						
lit	.0468977	.0078084	6.01	0.000	.0315894	.062206
GDP_pc	-2.12e-07	7.28e-06	-0.03	0.977	-.0000145	.0000141
post_comm	(omitted)					
year	.2384537	.0071899	33.17	0.000	.2243581	.2525493
_cons	-413.1269	13.79678	-29.94	0.000	-440.1753	-386.0785
sigma_u	8.8759641					
sigma_e	2.5231168					
rho	.92523542	(fraction of variance due to u_i)				

F test that all u_i=0: F(132, 4533) = 127.85 Prob > F = 0.0000

Hausman test

chi2(2) = (b-B)' [(V_b-V_B)^(-1)] (b-B) = 301.07
 Prob>chi2 = 0.0000

Rozwiązanie:

1. Badacz użył w regresji MNK warstwowego estymatora odpornego, ponieważ w przypadku regresji na panelu można spodziewać się niediagonalnej macierzy wariancji i kowariancji ze względu na występowanie efektów indywidualnych

IMIĘ NAZWISKO.....

2. Estymatorem efektywnym jest w tym przypadku estymator efektów losowych, ponieważ z wielkości statystyki Breuscha-Pagana 43672.49 [0.0000] wynika, że efekty losowe są istotne w modelu, macierz wariancji kowariancji łącznego błędu losowego jest niediagonalna, a w takim przypadku estymator SUMNK, którego szczególnym przypadkiem jest estymator efektów losowych, jest efektywniejszy od estymatora MNK.
3. Podstawą wyboru odpowiedniego estymatora w rozpatrywanym przypadku powinien być wynik testu Hausmana. Hipotezą zerową w tym teście jest warunek konieczny dla zgodności estymatora efektów losowych to jest brak korelacji między efektem indywidualnym a zmiennymi objaśniającymi $Cov(u_i, \mathbf{X}_i) = 0$. W naszym przypadku hipoteza ta jest odrzucana 301.07 [0.0000] a tym samym jedynym zgodnym estymatorem jest estymator efektów stałych.
4. Za pomocą estymatora efektów stałych nie jest możliwe oszacowanie wpływu zmiennych, które nie zmieniają się w czasie. Zmienna zerojedynkowa oznaczająca, że dany kraj był bądź jest krajem komunistycznym nie zmienia się w czasie.
5. Wielkość R^2_{within} oznacza, że 58% zróżnicowania wewnątrz obiektowego (to jest zmian długości oczekiwanego życia dla danego kraju) udało się wyjaśnić za pomocą zróżnicowania zmiennych objaśniających dla tego kraju. Wielkość $R^2_{between}$ oznacza, że 45% zróżnicowania długości trwania życia między krajami udało się wyjaśnić za pomocą różnic w wielkościach zmiennych objaśniających pomiędzy krajami, $R^2_{overall}$ oznacza, że 30% całkowitej zmienności zmiennej zależnej udało się wyjaśnić zmiennością zmiennych niezależnych.
6. Wzrost poziomu alfabetyzacji o 1 punkt procentowy wydłuża średnią długość życia o 0.04 roku, średnia długość życia wydłuża się o 0.23 roku w każdym kolejnym okresie badanym.
7. W modelu efektów stałych nie można pominąć charakterystyk indywidualnych krajów ponieważ są one istotne co wnioskujemy z wyniku testu, że wszystkie $u_i = 0$ (statystyka 127.85 [0.0000]) a zarazem wiemy z wyniku testu Hausmana, że efekty indywidualne są skorelowane ze zmiennymi objaśniającymi. Pominięcie efektów indywidualnych wywołałoby w tym przypadku pojawienie się problemu endogeniczności.
8. Ponieważ w przypadku liniowego modelu efektów nieobserwowalnych mamy dwa składniki losowe: efekt indywidualny u_i oraz błąd czystolosowy ε_i . Oszacowane odchylenia standardowe odpowiadają odchyleniom standardowym u_i oraz ε_i .