Egzamin z ekonometrii 17.06.2014

Pytania teoretyczne

1. Opisz procedurę testowania stacjonarności za pomocą rozszerzonego testu Dickey-Fullera (ADF).
2. Jakie są podobieństwa i różnice między ocenzurowaną próbą a przypadkiem, kiedy próba zawierają 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 Mamy następujący model:

$$x_t = \alpha_1 + x_{t-1} + \varepsilon_{1t} \tag{1}$$
$$y_t = \beta_1 + \beta_2 y_{t-1} + \beta_3 x_{t-1} + \varepsilon_{2t} \tag{2}$$

gdzie $\boldsymbol{\varepsilon}_t = [\varepsilon_{1t}, \varepsilon_{2t}] \sim IID(\mathbf{0}, \boldsymbol{\Sigma})$. ($IID$ jest skrótem od **I**ndependent and **I**dentically **D**istributed)

1. Wykaż, że zmienna x_t jest $I(1)$
2. Wykaż, jeśli y_t i x_t są skointegrowane to w mechanizmie korekty błędem:
   (a) relacja kointegrująca ma postać $y_{t-1} - \frac{\beta_3}{1-\beta_2} x_{t-1} \sim I(0)$
   (b) współczynnik szybkości dostosowań jest równy $\beta_2 - 1$
3. Badacz oszacował najpierw równanie (2) za pomocą MNK i otrzymał oszacowania $\widehat{\beta}_1$, $\widehat{\beta}_2$, $\widehat{\beta}_3$.
   (a) Czy stosując standardowe statystyki t badacz może poprawnie zweryfikować istotność oszacowanych parametrów β_1, β_2, β_3? Odpowiedź uzasadnij.
   (b) Badacz sformułował oszacowania współczynnika przy x_{t-1} w zależności kointegrującej jako $\frac{\widehat{\beta}_3}{1-\widehat{\beta}_2}$ a współczynnika dostosowań jako $\widehat{\beta}_2 - 1$. Czy za pomocą takiej procedury otrzyma zgodne estymatory parametrów mechnizmu korekty błędem?

Rozwiązanie:

1. Zmienna x_t jest błądzeniem przypadkowym z dryfem:

$$\begin{aligned} x_t &= \alpha_1 + x_{t-1} + \varepsilon_{1t} = 2\alpha_1 + x_{t-2} + \varepsilon_{1t} + \varepsilon_{1t-1} \\ &= t\alpha_1 + \sum_{s=1}^{t} \varepsilon_s \end{aligned}$$

   i

$$\begin{aligned} \mathrm{E}(x_t) &= t\alpha_1 \\ \mathrm{Var}(x_t) &= t\sigma_1^2 \end{aligned}$$

   co implikuje, że zmienna ta nie jest ani stacjonarna, bo nie ma stałej w czasie wartości oczekiwanej ani trendostacjonarna, ponieważ zmienia się w czasie jej wariancja. Jednak pierwsze różnice zmiennej x_t są równe

$$\Delta x_t = \alpha_1 + \varepsilon_{1t}$$

   i są stacjonarne, ponieważ ε_{1t} ma z założenia stałą wartość oczekiwaną i wariancję.

IMIĘ NAZWISKO.....

(a) Zauważmy, że odejmując stronami y_{t-1} od równania (2) otrzymujemy:

$$\Delta y_t = \beta_1 + (\beta_2 - 1)\, y_{t-1} + \beta_3 x_{t-1} + \varepsilon_{2t}$$

i przekształcając

$$\Delta y_t = \beta_1 + (\beta_2 - 1)\left(y_{t-1} - \frac{\beta_3}{1-\beta_2} x_{t-1}\right) + \varepsilon_{2t}$$

a więc model w postaci mechanizmu korekty błędem. Relacja kointegrująca ma rzeczywiście postać $y_{t-1} - \frac{\beta_3}{1-\beta_2} x_{t-1} \sim I(0)$

(b) z poprzednio wyprowadzonego równania widać że współczynnik szybkości dostosowań jest równy $\beta_2 - 1$

(a) Taka procedura testowania jest nieprawidłowa, ponieważ w regresji MNK dla modelu (2) pojawi się problem regresji pozornej związany z niestacjonarnością zmiennych użytych w tej regresji.

(b) Uzyskany w ten sposób estymator będzie statystycznie poprawny w tym sensie, że będzie zgodny. Wynika to z tego, że estymator MNK w modelu z niestacjonarnymi zmiennymi jest zgodny (a nawet superzgodny).

ZADANIE 2 Na podstawie wyników badań Polskiego Generalnego Sondażu Społecznego z lat 1999, 2002, 2005, 2008, 20010 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 śrdnim 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
                                                  LR chi2(4)      =     547.03
                                                  Prob > chi2     =     0.0000
Log likelihood = -7877.3723                       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
```

IMIĘ NAZWISKO.....

EFEKTY CZĄSTKOWE

	Delta-method				
	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 pseudo-R^2 liczebnościowego. Sprawdź, czy wszystkie zmienne w modelu są łącznie istotne.
4. Zbadaj poprawność formy funkcyjnej modelu.
5. Zinterpretuj znak parametru oszacowanego dla zmiennej pgssyear.
6. Zintepretuj uzyskaną wielkość efektu cząstkowego dla zmiennej płeć.
7. Badany model powtórnie oszacowano po dodaniu zmiennych zero jedynkowych związanych z województwem, w którym mieszka respondent. Uzyskana wielkość logarytmu funkcji wiarygodności wyniosła -7822.3119. Zweryfikuj hipotezę łączną, że poziom wykształcenia nie ma związku województwem zamieszkania - wskaż wartość krytyczną, którą należy w tym przypadku użyć.

   Podpowiedź: $\chi^2_{0.95}(13) = 22.36$, $\chi^2_{0.95}(14) = 23.68$, $\chi^2_{0.95}(15) = 24.00$

Rozwiązanie:

1. Zmienna ukryta

$$y_i^* = \boldsymbol{x}_i\boldsymbol{\beta} + \varepsilon_i$$
$$\varepsilon_i \sim N(0,1)$$

   i poszczególne obserwacje są niezależne. Obserwujemy y, który powstaje w sposób następujący:

$$\begin{array}{lll} y_i = 0 & \text{jeśli} & y_1^* \leq \alpha_1 \\ y_i = 1 & \text{jeśli} & \alpha_1 < y_i^* \leq \alpha_2 \\ \vdots & & \vdots \\ y_i = J & \text{jeśli} & y_i^* > \alpha_J \end{array}$$

IMIĘ NAZWISKO.....

gdzie $\alpha_1 < \alpha_2 < \ldots < \alpha_J$ są nieznane.

2. Zmienna zależna jest zmienną dyskretną o dobrze zdefiniowanym porządku i znanej liczbie możliwych alternatyw (3 alternatywy). W tym przypadku nie możemy stosować modelu probitowego (więcej niż dwie alternatywy) nie powinniśmy też stosować regresji liniowej, ponieważ chcemy wyjaśnić prawdopodobieństwo alternatyw (wartości dopasowane z regresji liniowej będą trudne do zintepretowania, mogą być np. ujemne)

3. Wartość statystyki pseudo-R^2 McKalveya i Zavoiny oznacza, że model wyjaśnia 9.2% odrzucamy zmienności zmiennej ukrytej. Wartość pseudo-R^2 liczebnościowego oznacza, że 57.7% przypadków model daje prawidłowe przewidywania poziomu wykształcenia, ale z wielkości skorygowanego pseudo-R^2 liczebnościowego wynika, że jedynie w 2.4% te trafne prognozy można powiązać ze zmianami wielkości zmiennych objaśniających. Na postawie statystyki LR odrzucamu hipotezę o tym, że wszystkie zmienne w modelu są niestotne $[547.03, 0.000 < 0.05]$.

4. Wynik testu typu związku sugeruje, że forma funkcyjna jest niepoprawna - wielkość statystyki z dla _hatsq 6.73 $[0.000 < 0.5]$

5. W modelu istotne okazały się zmienne: kobieta $[9.04, 0.000 < 0.05]$, pgssyear $[10.22, 0.000 < 0.05]$, wiek $[6.70, 0.000 < 0.05]$, wiek2 $[-9.90, 0.000 < 0.05]$.

6. Dodatni znak przy zmiennej pgssyer oznacza, że dla kolejnych lat badania rosło (ceteris paribus) prawdopodobieństwo że respondent ma wyższe wykształcenie i maleje prawdopodobieństwo, że ma tylko wykształcenie podstawowe.

7. Oszacowana wielkość efektu cząstkowego oznacza, że kobiety mają, (ceteris paribus) 4.6 pp wyższe p-stwo posiadania średniego wykształcenia niż mężczyźni.

8. Statystyka testu LR ma postać: .

$$LR = 2\left(-7822.3119 + 7877.3723\right) = 110.12 > \chi^2_{0.95}(15) = 24.00$$

Testujemy zerowość 15 współczynników (jeden poziom zmiennej dyskretnej został usunięty jako bazowy) a więc właściwą wartością krytyczną jest $\chi^2_{0.95}(15)$. Odrzucamy hipotezę o tym, że łącznie wszystkie współczynniki są nieistotne. Województwo zamieszkania istotnie wpływa na wykształcenie respondentów.

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 uwględnić postęp medycyny. Dodatkowo do regresji włączono zmienną zerojedynkową $post\_comm$ oznaczająca, że danych 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 $post\_comm$?

IMIĘ NAZWISKO.....

5. Zintepretuj wielkości wszystkich trzech statystyk R^2 uzyskanych dla estymatora efektów stałych.
6. Zintepretuj 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 ($sigma\_u, sigme\_e$)?

```
Linear regression
------------------------------------------------------------------------------
             |               Robust
        life |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         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
------------------------------------------------------------------------------
        life |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         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
------------------------------------------------------------------------------
        life |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         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
```

IMIĘ NAZWISKO.....

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

Rozwiązanie:

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

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 $\mathrm{Cov}\left(u_i, \boldsymbol{X}_i\right) = 0$. 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.