Mikroekonometria 4

Mikołaj Czajkowski
Wiktor Budziński

Endogeniczność – regresja liniowa

- W regresji liniowej estymujemy następujące równanie:

$$Y_i = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i$$

- KMRL zakłada, że wszystkie zmienne objaśniające są egzogeniczne tj. $E(\varepsilon|\mathbf{X}) = 0$
- Jeżeli ten warunek jest niespełniony mówimy o endogeniczności zmiennych objaśniających
  - Złamanie tego założenia ma poważne konsekwencje – estymator MNK przestaje być zgodny
- Może to wynikać z różnych przyczyn:
  - Korelacja ze zmienną pominiętą
  - Sprzężenie zwrotne między zmienna objaśnianą i objaśniającą

Metoda zmiennych instrumentalnych

- Rozwiązanie – zastosowanie zmiennych instrumentalnych
  - Znalezienie zmiennych, które są silnie skorelowane ze zmienną, którą podejrzewamy o endogeniczność, ale nie są skorelowane z błędem losowym
  - Nie jest to zadanie łatwe …
- Podstawowym estymatorem jest tzw. Dwustopniowa Metoda Najmniejszych Kwadratów (2MNK)

Dwustopniowa Metoda Najmniejszych Kwadratów

- W pierwszym kroku liczymy regresję, w której wyjaśniamy wszystkie zmienne z podstawowego modelu zmiennymi instrumentalnymi

$$\mathbf{B} = \left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X}$$

- Wśród zmiennych instrumentalnych $\mathbf{Z}$ mogą być egzogeniczne zmienne z $\mathbf{X}$
- W drugim kroku wyjaśniamy $\mathbf{Y}$ przy użyciu wartości dopasowanych z pierwszych regresji $\mathbf{X} = \mathbf{ZB}$:

$$\boldsymbol{\beta}_{2MNK} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$$

- Estymator jest zgodny jeżeli instrumenty są nieskorelowane z błędem losowym
- Aby model skonwergował potrzebujemy co najmniej tyle samo zmiennych instrumentalnych co zmiennych endogenicznych

Zadanie 1. Korzystając z danych me.twins.dta przeprowadź regresję wyjaśniającą jak posiadanie więcej niż dwójki dzieci wpływa na dochód kobiet

1. Przeprowadź zwykłą regresję liniową, w której logarytm dochodu jest objaśniany przez to czy, kobieta ma więcej niż dwójkę dzieci (*morekids*), wiek (*agem*), wiek przy pierwszym porodzie (*agefstm*),oraz zmienne mówiące o kolorze skóry (*blackm, hispm, othracem*)
2. Porównaj wyniki z estymatorem 2MNK, gdzie zmienna *morekids* jest endogeniczna, a jako instrument wykorzystana jest zmienna *twins*. Czy wyniki różnią od tych uzyskanych z MNK?
3. Sprawdź czy wykorzystane instrumenty są dobrymi predyktorami zmiennej *morekids*
4. Porównaj wyniki z modelem, w którym wykorzystany jest instrument *samesex*

Testowanie endogeniczności

- Popularne są dwa testy na endogeniczność

1. Test Hausmana
   - Jeżeli w modelu nie ma endogeniczności to zastosowanie estymatora 2MNK nie powinno specjalnie zmienić oszacowań parametrów
   - H_0: endogeniczność nie występuje
   - Jeśli H_0 spełnione – estymator 2MNK jest nieefektywny, ale nie powinien powodować niezgodności
   - Statystyka testu

$$H = \left(\boldsymbol{\beta}_{2MNK} - \boldsymbol{\beta}_{MNK}\right)' \left(\mathbf{V}_{2MNK} - \mathbf{V}_{MNK}\right)^{-1} \left(\boldsymbol{\beta}_{2MNK} - \boldsymbol{\beta}_{MNK}\right)$$

     - Statystyka ma rozkład chi-kwadrat z liczbą stopni swobody równą liczbie parametrów

Testowanie endogeniczności

2. Test Wu (i Hausmana)
   - H_0: endogeniczność nie występuje
   - Zakłada się, że mamy następujący model

$$Y_i = \mathbf{X}_i\boldsymbol{\beta} + \alpha X_i^e + \rho v_i + \varepsilon_i$$

   - Gdzie v_i jest zmienną, której nie obserwujemy i która jest skorelowana z X_i^e
   - Chcemy zbadać istotność tej zmiennej tj. czy $\rho = 0$
   - Robimy to dwu stopniowo
     1. Estymujemy $X_i^e = \mathbf{Z_i}\boldsymbol{\theta} + \mathbf{X}_i\boldsymbol{\beta} + u_i$, z czego liczymy reszty $\hat{u}_i$
     2. Estymujemy $Y_i = \mathbf{X}_i\boldsymbol{\beta} + \alpha X_i^e + \rho\hat{u}_i + \varepsilon_i$, i testujemy $\rho = 0$
   - Statystyka ma rozkład F

Zadanie 1. – c.d.

5. Przetestuj model pod kątem występowania endogeniczności
   - Wykorzystaj test Hausmana
   - Wykorzystaj test Wu

```
► Uwaga: Aby uodpornić testy na heteroskedastyczność można we
  wszystkich regresjach dodać opcję vce(robust)
```

Słabe instrumenty

- Pomimo, że estymator 2MNK jest zgodny w nieskończonej próbie, w skończonej próbie jest on obciążony
  - Więcej zmiennych instrumentalnych może powodować większe obciążenie
- Kiedy zmienne instrumentalne są słabo skorelowane ze zmienną endogeniczną mówimy o problemie 'słabych instrumentów' (ang. *weak instruments*)
  - Słabe instrumenty mogą również zwiększać obciążenie

Słabe instrumenty

- Mniej formalne testowanie czy instrumenty są 'słabe' można przeprowadzić na dwa sposoby
  1. Licząc korelacje zmiennej endogenicznej ze zmiennymi instrumentalnymi
  2. Sprawdzając łączną istotność instrumentów w regresji pomocniczej w 2MNK
     - Reguła kciuka mówi, że jeśli statystyka F jest mniejsza niż 10 to instrumenty są słabe

6. Sprawdź czy instrumenty *twins* i *samesex* są słabe.

Słabe instrumenty

- Formalny test na słabość instrumentów można przeprowadzić wykorzystując procedurę Stock i Yogo
  - Tak jak wcześniej wykorzystują oni test F na łączną istotność instrumentów
  - Jeżeli jest więcej niż jedna zmienna endogeniczna wykorzystywane są wartości własne macierzowego odpowiednika statystyki F
  - W tej procedurze wartości krytyczne są jednak inne niż w standardowym teście F
    - Liczone na dwa sposoby

Słabe instrumenty

- Jeżeli mamy o co najmniej 2 instrumenty więcej niż zmiennych endogenicznych to wartość krytyczna może zostać policzona na podstawie
  - Największego obciążenia (relatywnie do MNK) jakie jesteśmy w stanie zaakceptować
  - Liczby zmiennych instrumentalnych
  - Liczby zmiennych endogenicznych
- Dla dowolnej liczby zmiennych instrumentalnych wartość krytyczna może zostać policzona na podstawie
  - Największego obciążenia testu Walda na istotność zmiennych endogenicznych jakie jesteśmy w stanie zaakceptować
  - Liczby zmiennych instrumentalnych + liczby zmiennych endogenicznych

Słabe instrumenty

7. Przeprowadź test na słabość instrumentów.

Słabe instrumenty

- Problem słabych instrumentów można próbować rozwiązać
  - Stosując inny estymator niż 2MNK
    - LIML (Limited Information Maximum Likelihood)
    - GMM (Generalized Method of Moments)
  - Wykorzystując bootstrap i JIVE
  - Znajdując inne instrumenty

Praca domowa ME.4 (grupy 2-osobowe)

1. Wykorzystaj zbiór danych me.wagehw.dta aby sprawdzić wpływ lat edukacji na logarytm płac
   1. Policz regresję w której logarytm płacy jest wyjaśniany przez wybrane zmienne. (Uwaga: nie wykorzystuj edukacji rodziców oraz tego czy respondent mieszkał blisko college'u)
   2. Istnieje obawa, że liczba lat edukacji jest endogeniczna przez to, że nie kontrolujemy np. inteligencji. Przygotuj dwa modele 2MNK: jeden, w którym za zmienne instrumentalne przyjmiesz wykształcenie rodziców (*fatheduc* i *motheduc*) oraz drugi w którym za zmienne instrumentalne przyjmiesz czy respondent mieszkał blisko college'u (*nearc2* i *nearc4*).
      1. Wyjaśnij dlaczego te zmienne mogą być dobrymi zmiennymi instrumentalnymi
      2. Przetestuj endogeniczność dla obu modeli testami Hausmana i Wu
      3. Sprawdź czy nie występuje problem słabych instrumentów – wykorzystaj zarówno metody formalne jak i mniej formalne

Proszę pamiętać o losowej podróbie:

```
set seed 10+"Nr indeksu"
sample 90
```

Regresja kwantylowa

- W KMRL modelujemy warunkową średnią zmiennej objaśnianej:

$$E\left(y_i \mid \mathbf{X}_i\right) = \mu\left(\mathbf{X}_i\right)$$

- Pokazaliśmy, że można modelować równocześnie również jej drugi moment, to jest warunkową wariancję: $\operatorname{var}\left(y_i \mid \mathbf{Z}_i\right) = \sigma^2\left(\mathbf{Z}_i\right)$
  - W przypadku rozkładu normalnego, te dwa momenty w sposób jednoznaczny definiują rozkład zmiennej objaśnianej
  - W praktyce, zmienne ciągłe mogą mieć różne inne rozkłady
  - Czasem chcielibyśmy modelować inne charakterystyki rozkładu, aby lepiej zrozumieć co się dzieje w danych
  - Rozwiązanie – regresja kwantylowa

Regresja kwantylowa

- Kwantylem rzędu τ nazywamy taką wartość λ_τ, dla której zachodzi: $F(\lambda_\tau) = \tau$
  - Taka wartość, że wartości mniejsze lub równe od niej przyjmowane są z prawdopodobieństwem τ, a większe z prawdopodobieństwem $1-\tau$
- Kwantyl można policzyć korzystając ze wzoru: $\lambda_\tau = F^{-1}(\tau)$
- W regresji kwantylowej chcemy zobaczyć, jak różne kwantyle zmiennej objaśnianej zależą od wybranych zmiennych objaśniających
  - Definiuje się tzw. kwantyle warunkowe: $Q_\tau(\mathrm{y}_i \mid \mathbf{X}_i) = F_{y_i}^{-1}(\tau \mid \mathbf{X}_i)$
  - Zazwyczaj, analogicznie jak w regresji liniowej, zakłada się zależność liniową: $Q_\tau(\mathrm{y}_i \mid \mathbf{X}_i) = \mathbf{X}_i\boldsymbol{\beta}_\tau$
  - Uwaga: w tej notacji parametry mogą być różne dla każdego kwantyla

Regresja kwantylowa – estymacja

- Najprostszym wariantem regresji kwantylowej jest tzw. regresja na medianie
  - Staramy się znaleźć postać funkcyjną warunkowej mediany
  - Parametry znajdujemy minimalizując następującą funkcję celu:

$$Q\left(\boldsymbol{\beta}_{0,5}\right) = 0,5\sum_{i=1}^{N}\left|y_i - \mathbf{X}_i\boldsymbol{\beta}_{0,5}\right|$$

  - Bardziej ogólnie, aby oszacować parametry dowolnego kwantyla minimalizujemy funkcję:

$$Q\left(\boldsymbol{\beta}_{\tau}\right) = \sum_{i:y_i \geq \mathbf{X}_i\boldsymbol{\beta}_{\tau}} \tau\left|y_i - \mathbf{X}_i\boldsymbol{\beta}_{\tau}\right| + \sum_{i:y_i \leq \mathbf{X}_i\boldsymbol{\beta}_{\tau}} \left(1-\tau\right)\left|y_i - \mathbf{X}_i\boldsymbol{\beta}_{\tau}\right|$$

  - W literaturze model ten jest również nazywany estymatorem LAD (ang. *Least Absolute Deviations*)
  - Do optymalizacji używa się metod programowania liniowego (metoda simpleks)

Zadanie 2. Wykorzystując dane me.medexp3.dta przygotuj model regresji kwantylowej

1. Przygotuj model regresji kwantylowej w którym logarytm wydatków na leki jest objaśniany przez to, czy ktoś posiada dodatkowe ubezpieczenie od pracodawcy (*hi_empun*), liczbę chorób przewlekłych (*totchr*), wiek (*age*), płeć (*female*), kolor skóry (*blhisp*) oraz logarytm dochodu (*linc*)
   - Domyślnie Stata liczy regresję na medianie
   - Policz model regresji kwantylowej dla kwantyli innych rzędów np. 0,25 i 0,75

Zalety regresji kwantylowej

- Brak założeń na temat rozkładu zmiennej zależnej
- Można modelować cały rozkład zmiennej, a nie tylko wybrane momenty
- Łatwość wykrycia heteroskedastyczności
- Bardziej odporna na występowanie outlierów niż KMRL dla średniej
- Kwantyl zmiennej przekształconej przy pomocy ściśle rosnącej funkcji $g(x)$ jest równy przekształconemu kwantylowi oryginalnej zmiennej:

$$Q_\tau\left(g\left(y_i\right)|\mathbf{X}_i\right)=g\left(Q_\tau\left(\mathrm{y}_i|\mathbf{X}_i\right)\right)$$

Wady regresji kwantylowej

- Wzory analityczne na macierz kowariancji estymatora regresji kwantylowej są skomplikowane, a przez to rzadko stosowane
  - W praktyce stosuje się metody symulacyjne np. bootstrap, przez co model liczy się dłużej niż np. regresja liniowa
  - Współcześnie nie jest to duży problem