Mikroekonometria
12

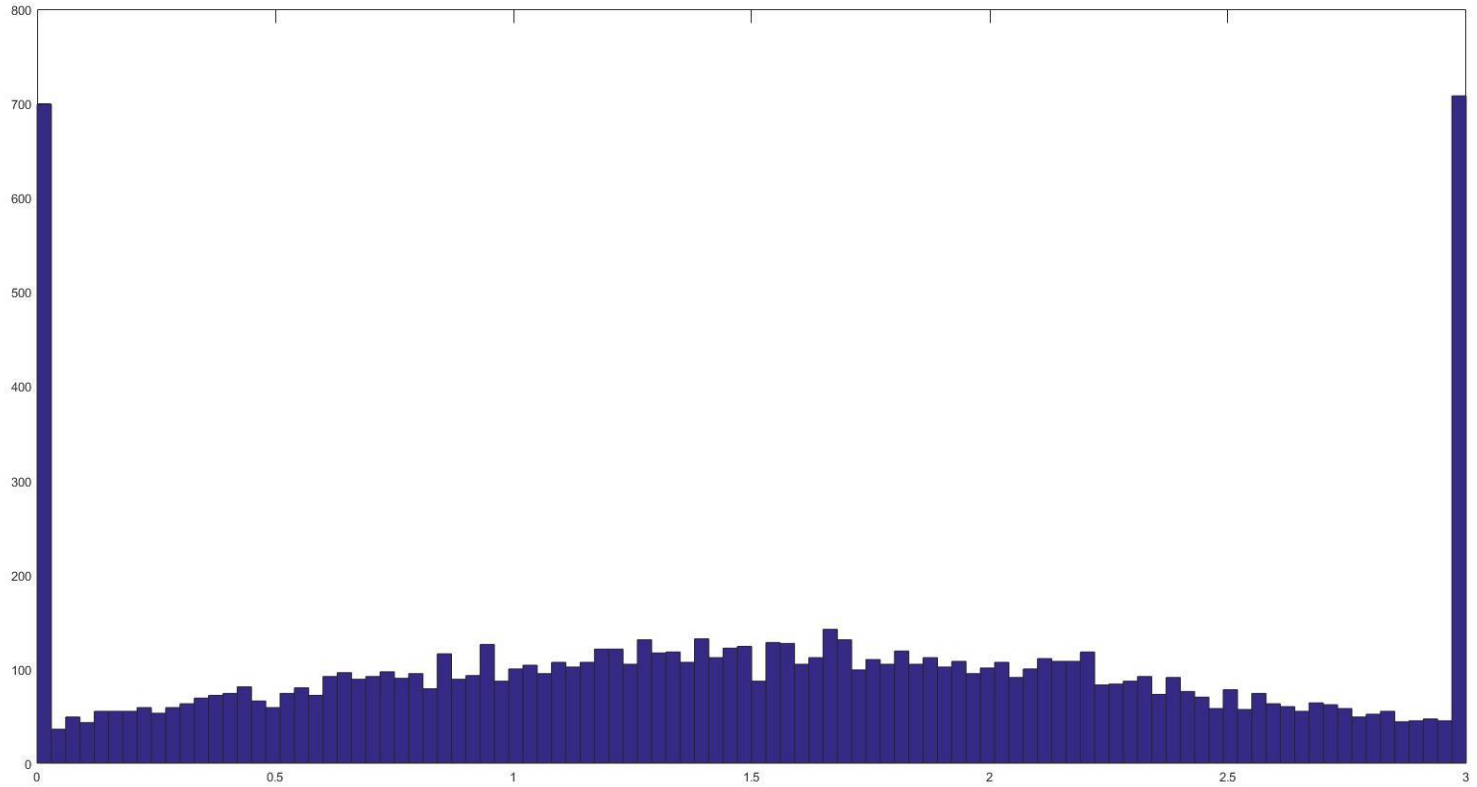
Mikołaj Czajkowski
Wiktor Budziński

Zmienna ocenzurowana

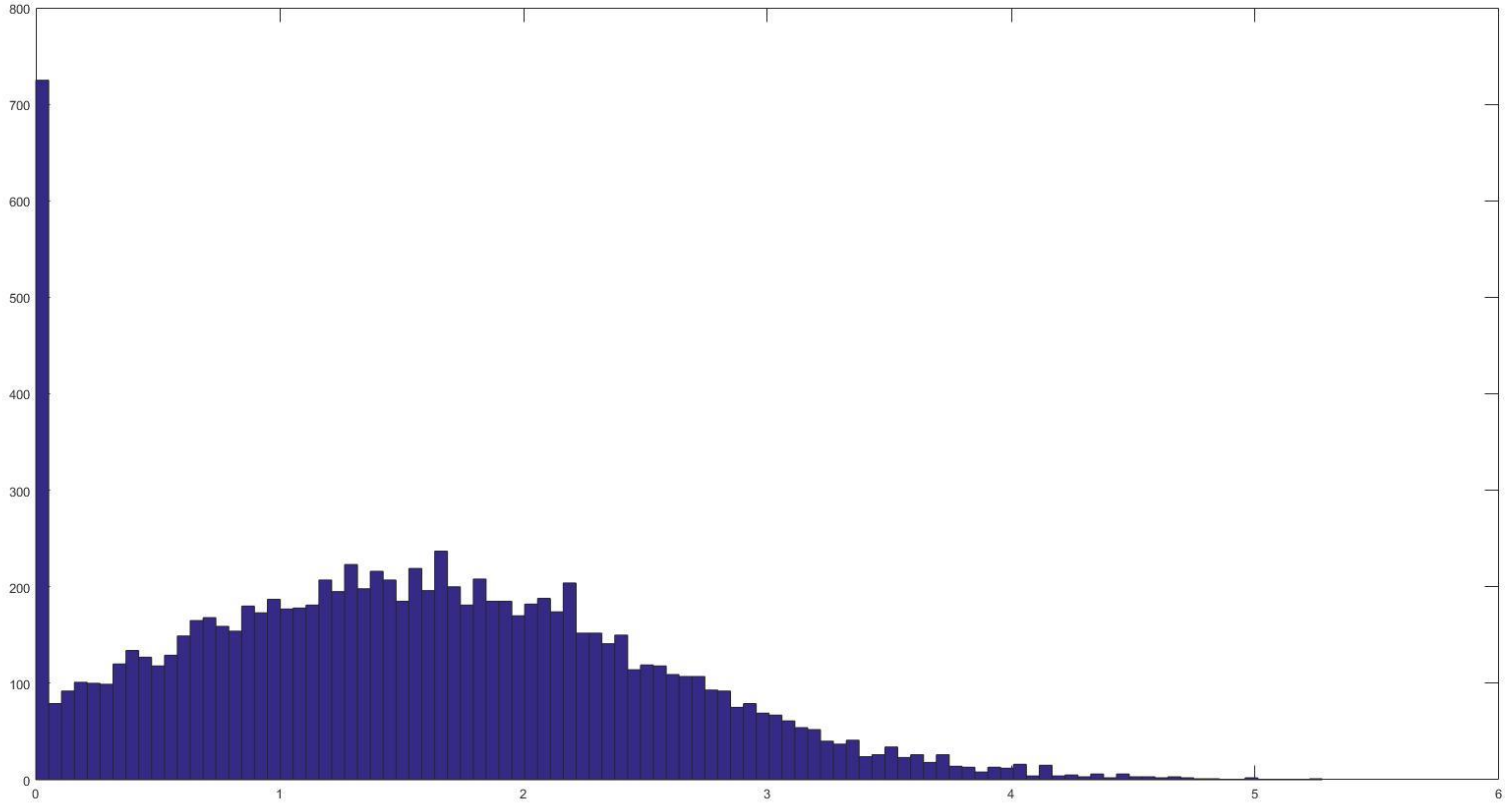
- O ocenzurowanej zmiennej mówimy, gdy dla pewnych obserwacji nie mamy pełnej informacji o wartościach jakie ta zmienna przyjmuje
  - Np. obserwujemy jej wartości jedynie w jakimś przedziale $[a,b]$
    - Dla wartości poza tym przedziałem widzimy jedynie wartości brzegowe
- W takim przypadku, oszacowania MNK oraz MNW, które ignorują ocenzurowanie są obciążone
- Matematycznie można zapisać tę zależność używając funkcji wskaźnikowej (analogicznie jak w modelach dla zmiennych binarnych)
  - Zakładamy, że istnieje pewna ciągła zmienna y^*, której w pełni nie obserwujemy
  - To co obserwujemy to ocenzurowana zmienna y:

$$\begin{cases} y = y^* & y^* \in [a,b] \\ y = a & y^* \leq a \\ y = b & y^* \geq b \end{cases}$$

Przykład – zmienna ocenzurowana dwustronnie



Przykład – zmienna ocenzurowana jednostronnie



Zmienne ocenzurowane – kiedy się pojawiają

- Czasem dane celowo zbierane w postaci ocenzurowanej
  - Mniejszy błąd pomiaru, niż przy pytaniu otwartym
  - Mniej bezpośredni sposób pytania o drażliwe rzeczy
  - Np. pytanie respondenta ankiety o dochód – prośba o wskazanie jednego z przedziałów
- Tych samych modeli używa się do analizy danych, w których mogą występować rozwiązania brzegowe
  - Formalnie nie są to zmienne ocenzurowane
  - Np. wydatki na leki – histogram zmiennej zależnej często przypomina wtedy ten z poprzedniego slajdu

Model Tobitowy (Tobin, 1958)

- Podstawowy model będący pewnym uogólnieniem regresji liniowej
- Model funkcji wskaźnikowej:

$$\begin{cases} y = y^* & y^* > 0 \\ y = 0 & y^* \le 0 \end{cases}, \quad y^* = \mathbf{X}\boldsymbol{\beta} + \varepsilon, \quad \varepsilon \sim N(0, \sigma)$$

- Estymacja metodą największej wiarygodności:

$$L_i = \begin{cases} \frac{1}{\sigma}\phi\left(\frac{y_i - \mathbf{X}_i\boldsymbol{\beta}}{\sigma}\right) & y_i > 0 \\ 1 - \Phi\left(\frac{\mathbf{X}_i\boldsymbol{\beta}}{\sigma}\right) & y_i = 0 \end{cases}$$

  - gdzie $\phi(\cdot), \Phi(\cdot)$ to gęstość i dystrybuanta standardowego rozkładu normalnego
- Analogiczny estymator można skonstruować dla dowolnego rozkładu zmiennej i dowolnego punktu cenzurowania

Zadanie 1. Wydatki na leki

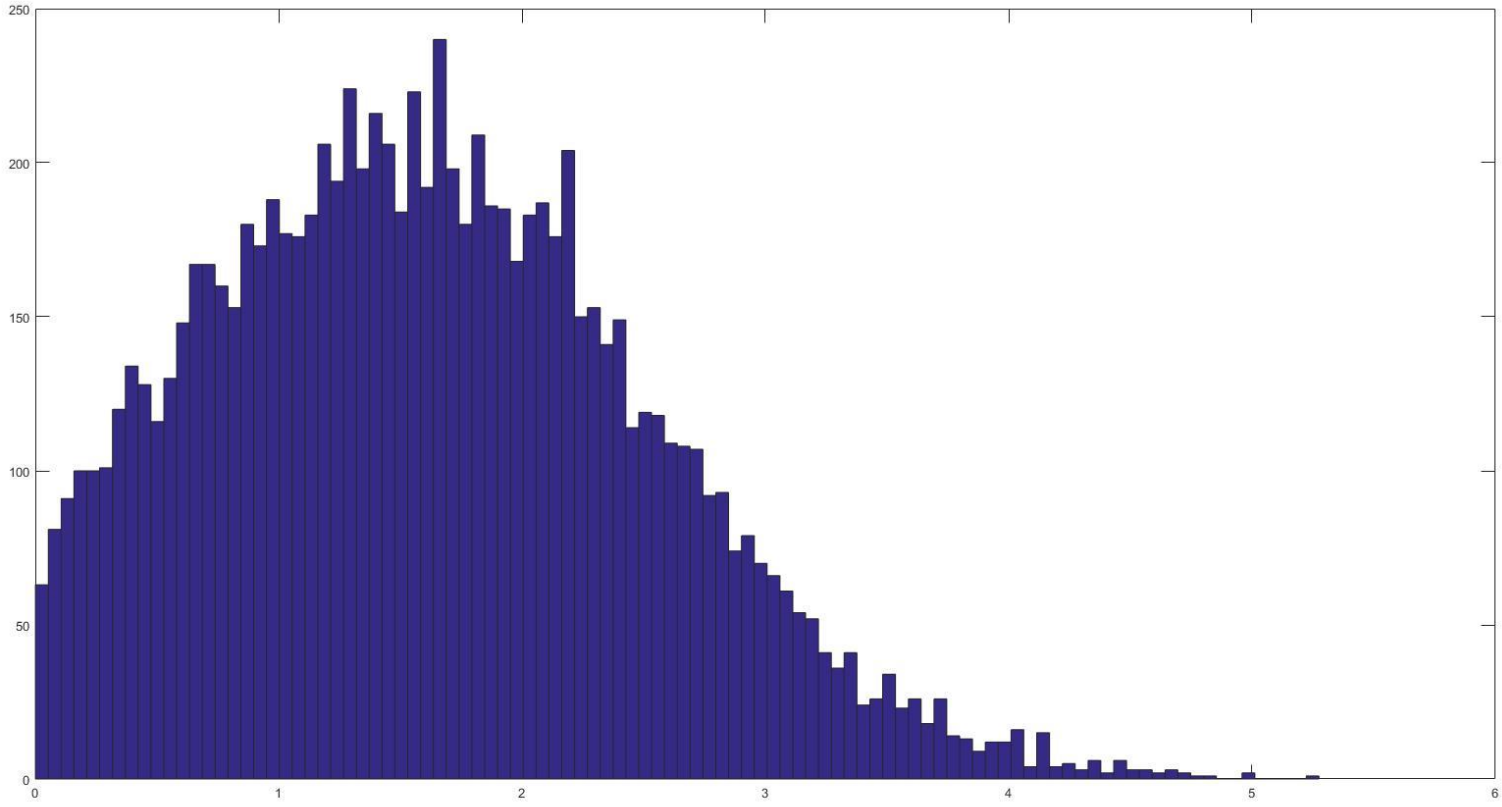
1. Wczytaj zbiór danych me.usahealth.dta
2. Narysuj histogram zmiennej objaśnianej *med* oraz oblicz jaki procent respondentów nie ponosił żadnych wydatków na leki
3. Przeprowadź zwykłą regresję liniową, oraz regresję tobitową, aby wyjaśnić co wpływa na wysokość wydatków na leki. Czy wyniki się różnią? Jak należy je interpretować?

Obcięcie próby

- Obcięciem próby nazywamy sytuację, w której nie posiadamy obserwacji dla pewnych wartości zmiennej objaśnianej
  - Takie obserwacje nie są ocenzurowane – po prostu ich nie ma
- Oszacowania MNK są obciążone
- Jeżeli mamy rozkład obcięty dwustronnie to jego gęstość można zapisać w następujący sposób:

$$f(x) = \begin{cases} \dfrac{f(x)}{F(b) - F(a)} & \text{dla} \quad a < x < b \\ 0 & \text{dla} \quad x \notin [a, b] \end{cases}$$

Przykładowy kształt rozkładu obciętego



Obcięcie próby

- Estymacja metodą największej wiarygodności:

$$L_i = f\left(y_i \mid a < y_i < b, \mathbf{X}_i, \boldsymbol{\beta}\right)$$

  - Np. dla rozkładu normalnego:

$$L_i = \frac{\frac{1}{\sigma}\phi\left(\frac{y_i - \mathbf{X}_i\boldsymbol{\beta}}{\sigma}\right)}{\Phi\left(\frac{b - \mathbf{X}_i\boldsymbol{\beta}}{\sigma}\right) - \Phi\left(\frac{a - \mathbf{X}_i\boldsymbol{\beta}}{\sigma}\right)}$$

  - gdzie $\phi(\cdot), \Phi(\cdot)$ to gęstość i dystrybuanta standardowego rozkładu normalnego
  - Analogiczny estymator można skonstruować dla dowolnego rozkładu zmiennej i dowolnego punktu obcięcia

Zadanie 2. Indeks osiągnięć

1. Wczytaj zbiór danych me.achiv.dta
2. Sprawdź jak znajomości języków i typ studiów wpływa na indeks osiągnięć. Wykorzystaj regresję liniową oraz model z obcięciem próby dla poziomu 40.

Selekcja próby

- Selekcja próby jest problemem podobnym do ocenzurowania – dla części obserwacji nie znamy wartości zmiennej objaśnianej
  - Tutaj zakładamy jednak, że to zjawisko zależy od decyzji badanych podmiotów
- Potrzebujemy wszystkich obserwacji dla zmiennych objaśniających (nawet tych w których brakuje zmiennej objaśnianej)
- Model selekcji próby, nazywany czasem modelem Heckmana, jest modelem dwuwymiarowym

Selekcja próby

- Model składa się z dwóch równań: równania selekcji oraz równania regresji
- Szukamy związku między zmiennymi objaśniającymi $\mathbf{X}_1$ a zmienną objaśnianą y_1, której wartości nie obserwujemy dla niektórych jednostek
- Zakładamy liniowy związek: $y_1 = \mathbf{X}_1\boldsymbol{\beta} + \varepsilon$
- W równaniu selekcji zakładamy, że to, czy obserwujemy wartości zmiennej objaśnianej zależy od funkcji wskaźnikowej y_2^*
- Analogicznie jak w modelach dla zmiennej binarnej, zakładamy, że jeżeli $y_2^* > 0$ to obserwujemy wartości y_1, w przeciwnym wypadku ich nie obserwujemy
- Dodatkowo zakładamy: $y_2^* = \mathbf{X}_2\boldsymbol{\alpha} + \omega$

Selekcja próby

- Problem selekcji próby pojawia się kiedy błędy losowe ε i ω są skorelowane
- Aby je modelować zakłada się, że pochodzą z dwuwymiarowego rozkładu normalnego: $(\varepsilon, \omega) \sim \mathrm{BN}(\mathbf{0}, \Sigma)$
- Z macierzą kowariancji:

$$\Sigma = \begin{bmatrix} \sigma^2 & \sigma\rho \\ \sigma\rho & 1 \end{bmatrix}$$

- Estymacja modelu:
  - Dwustopniowa (Heckman)
  - Jednostopniowa (jednoczesna) – MNW

Selekcja próby – estymacja dwustopniowa

- W pierwszym kroku liczymy model probitowy na zmiennej y_2, zdefiniowanej jak w modelach binarnych:

$$y_2 = \begin{cases} 1 & \text{dla } y_2^* > 0 \\ 0 & \text{dla } y_2^* \leq 0 \end{cases}$$

- Mając wartości dopasowane dla takiego modelu liczymy tzw. odwrotność ilorazu Millsa:

$$\lambda\left(\mathbf{X}_2\boldsymbol{\alpha}\right) = \frac{\phi\left(\mathbf{X}_2\boldsymbol{\alpha}\right)}{\Phi\left(\mathbf{X}_2\boldsymbol{\alpha}\right)}$$

- Następnie liczymy regresję:

$$y_1 = \mathbf{X}_1\boldsymbol{\beta} + \sigma\rho\lambda\left(\mathbf{X}_2\boldsymbol{\alpha}\right) + \varepsilon$$

Selekcja próby – estymacja dwustopniowa

- Testując hipotezę: $\rho\sigma = 0$, możemy sprawdzić czy problem selekcji próby faktycznie występuje
- W metodzie dwustopniowej, aby poprawnie zidentyfikować model w wektorze $\mathbf{X}_2$ powinna być chociaż jedna zmienna spoza wektora $\mathbf{X}_1$

Zadanie 3. Płace kobiet

1. Wczytaj zbiór danych me.femlab.dta
2. Użyj modelu Heckmana, aby wyjaśnić co wpływa na płace kobiet. Czy wszystkie parametry mają oczekiwane znaki?
3. Wypróbuj model wyjaśniający logarytm płac kobiet
4. Zinterpretuj wyniki

Praca domowa ME.12 (grupy 2-3-osobowe)

1. Wykorzystując zbiór me.usahealth.dta przeanalizuj co determinuje wydatki medyczne (zmienna *med*)
2. Wykorzystaj model selekcji próby, w którym zakładamy, że w równaniu selekcji modelujemy czy ktoś ma dodatnie wydatki medyczne czy nie. Czy występuje problem selekcji próby?
   - Uzasadnij wybór dodatkowych zmiennych do równania selekcji
3. Zinterpretuj wyniki
4. Porównaj wyniki z modelem, w którym wyjaśniany jest logarytm wydatków medycznych
5. Porównaj wyniki z modelem Tobitowym i zwykłą regresją liniową

```
set seed 10+"Nr indeksu"
sample 90, by(coins)
```