Mikroekonometria 10

Mikołaj Czajkowski
Wiktor Budziński

Jak analizować dane o charakterze uporządkowanym?

- Dane o charakterze uporządkowanym
  - Wybór jednej z wielkości na uporządkowanej skali
  - Skala nie ma interpretacji absolutnej, tylko uporządkowaną
- Przykłady:
  - Oceny konsumenckie produktów (IMDB, Amazon, etc.)
  - Skala Likerta (stopień braku zgody / zgody z określonymi stwierdzeniami)
  - Ratingi kredytowe (S&P, Moody's, Fitch)
  - Kolejność zawodników w turnieju, zawodach

Wybór uporządkowany

- Model funkcji wskaźnikowej / użyteczności losowej

$$y_i^* = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i$$

  - **X** – charakterystyki konsumenta
    - Zwykle – liniowa specyfikacja funkcji
  - ε – składnik losowy, nieobserwowalne indywidualne idiosynkratyczności
    - Rozkład normalny – uporządkowany probit
    - Rozkład logistyczny – uporządkowany logit
  - Obserwujemy *J* różnych ocen
    - Interesuje nas estymacja β oraz *J*-1 progów, które kategoryzują (cenzurują) y^* na y

$$-\infty < y_i^* < +\infty$$

$$y_i = 1,\ldots,J$$

$$y_i = j \quad dla \quad \alpha_{j-1} < y_i^* < \alpha_j$$

Wybór uporządkowany

- Obserwujemy

$$
\begin{aligned}
y_i &= 1 \quad \text{dla} \quad y_i^* \le \alpha_1 \\
y_i &= 2 \quad \text{dla} \quad \alpha_1 < y_i^* \le \alpha_2 \\
&\ldots \\
y_i &= J \quad \text{dla} \quad y_i^* > \alpha_{J-1}
\end{aligned}
$$

- Jeśli w modelu (**X**) jest stała, to tak jakby $\alpha_1 = 0$ (normalizacja)

Wybór uporządkowany

- Model jest nieliniowy …

$$P\left(y_i = 1 \mid \mathbf{X}_i\right) = F\left(\alpha_1 - \mathbf{X\beta}\right)$$
$$P\left(y_i = 2 \mid \mathbf{X}_i\right) = F\left(\alpha_2 - \mathbf{X\beta}\right) - F\left(\alpha_1 - \mathbf{X\beta}\right)$$
$$\ldots$$
$$P\left(y_i = J \mid \mathbf{X}_i\right) = 1 - F\left(\alpha_{J-1} - \mathbf{X\beta}\right)$$

  - … ponieważ F (dystrybuanta rozkładu normalnego, logistycznego, …) jest funkcją nieliniową
  - Prawdopodobieństwa muszą być dodatnie, więc:

$$\alpha_1 < \alpha_2 < \ldots < \alpha_{J-1}$$

Wybór uporządkowany – funkcja ML

- Z prawdopodobieństw – funkcja największej wiarygodności

$$\ln L = \sum_{i=1}^{N} \ln\left(P\left(y_i\right)\right)$$

  - Suma logarytmów prawdopodobieństw wybranych wartości
- I dalej estymacja normalnie

Zadanie 1. Analiza odpowiedzi na pytania światopoglądowe dotyczące Morza Bałtyckiego

- Ahtiainen et al. (2013) – badanie reprezentatywnej próby mieszkańców 9 krajów nadbałtyckich (*n* = 9627)
- Wśród pytań – pytania światopoglądowe, m.in.:
  - *Martwi mnie stan środowiska Morza Bałtyckiego*
  - *Ja także wpływam na stan środowiska Morza Bałtyckiego*
- Skala odpowiedzi:
  - *1 – zdecydowanie się nie zgadzam*
  - *2 – raczej się nie zgadzam*
  - *3 – trudno powiedzieć*
  - *4 – raczej się zgadzam*
  - *5 – zdecydowanie się zgadzam*

Zadanie 1. Analiza odpowiedzi na pytania światopoglądowe dotyczące Morza Bałtyckiego

1. Wykorzystaj zbiór me.baltic.dta do przeanalizowania, jakie charakterystyki respondentów pozwalają wyjaśnić ich odpowiedzi na pytanie o to czy stan środowiska Bałtyku ich martwi (*envw*)
2. Zinterpretuj wyniki jakościowo
3. Zinterpretuj wyniki ilościowo

Jak analizować dane o liczbie wystąpień jakiegoś zjawiska?

- Dane o liczbie zdarzeń (ang. *count data*)
  - Zmienna objaśniana przyjmuje wartości całkowite (0,1,2,…)
  - Liczby mają bezpośrednią interpretację
- Przykłady
  - Liczba wizyt u lekarza, w parku narodowym, na basenie
  - Liczba dzieci, zachorowań, aresztowań, zabójstw w danym okresie / na jednostce powierzchni
  - Liczba wadliwych sztuk w procesie produkcyjnym

Regresja Poissona

- Potrzebna metoda, która uwzględni charakter zmiennej zależnej
- Regresja Poissona
  - Zmienna zależna y_i traktowana jak zmienna losowa o rozkładzie Poissona
  - $P\left(Y = y_i \mid \mathbf{X}_i\right) = \dfrac{\exp\left(-\lambda_i\right)\lambda_i^{y_i}}{y_i!}$
  - $\ln\left(\lambda_i\right) = \mathbf{X}_i\boldsymbol{\beta}$
  - Oczekiwana liczba zdarzeń w okresie

$$\mathrm{E}\left(y_i \mid \mathbf{X}_i\right) = \mathrm{var}\left(y_i \mid \mathbf{X}_i\right) = \lambda_i = \exp\left(\mathbf{X}_i\boldsymbol{\beta}\right)$$

    - Silne założenie modelu: średnia = wariancji rozkładu – wrócimy do tego

Regresja Poissona – estymacja

- Model Poissona można estymować za pomocą regresji nieliniowej, ale prościej za pomocą MNW
  - Funkcja LL – suma logarytmów prawdopodobieństw zaobserwowanych ilości

$$\ln L = \sum_{i=1}^{N} \left(-\lambda_i + y_i \mathbf{X}_i \boldsymbol{\beta} - \ln(y_i!) \right)$$

  - Gradient

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} \mathbf{X}_i \left(y_i - \lambda_i \right)$$

  - Hesjan

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = -\sum_{i=1}^{N} \lambda_i \mathbf{X}_i \mathbf{X}_i'$$

  - Hesjan ujemnie określony dla wszystkich β i X
  - Optymalizacja metodą Newtona

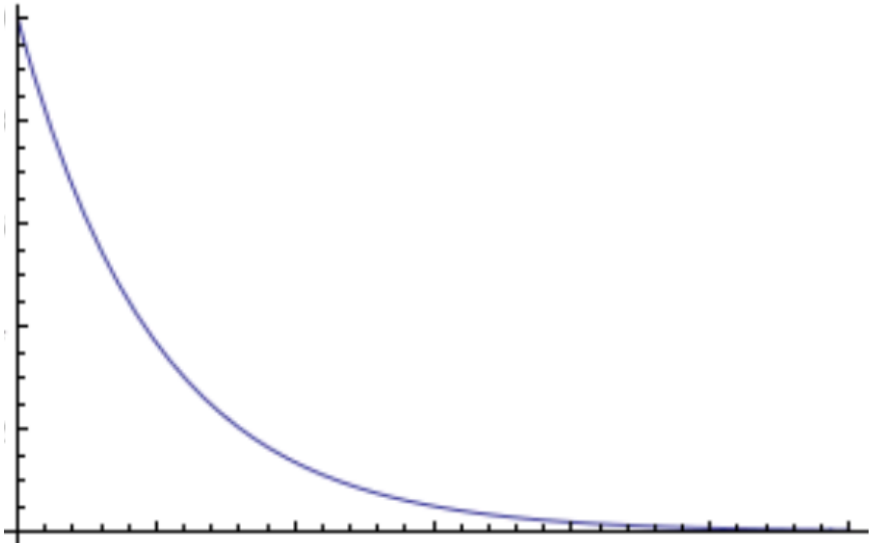
Zadanie 2. Liczba podróży nad Morze Bałtyckie

- Ahtiainen et al. (2013) – badanie reprezentatywnej próby mieszkańców 9 krajów nadbałtyckich (*n* = 9627)
- Pytania dotyczące liczby podróży nad Morze Bałtyckie w ciągu ostatnich 12 miesięcy i szczegółów ostatniej podróży
  - Dystans, środek transportu, czas, …

1. Wykorzystaj zbiór me.baltic.dta do przeprowadzenia regresji Poissona liczby wizyt nad morze (*TRIPS*), wyjaśniając je stałą specyficzną dla kraju i kosztem podróży (*TC_km*)
   - Zinterpretuj wyniki
   - Jaka jest nadwyżka konsumenta wynikająca z możliwości wizyt nad Morzem Bałtyckim?

Nadwyżka konsumenta

- Oczekiwana liczba podróży jest funkcją m.in. kosztu podróży
  - $\mathrm{E}(y_i \mid \mathbf{x}_i) = \exp(\mathbf{x}_i \boldsymbol{\beta})$
- Funkcja popytu
  - Parametr przy koszcie jest ujemny, więc funkcja $y = \exp(-x)$ ma taki kształt:



Nadwyżka konsumenta

- Funkcja popytu dana przez $\mathrm{E}(y_i \mid \mathbf{X}_i) = \exp(\mathbf{X}_i\boldsymbol{\beta})$
- Nadwyżka konsumenta to:

$$\mathrm{CS} = \int_{TC}^{+\infty} \exp(\beta x)\,dx = -\frac{\exp(\beta TC)}{\beta}$$

- Nadwyżka konsumenta na jedną wizytę jest więc dana przez:

$$CS_{per\,trip} = \frac{\mathrm{CS}}{\mathrm{E}(y_i \mid \mathbf{X}_i)} = -\frac{1}{\beta}$$

  - Minus odwrotność parametru dla kosztu podróży

2. Oszacuj nadwyżkę konsumenta biorąc pod uwagę także koszt alternatywny czasu podróży (*TC_time*)

Regresja Poissona – ekwidyspersja

- Jednym z założeń modelu – ekwidyspersja
  - Średnia = wariancja rozkładu $\mathrm{E}(y_i \mid \mathbf{X}_i) = \mathrm{var}(y_i \mid \mathbf{X}_i) = \lambda_i = \exp(\boldsymbol{\beta}'\mathbf{X}_i)$
- To założenie może w praktyce nie być spełnione
  - Test nadmiernej dyspersji:

    $H_0 : \mathrm{var}(y_i) = E(y_i)$

    $H_1 : \mathrm{var}(y_i) = E(y_i) + \alpha^2 g(E(y_i))$
  - Prosta regresja liniowa objaśniająca z_i (wariancja minus średnia) przez w_i (średnia):

$$z_i = \frac{\left(y_i - \hat{\lambda}_i\right)^2 - y_i}{\hat{\lambda}_i} \qquad w_i = \frac{g\left(\hat{\lambda}_i\right)}{\hat{\lambda}_i} \qquad g\left(\hat{\lambda}_i\right) = \hat{\lambda}_i \text{ lub } g\left(\hat{\lambda}_i\right) = \hat{\lambda}_i^2$$

    $\hat{\lambda}_i$ – średnia przewidywana przez model
    - Sprawdzamy istotność współczynnika w regresji bez stałej

3. Sprawdź czy regresja Poissona jest w naszym przypadku uzasadniona

Model ujemny dwumianowy

- Model ujemny dwumianowy – rozszerzenie modelu Poissona, polegające na wprowadzeniu dodatkowego składnika losowego (nieobserwowalnej heterogeniczności) do średniej
  - $\ln\left(\mathrm{E}\left(y_i \mid \mathbf{X}_i\right)\right) = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i = \ln\lambda_i + \ln u_i$
- y_i (pod warunkiem x_i i u_i) ma rozkład Poissona z warunkową średnią i wariancją μ_i

$$f\left(y_i \mid \mathbf{X}_i, u_i\right) = \frac{\exp\left(-\lambda_i u_i\right)\left(\lambda_i u_i\right)^{y_i}}{y_i!}$$

- Bezwarunkowy rozkład $f\left(y_i \mid \mathbf{X}_i\right)$

$$f\left(y_i \mid \mathbf{X}_i\right) = \int_0^{+\infty} \frac{\exp\left(-\lambda_i u_i\right)\left(\lambda_i u_i\right)^{y_i}}{y_i!} g\left(u_i\right) du_i$$

Model ujemny dwumianowy

- Funkcja gęstości u_i determinuje bezwarunkowy rozkład y_i
- Wygodnie założyć rozkład gamma, $E(u_i)=1$

$$g(u_i)=\frac{\theta^\theta}{\Gamma(\theta)}\exp(-\theta u_i)u_i^{\theta-1}$$

- Wtedy bezwarunkowy rozkład y_i dany jest przez

$$f(y_i|\mathbf{x}_i)=\int_0^{+\infty}\frac{\exp(-\lambda_i u_i)(\lambda_i u_i)^{y_i}}{y_i!}\frac{\theta^\theta u_i^{\theta-1}\exp(-\theta u_i)}{\Gamma(\theta)}du_i$$

$$=\frac{\theta^\theta\lambda_i^{y_i}}{\Gamma(y_i+1)\Gamma(\theta)}\int_0^{+\infty}\exp\left((-\lambda_i+\theta)u_i\right)u_i^{\theta+y_i-1}du_i$$

$$=\frac{\theta^\theta\lambda_i^{y_i}\Gamma(\theta+y_i)}{\Gamma(y_i+1)\Gamma(\theta)(\lambda_i+\theta)^{\theta+y_i}}$$

Model ujemny dwumianowy

- Warunkowa średnia λ_i
- Warunkowa wariancja $\lambda_i\left(1+\left(1/\theta\right)\lambda_i\right)$
- Uwagi:
  - Ekwidyspersję można przetestować *ex post* jako $1/\theta = 0$
  - Możliwe inne specyfikacje u_i – np. rozkład normalny

Zadanie 2. Liczba podróży nad Morze Bałtyckie

4. Skonstruuj model regresji ujemnej dwumianowej
   - Czy restrykcja $1/\theta = 0$ jest uzasadniona?
   - Czy zmieniły się oszacowania CS?

Model ujemny dwumianowy – overdyspersja

- Jednym z możliwych rozszerzeń modelu – modelowanie determinant overdyspersji
  - Heterogeniczność średniej i wariancji zawsze ważna dla danych mikroekonomicznych
  - Parametr dyspersji θ wyłapuje ogólne skalowanie rozkładu (średnia vs. wariancja)
  - Ogólnie wariancja to $\text{var}\left(y_i \mid X_i\right) = \lambda_i\left(1+\left(1/\theta\right)\lambda_i\right)$
  - Zróbmy więc $\theta_i = \theta \exp\left(\mathbf{Z}_i\boldsymbol{\delta}\right)$
    - Teraz θ (a więc ogólniej – wariancja) funkcją obserwowalnych zmiennych charakteryzujących respondentów **Z**

5. Skonstruuj model regresji ujemnej dwumianowej z różnym poziomem overdyspersji dla różnych krajów
   - De facto jest to próba lepszego dopasowania rozkładów – trochę inny rozkład liczby wycieczek dla każdego kraju

Zero inflated models

- Częstym problemem jest duża liczba obserwacji przyjmująca wartość 0
  - Znane rozkłady, takie jak Poisson czy ujemny dwumianowy, nie przewidują ponadproporcjonalnej ilości obserwacji 0, więc źle pasują
- Możliwym rozwiązaniem są tzw. *Zero Inflated Models*
  - Załóżmy, że mamy dwa typy konsumentów
    - Uczestników rynku – robią liczbę wycieczek zależną od kosztu (w tym możliwe 0 wycieczek)
    - Nieuczestników rynku – niezależnie od kosztu i tak nie pojadą nad Bałtyk
  - Oba segmenty powinny być modelowane osobno, a nie jako jeden ciągły rozkład

Zero inflated models

- 0 w danych może się pojawić z dwóch powodów
  - Ktoś jest uczestnikiem rynku, ale tak się zdarzyło, że w ostatnim roku nie pojechał ani razu
  - Ktoś nie jest uczestnikiem rynku, więc pojechał 0 razy
- Prawdopodobieństwo pojechania *k* razy będzie wtedy dane przez

$$P(Y=y_i \mid \mathbf{X}_i, \mathbf{Z}_i) = \begin{cases} p_i(\mathbf{Z}_i) + (1-p_i(\mathbf{Z}_i))F(y_i \mid \mathbf{X}_i) & \text{jeśli} \quad y_i = 0 \\ (1-p_i(\mathbf{Z}_i))F(y_i \mid \mathbf{X}_i) & \text{jeśli} \quad y_i \neq 0 \end{cases}$$

  - Gdzie $p_i(\mathbf{Z}_i)$ to prawdopodobieństwo bycia poza rynkiem, zazwyczaj modelowane binarnym logitem albo probitem
  - $F(y_i \mid \mathbf{X}_i)$ to p-stwo zaobserwowania danej liczby zdarzeń, opisane np. modelem Poissona albo Ujemnym dwumianowym

Zero inflated models

6. Dokonaj estymacji modelu Zero Inflated Negative Binomial
   - Porównaj dopasowanie do danych ze wcześniejszymi modelami

Praca domowa ME.10 (grupy 2 lub 3-osobowe)

1. Wykorzystaj projekt me.baltic.dta do przeanalizowania, jakie charakterystyki respondentów pozwalają wyjaśnić ich ocenę własnego wpływu na środowisko Bałtyku (*ienv*)
   1. Wybierz najlepszą, Twoim zdaniem, specyfikację
   2. Zinterpretuj wyniki używając interpretacji jakościowej oraz ilościowej (wykorzystując efekty krańcowe)

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

2. Wykorzystując projekt me.usahealth.dta skonstruuj model liczności zdarzeń objaśniający liczbę wizyt u lekarza (*mdu*)
   1. Wybierz najlepszą, Twoim zdaniem, postać funkcyjną i specyfikację
      - Uwzględnij zmienne socjodemograficzne, wskaźniki stanu zdrowia oraz udział własny w kosztach opieki medycznej
   2. Zinterpretuj wyniki używając interpretacji jakościowej oraz ilościowej (wykorzystując efekty krańcowe)

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