

**Model selekcji próby Heckmana**

Modele zmiennych ograniczonych

Justyna Zbiegień  
 107961

# Opis problemu

Opisanym problem jest rosnąca liczba rozwodów w Iraku i czynniki wpływające na podjęcie decyzji o zakończeniu małżeństwa. Liczba rozwodów w Iraku rośnie z roku na rok w szybkim tempie, w 2004 roku liczba sięgnęła 28 689, natomiast w 2016 roku było to już 97 273 rozwodów. Łatwo można zauważyć, że coraz więcej osób decyduje się na taki krok, w tym wiele osób rozwodzi się wielokrotnie.

Problem dotyczący rosnącej liczby rozwodów opisali w swym artykule Meshal Harbi Odah, Ali Sadig Mohommed Bager oraz Bahr Kadhim Mohammed zatytułowanym „Studying the determinants of divortiality in Iraq. A two-stage estimation model with tobit regression”, który ukazał się w 2018 roku w czasopiśmie pt. “International Journal of Applied Mathematics & Statistical Sciences”.

# Opis metod i zmiennych

W badaniu użyto modelu selekcji próby Heckmana, w którym to w pierwszej kolejności oszacowano model probitowy, a w drugim etapie użyto MNK do oszacowania ostatecznej wersji modelu – w ten sposób tworząc zarówno równanie selekcji oraz regresji. Badanie przeprowadzono w 2017 roku w Iraku na próbie 127 osób (93 mężczyzn oraz 34 kobiety).

Za zmienną objaśnianą w modelu przyjęto liczbę mówiącą o tym ile raz dana osoba była rozwiedziona. Respondenci udzielili odpowiedzi z zakresu 0 – 5, jednak nie otrzymali górnego ograniczenia. Jako zmienne objaśniające przyjęto 11 czynników, które przedstawia poniższa tabela:

Tabela 1: Opis zmiennych

|  |  |
| --- | --- |
| Zmienna | Opis |
| AGE | Wiek, w którym zawarto małżeństwo |
| GEN | Płeć (1-M, 0-K) |
| EDU | Najwyższy ukończony poziom edukacji (0-Wyższy, 1-Poniżej wyższego) |
| OCC | Sytuacja zawodowa małżonka/-ki (1-Zatrudniony, 0-W przeciwnym wypadku) |
| INC | Dochód całej rodziny (USD) |
| CHILD | Liczba dzieci |
| YEARS | Liczba lat małżeństwa |
| TYPE | Rodzaj małżeństwa (1-Tradycyjne, 0-W przeciwnym wypadku) |
| NOT CHILD | Czy posiada dzieci (0-Nie, 1-Tak) |
| INFID | Czy kiedykolwiek dokonano zdrady na małżonku/-ce (1-Tak, 2-Nie) |
| SOCIA | Czy małżonek/-ka posiada konto w mediach społecznościowych (1-Tak, 0-Nie) |

# Zadania

Poniższe tabele przedstawiają dwa etapy stworzonych modeli na podstawie powyższych zmiennych:

Tabela 2: Model Probitowy - etap 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Zmienna | Coef. | Std. Err. | t-Student | *P > |t|* |
| AGE | 0.1444365 | 0.0233873 | 6.176 | 6.58e-10 |
| GEN | -0.0387082 | 0.3653799 | -0.106 | 0.9156 |
| EDU | -0.2222043 | 0.3390066 | -0.655 | 0.5122 |
| OCC | 0.2463305 | 0.4116445 | 0.598 | 0.5496 |
| INC | 0.0004454 | 0.0002104 | 2.117 | 0.0343 |
| CHILD | -0.3887768 | 0.2210644 | -1.759 | 0.0786 |
| YEARS | -0.0675317 | 0.0525393 | -1.285 | 0.1987 |
| TYPE | -0.0537258 | 0.3434738 | -0.156 | 0.8757 |
| NOTCHILD | -0.0863410 | 0.4886758 | -0.177 | 0.8598 |
| INFID | -0.2398445 | 0.4624700 | -0.519 | 0.6040 |
| SOCIA | -2.8531856 | 0.6864336 | -4.157 | 3.23e-05 |
| Sigma | 0.5157508 | 0.0792613 | 6.507 | 7.67e-11 |

**Log-likelihood: -206.4877**

Tabela 3: Model estymowany MNK - etap 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Zmienna | Coef. | Std. Err. | t-Student | *P > |t|* |
| AGE | -0.129 | 0.018 | -7.1667 | 0.000 |
| GEN | 0.2120 | 0.0161 | 13.1677 | 0.0000 |
| EDU | -0.1358 | 0.0336 | -4.0416 | 0.000 |
| OCC | -0.258 | 0.3947 | -0.65366 | 0.4363 |
| INC | 521.362 | 18.075 | 28.8447 | 0.000 |
| CHILD | 0.2297 | 0.4059 | 0.5659 | 0.8637 |
| YEARS | -0.010 | 0.105 | -0.098 | 0.922 |
| TYPE | 0.157 | 0.216 | 0.726 | 0.215 |
| NOT CHILD | -0.271 | 0.0168 | -19.4702 | 0.000 |
| INFID | 0.301 | 0.028 | 10.075 | 0.000 |
| SOCIA | 0.273 | 0.0167 | 16.347 | 0.000 |
| \_SE1 | -0.4255811 | 0.5884499 |  |  |

**Pseudo R2: 0.901**

Na podstawie powyższych tabel rozwiąż poniższe zadania:

1. Czemu autorzy zastosowali model selekcji próby Hackmana zamiast modelu tobitowego? Uzasadnij.
2. Zinterpretuj wynik *Log-likelihood*.
3. Zinterpretuje wynik pseudo R2. Czy jest to wysoki wynik?
4. Które ze zmiennych finalnie uznano za najbardziej istotne?
5. Zinterpretuj oszacowany parametr dla zmiennych INC oraz NOT CHILD.
6. Ile razy średnio rozwiedzie się osoba gdy:
   * Będzie miała 25 lat;
   * Będzie mężczyzną;
   * Ukończy stopień wyższy edukacji;
   * Będzie żyła w gospodarstwie domowym o dochodzie 2000 USD;
   * Będzie posiadać jedno dziecko;
   * Nigdy nie zdradziła partnera/-ki;
   * Będzie posiadała partnera mającego konto w mediach społecznościowych?

Pomiń zmienne nieistotne.

# Rozwiązania do zadań

Rozwiązania do kolejnych podpunktów:

1. Model tobitowy jest stosowany do wartości ocenzurowanych, w tym wypadku z dołu (0). Jednak w tym wypadku nie mamy wartości nieznanych (nie ma przypadku, że ktoś udzielił odpowiedzi -1, a model przyjął odpowiedź jako 0), są to zwyczajnie poprawne wartości, jednak które należy odrzucić w celu poprawnego oszacowania modelu. Dlatego najlepiej użyć modelu selekcji, by odpowiedzi równe 0 „odfiltrować”.
2. *Log-likelihood* jest logarytmem prawdopodobieństwa. Służy do oceny dopasowania modelu. Pożądana jest jak największa wartość, jednak trzeba pamięć że może ona przyjmować wartości z przedziału -∞ do ∞. Tutaj wartość funkcji jest dość niska, co może sugerować na słabe dopasowanie modelu do danych.
3. W modelu zmiennych ograniczonych nie można używać zwykłego R2 do oceny modelu, dlatego użyto pseudo R2, które wylicza się ze wzoru , gdzie *ln*Lfit to logarytm funkcji wiarygodności modelu, który estymujemy, a *ln*L0 to logarytm funkcji wiarygodności modelu tylko z wyrazem wolnym. W tym modelu pseudo R2 wyniosło ponad 90%, co jest wartością niespotykanie dobrą. W takich modelach jest tendencja otrzymywanie niskich wartości dla tego współczynnika. W tym wypadku współczynnik mówi nam, że ponad 90% rozwodów jest tłumaczona dobranymi zmiennymi.
4. Spośród dobranych 11 zmiennych, tylko 7 uznano za istotne, a są to: wiek, płeć, poziom edukacji, dochód, zmienna binarna dotycząca posiadania dzieci, zmienna binarna dotycząca zdrady i aktywność partnera/-ki w mediach społecznościowych.
5. Oficjalny model można interpretować jak zwykłą regresję liniową estymowaną MNK, jednak wskazuje on wpływ danej zmiennej na zmienną objaśnianą a nie rzeczywistą jej wartość. Zmienna INC dotycząca dochodów ma największy wpływ, ponieważ każdy 1 dodatkowy dolar (USD), który otrzymuje rodzina zwiększa prawdopodobieństwo rozwodu – dodając do „wpływu” 521.362. Zmienna NOT CHILD ma negatywny wpływ na rozwód, oznacza to, że jeśli respondent posiada co najmniej jedno dziecko jest mniejsza szansa na rozwód o -0.271 jednostki.
6. Średnią liczbę rozwodów liczmy z modelu 1 pomijając zmienne nieistotne (lata w małżeństwie, typ małżeństwa, liczba posiadanych dzieci, stan zatrudnienia partnera/-ki). Można to obliczyć ze wzoru:

co po podstawieniu wartości z zadania wygląda następująco:

Ostatecznie:

Taka osoba rozwiedzie się średnio 1.6 razy, zaokrąglając – 2 razy w życiu.

# Literatura

Odah M.H., Bager A.S.M., Bahr M.K. (2018), *Studying the determinants of divortiality in Iraq. A two-stage estimation model with tobit regression*, IASET

Bazyl M., Gruszczyński M. i inni (2012), *Mikroekonometria. Modele i metody analizy danych indywidualnych,* Wolters Kluwer

Gruszczyński M., Przekop D. i inni (2015), *Zbiór zadań z mikroekonometrii. Modele i metody analizy danych indywidualnych*, Wolters Kluwer

Spis treści

[Opis problemu 1](#_Toc73455531)

[Opis metod i zmiennych 1](#_Toc73455532)

[Zadania 2](#_Toc73455533)

[Rozwiązania do zadań 3](#_Toc73455534)

[Literatura 5](#_Toc73455535)