ZMIENNE ZEROJEDYNKOWE I KATEGORYCZNE

EKONOMETRIA WNE 2022/23

Sebastian Zalas

24 listopada 2022

WPROWADZENIE

- W dotychczas analizowanych modelach, zmienne zależne i niezależne miały interpretację *ilościową*

  przykłady: wynagrodzenie, średnia ocen, cena domów etc.
- w pracy empirycznej należy również uwzględniać czynniki *jakościowe*
  - płeć lub rasa
  - sektor w którym operuje firma (przemysł, usługi etc.)
  - region
- dziś zajmiemy się **zależnymi** zmiennymi jakościowymi

WPROWADZENIE

- W dotychczas analizowanych modelach, zmienne zależne i niezależne miały interpretację *ilościową*

  przykłady: wynagrodzenie, średnia ocen, cena domów etc.
- w pracy empirycznej należy również uwzględniać czynniki *jakościowe*
  - płeć lub rasa
  - sektor w którym operuje firma (przemysł, usługi etc.)
  - region
- dziś zajmiemy się **zależnymi** zmiennymi jakościowymi

WPROWADZENIE

- W dotychczas analizowanych modelach, zmienne zależne i niezależne miały interpretację *ilościową*

  przykłady: wynagrodzenie, średnia ocen, cena domów etc.
- w pracy empirycznej należy również uwzględniać czynniki *jakościowe*
  - płeć lub rasa
  - sektor w którym operuje firma (przemysł, usługi etc.)
  - region
- dziś zajmiemy się **zależnymi** zmiennymi jakościowymi

WPROWADZENIE

- W dotychczas analizowanych modelach, zmienne zależne i niezależne miały interpretację *ilościową*

  przykłady: wynagrodzenie, średnia ocen, cena domów etc.
- w pracy empirycznej należy również uwzględniać czynniki *jakościowe*
  - płeć lub rasa
  - sektor w którym operuje firma (przemysł, usługi etc.)
  - region
- dziś zajmiemy się **zależnymi** zmiennymi jakościowymi

ZMIENNE BINARNE

- czynniki jakościowe można wyrazić jako **zmienną zero-jedynkową** (**zmienną binarną**).
- w takich przypadkach należy zdecydować jaka cecha ma przyjąć wartość 1 (a jaka 0)
- Przykład: analizujemy zależność między wynagrodzeniem a płcią. Płeć możemy zakodować jako:
  1. gender (1 kobieta) (0 mężczyzna)
  2. female (1 kobieta) (0 mężczyzna)
  3. male (0 kobieta) (1 mężczyzna)
- Który sposób jest najlepszy?

ZMIENNE BINARNE II

- *gender* jest nieintuicyjne: nie domyślamy się co oznacza wartość 1
- *male* lub *female* to odpowiedni wybór, w zależności od pytania które stawiamy.
- Dlaczego zmienne zerojedynkowe przyjmują wartości 0 i 1?
  - wartości te są arbitralne, jakiekolwiek dwie wartości mogłyby opisać cechę jakościową.
  - stosowanie zera i jedynki opłaca się, ponieważ w regresji taka zmienna zyskuje intuicyjną interpretację

ZMIENNE BINARNE II

- *gender* jest nieintuicyjne: nie domyślamy się co oznacza wartość 1
- *male* lub *female* to odpowiedni wybór, w zależności od pytania które stawiamy.
- Dlaczego zmienne zerojedynkowe przyjmują wartości 0 i 1?
  - wartości te są arbitralne, jakiekolwiek dwie wartości mogłyby opisać cechę jakościową.
  - stosowanie zera i jedynki opłaca się, ponieważ w regresji taka zmienna zyskuje intuicyjną interpretację

ZMIENNE BINARNE - PRZYKŁAD

Rozważmy model objaśniający płace:

$$wage = \beta_0 + \delta\, female + \beta_1 educ + u \tag{1}$$

- $female = 1$ jeśli osoba jest kobietą, $female = 0$ jest osoba jest mężczyzną.
- dlatego $female$ mierzy różnicę w zarobakch między kobietami i mężczyznami, przy takim samym poziomie edukacji:

$$\delta = \mathbb{E}[wage \mid female = 1, educ] - \mathbb{E}[wage \mid female = 1, educ]$$

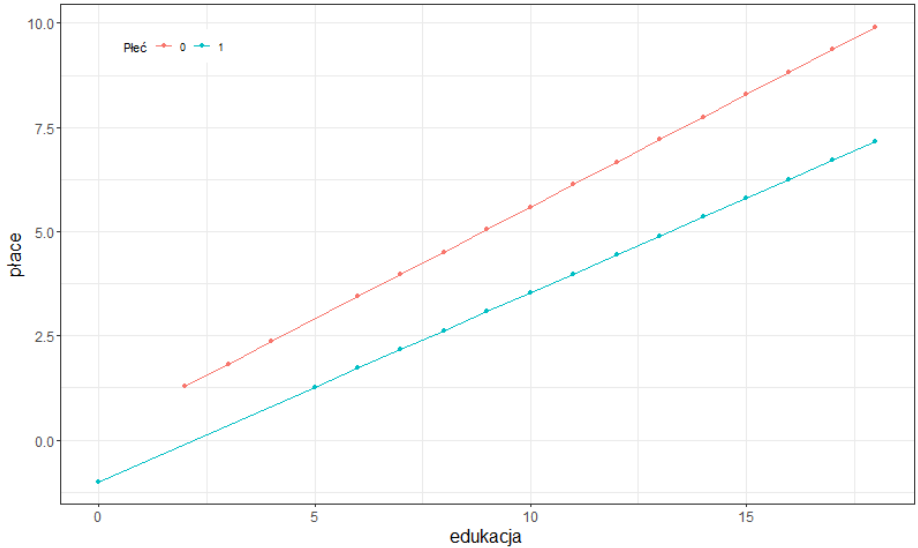
ZMIENNE BINARNE - PRZYKŁAD

Używając danych `wage1` oszacowano model objaśniający płace (1):

$$wage = 0.6228 - 2.2734\, female + 0.5065 educ + u$$

- interpretacja: kobiety zarabiają średnio $2.2734 na godzinę mniej niż mężczyźni z takim samym wykształceniem.
- mężczyźni są poziomem odniesienia dal kobiet.
- w przypadku modelu z większą liczbą zmiennych objaśniających, interpretacja jest analogiczna

ZMIENNE BINARNE - PRZYKŁAD



ZMIENNE BINARNE - LOGARYTM ZMIENNEJ ZALEŻNEJ

Model objaśniający *logarytm* płac:

$$\log(wage) = \beta_0 + \delta\, female + \beta_1 educ + u$$

$\Rightarrow$ Jak zinterpretować $\hat{\delta}$?

zapiszmy modele dla kobiet i mężczyzn:

$$\log(wage_F) = \beta_0 + \delta * 1_{\{female==1\}} + \beta_1 educ + u$$

$$\log(wage_M) = \beta_0 + \delta * 1_{\{female==0\}} + \beta_1 educ + u$$

odejmijmy stronami, weźmy exponent

$$\log(wage_F) - \log(wage_M) = \delta$$

$$\frac{wage_F}{wage_M} = e^{\delta}$$

ZMIENNE BINARNE - LOGARYTM ZMIENNEJ ZALEŻNEJ

Przekształćmy w taki sposób, aby interpretować zmienną zależną w kategoriach zmiany procentowej:

$$\frac{wage_F}{wage_M} - 1 = e^{\delta} - 1$$

$$100 * (\frac{wage_F}{wage_M} - 1) = 100 * (e^{\delta} - 1)$$

$\Rightarrow$ kobiety zarabiają średnio mniej/więcej niż mężczyźni o $100 * (e^{\delta} - 1)$ procent, ceteris paribus.
$\Rightarrow$ do interpretacji możemy użyć także przybliżenia $e^{\delta} \approx \delta + 1$ rozwinięciem szeregu Taylora.
Wtedy:

$$100 * (\frac{wage_F}{wage_M} - 1) = 100\delta$$

ZMIENNE BINARNE - LOGARYTM ZMIENNEJ ZALEŻNEJ - PRZYKŁAD

Oszacowano model objaśniający *logarytm* płac:

$$\log(wage) = 0.8263 - 0.3609\, female + 0.0772 educ + u$$

Interpretacja:

- $100 * (e^{\delta} - 1) = -30.29272$
- $100\delta = -36.08654$
- kobiety zarabiają średnio mniej niż mężczyźni o 30 procent (36 procent), ceteris paribus
- przybliżone oszacowanie efektu pozostanie takie samo przy zamianie poziomu odniesienia
- ten sposób interpretacji obowiązuje w modelu potęgowym

ZMIENNE KATEGORYCZNE

- zmienne jakościowe mogą opisywać więcej niż dwa poziomy $\Rightarrow$ kolor samochodów (1=czarny, 2=czerwony, 3=żółty)
- aby te cechy uwzględnić w regresji, każdy poziom należy zakodować jako oddzielną zmienną binarną.
- należy wybrać jedną grupę jako poziom odniesienia i nie uwzględniać go regresji

ZMIENNE KATEGORYCZNE - PRZYKŁAD

- przykład: model objaśniający płacę za pomocą statusu matrymonialnego (dla kobiet i mężczyzn) oraz edukacji.
- *Wage premium*: osoby w małżeństwie zarabiają więcej, poniższy model eksploruje to zjawisko w zależności od płci:

$$\begin{aligned}\log(wage) = 0.3213781 &+ 0.2126757 marr\_male - 0.1982676 marr\_fem \\ &- 0.1103502 sing\_fem + 0.0789103 educ + 0.0268006 ex\ per \\ &- 0.0005352 ex\ persq + 0.0290875 tenure - 0.0005331 tenursq + u\end{aligned}$$

- poziom odniesienia: nieżonaci mężczyźni (single)
- żonaci mężczyźni zarabiają średnio o $\approx$21% więcej niż single
- zamężne kobiet zarabiają średnio o $\approx$20% mniej niż single

ZMIENNE KATEGORYCZNE UPORZĄDKOWANE

- uporządkowanie: kolejność poziomów zmiennej zależnej
- tak jak w przypadku zwykłej zmiennej kategorycznej, należy utworzyć zmienne binarne
- jako poziom odniesienia można wybrać np.: najniższy poziom
- przykład: edukacja (1=podstawowe, 2=średnie, 3=wyższe)
- ranking szkół: utworzenie tylu zmiennych zero-jedynkowych może być niemożliwe! (brak stopni swobody)

ZMIENNE KATEGORYCZNE UPORZĄDKOWANE - PRZYKŁAD

- przeanalizujmy związek między medianą początkowych zarobków dla absolwentów szkół prawnych a rankingiem szkoły
- można podzielić ranking na grupy: *to p*10, *r*11_25, *r*26_40, *r*41_60, *r*61_100
- poziom odniesienia: szkoły z rankingiem poniżej 100

$$\log(sal\,ar\,y) = 9.17 + .700 to\,p10 + .594 r11\_25 + .375 r26\_40 + .263 r41\_60 + 1.132 r61\_100 + .0057 LSAT + .014 GPA + .036 \log(l\,ibvol) + .0008 \log(cost)$$

- interpretacja: absolwenci szkół z top10 rankingu mają ($e^{0.7} - 1 = 1.014$) medianę zarobków wyższą o ponad 100%, przy pozostałych czynnikach niezmienionych

Pytania? Wątpliwości?

Dziękuję!

e: s.zalas@uw.edu.pl