Zmienne binarne i kategoryczne

Ekonometria WNE

Sebastian Zalas

University of Warsaw
s.zalas@uw.edu.pl

Wprowadzenie

- Dotychczas zmienne 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

- Dotychczas zmienne 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

- Dotychczas zmienne 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

- *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

- *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

Rozważmy model objaśniający płace:

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

gdzie $female = 1 \Rightarrow$ kobieta, $female = 0 \Rightarrow$ mężczyzna

Co mierzy δ?

$$\mathbb{E}[wage \mid female = 1, educ] = \beta_0 + \delta \times 1 + \beta_1 educ + \varepsilon$$

$$\mathbb{E}[wage \mid female = 0, educ] = \beta_0 + \delta \times 0 + \beta_1 educ + \varepsilon$$

Po odjęciu stronami otrzymujemy:

$$\delta = \mathbb{E}[wage \mid female = 1, educ] - \mathbb{E}[wage \mid female = 0, educ]$$

δ = różnica średnich płac między kobietami i mężczyznami, przy takim samym poziomie edukacji

Zmienne binarne

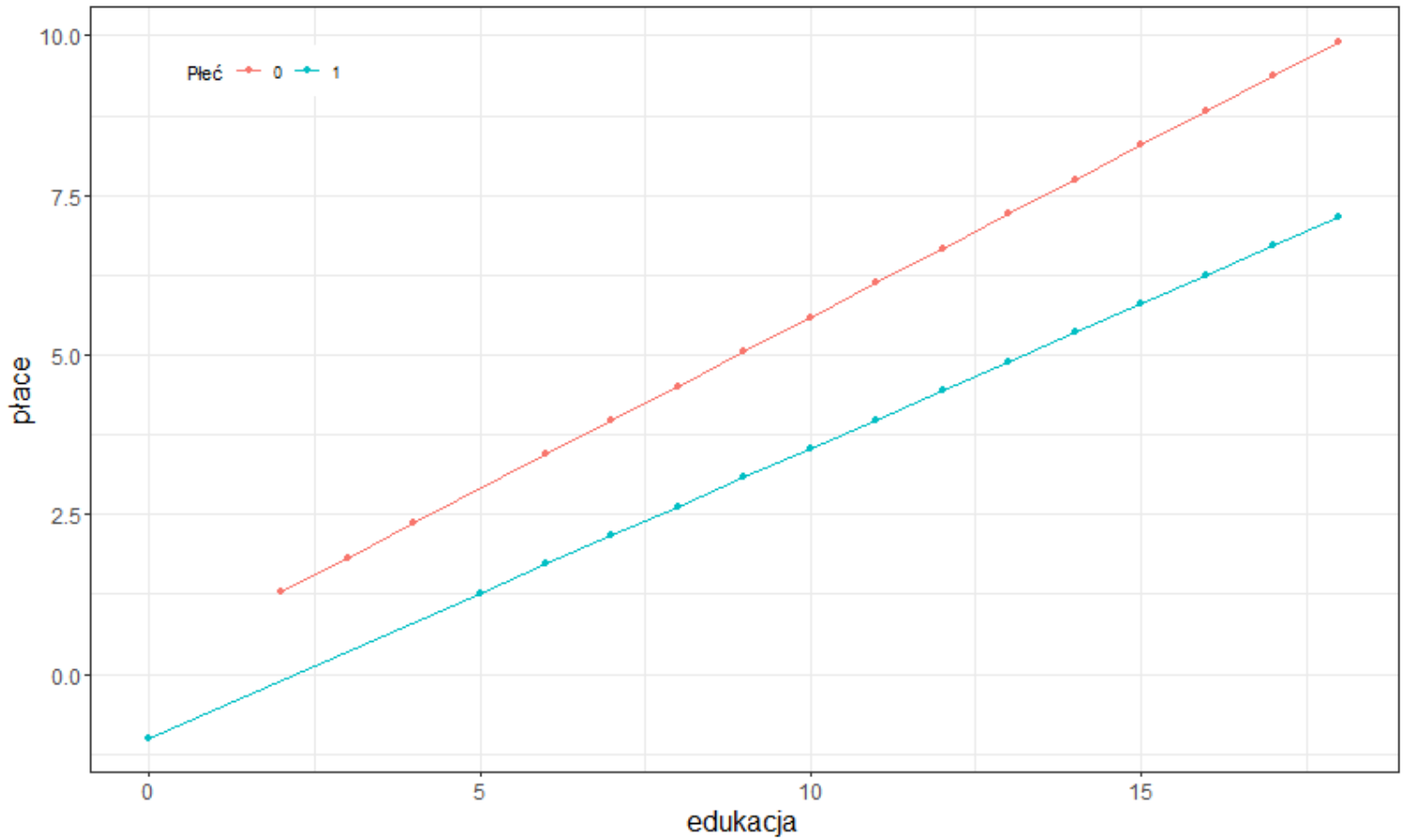
Używając danych `wage1` oszacowano model objaśniający płace (1):

$$wage = 0.62 - 2.27\ female + 0.51\ educ + e$$

- interpretacja $\delta = 2.27$:
  Kobiety zarabiają średnio o \$2.27 na godzinę mniej niż mężczyźni przy niezmienionym poziomie wykształcenia

- mężczyźni są poziomem odniesienia dla kobiet.
- w przypadku modelu z większą liczbą zmiennych objaśniających, interpretacja jest analogiczna

Zmienne binarne



Zmienne binarne - $log(y)$

Model objaśniający *logarytm* płac:

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

Co mierzy δ? Zapiszmy modele dla kobiet i mężczyzn:

$$\mathbb{E}[\log(wage)|female = 1, educ] = \beta_0 + \delta \times 1 + \beta_1 educ + \varepsilon$$
$$\mathbb{E}[\log(wage)|female = 0, educ] = \beta_0 + \delta \times 0 + \beta_1 educ + \varepsilon$$

odejmijmy stronami

$$\delta = \mathbb{E}[\log(wage)|female = 1, educ] - \mathbb{E}[\log(wage)|female = 0, educ]$$

teraz mamy logarytmy!

$$e^{\delta} = \frac{wage_{female=1}}{wage_{female=0}}$$

Zmienne binarne - $log(y)$

Zmiana procentowa:

$$\frac{wage_{female=1}}{wage_{female=0}} - 1 = e^{\delta} - 1$$

$$\frac{wage_{female=1} - wage_{female=0}}{wage_{female=0}} \times 100 = 100 \times (e^{\delta} - 1)$$

Interpretacja $\delta \Rightarrow$ kobiety zarabiają średnio mniej/więcej niż mężczyźni o $100(e^{\delta} - 1)$ procent, przy innych czynnikach niezmienionych

Do interpretacji możemy użyć także przybliżenia $e^{\delta} \approx \delta + 1$ (szeregu Taylora):

$$\frac{wage_{female=1} - wage_{female=0}}{wage_{female=0}} \times 100 = 100\,\delta$$

Zmienne binarne - $log(y)$

Oszacowano model objaśniający *logarytm* płac:

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

Interpretacja:

- $100 * (e^{\delta} - 1) = -30.29272$
- $100\delta \approx -36.08654$
- kobiety zarabiają średnio mniej niż mężczyźni o 30 procent (36 procent), przy innych czynnikach niezmienionych
- przybliżone oszacowanie efektu pozostanie takie samo przy zamianie poziomu odniesienia

Zmienne kategoryczne

- zmienne jakościowe mogą opisywać więcej niż dwa poziomy:
  - sektory: przemsył, usługi, rolnictwo..
  - zawody
  - wykształcenie: podstawowe, średnie, wyższe

- aby te cechy uwzględnić w regresji, każdy poziom należy zakodować jako oddzielne zmienne binarne

- należy wybrać jedną grupę jako poziom odniesienia i nie uwzględniać go regresji

Zmienne kategoryczne

- model objaśniający płace za pomocą statusu matrymonialnego (dla kobiet i mężczyzn) oraz edukacji i doświadczenia:

$$\begin{aligned}\log(wage) = 0.32 &+ 0.21 married\_male - 0.19 married\_fem \\ &- 0.11 single\_fem + 0.079 educ + 0.027 exper \\ &- 0.0005 exper^2 + 0.02 tenure - 0.0005 tenure^2 + e\end{aligned}$$

- poziom odniesienia? nieżonaci mężczyźni (single)
  - żonaci mężczyźni zarabiają średnio o $\approx$21% więcej niż single, przy pozostałych czynnikach niezmienionych
  - zamężne kobiet zarabiają średnio o $\approx$20% mniej niż single, przy pozostałych czynnikach niezmienionych

Zmienne kategoryczne uporządkowane

- szczególny przypadek zmiennej kategorycznej: tu kolejność poziomów zmiennej ma znaczenie
- jako poziom odniesienia można wybrać np.: najniższy poziom
- przykłady:
  - edukacja (1=podstawowe, 2=średnie, 3=wyższe)
  - ranking szkół

Zmienne kategoryczne uporządkowane

- medianą początkowych zarobków absolwentów szkół prawnych a ranking szkoły
- można podzielić ranking na grupy: *top*10, *r*11_25, *r*26_40, *r*41_60, *r*61_100

$$\begin{aligned}\log(salary) = 9.17 &+ 0.700top10 + 0.594r11\_25 + 0.375r26\_40 \\ &+ 0.263r41\_60 + 1.132r61\_100 + 0.0057LSAT \\ &+ 0.014GPA + 0.036\log(libvol) + 0.0008\log(cost) + e\end{aligned}$$

- interpretacja: absolwenci topowych dziesięciu szkół mają ($e^{0.7} - 1 = 1.014$) medianę początkowych zarobków wyższą o ponad 100% w porównaniu z absolwentami szkół o rankingu poniżej setnego, przy pozostałych czynnikach niezmienionych

Podsumowanie

- zmienne binarne/kategoryczne pozwalają uwzględnić cechy jakościowe
- zmienne kategoryczne sprowadzamy do zbiory zmiennych binarnych
- uważamy na poziom odniesienia i logarytmy

Pytania? Wątpliwości?

Dziękuję!

e: s.zalas@uw.edu.pl