Zmienne dyskretne
Kontrasty

Stanisław Cichocki

Natalia Nehrebecka

Wykład 7

Plan wykładu

- 1. Interpretacja parametrów przy zmiennych dyskretnych
- 2. Kontrasty: kontrasty w odchyleniach, efekty progowe

Plan wykładu

- 1. Interpretacja parametrów przy zmiennych dyskretnych
- 2. Kontrasty: kontrasty w odchyleniach, efekty progowe

Zmienne dyskretne

- Nieco bardziej skomplikowana jest sytuacja, gdy mamy do czynienia ze zmienną dyskretną która przyjmuje więcej niż 2 wartości.

- np. **wykształcenie** (1 – podstawowe, 2 – średnie, 3 - wyższe)

- W tym przypadku do każdego poziomu **s** zmiennej dyskretnej $\mathbf{X_i}$ musimy przypisać jedną zmienną zero-jedynkową **Ds,i**

$$D_{s,i} = 1 \text{ gdy } X_i = s$$
$$D_{s,i} = 0 \text{ gdy } X_i \neq s \text{ dla } s = 1,2,...,S$$

Przykład

$$wykształcenie_i = \begin{cases} 1 & podstawowe \\ 2 & średnie \\ 3 & wyższe \end{cases}$$

$$podstawowe_i = \begin{cases} 1 & podstawowe \\ 0 & w\ p.p. \end{cases}$$

$$średnie_i = \begin{cases} 1 & średnie \\ 0 & w\ p.p. \end{cases}$$

$$wyższe_i = \begin{cases} 1 & wyższe \\ 0 & w\ p.p. \end{cases}$$

Zmienne dyskretne

- Za poziom bazowy uznajemy jeden z poziomów (np. poziom 1), i zmienną zero-jedynkową związaną z tym poziomem usuwamy z modelu ze stałą.

- Np. dla zmiennej wykształcenie
- Poziom bazowy : wykształcenie podstawowe

$$placa_i = \beta_1 + \beta_2 średnie_i + \beta_3 wyzsze_i + \varepsilon_i$$

- Dlaczego?

  Nie jest możliwe, by w modelu była jednocześnie stała i wszystkie zmienne zero-jedynkowe (dla każdego poziomu zmiennej dyskretnej), ponieważ macierz X^TX byłaby osobliwa!

Interpretacja parametrów przy zmiennych dyskretnych

- Interpretacja współczynników w modelu z wieloma zmiennymi 0-1 (zmiennymi dyskretnymi) jest analogiczna jak w przypadku modelu z jedną tylko taką zmienną:

  dany współczynnik opisuje różnicę między oczekiwaną wartością zmiennej **y** dla respondenta o charakterystyce bazowej i dla respondenta o charakterystyce s.

Przykład

Modelujemy płace za pomocą płci, wieku i wykształcenia:

Zmienna	Współczynniki
Płeć	-0,278
Wiek	0,078
Wykszt. średnie	-0,273
Wykszt. średnie zawodowe	-0,273
Wykszt. zawodowe	-0,444
Wykszt. podstawowe	-0,571
Stała	6,64

Dlaczego rozkodowujemy zmienne dyskretne

$$\ln(placa_i) = \beta_1 + \beta_2 plec_i + \beta_3 wyksztalcenie + \varepsilon_i$$

$$\ln(placa_i) = \beta_1 + \beta_2 plec_i + \beta_3 wyksztalcenie + \beta_4 wojewodztwo + \varepsilon_i$$

Plan wykładu

- 1. Interpretacja parametrów przy zmiennych dyskretnych
- 2. Kontrasty: kontrasty w odchyleniach, efekty progowe

Kontrasty w odchyleniach

- Jeśli jednym z celów badania jest zidentyfikowanie poziomów zmiennej dyskretnej, których **wpływ wyróżnia się znacząco od wpływu pozostałych poziomów**, wtedy celowe jest użycie tak zwanych kontrastów w odchyleniach.

Przykład – kontrasty w odchyleniach

W modelu będziemy uzależniać dochód od wieku, płci oraz zmiennej województwo (16 poziomów):

1 Dolnośląskie
2 Kujawsko-pomorskie
3 Lubelskie
4 Lubuskie
5 Łódzkie
6 Małopolskie
7 Mazowieckie
8 Opolskie
9 Podkarpackie
10 Podlaskie
11 Pomorskie
12 Śląskie
13 Świętokrzyskie
14 Warmińsko-mazurskie
15 Wielkopolskie
16 Zachodniopomorskie

Kontrasty w odchyleniach

- Krok 1: tworzymy 16 zmiennych zerojedynkowych odpowiadających zmiennej województwo:

$$D_{s,i} = \begin{cases} 1 & \text{dla woj} = \text{j} \\ 0 & \text{dla woj} \neq \text{j} \end{cases}$$

Dla s = 1,...,16

- Krok 2: Następnie definiujemy zmienne:

$$D^{*}_{s,i} = D_{s,i} - D_{1,i} \quad dla \quad s = 2,..,16$$

Kontrasty w odchyleniach

- Krok 3: Zapisujemy regresje:

$$placa_i = \beta_1 wiek_i + \beta_2 plec_i + \gamma_0^* + \gamma_2^* D_{2,i}^* + \ldots + \gamma_{16}^* D_{16,i}^* + \varepsilon_i$$

- **W jaki sposób można interpretować parametry przy zmiennych $D^*_{s,i}$.**
- Dla każdej obserwacji zachodzi:

$$D_{1,i} + \ldots + D_{16,i} = 1$$

$$placa_i = \beta_1 wiek_i + \beta_2 plec_i + \gamma_0^*(D_{1,i} + \ldots + D_{16,i}) + \gamma_2^*(D_{2,i} - D_{1,i}) + \ldots + \gamma_{16}^*(D_{16,i} - D_{1,i}) + \varepsilon_i$$

$$placa_i = \beta_1 wiek_i + \beta_2 plec_i + \underbrace{(\gamma_0^* - \gamma_2^* - \ldots - \gamma_{16}^*)}_{\gamma_1} D_{1,i} + \underbrace{(\gamma_0^* + \gamma_2^*)}_{\gamma_2} D_{2,i} + \ldots + \underbrace{(\gamma_0^* + \gamma_{16}^*)}_{\gamma_{16}} D_{16,i} + \varepsilon_i$$

Kontrasty w odchyleniach

- Przekształciliśmy model do modelu bez stałej.
- Sumujemy parametry przy zmiennych zerojedynkowych dotyczących województwa:

$$\sum_{s=1}^{16} \gamma_s = 16\gamma_0^* \Rightarrow \gamma_0^* = \frac{\sum_{s=1}^{16} \gamma_s}{16}$$

- Czyli **stała w modelu jest średnią z parametrów dla poszczególnych zmiennych dotyczących województwa**.

Kontrasty w odchyleniach

- Pozostaje nadanie interpretacji parametrom przy zmiennych $D^*_{s,i}$:

$$\gamma_2 = \gamma_0^* + \gamma_2^* \Rightarrow \gamma_2^* = \gamma_2 - \gamma_0^*$$
$$\vdots$$
$$\gamma_{16} = \gamma_0^* + \gamma_{16}^* \Rightarrow \gamma_{16}^* = \gamma_{16} - \gamma_0^*$$

- Czyli parametry γ_s^* można interpretować jako **odchylenia parametrów dla poszczególnych poziomów województwa od średniej z tych parametrów**.
- Trzeba jeszcze wyznaczyć odchylenie od średniej dla poziomu bazowego :

$$\gamma_1 = \gamma_0^* - \gamma_2^* - \dots \gamma_{16}^* \Rightarrow \gamma_1 - \gamma_0^* = -\gamma_2^* - \dots \gamma_{16}^*$$

Przykład – kontrasty w odchyleniach

Płaca i miejsce zamieszkania: kontrasty w odchyleniach

log(placa)	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
_woj_2	-.0258665	.0268517	-0.96	0.335	-.0785046	.0267717
_woj_3	-.0749633	.0280217	-2.68	0.007	-.129895	-.0200316
_woj_4	-.0001867	.0368011	-0.01	0.996	-.0723291	.0719557
_woj_5	-.0717755	.0238393	-3.01	0.003	-.1185085	-.0250425
_woj_6	-.012634	.0218834	-0.58	0.564	-.0555327	.0302647
_woj_7	.2557709	.0166333	15.38	0.000	.2231642	.2883777
_woj_8	-.0027719	.0366859	-0.08	0.940	-.0746884	.0691446
_woj_9	-.0500334	.0272721	-1.83	0.067	-.1034957	.003429
_woj_10	-.1031224	.0345293	-2.99	0.003	-.1708112	-.0354337
_woj_11	.0841202	.0265058	3.17	0.002	.0321601	.1360804
_woj_12	.0839597	.0168495	4.98	0.000	.0509291	.1169903
_woj_13	-.0096191	.0372951	-0.26	0.796	-.0827298	.0634915
_woj_14	-.0930655	.0341943	-2.72	0.007	-.1600977	-.0260334
_woj_15	-.0062165	.0229601	-0.27	0.787	-.0512258	.0387928
_woj_16	.0280367	.034522	0.81	0.417	-.0396379	.0957113
_Iplec_2	-.1706226	.0126903	-13.45	0.000	-.1954997	-.1457455
wiek	.0121618	.0006334	19.20	0.000	.0109201	.0134035
_cons	7.135958	.0272595	261.78	0.000	7.082521	7.189396

$$\gamma_1 - \gamma_0^* = -\gamma_2^* - \ldots \gamma_{16}^* = -0{,}002$$ dla woj. Dolnośląskiego

Efekty progowe

- Stosowane do **zmiennych dyskretnych o uporządkowanych kategoriach** (rosnąco lub malejąco).
- Przy standardowym rozkodowaniu zmiennej dyskretnej na zmienne zerojedynkowe, kategorie wprowadzone do modelu interpretuje się względem kategorii w modelu nieuwzględnionej (bazowej).
- Niewiadomo natomiast jak zmienia się poziom analizowanego zjawiska przy przejściu z jednej kategorii wprowadzonej do modelu do drugiej.
- Na taką interpretację pozwalają efekty progowe.

Efekty progowe

- Sposób zdefiniowania zmiennych zerojedynkowych zależy od tego, czy uporządkowanie zmiennej dyskretnej jest rosnące, czy malejące.

- W przypadku **porządku rosnącego** zmienne zerojedynkowe zdefiniowane są następująco:

$$D^{+}_{s,i} = \begin{cases} 1 & \text{dla } z_i \geq s \\ 0 & \text{dla } z_i < s \end{cases}$$

Dla s = 2,...,S

- W przypadku **porządku malejącego** zmienne zerojedynkowe zdefiniowane są następująco:

$$D^{-}_{s,i} = \begin{cases} 1 & \text{dla } z_i \leq s \\ 0 & \text{dla } z_i > s \end{cases}$$

Dla s = 1,...,S-1

Przykład – efekty progowe

miasto	Freq.	Percent	Cum.
1 - wies	323	29.82	29.82
2 - miasto do 25tyś	194	17.91	47.74
3 - miasto od 25tyś do 250tyś	356	32.87	80.61
4 - miasto powyżej 250tyś	210	19.39	100.00
Total	1,083	100.00	

```
generate miasto_male = (miasto > 1)

generate miasto_srednie = (miasto > 2)

generate miasto_duze = (miasto > 3)
```

Przykład – efekty progowe

```
. generate miasto_male = (miasto > 1)

. generate miasto_srednie = (miasto > 2)

. generate miasto_duze = (miasto > 3)


. regres dochod wiek wiek_2 miasto_male miasto_srednie miasto_duze
```

Source	SS	df	MS		
Model	23872603.5	5	4774520.71	Number of obs	1083
Residual	723608532	1077	671874.217	F(5, 1077)	7.11
				Prob > F	0.0000
				R-squared	0.0319
				Adj R-squared	0.0274
Total	747481135	1082	690832.842	Root MSE	819.68

dochod	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
wiek	37.8833	16.01033	2.37	0.018	6.468336	69.29827
wiek_2	-.4486477	.2039518	-2.20	0.028	-.8488356	-.0484597
miasto_male	**158.2807**	74.50027	2.12	0.034	12.0986	304.4629
miasto_srednie	**107.7085**	73.16483	1.47	0.141	-35.85331	251.2702
miasto_duze	**79.57117**	71.45687	1.11	0.266	-60.63929	219.7816
_cons	-119.8138	303.7319	-0.39	0.693	-715.7871	476.1596

Pytania teoretyczne

1. Dlaczego zmienną zależną rozkodowywujemy na zmienne zerojedynkowe?
2. Dlaczego w modelu nie powinno się umieszczać stałej i wszystkich zmiennych zero-jedynkowych, związanych z poziomami zmiennej dyskretnej?
3. Porównać zastosowania znanych kontrastów ze standardowym sposobem rozkodowywania zmiennej dyskretnej.

Dziękuję za uwagę