Kontrasty
Interakcje
Przybliżanie modeli nieliniowych

Stanisław Cichocki

Natalia Nehrebecka

Wykład 8

Plan wykładu

- 1. Kontrasty: efekty progowe, kontrasty w odchyleniach
- 2. Interakcje
- 3. Przybliżanie modeli nieliniowych:
  - Model wielomianowy

Plan wykładu

- 1. Kontrasty: efekty progowe, kontrasty w odchyleniach
- 2. Interakcje
- 3. Przybliżanie modeli nieliniowych:
  - Model wielomianowy

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

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

Plan wykładu

- 1. Kontrasty: efekty progowe, kontrasty w odchyleniach
- 2. Interakcje
- 3. Przybliżanie modeli nieliniowych:
  - Model wielomianowy

Modele z interakcjami

- W standardowym modelu liniowym zakładamy, że wpływ poszczególnych zmiennych niezależnych na oczekiwaną wartość zmiennej niezależnej jest **addytywny**.

- W ramach modelu liniowego można także uwzględnić efekt krzyżowego wzmacniania się efektów poszczególnych zmiennych.

- Efekt ten zachodzi, gdy siła oddziaływania jednej zmiennej niezależnej jest uwarunkowana wielkością innych zmiennych niezależnych.

- Ten efekt można uwzględnić, wstawiając do modelu iloczyny zmiennych (interakcje).

Interakcje między zmiennymi zerojedynkowymi

- Interakcje między zmiennymi zerojedynkowymi bierzemy pod uwagę, jeśli wpływ poszczególnych zmiennych nie jest addytywny.

- **Sytuacja taka może wystąpić, jeśli pewne kombinacje charakterystyk jakościowych wpływają na zmienną zależną bardziej lub mniej, niż wynikałoby z wpływu poszczególnych zmiennych.**

- Np.
- Zmienna zależna: dochód
- Zmienne niezależna płeć, wykształcenie, interakcja: płećXwykształcenie
- Do modelu wprowadzamy interakcje, ponieważ spodziewamy się, iż wpływ zmiennej oznaczającej wykształcenie zależy od płci.

INTERAKCJE MIĘDZY ZMIENNYMI DYSKRETNYMI - WYKSZTAŁCENIE I PŁEĆ

dochod - zmienna zależna,
wiek, wiek_2 oraz interakcje miedzy wykształceniem i płcią - zmienne niezależne

```
xi: regress dochod wiek wiek_2  i.plec*i.wykształcenie
```

```
      Source |       SS       df       MS              Number of obs =    1083
-------------+------------------------------           F(  7,  1075) =   18.83
       Model |  81648217.6     7  11664031.1           Prob > F      =  0.0000
    Residual |   665832918  1075  619379.458           R-squared     =  0.1092
-------------+------------------------------           Adj R-squared =  0.1034
       Total |   747481135  1082  690832.842           Root MSE      =  787.01

------------------------------------------------------------------------------
      dochod |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        wiek |   36.38318   15.39846     2.36   0.018     6.168745    66.59762
      wiek_2 |  -.4049352   .1962222    -2.06   0.039    -.7899572   -.0199131
    _Iplec_1 |  -144.4044   143.4615    -1.01   0.314    -425.9008    137.0919
_Iwykształ~2 |   274.2703   105.1538     2.61   0.009     67.94046    480.6002
_Iwykształ~3 |   1040.998   137.1701     7.59   0.000     771.8461    1310.149
_IpleXwyk_~2 |  -143.4455   153.4394    -0.93   0.350    -444.5201    157.6292
_IpleXwyk_~3 |   -682.341   197.7395    -3.45   0.001     -1070.34   -294.3418
       _cons |  -121.1625   300.6773    -0.40   0.687    -711.1435    468.8184
------------------------------------------------------------------------------
```

Interakcje między zmiennymi dyskretnymi i ciągłymi

- Wprowadzenie do modelu interakcji pomiędzy zmiennymi dyskretnymi i ciągłymi ma sens, jeśli **wpływ pewnej zmiennej niezależnej ciągłej na zmienną zależną zależy od poziomów zmiennej dyskretnej**.

INTERAKCJE MIĘDZY ZMIENNĄ CIĄGŁA I DYSKRETNĄ - WIEK I MIEJSCE ZAMIESZKANIA

interakcje między zmienną miasto a wiekiem

```
xi: regress dochod i.miasto_1*wiek
```

Source	SS	df	MS		
Model	21268278.5	7	3038325.5	Number of obs	= 1083
Residual	726212857	1075	675546.843	F(7, 1075)	= 4.50
Total	747481135	1082	690832.842	Prob > F	= 0.0001
				R-squared	= 0.0285
				Adj R-squared	= 0.0221
				Root MSE	= 821.92

dochod	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
_Imiasto_1_2	28.34615	296.4254	0.10	0.924	-553.2919	609.9842
_Imiasto_1_3	53.41383	249.026	0.21	0.830	-435.2183	542.046
_Imiasto_1_4	135.6545	283.6069	0.48	0.633	-420.8315	692.1404
wiek	-.4870689	4.569159	-0.11	0.915	-9.452549	8.478412
_ImiaXwiek_2	**3.588019**	7.603214	0.47	0.637	-11.3308	18.50684
_ImiaXwiek_3	**5.698882**	6.355967	0.90	0.370	-6.772626	18.17039
_ImiaXwiek_4	**5.396286**	7.063888	0.76	0.445	-8.464285	19.25686
_cons	641.7219	175.9821	3.65	0.000	296.4145	987.0292

Plan wykładu

- 1. Kontrasty: efekty progowe, kontrasty w odchyleniach
- 2. Interakcje
- 3. Przybliżanie modeli nieliniowych:
  - Model wielomianowy

Modele wielomianowe

- Nieliniowa zależność między ***y*** a ***x*** można przybliżyć za pomocą modelu liniowego stosując model:
- **1. Model wielomianowy**

$$y_i = \beta_0 + x_i\beta_1 + x_i^2\beta_2 + \ldots + x_i^k\beta_K + \varepsilon_i$$

- Przy większej liczbie zmiennych objaśniających wstawia się do modelu ich kwadraty i iloczyny

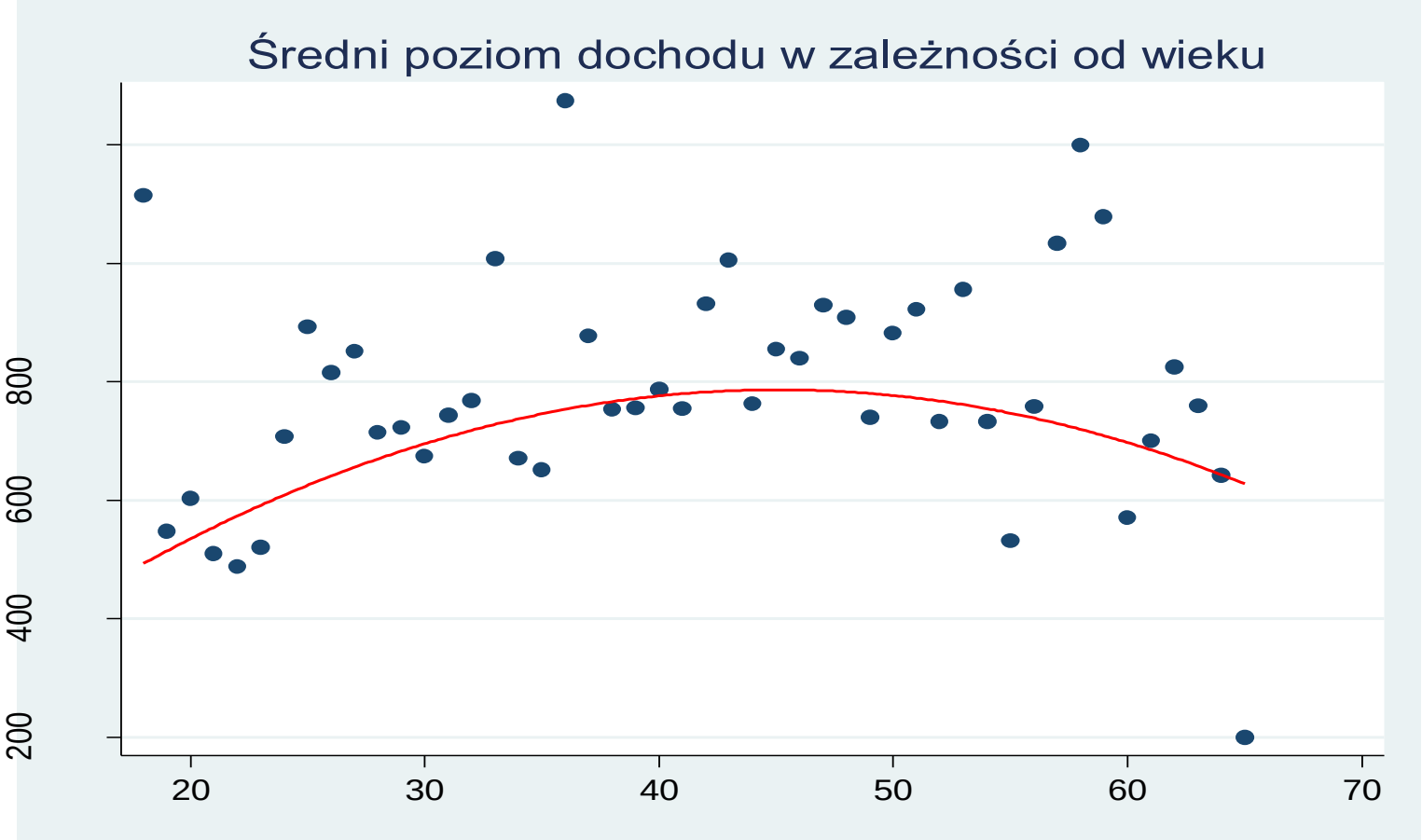
INNE FORMY FUNKCYJNE MODELU ZE WZGLĘDU NA WIEK - WIELOMIAN STOPNIA II

```
. regress dochod wiek wiek_2 plec srednie wyzsze

      Source |       SS       df       MS              Number of obs =    1083
-------------+------------------------------           F(  5,  1077) =   22.98
       Model |  72048793.8     5  14409758.8            Prob > F      =  0.0000
    Residual |   675432341  1077  627142.378            R-squared     =  0.0964
-------------+------------------------------           Adj R-squared =  0.0922
       Total |   747481135  1082  690832.842            Root MSE      =  791.92

------------------------------------------------------------------------------
      dochod |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        wiek |   36.06131   15.48328     2.33   0.020     5.680494    66.44212
      wiek_2 |  -.3998842   .1973767    -2.03   0.043    -.7871707   -.0125977
        plec |  -338.0671   48.25867    -7.01   0.000    -432.7588   -243.3755
     srednie |   208.5538   77.72619     2.68   0.007     56.04182    361.0657
      wyzsze |   708.2862   99.55596     7.11   0.000     512.9406    903.6318
       _cons |  -26.64989   298.3288    -0.09   0.929    -612.0215    558.7217
------------------------------------------------------------------------------
```

INNE FORMY FUNKCYJNE MODELU ZE WZGLĘDU NA WIEK - WIELOMIAN STOPNIA II



Dziękuję za uwagę