

Binarne zmienne zależne

cz. I

Stanisław Cichocki

Natalia Nehrebecka

Plan zajęć

1. Wstęp
 - a) Binarne zmienne zależne
 - b) Interpretacja ekonomiczna
 - c) Interpretacja współczynników

2. **Liniowy model prawdopodobieństwa**
 - a) Interpretacja współczynników

3. **Probit**
 - a) Interpretacja współczynników
 - b) Miary dopasowania
 - c) Diagnostyka

4. **Logit**
 - a) Interpretacja współczynników
 - b) Miary dopasowania
 - c) Diagnostyka

Plan zajęć

1. Wstęp
 - a) Binarne zmienne zależne
 - b) Interpretacja ekonomiczna
 - c) Interpretacja współczynników

2. Liniowy model prawdopodobieństwa
 - a) Interpretacja współczynników

3. Probit
 - a) Interpretacja współczynników
 - b) Miary dopasowania
 - c) Diagnostyka

4. Logit
 - a) Interpretacja współczynników
 - b) Miary dopasowania

Binarne zmienne zależne

- ▶ zmienne zero-jedynkowe nazywane są także **zmiennymi binarnymi**
- ▶ zmienna zero-jedynkowa opisuje stan w jakim znajdują się badane obiekty
 - przy kodowaniu zmiennej przyjmuje się konwencję:
 - 0 - porażka
 - 1 - sukces
- ▶ czym jest sukces a czym porażka zależy od badania

Binarne zmienne zależne

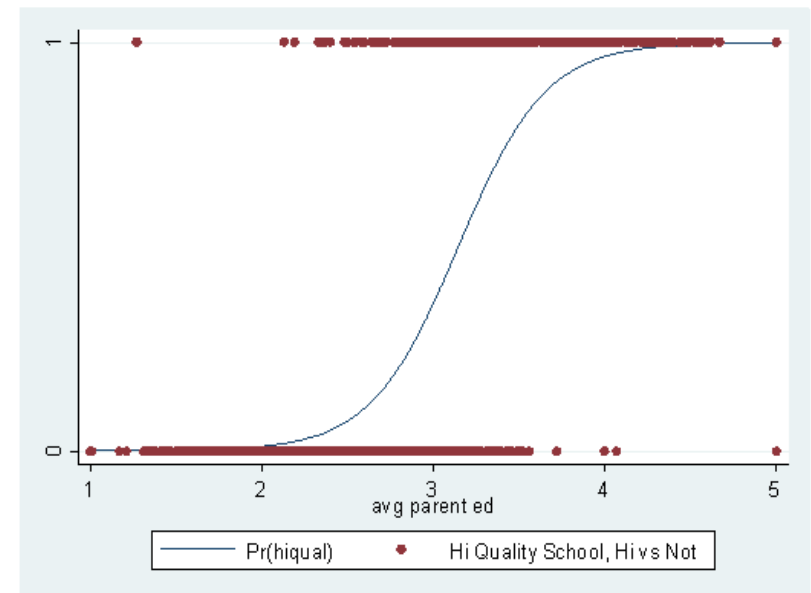
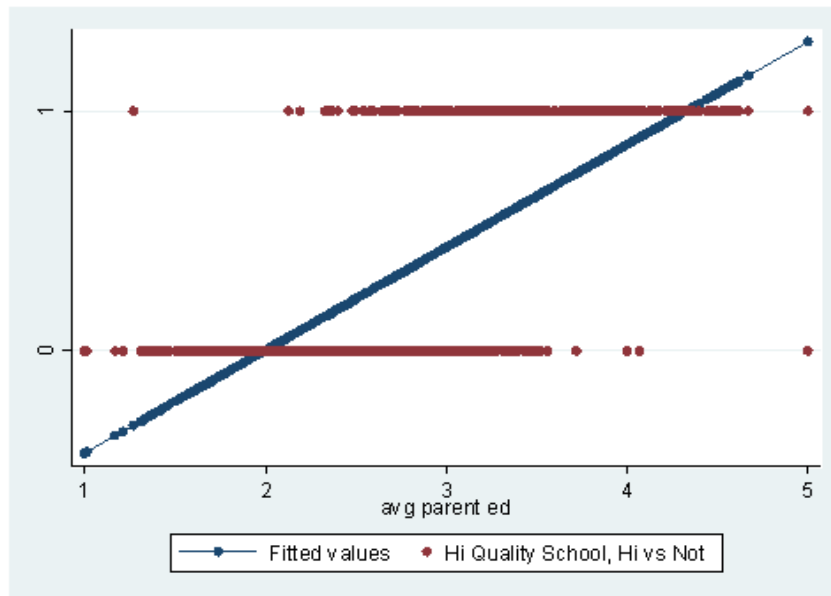
- ▶ w modelach dla zmiennych binarnych wyjaśniamy, za pomocą zmiennych niezależnych, prawdopodobieństwa stanów
- ▶ aby sformułować model musimy opisać **prawdopodobieństwa zdarzeń** dla poszczególnych obserwacji na największym poziomie ogólności:

$$\Pr(y_i | \mathbf{x}_i) = \begin{cases} 1 - p(\mathbf{x}_i) & \text{dla } y_i = 0 \\ p(\mathbf{x}_i) & \text{dla } y_i = 1 \end{cases}$$

- gdzie $p(\mathbf{x}_i)$ opisuje zależność prawdopodobieństwa sukcesu od wielkości zmiennych niezależnych x_i
 - W jaki sposób powinniśmy dobrać funkcję $p(\mathbf{x}_i)$, która opisuje zależność prawdopodobieństwa sukcesu od wielkości zmiennych niezależnych x_i ?
 - Oczywiście, funkcja ta powinna mieć wartości w przedziale $[0,1]$.

Binarne zmienne zależne

- modele dla zmiennych binarnych nie są modelami liniowymi



Interpretacja ekonomiczna

- ▶ y^*
 - jest użytecznością lub zyskiem netto
 - jest nieobserwowalne
 - jest losowe

$$y^* = x_i\beta + \varepsilon_i$$

gdzie ε ma symetryczną dystrybuantę z $E(\varepsilon) = 0$

- Parametry modelu szacujemy na podstawie obserwowalnych wyborów y , na które zdeterminowane są nieobserwowalnym y^* .

$$y_i = \begin{cases} 0 & \text{dla } y_i^* \leq 0 \\ 1 & \text{dla } y_i^* > 0 \end{cases}$$

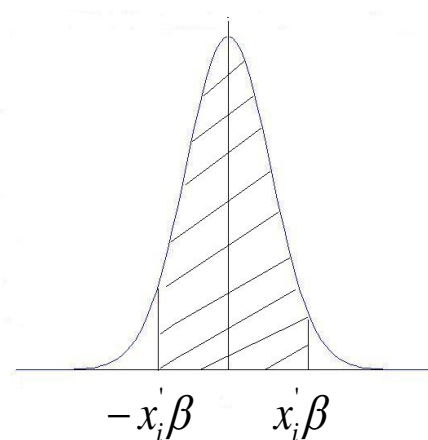
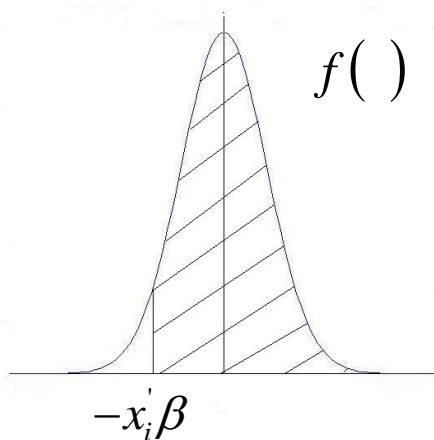
Interpretacja ekonomiczna

- ▶ Prawdopodobieństwo sukcesu:

$$\begin{aligned} Pr\{y_i = 1\} &= Pr\{y_i^* > 0\} = Pr\{x_i\beta + \varepsilon_i > 0\} = Pr\{\varepsilon_i > -x_i\beta\} = \\ &= Pr\{\varepsilon_i < x_i\beta\} = F(x_i\beta) \end{aligned}$$

- gdzie $F(\cdot)$ jest dystrybuantą ε

Np.



Interpretacja współczynników

- ▶ Wartość oczekiwana zmiennej zależnej w modelu dla zmiennej binarnej jest równa:

$$E(y_i | x_i) = 0 \cdot [1 - F(x_i\beta)] + 1 \cdot F(x_i\beta) = F(x_i\beta)$$

- ▶ Ponieważ przyjęta jest konwencja, że 1 oznacza sukces a 0 porażkę więc, wartość oczekiwana zmiennej zależnej w modelu dla zmiennej binarnej równa jest prawdopodobieństwu sukcesu.
- ▶ Efekt cząstkowy jest równy:

$$\frac{\partial E(y|x)}{\partial x_k} = \frac{\partial F(x_i\beta)}{\partial x_k} = f(x_i\beta)\beta_k$$

Interpretacja współczynników

- ▶ Efekt cząstkowy możemy interpretować jako **wpływ jednostkowej zmiany zmiennej niezależnej na wielkość prawdopodobieństwa sukcesu**.
- ▶ Wielkość efektu cząstkowego zależy od wielkości x_i dla którego jest on liczony.
 - Zazwyczaj efekty cząstkowe liczymy **dla średnich wartości zmiennych niezależnych**.
- ▶ W przypadku zmiennych zerojedynkowych efekt krańcowy wyznaczamy jako różnicę między prawdopodobieństwem sukcesu dla wartości 1 i 0.

Plan zajęć

1. Wstęp
 - a) Binarne zmienne zależne
 - b) Interpretacja ekonomiczna
 - c) Interpretacja współczynników

2. Liniowy model prawdopodobieństwa
 - a) Interpretacja współczynników

3. Probit
 - a) Interpretacja współczynników
 - b) Miary dopasowania
 - c) Diagnostyka

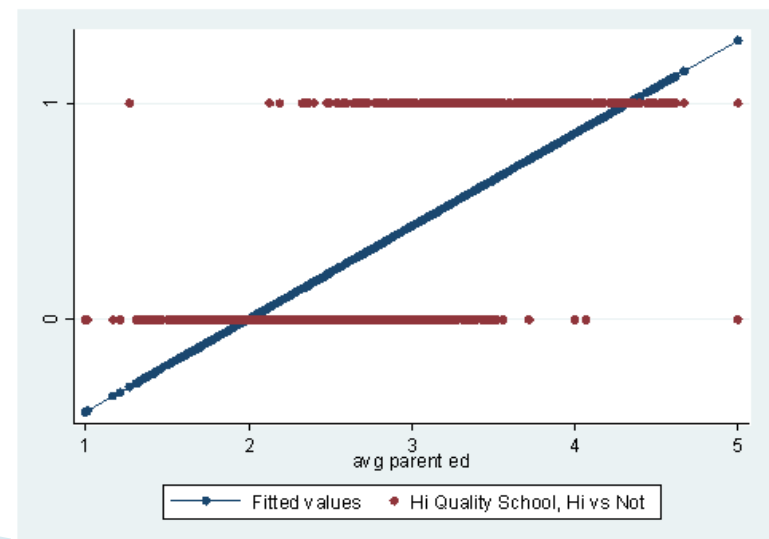
4. Logit
 - a) Interpretacja współczynników
 - b) Miary dopasowania

Liniowy model prawdopodobieństwa (LPM)

- ▶ rozpatrujemy różne formy funkcyjne dla zależności między prawdopodobieństwem sukcesu a wielkością zmiennych objaśniających
- ▶ najprostszy model \longrightarrow **liniowy model prawdopodobieństwa (LPM)**

$$p(x_i) = F(x_i\beta) = x_i\beta$$

- ▶ gdzie:
 - $p(x_i)$ - prawdopodobieństwo sukcesu
 - $F(\)$ - dystrybuanta



Liniowy model prawdopodobieństwa (LPM)

- ▶ liniowy w tym sensie, że warunkowa wartość oczekiwana jest dana funkcją liniową:

$$E(y_i | x_i) = x_i \beta$$

- ▶ Liniowy model prawdopodobieństwa sprowadza się do estymacji Metodą Najmniejszych Kwadratów równania:

$$y_i = x_i \beta + \varepsilon_i$$

Liniowy model prawdopodobieństwa (LPM)

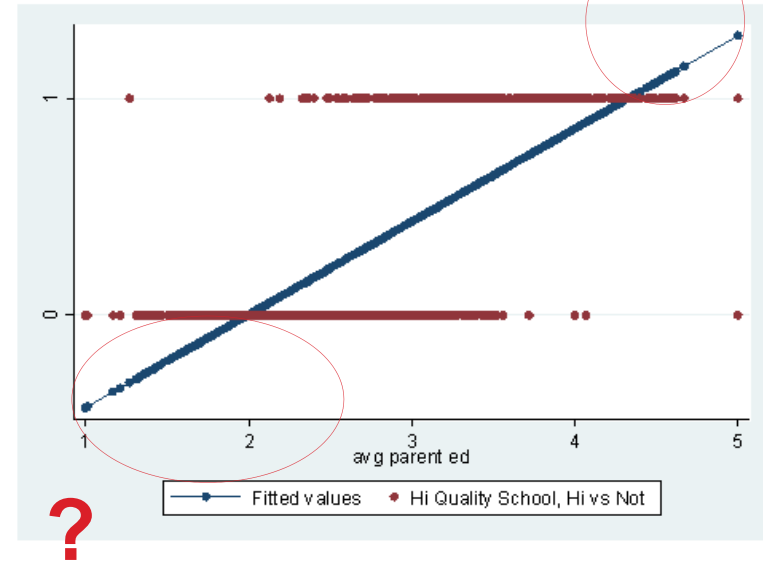
?

▶ zalety LPM:

- a) łatwy do wyestymowania
- b) współczynniki = efekty cząstkowe

▶ wady LPM:

- a) wartości dopasowane spoza $[0;1]$ **→** nieinterpretowalne (*prawdopodobieństwo sukcesu*)
- b) Błędy losowe w równaniu regresji będą heteroskedastyczne



Liniowy model prawdopodobieństwa (LPM)

- ▶ problemu z punktu a) nie da się w prosty sposób rozwiązać
- ▶ w przypadku heteroscedastyczności stosujemy *UMNK* lub odporne macierze wariancji i kowariancji

Liniowy model prawdopodobieństwa (LPM)

Source	SS	df	MS			
Model	69.3608203	18	3.85337891	Number of obs =	4877	
Residual	985.092737	4858	.202777426	F(18, 4858) =	19.00	
Total	1054.45356	4876	.216253806	Prob > F =	0.0000	
				R-squared =	0.0658	
				Adj R-squared =	0.0623	
				Root MSE =	.45031	

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
statur	.0181616	.0030859	5.89	0.000	.0121119	.0242113
statemb	.0012594	.0002039	6.18	0.000	.0008597	.001659
age	.004297	.0007798	5.51	0.000	.0027682	.0058258
tenure	.0054326	.0012124	4.48	0.000	.0030558	.0078095
slack	.1260833	.0142071	8.87	0.000	.098231	.1539356
abol	-.0082705	.0248301	-0.33	0.739	-.0569486	.0404077
seasonal	.0560999	.035809	1.57	0.117	-.014102	.1263017
nwhite	.0193939	.0186845	1.04	0.299	-.0172361	.056024
school12	-.0096339	.0167537	-0.58	0.565	-.0424787	.023211
male	-.0388263	.0177931	-2.18	0.029	-.0737088	-.0039437
smsa	-.0352754	.0140208	-2.52	0.012	-.0627624	-.0077883
married	.0500559	.0161317	3.10	0.002	.0184306	.0816813
dkids	-.0187618	.0167546	-1.12	0.263	-.0516085	.0140848
dykids	.0358303	.0195493	1.83	0.067	-.0024952	.0741557
yrdispl	-.0130974	.0030688	-4.27	0.000	-.0191138	-.0070811
rr	.6062438	.3842868	1.58	0.115	-.1471322	1.35962
rr2	-1.010374	.4811835	-2.10	0.036	-1.953711	-.0670365
head	-.0386771	.0165195	-2.34	0.019	-.0710629	-.0062914
_cons	.1279247	.0882052	1.45	0.147	-.0449975	.3008469

Liniowy model prawdopodobieństwa (LPM)

a) Test White'a na heteroscedastyczność

```
imtest, white
```

```
White's test for Ho: homoskedasticity
```

```
against Ha: unrestricted heteroskedasticity
```

```
chi2(173) = 458.73
```

```
Prob > chi2 = 0.0000
```

b) Test Breucha-Pagana:

```
hettest, rhs
```

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
```

```
Ho: Constant variance
```

```
Variables: stateur statemb age tenure slack abol seasonal nwhite
```

```
school12 male smsa married dkids dykids yrdispl rr rr2 head
```

```
chi2(18) = 84.68
```

```
Prob > chi2 = 0.0000
```

Liniowy model prawdopodobieństwa (LPM)

Linear regression

Number of obs = 4877
 F(18, 4858) = 20.77
 Prob > F = 0.0000
 R-squared = 0.0658
 Root MSE = .45031

y	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
stateur	.0181616	.0029548	6.15	0.000	.0123688	.0239544
statemb	.0012594	.0002038	6.18	0.000	.0008599	.0016588
age	.004297	.0007668	5.60	0.000	.0027936	.0058004
tenure	.0054326	.0011656	4.66	0.000	.0031476	.0077176
slack	.1260833	.0142299	8.86	0.000	.0981863	.1539802
abol	-.0082705	.0259736	-0.32	0.750	-.0591906	.0426496
seasonal	.0560999	.0373422	1.50	0.133	-.0171078	.1293075
nwhite	.0193939	.0184205	1.05	0.292	-.0167186	.0555064
school12	-.0096339	.016992	-0.57	0.571	-.0429459	.0236782
male	-.0388263	.0178936	-2.17	0.030	-.0739059	-.0037466
smsa	-.0352754	.0139739	-2.52	0.012	-.0626705	-.0078803
married	.0500559	.0163331	3.06	0.002	.0180358	.0820761
dkids	-.0187618	.0168417	-1.11	0.265	-.0517793	.0142556
dykids	.0358303	.0196643	1.82	0.069	-.0027207	.0743813
yrdispl	-.0130974	.0030852	-4.25	0.000	-.0191459	-.0070489
rr	.6062438	.3957315	1.53	0.126	-.169569	1.382057
rr2	-1.010374	.4951863	-2.04	0.041	-1.981163	-.0395846
head	-.0386771	.0166618	-2.32	0.020	-.0713418	-.0060125
_cons	.1279247	.0882096	1.45	0.147	-.0450061	.3008555

Liniowy model prawdopodobieństwa (LPM)

```
test abol seasonal nwhite school12 dkids dykids rr
```

```
( 1)  abol = 0  
( 2)  seasonal = 0  
( 3)  nwhite = 0  
( 4)  school12 = 0  
( 5)  dkids = 0  
( 6)  dykids = 0  
( 7)  rr = 0
```

```
F( 7, 4858) = 1.50  
Prob > F = 0.1621
```

Liniowy model prawdopodobienstwa (LPM)

Linear regression

Number of obs = 4877
 F(11, 4865) = 32.57
 Prob > F = 0.0000
 R-squared = 0.0637
 Root MSE = .45048

	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
stateur	.0181462	.0029481	6.16	0.000	.0123665	.0239259
statemb	.001258	.0002022	6.22	0.000	.0008617	.0016543
age	.0039118	.0007047	5.55	0.000	.0025304	.0052933
tenure	.0055123	.0011489	4.80	0.000	.0032601	.0077646
slack	.1247634	.0131511	9.49	0.000	.0989813	.1505456
male	-.0395063	.0176193	-2.24	0.025	-.0740481	-.0049644
smsa	-.0344193	.0139264	-2.47	0.013	-.0617214	-.0071172
married	.0500746	.0142392	3.52	0.000	.0221593	.07799
yrdispl	-.0128674	.0030768	-4.18	0.000	-.0188994	-.0068355
rr2	-.2527459	.0860807	-2.94	0.003	-.421503	-.0839887
head	-.0376671	.0161592	-2.33	0.020	-.0693465	-.0059878
_cons	.2532724	.048884	5.18	0.000	.1574377	.349107

Liniowy model prawdopodobieństwa (LPM)

- Wada LMP - wartości dopasowane nie znajdują się w przedziale [0,1]

```
predict wartosci_dop, xb
```

```
list y wartosci_dop if wartosci_dop > 1 | wartosci_dop < 0
```

```
+-----+
| y  wartosci_dop |
+-----+
 48. | 1    1.03823  |
159. | 1    1.202082 |
595. | 1    1.077022 |
842. | 1    1.037514 |
1194. | 1    1.097218 |
+-----+
1487. | 1    1.014822 |
1738. | 1     1.0229  |
1789. | 1    1.060014 |
1936. | 1    1.090258 |
1940. | 1    1.054053 |
+-----+
1970. | 1    1.092681 |
2166. | 1     1.05898  |
2537. | 1    1.100604 |
2598. | 1    1.082886 |
2607. | 1     1.01655  |
+-----+
2620. | 1    1.003613 |
2623. | 1    1.062682 |
2710. | 1    1.028138 |
2964. | 1    1.048522 |
3260. | 1    1.011603 |
+-----+
3916. | 1    1.065829 |
4150. | 1    1.035251 |
4203. | 1    1.015945 |
4312. | 1    1.026004 |
4763. | 1    1.019101 |
+-----+
```

LPM - Interpretacja współczynników

- ▶ w liniowym modelu prawdopodobieństwa interpretuje się współczynniki
- ▶ interpretacja:
 - a) **dla zmiennych objaśniających ciągłych:** wpływ jednostkowej zmiany zmiennej niezależnej na wielkość prawdopodobieństwa sukcesu
 - b) **dla zmiennych objaśniających zero-jedynkowych:** różnica między prawdopodobieństwem sukcesu dla zmiennej zero-jedynkowej równej 0 i równej 1

Liniowy model prawdopodobieństwa (LPM)

Linear regression

Number of obs = 4877
 F(11, 4865) = 32.57
 Prob > F = 0.0000
 R-squared = 0.0637
 Root MSE = .45048

	y	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
stateur		.0181462	.0029481	6.16	0.000	.0123665	.0239259
statemb		.001258	.0002022	6.22	0.000	.0008617	.0016543
age		.0039118	.0007047	5.55	0.000	.0025304	.0052933
tenure		.0055123	.0011489	4.80	0.000	.0032601	.0077646
slack		.1247634	.0131511	9.49	0.000	.0989813	.1505456
male		-.0395063	.0176193	-2.24	0.025	-.0740481	-.0049644
smsa		-.0344193	.0139264	-2.47	0.013	-.0617214	-.0071172
married		.0500746	.0142392	3.52	0.000	.0221593	.07799
yrdispl		-.0128674	.0030768	-4.18	0.000	-.0188994	-.0068355
rr2		-.2527459	.0860807	-2.94	0.003	-.421503	-.0839887
head		-.0376671	.0161592	-2.33	0.020	-.0693465	-.0059878
_cons		.2532724	.048884	5.18	0.000	.1574377	.349107

Pytania teoretyczne

1. Co jest modelowane w przypadku jeśli zmienna zależna jest zmienna binarną (zero-jedynkową)? Wyjaśnić, jaka jest relacja między zmienną obserwowalną a ukrytą w przypadku modeli dla zmiennych binarnych i jak tę relację można uzasadnić na bazie teorii ekonomii.
2. Jakie są wady liniowego modelu prawdopodobieństwa? Odpowiedz uzasadnić.

Dziękuję za uwagę