

# **Binarne zmienne zależne**

## **cz. IV**

**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

# Logit

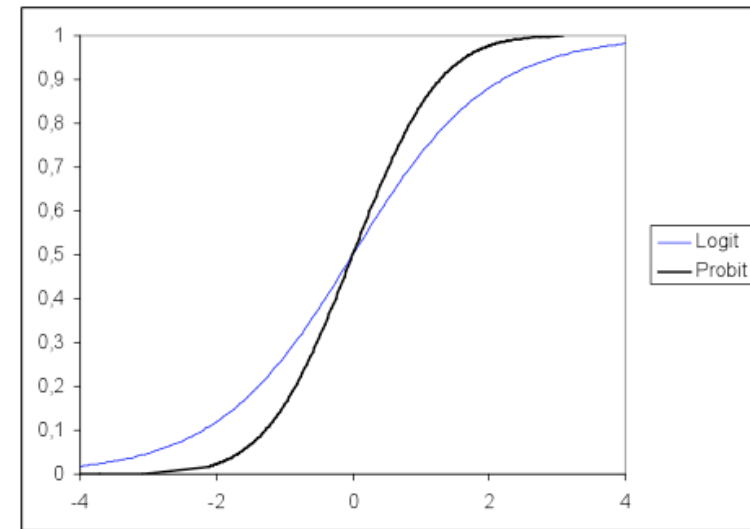
- ▶ W **modelu logitowym** zakładamy, że  $F()$  jest dystrybuantą rozkładu logistycznego
- ▶ Założenia modelu logitowego:
  - Obserwacje są niezależne
  - Rozkład warunkowy:

$$Pr(y_i|x_i) = \begin{cases} \Lambda(x_i\beta) & \text{dla } y_i = 1 \\ 1 - \Lambda(x_i\beta) & \text{dla } y_i = 0 \end{cases}$$

- gdzie:  $\Lambda(x_i\beta) = \frac{e^{x_i\beta}}{1+e^{x_i\beta}}$

- ▶ Logarytm funkcji wiarygodności:

$$l(\beta) = \sum_{i=1}^n \left[ (1 - y_i) \ln \left( \frac{1}{1+e^{x_i\beta}} \right) + y_i \ln \left( \frac{e^{x_i\beta}}{1+e^{x_i\beta}} \right) \right]$$



# Logit

- ▶ Wartość oczekiwana:

$$E(y_i|x_i) = 1 \cdot \Lambda(x_i\beta) + 0 \cdot (1 - \Lambda(x_i\beta)) = \Lambda(x_i\beta)$$

- ▶ Efekt cząstkowy dla zmiennej  $x_k$ :

$$\frac{\partial E(y|x)}{\partial x_k} = \frac{\partial \Lambda(x_i\beta)}{\partial x_k} = \Lambda'(x_i\beta)\beta_k = \Lambda(x_i\beta)(1 - \Lambda(x_i\beta))\beta_k$$

# Logit - Interpretacja współczynników

- ▶ nie interpretuje się współczynników w modelu logitowym
- ▶ interpretuje się efekty cząstkowe (*krańcowe*):
  - a) dla zmiennych objaśniających ciągłych:
    - wpływ jednostkowej zmiany zmiennej niezależnej na wielkość prawdopodobieństwa sukcesu;
      - efekty cząstkowe dla zmiennych objaśniających ciągłych liczymy zwykle dla średnich wartości tych zmiennych (efekty cząstkowe zależą od wielkości zmiennych objaśniających)
  - b) dla zmiennych objaśniających zero-jedynkowych:
    - różnica między prawdopodobieństwem sukcesu dla zmiennej zero-jedynkowej równej 0 i równej 1, przy pozostałych zmiennych ustalonych na poziomie średnich

# Logit - Interpretacja współczynników

- ▶ znak efektu cząstkowego dla danej zmiennej jest taki sam jak znak współczynnika przy tej zmiennej
- ▶ możemy zatem interpretować znaki przy współczynnikach:
  - dodatni znak  $\Rightarrow$  zmienna wpływa dodatnio na prawdopodobieństwo sukcesu
  - ujemny znak  $\Rightarrow$  zmienna wpływa ujemnie na prawdopodobieństwo sukcesu

# Logit

Logistic regression

Number of obs = 4877

LR chi2(11) = 324.97

Prob > chi2 = 0.0000

Pseudo R2 = 0.0534

Log likelihood = -2880.5412

y	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
<b>stateur</b>	<b>.0953049</b>	<b>.0158821</b>	<b>6.00</b>	<b>0.000</b>	<b>.0641765</b>	<b>.1264332</b>
statemb	.0061076	.0009994	6.11	0.000	.0041489	.0080663
age	.0195309	.0036723	5.32	0.000	.0123333	.0267286
tenure	.0308234	.0065936	4.67	0.000	.0179001	.0437467
slack	.6091688	.0659404	9.24	0.000	.479928	.7384096
<b>male</b>	<b>-.1876878</b>	<b>.0860323</b>	<b>-2.18</b>	<b>0.029</b>	<b>-.356308</b>	<b>-.0190676</b>
smsa	-.1664986	.0693336	-2.40	0.016	-.30239	-.0306072
married	.2403562	.0687984	3.49	0.000	.1055137	.3751986
yrdispl	-.0615481	.0149413	-4.12	0.000	-.0908324	-.0322638
rr2	-1.151087	.4204778	-2.74	0.006	-1.975208	-.3269655
head	-.1879037	.0781542	-2.40	0.016	-.3410832	-.0347242
_cons	-1.373795	.255853	-5.37	0.000	-1.875258	-.8723321

# Logit

Marginal effects after logit

$$y = \Pr(y) \text{ (predict)}$$

$$= .6970697$$

variable	dy/dx	Std. Err.	z	P> z	[ 95% C.I. ]	X
<b>stateur</b>	<b>.0201249</b>	<b>.00334</b>	<b>6.02</b>	<b>0.000</b>	<b>.013575 .026675</b>	<b>7.51103</b>
statemb	.0012897	.00021	6.12	0.000	.000877 .001702	180.66
age	.0041242	.00077	5.33	0.000	.002608 .005641	36.13
tenure	.0065088	.00139	4.69	0.000	.003788 .00923	5.66414
slack*	.1273793	.01353	9.42	0.000	.10087 .153889	.476112
<b>male*</b>	<b>-.0388319</b>	<b>.01742</b>	<b>-2.23</b>	<b>0.026</b>	<b>-.072971 -.004693</b>	<b>.764199</b>
smsa*	-.034793	.01433	-2.43	0.015	-.062873 -.006713	.652655
married*	.0513508	.01485	3.46	0.001	.022251 .08045	.632766
yrdispl	-.0129967	.00315	-4.12	0.000	-.019175 -.006818	5.20361
rr2	-.2430676	.08878	-2.74	0.006	-.417064 -.069071	.20344
head*	-.0391275	.01603	-2.44	0.015	-.070547 -.007708	.680541

(\*) dy/dx is for discrete change of dummy variable from 0 to 1



# Iloraz szans

- ▶ **Szansa** zdefiniowana jako prawdopodobieństwo wystąpienia zdarzenia (*sukcesu*) w odniesieniu do zdarzenia przeciwnego (*porażki*)

$$Odds(x_i) = \frac{\Pr(y_i = 1|x_i)}{\Pr(y_i = 0|x_i)} = \frac{e^{x_i\beta}}{1 + e^{x_i\beta}} = e^{x_i\beta}$$

- ▶ **Iloraz szans** mówi ile więcej prawdopodobne jest (w odniesieniu do szansy), że określone zdarzenie wystąpi w jednej grupie w odniesieniu do tego samego zdarzenia w innej grupie

$$\frac{Odds(x_i)}{Odds(x_j)} = e^{(x_i - x_j)\beta} = e^{\Delta x\beta}$$

# Logit – Iloraz szans

Logistic regression

Number of obs = 4877  
 LR chi2(11) = 324.97  
 Prob > chi2 = 0.0000  
 Pseudo R2 = 0.0534

Log likelihood = -2880.5412

*exp(.0953049)*

	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
stateur	1.099994	.0174702	6.00	0.000	1.066281 1.134774
statemb	1.006126	.0010055	6.11	0.000	1.004157 1.008099
<b>age</b>	<b>1.019723</b>	<b>.0037448</b>	<b>5.32</b>	<b>0.000</b>	<b>1.01241 1.027089</b>
tenure	1.031303	.0068	4.67	0.000	1.018061 1.044718
slack	1.838902	.1212579	9.24	0.000	1.615958 2.092605
male	.8288735	.0713099	-2.18	0.029	.7002569 .9811131
smsa	.846624	.0586995	-2.40	0.016	.7390498 .9698565
married	1.271702	.0874911	3.49	0.000	1.111281 1.45528
yrdispl	.9403077	.0140494	-4.12	0.000	.9131708 .9682512
rr2	.3162928	.1329941	-2.74	0.006	.1387324 .7211086
<b>head</b>	<b>.8286945</b>	<b>.064766</b>	<b>-2.40</b>	<b>0.016</b>	<b>.7109997 .9658717</b>

# Różnice między probitem a logitem

- ▶ Różnica związana jest z przyjętą formą funkcyjną dystrybuanty  $F()$ .
- ▶ Interpretacja efektów cząstkowych jest identyczna.

## ▶ logit

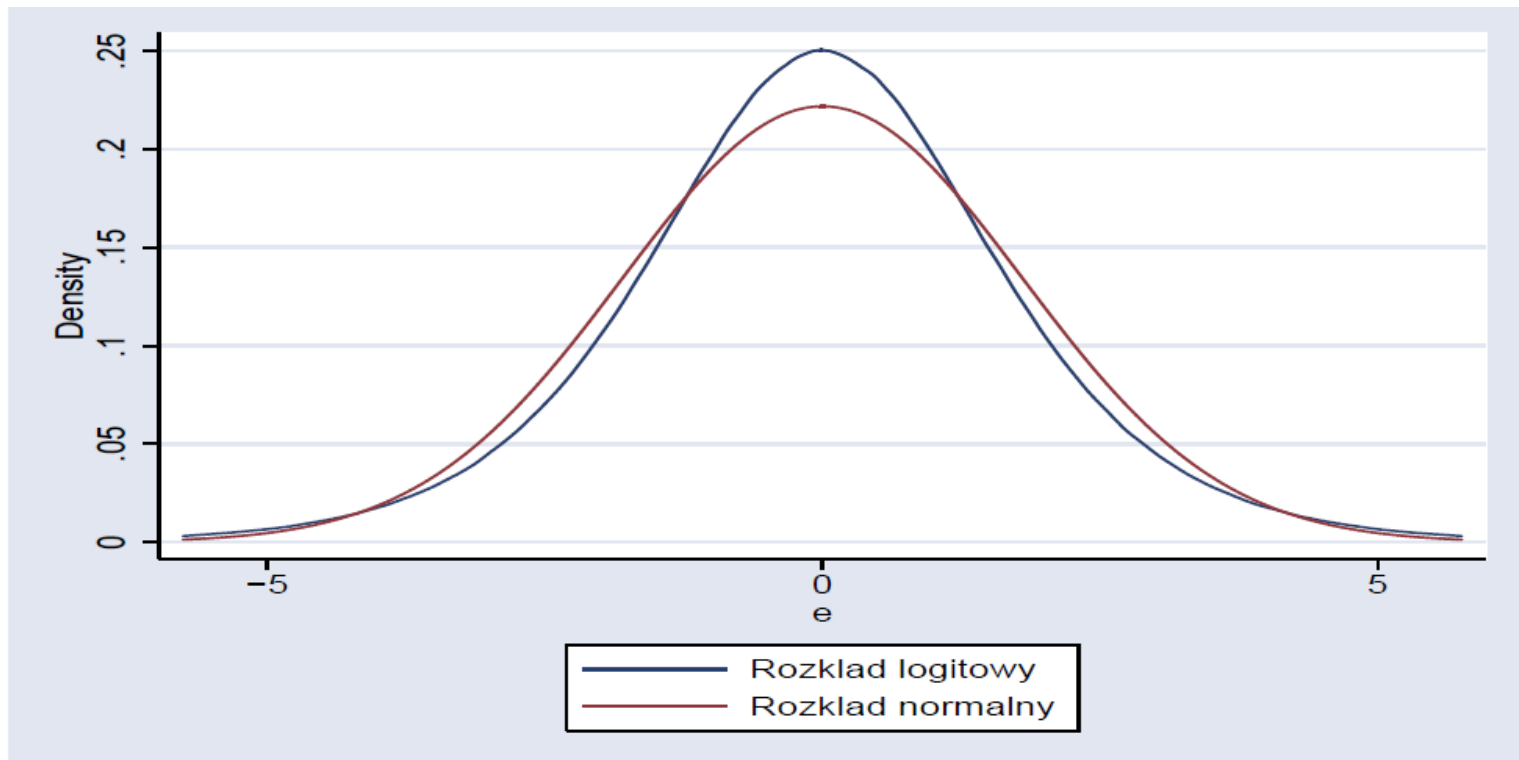
variable	dy/dx
stateur	.0201249
statemb	.0012897
age	.0041242
tenure	.0065088
slack*	.1273793
male*	-.0388319
smsa*	-.034793
married*	.0513508
yrdispl	-.0129967
rr2	-.2430676
head*	-.0391275

## ▶ probit

variable	dy/dx
stateur	.019886
statemb	.0012923
age	.0041137
tenure	.0060708
slack*	.1273458
male*	-.0393392
smsa*	-.033908
married*	.0510228
yrdispl	-.0130458
rr2	-.246271
head*	-.0382403

# Różnice między probitem a logitem

- ▶ Oba rozkłady prawdopodobieństwa są symetryczne jednak rozkład logistyczny ma nieco grubsze ogony.



# Różnice między probitem a logitem

- ▶ W związku z tym istotne różnice między modelami będą powstawać dla prób, o nikłym odsetku odpowiedzi 0 albo odpowiedzi 1 i bardzo zróżnicowanych zmiennych niezależnych.
- ▶ Dla  $\bar{x}\beta$  bliskiego 0 funkcja gęstości :
  - $f_{probit}(0) = \frac{1}{\sqrt{2\pi}} \approx 0.4$
  - $f_{logit}(0) \approx 0.25$
  - $f_{LPM}(0) = 1$
  - Przybliżona relacja między współczynnikami *logitu* i *probitu* będzie w przybliżeniu równa

$$f_{probit}(0)\beta_{probit} \approx f_{logit}(0)\beta_{logit}$$

$$\frac{\beta_{logit,i}}{\beta_{probit,i}} \approx \frac{0,4}{0,25} = 1,6$$

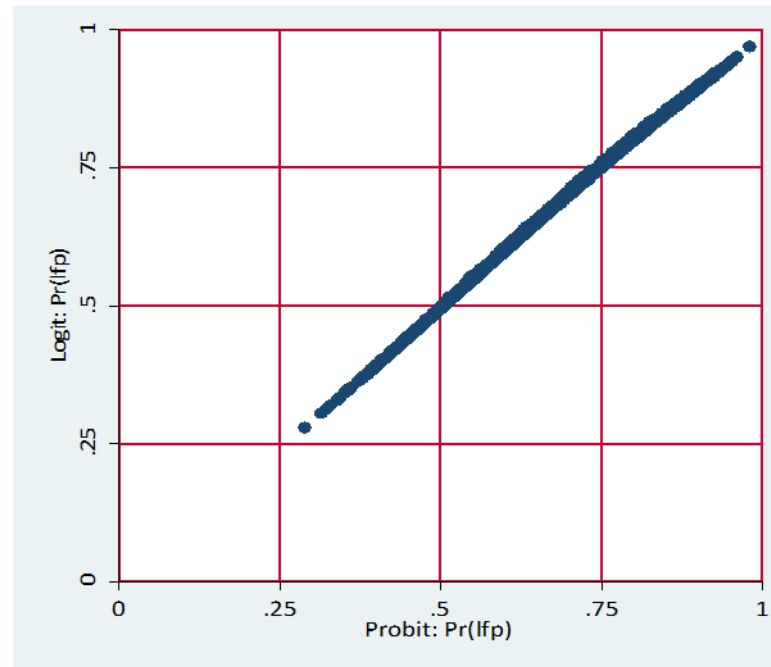
# Różnice między probitem a logitem

- ▶ Nie ma dobrej statystyki, która mogłaby posłużyć do wyboru między tymi modelami.
  - W praktyce wybieramy ten, który jest analitycznie bardziej wygodny.
  - Kierujemy się także jakością dopasowania oraz wynikami testów diagnostycznych.

# Różnice między probitem a logitem

```
pwcorr prlogit prprobit
```

	prlogit	prprobit
prlogit	1.0000	
prprobit	0.9996	1.0000



# Pytania teoretyczne

1. Co to są ilorazy szans i dlaczego w kontekście modelu logitowego lepiej jest używać ilorazów szans niż efektów krańcowych.
2. Na czym polega różnica między LPM, logitem i probitem? Jakich statystyk można użyć by sprawdzić, który z tych modeli jest prawidłowy?



**Dziękuję za uwagę**