Binarne zmienne zależne cz. III

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

Miary dopasowania

- tablica klasyfikacyjna:

		Zaobserwowane		
Prognozowane		0	1	**Razem**
	0	n_{00}	n_{01}	$n_{00} + n_{01}$
	1	n_{10}	n_{11}	$n_{10} + n_{11}$
	Razem	$n_{00} + n_{10}$	$n_{01} + n_{11}$	n

Miary dopasowania

- tablica klasyfikacyjna może być użyta do zdefiniowania jeszcze 2 miar:

a) **wrażliwość** – prawdopodobieństwo przewidzenia sukcesu dla obserwacji, dla której zaobserwowano sukces

$$wrazliwość = \Pr(\hat{p}_i \geq p^* | y_i = 1) \approx \frac{n_{11}}{n_{01}+n_{11}}$$

Miary dopasowania

b) **specyficzność** – prawdopodobieństwo przewidzenia porażki dla obserwacji, dla której zaobserwowano porażkę

$$specyficzność = \Pr(\hat{p}_i < p^* | y_i = 0) \approx \frac{n_{00}}{n_{00} + n_{10}}$$

Miary dopasowania

Probit model for y

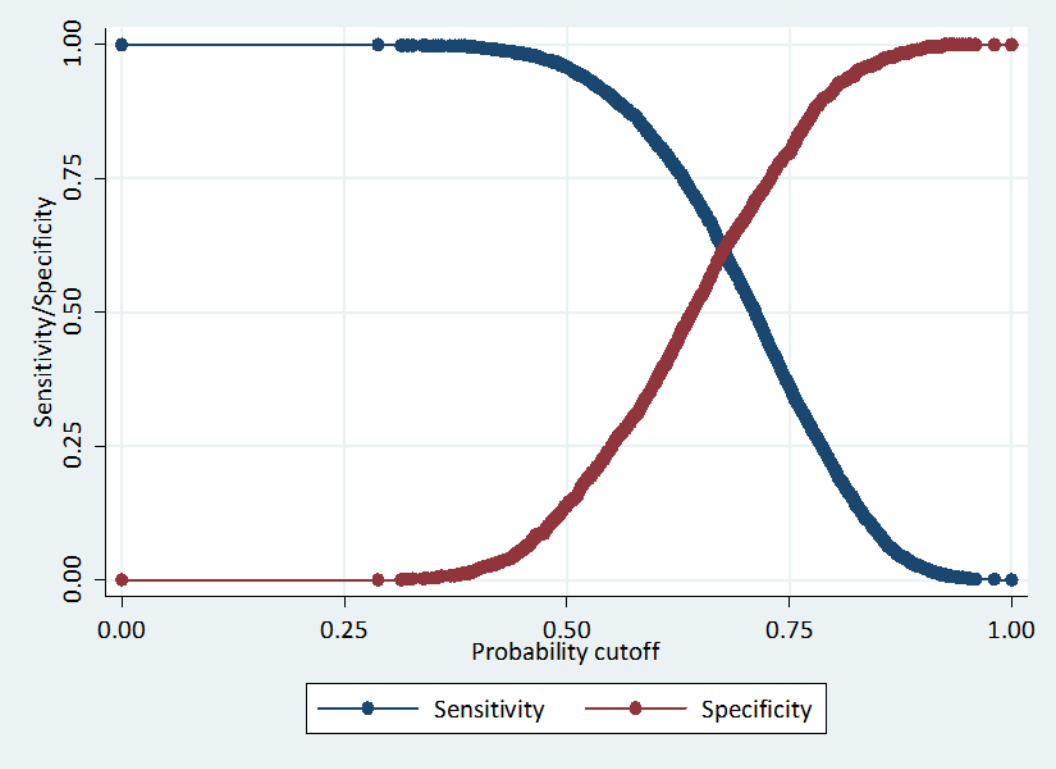
Classified	True D	True ~D	Total
+	3197	1330	4527
-	138	212	350
Total	3335	1542	4877

Classified + if predicted Pr(D) >= .5
True D defined as y != 0

Sensitivity	**Pr(+| D)**	**95.86%**
Specificity	**Pr(-|~D)**	**13.75%**
Positive predictive value	Pr(D| +)	70.62%
Negative predictive value	Pr(~D| -)	60.57%
False + rate for true ~D	Pr(+|~D)	86.25%
False - rate for true D	Pr(-| D)	4.14%
False + rate for classified +	Pr(~D| +)	29.38%
False - rate for classified -	Pr(D| -)	39.43%
Correctly classified		**69.90%**

Miary dopasowania

- wrażliwość i specyficzność



Miary dopasowania

Probit model for y

Classified	True D	True ~D	Total
+	2015	577	2592
-	1320	965	2285
Total	3335	1542	4877

Classified + if predicted Pr(D) >= .68
True D defined as y != 0

Sensitivity	**Pr(+| D)**	**60.42%**
Specificity	**Pr(-|~D)**	**62.58%**
Positive predictive value	Pr(D| +)	77.74%
Negative predictive value	Pr(~D| -)	42.23%
False + rate for true ~D	Pr(+|~D)	37.42%
False - rate for true D	Pr(-| D)	39.58%
False + rate for classified +	Pr(~D| +)	22.26%
False - rate for classified -	Pr(D| -)	57.77%
Correctly classified		**61.10%**

Miary dopasowania

- dodatkowo:

c) **1- wrażliwość** – prawdopodobieństwo przewidzenia porażki dla obserwacji, dla której **zaobserwowano sukces**

$$1 - wrazliwość = \frac{n_{01}}{n_{01} + n_{11}}$$

Miary dopasowania

d) **1- specyficzność** – prawdopodobieństwo przewidzenia sukcesu dla obserwacji, dla której **zaobserwowano porażkę**

$$1 - specyficzność = \frac{n_{10}}{n_{00}+n_{10}}$$

ROC (Receiver operating characteristic)

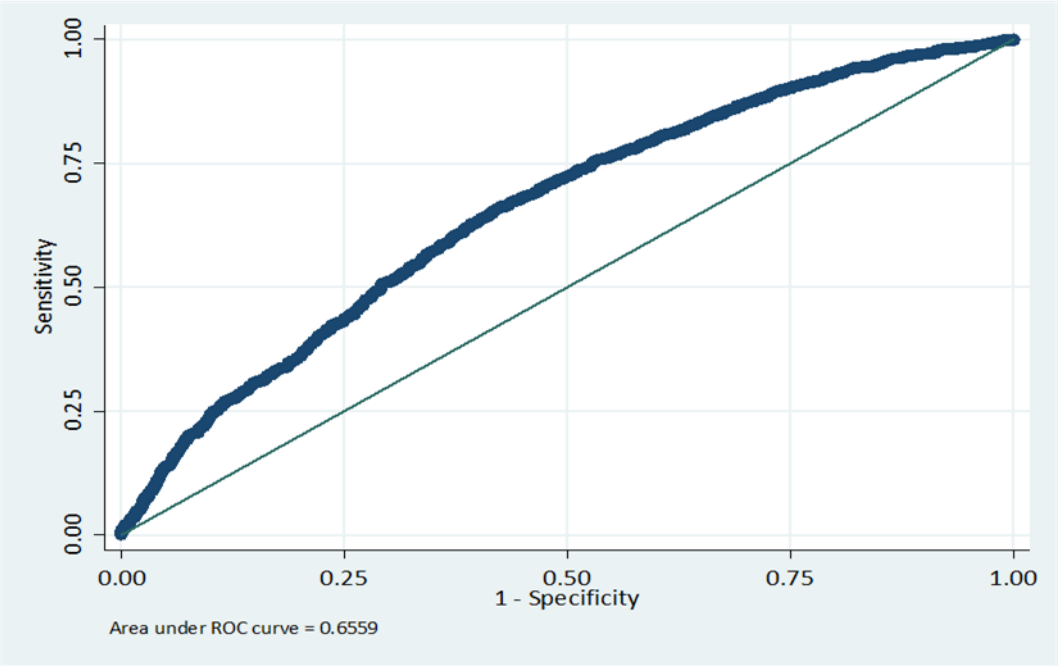
- Krzywa ROC obrazuje zależność pomiędzy wrażliwością ($TR - True\ Positives$)

$$\boldsymbol{TR} = wrazliwość = \Pr(\hat{p}_i \geq p^* | y_i = 1)$$

- i prawdopodobieństwem uzyskania fałszywych przewidywanych sukcesów ($FP - False\ Positives$)

$$\boldsymbol{FP} = 1 - specyficzność = \Pr(\hat{p}_i \geq p^* | y_i = 0)$$

ROC (*Receiver operating characteristic*)



- Im lepiej nasz model przewiduje, tym bardziej krzywa ROC odgięta jest w kierunku górnego rogu rysunku.
- Pole pod krzywą używane jest jako miara jakości dopasowania modelu.
  - $AUROC = 0.6559\ (\max 1)$

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

Diagnostyka

- **Testy na poprawność formy funkcyjnej**
  - W przypadku modeli z binarną zmienną objaśnianą zastosowanie znajduje $linktest$, który jest uogólnieniem i jednocześnie słabszą wersją testu $RESET$.
  - Test ten polega na przeprowadzeniu modelu probitowego y_i na stałą, $\hat{y}_i^*$ oraz $(\hat{y}_i^*)^2$.
  - Istotny współczynnik przy $(\hat{y}_i^*)^2$ powoduje odrzucenie hipotezy o poprawności formy funkcyjnej modelu.

Diagnostyka

- **Testy na poprawność formy funkcyjnej**

```
Iteration 0:   log likelihood =  -3043.028
Iteration 1:   log likelihood = -2879.6536
Iteration 2:   log likelihood = -2879.0342
Iteration 3:   log likelihood = -2879.0339
Iteration 4:   log likelihood = -2879.0339

Probit regression                                 Number of obs   =       4877
                                                  LR chi2(2)      =     327.99
                                                  Prob > chi2     =     0.0000
Log likelihood = -2879.0339                       Pseudo R2       =     0.0539

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        _hat |    1.28044   .1331964     9.61   0.000      1.01938      1.5415
      _hatsq |  -.2822587   .1200489    -2.35   0.019    -.5175501   -.0469672
       _cons |  -.0374954   .0365824    -1.02   0.305    -.1091955    .0342047
------------------------------------------------------------------------------
```

Diagnostyka

- **Test jakości dopasowania (*goodness of fit test*)**
  - *Test jest oparty na spostrzeżeniu, że jeżeli forma funkcyjna jest prawidłowa, to niezależnie od sposobu podziału próby na podpróbki, oszacowania stałych w modelu dla poszczególnych podpróbek nie powinny się istotnie różnić.*
  - Odrzucenie hipotezy zerowej o równości stałej w podpróbkach prowadzi do wniosku o niepoprawnej formie funkcyjnej modelu.
    - **Test Pearsona**
    - Podpróbki zdefiniowane jako wszystkie możliwe kombinacje zmiennych niezależnych. Powinno się go używać, kiedy takich grup (*covariate patterns*) jest znacząco mniej niż badanych obserwacji.
    - **Test Hosmera-Lemenshowa** stosuje się, kiedy liczba *covariate patterns* jest duża. Dzieli on obserwacje na grupy według kwantyli prawdopodobieństwa sukcesu przewidzianego przez model.

Diagnostyka

- **Wersja Pearsona**
- H_0: poprawna forma funkcyjna

```
Probit model for y, goodness-of-fit test

      number of observations =      4877
number of covariate patterns =      4871
          Pearson chi2(4859) =      4902.61
                 Prob > chi2 =         0.3271
```

- **Wersja Hosmera-Lemeshowa**

```
Probit model for y, goodness-of-fit test

          number of observations =      4877
              number of groups =        10
     Hosmer-Lemeshow chi2(8) =        21.39
                     Prob > chi2 =         0.0062
```

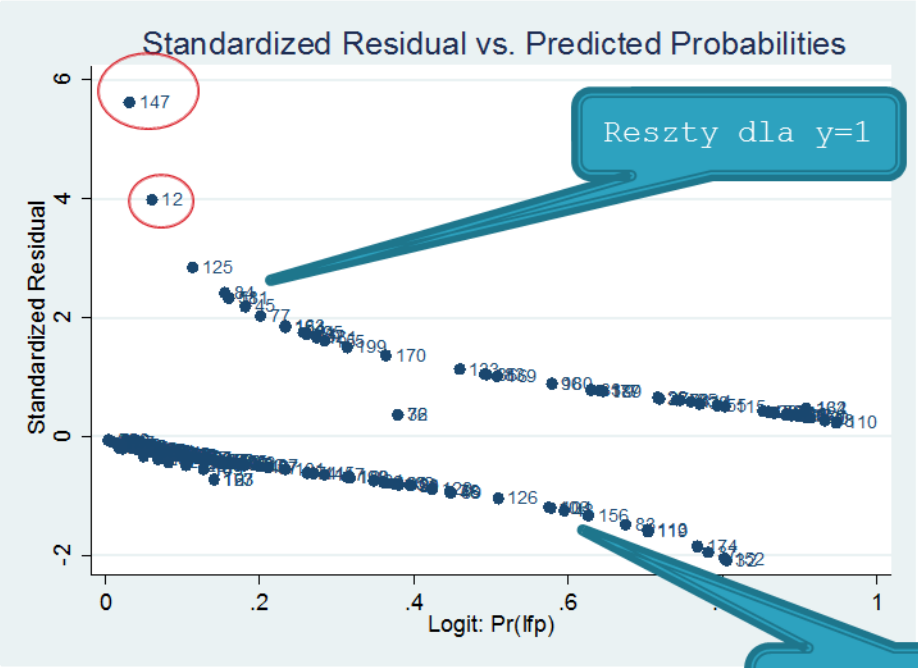
Diagnostyka

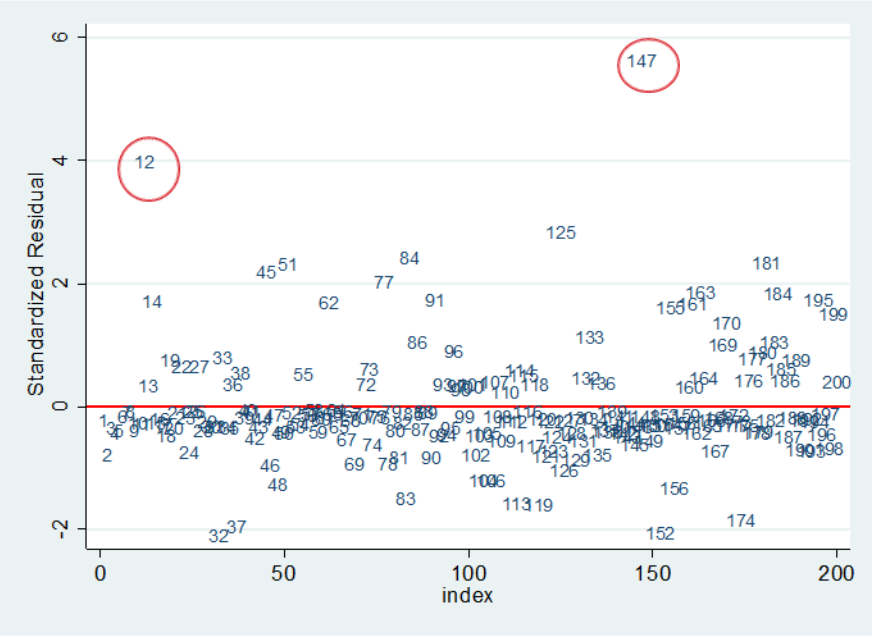
- **Współliniowość**

- $probit\ y\ x$
- $\boldsymbol{collin}$ // variance-inflation-factors

Diagnostyka - wykresy

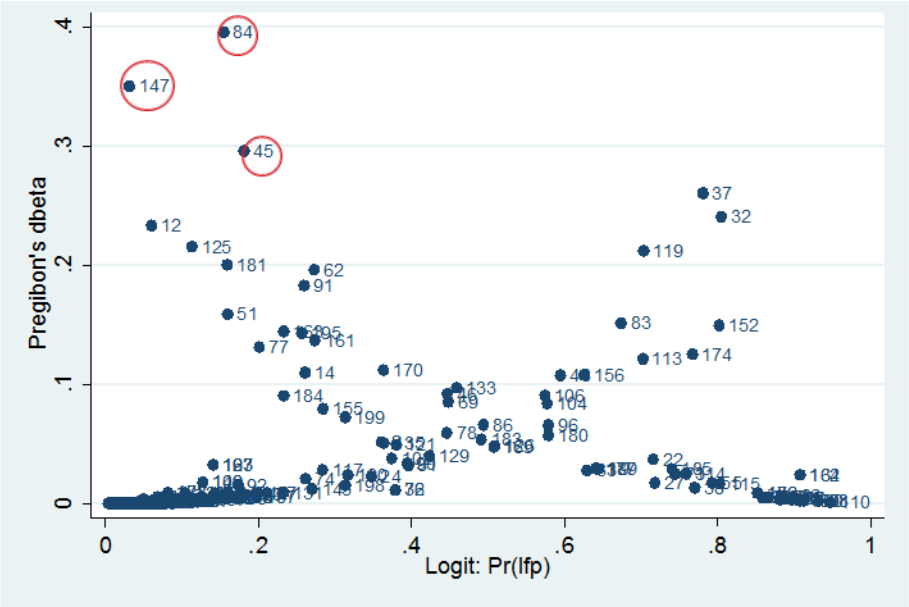
- Standardized Residual vs. Predicted Probabilities

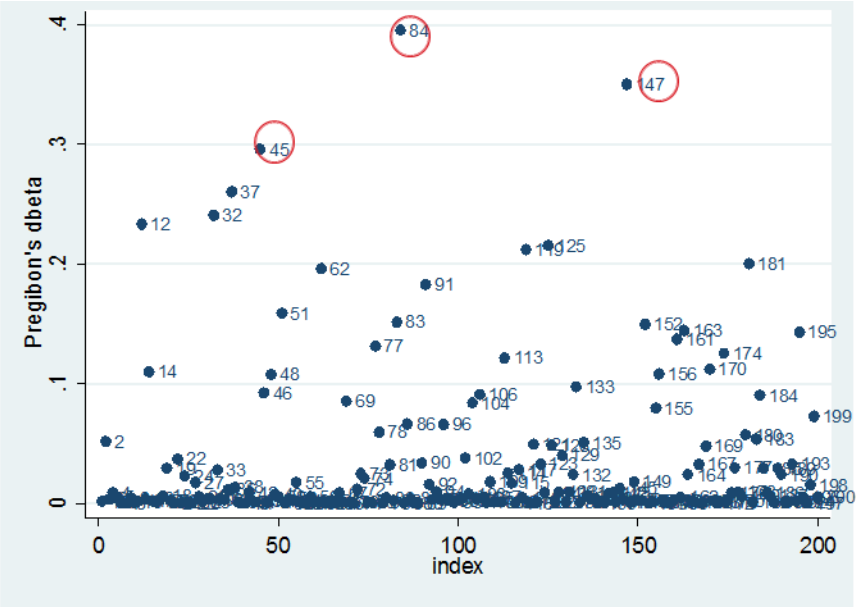




Diagnostyka - wykresy

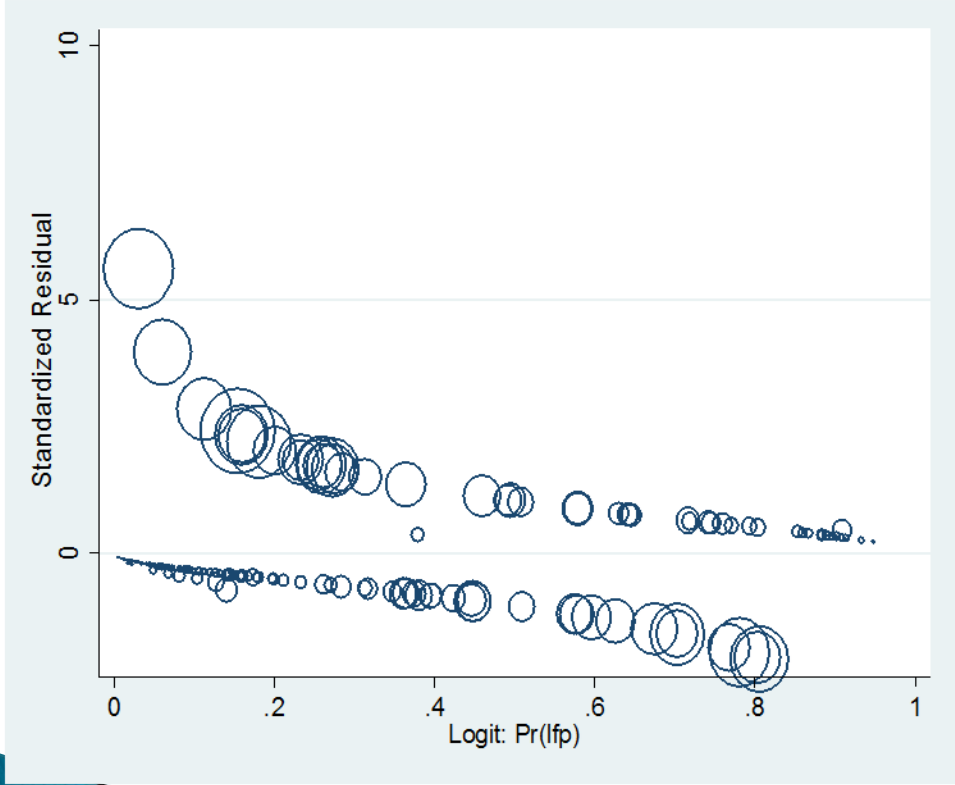
- Leverage residuals vs. predicted probabilities

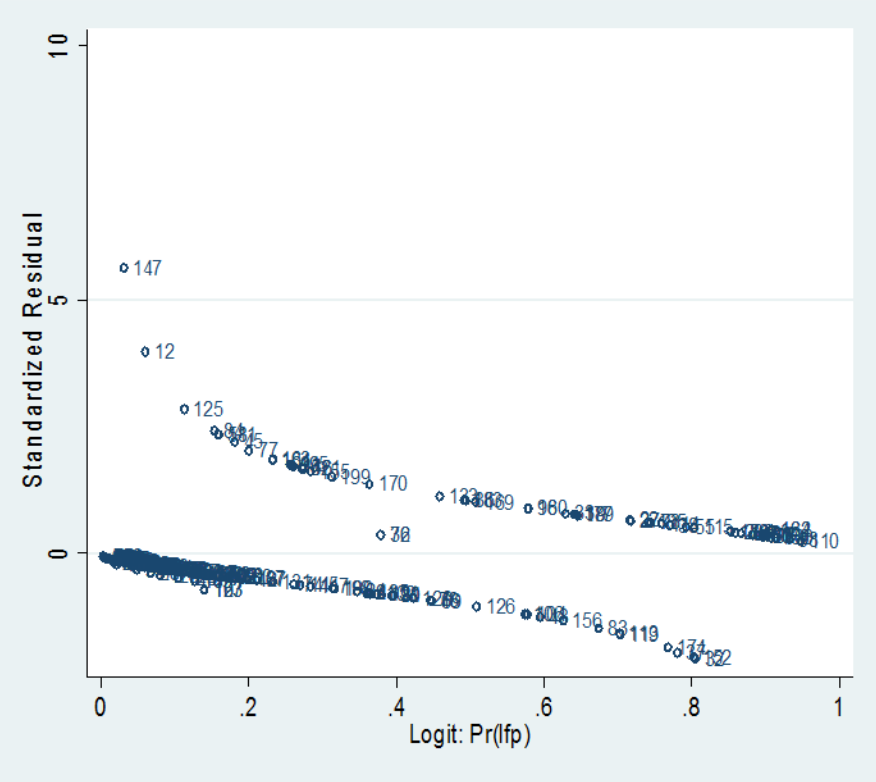




Diagnostyka - wykresy

- Residuals and Leverage



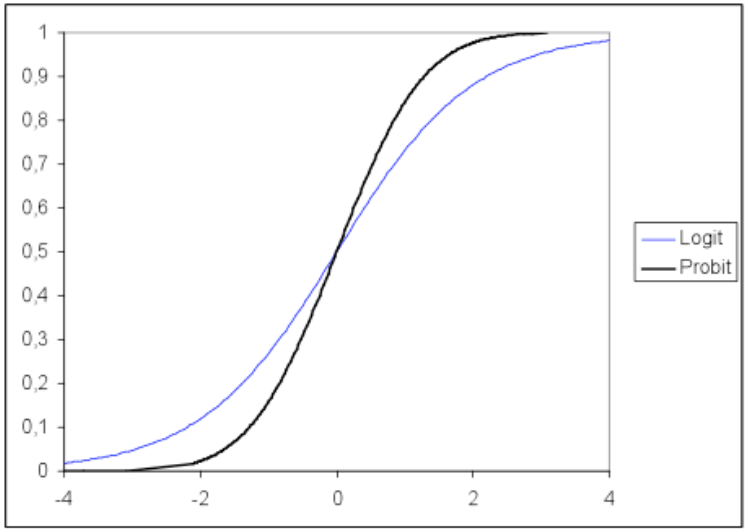


Plan zajęć

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:

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

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

- Logarytm funkcji wiarygodności:

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

Pytania teoretyczne

1. Wyjaśnić, czym jest wrażliwość i specyficzność modelu dla zmiennej binarnej i jak one zależą od progowego prawdopodobieństwa p^*.

Dziękuję za uwagę