



EKONOMETRIA

Temat wykładu:

Ekonometryczne modele specjalne

Prowadzący: dr inż. Zbigniew TARAPATA
e-mail: Zbigniew.Tarapata@isi.wat.edu.pl
http:// zbigniew.tarapata.akcja.pl/p_ekonometria/
tel.: 0-606-45-54-80

Plan wykładu

- Wprowadzenie do modeli specjalnych;
- Modele autoregresyjne jednorównaniowe z opóźnioną zmienną objaśnianą;
- Przykład budowy modelu autoregresyjnego;
- Modele ze zmiennymi binarnymi;
- Przykład budowy modelu ze zmiennymi binarnymi;

Wprowadzenie do modeli specjalnych

Do pewnej klasy modeli ekonometrycznych można zaliczyć modele, w których **odchodzi się od niektórych założeń dotyczących klasycznego modelu regresji**. Chodzi tu mianowicie o odejście od założenia o :

- braku autokorelacji składnika losowego;
- homoskedastyczności (braku zmienności w czasie wariancji składnika losowego);
- stabilności parametrów strukturalnych modelu;
- nielosowości zmiennych objaśniających.

W zasadzie każdy model, który odbiega od założeń klasycznego modelu regresji można nazwać **modelem specjalnym**.

Wprowadzenie do modeli specjalnych

Wśród różnego rodzaju modeli ekonometrycznych stosowanych w praktyce ważną rolę odgrywają **modele autoregresyjne (autokorelacyjne)**.

Cechą charakterystyczną tych modeli jest to, że nie określają one ilościowo związków zachodzących między zmienną endogeniczną (objaśnianą), a zmiennymi objaśniającymi. Możemy spotkać się w zasadzie z dwoma rodzajami modeli autoregresyjnych:

- z uwzględnieniem opóźnień zmiennej endogenicznej (objaśnianej);
- z uwzględnieniem opóźnień zmiennych objaśniających.

Modele autoregresyjne jednorównaniowe z opóźnioną zmienną objaśnianą

Modele uwzględniające zjawisko autokorelacji z uwzględnieniem opóźnień zmiennej objaśnianej można ogólnie zapisać w następującej postaci:

$$Y_t = f(Y_{t-1}, Y_{t-2}, \dots, Y_{t-p}, \xi_t) \quad (1)$$

W praktyce najczęściej przyjmuje się, że funkcja f jest liniowa. W przypadku funkcji liniowej prowadzi to do modelu postaci:

$$Y_t = \alpha_0 + \sum_{i=1}^p \alpha_i Y_{t-i} + \xi_t, \quad (2)$$

gdzie: Y_t – prognozowana zmienna w okresie (momencie) t ;

Y_{t-i} – prognozowana zmienna opóźniona w chwili $t-i$;

$\alpha_0, \alpha_1, \dots, \alpha_p$ – parametry modelu;

p – rząd autoregresji;

ξ_t – składnik losowy.

Modele autoregresyjne jednorównaniowe z opóźnioną zmienną objaśnianą

Parametry modelu (2) można szacować wykorzystując np. **metodę najmniejszych kwadratów**.

Zakładając, że parametry modelu (2) estymujemy w oparciu o n wartości szeregu czasowego $(y_t)_{t=1, \dots, n}$, wprowadzamy następujące oznaczenia:

y – wektor zaobserwowanych wartości zmiennej Y o rozmiarach $(n-p) \times 1$;

X – macierz o wymiarach $(n-p) \times (p+1)$, której pierwsza kolumna składa się z jedynek, a pozostałe z zaobserwowanych opóźnionych wartości zmiennej Y ;

α – wektor nieznanych parametrów o rozmiarach $(p+1) \times 1$;

ξ – wektor składników losowych o rozmiarach $(n-p) \times 1$.

Modele autoregresyjne jednorównaniowe z opóźnioną zmienną objaśnianą

$$y = \begin{bmatrix} y_{p+1} \\ y_{p+2} \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & y_p & y_{p-1} & \cdots & y_1 \\ 1 & y_{p+1} & y_p & \cdots & y_2 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & y_{n-1} & y_{n-2} & \cdots & y_{n-p} \end{bmatrix}, \alpha = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_p \end{bmatrix}, \xi = \begin{bmatrix} \xi_{p+1} \\ \xi_{p+2} \\ \vdots \\ \xi_n \end{bmatrix}$$

Przy tych oznaczeniach model ten można zapisać w równoważnej postaci:

$$y = X\alpha + \xi \quad (5)$$

Stosując metodę najmniejszych kwadratów (por. wzór (6) w wykładzie nr 5) otrzymujemy wektor wartości poszukiwanych parametrów (przy czym $X' = X^T$ – macierz transponowana macierzy X):

$$a = (X'X)^{-1} X'y \quad (6)$$

Modele autoregresyjne jednorównaniowe z opóźnioną zmienną objaśnianą

W wyniku estymacji parametrów przyjętego modelu uzyskujemy następującą jego postać:

$$\hat{y} = Xa + e \quad (7)$$

gdzie e to wektor reszt o rozmiarach $(n - p) \times 1$.

Macierz wariancji-kowariancji estymatorów parametrów modelu (2) dana jest wzorem (por. (12) w wykładzie nr 5):

$$D^2(a) = Var(a) = s_e^2(p) \cdot (X'X)^{-1} \quad (8)$$

przy czym **wariancja resztowa** $s_e^2(p)$ jest równa:

$$s_e^2(p) = \frac{e'e}{n - p - (p + 1)} \quad (9)$$

Modele autoregresyjne jednorównaniowe z opóźnioną zmienną objaśnianą

Z kolei współczynnik zbieżności wynosi:

$$\varphi^2 = \frac{\mathbf{e}'\mathbf{e}}{\sum_{t=p+1}^n (y_t - \bar{y})^2} \quad (10)$$

gdzie $\bar{y}$ to średnia arytmetyczna wartości szeregu od okresu (momentu) $t = p + 1$ do n , czyli:

$$\bar{y} = \frac{1}{n - p} \sum_{t=p+1}^n y_t \quad (10a)$$

Współczynnik determinacji R^2 możemy wyliczyć ze wzoru:

$$R^2 = 1 - \varphi^2 \quad (10b)$$

Modele autoregresyjne jednorównaniowe z opóźnioną zmienną objaśnianą

Przy budowaniu omawianych modeli występuje **problem ustalenia wartości parametru p** , czyli rzędu autoregresji.

Optymalną wartość p można ustalić wykorzystując np. metodę, w której podstawą wyboru jest funkcja:

$$SR(k) = \ln s_e^2(k) + \frac{k}{n} \ln n \quad (k = 0, 1, \dots, K) \quad (11)$$

gdzie: $s_e^2(k)$ - ocena wariancji składnika losowego modelu autoregresji rzędu k ;

K – maksymalny rząd autoregresji.

Korzystając z tej funkcji **parametr p wybieramy tak, aby spełniał warunek:**

$$SR(p) = \min_{k \in \{0, \dots, K\}} SR(k) \quad (12)$$

Modele autoregresyjne jednorównaniowe z opóźnioną zmienną objaśnianą

Zauważmy ważną rzecz:

- Aby wyliczyć wartości funkcji $SR(k)$ z (11) musimy znać wartość wariancji składnika resztowego $s_e^2(k)$ dla modelu autoregresji rzędu k ;
- Z kolei, aby znać wartość wariancji składnika resztowego należy najpierw dysponować modelem ☺ !

Zatem, aby określić optymalną wartość rzędu autoregresji z (12) należy wcześniej oszacować wszystkie modele autoregresji rzędu $k=0,1,...,K$!!!

Następnie z tych modeli (już oszacowanych !!!) wybieramy ten, dla którego wartość $SR(k)$ jest minimalna (zgodnie z (12)). I już !!!

Modele autoregresyjne jednorównaniowe z opóźnioną zmienną objaśnianą - Przykład

W tabeli 1 przedstawiono kwartalną wartość produkcji sprzedanej (w mln dol.) wyrobów z gumy i tworzyw sztucznych w Polsce w latach 1992-2000.

Na podstawie tych danych oszacować model autoregresyjny określony wzorem (2) oraz określić jego jakość.

Tabela 1

Rok \ Kwartał	1	2	3	4
	1	2	3	4
1992	298,3	296,2	345,1	362,8
1993	356,9	428,2	378,2	376,8
1994	373,5	418,5	474,7	516,1
1995	552,3	634,7	702,8	708,4
1996	635,7	754,4	809,9	752,8
1997	679,6	783,9	817,1	802,9
1998	801,3	911,2	960,0	895,8
1999	779,4	909,9	1 075,2	1 040,5
2000	877,1	989,0	1 041,0	949,1

Modele autoregresyjne jednorównaniowe z opóźnioną zmienną objaśnianą – Przykład, c.d.

Rozwiązanie. Ustalenie rzędu autoregresji. Wykorzystując model (2) należy ustalić rząd autoregresji p . W tym celu korzystamy ze wzoru (12); ocenę wariancji składnika losowego oszacowanych modeli autoregresji rzędu $k=1,...,8$ oraz wartości funkcji (11) przedstawiamy w tabeli 2.

k	1	2	3	4	5	6	7	8
$s_e^2(k)$	5466,3835	5742,4850	3181,9908	3288,0682	3123,3835	3072,1068	2615,6568	2791,2091
$SR(k)$	8,71	8,85	8,36	8,50	8,54	8,63	8,57	8,73

Tabela 2

Jak obliczyliśmy elementy tabeli 2? Np. $SR(k=2)$, $s_e^2(k=2)$

$$s_e^2(k=2) = \frac{\mathbf{e}'\mathbf{e}}{n-k-(k+1)} = \frac{\sum_{t=p+1}^n e_t^2}{36-2-3} = \frac{178017}{31} = 5742,48 \quad \text{patrz tabela 3 na następnej stronie}$$

$$SR(k=2) = \ln s_e^2(k) + \frac{k}{n} \ln n = \ln s_e^2(k=2) + \frac{2}{36} \ln 36 = \ln 5742,48 + \frac{1}{18} \ln 36 = 8,65 + 0,2 = 8,85$$

t	y_t	$y(t-1)$	$y(t-2)$	$y(t)^2$	$e(t)^2$
3	345,1	296,2	298,3	345,0409	0,003493
4	362,8	345,1	296,2	389,3016	702,3356
5	356,9	362,8	345,1	406,252	2435,615
6	428,2	356,9	362,8	401,2383	726,934
7	378,2	428,2	356,9	465,7207	7659,877
8	376,8	378,2	428,2	421,7588	2021,297
9	373,5	376,8	378,2	419,5546	2121,025
10	418,5	373,5	376,8	416,5388	3,846275
11	474,7	418,5	373,5	457,2439	304,7144
12	516,1	474,7	418,5	508,9996	50,41588
13	552,3	516,1	474,7	547,5572	22,49458
14	634,7	552,3	516,1	581,1268	2870,085
15	702,8	634,7	552,3	656,4532	2148,029
16	708,4	702,8	634,7	719,6895	127,4532
17	635,7	708,4	702,8	726,0375	8160,863
18	754,4	635,7	708,4	660,2811	8858,361
19	809,9	754,4	635,7	766,4543	1887,531
20	752,8	809,9	754,4	818,9554	4376,536
21	679,6	752,8	809,9	768,2656	7861,595
22	783,9	679,6	752,8	700,8826	6891,884
23	817,1	783,9	679,6	794,001	533,5636
24	802,9	817,1	783,9	826,0303	535,011
25	801,3	802,9	817,1	813,7876	155,939
26	911,2	801,3	802,9	812,0723	9826,31
27	960	911,2	801,3	911,8041	2342,161
28	895,8	960	911,2	957,8708	3852,781
29	779,4	895,8	960	900,6235	14695,13
30	909,9	779,4	895,8	793,9714	13439,45
31	1 075,20	909,9	779,4	910,0165	27285,6
32	1 040,50	1 075,20	909,9	1 062,21	471,3106
33	877,1	1 040,50	1 075,20	1 033,868	24576,25
34	989	877,1	1 040,50	885,1895	10776,63
35	1 041,00	989	877,1	983,5045	3305,738
36	949,1	1 041,00	989	1 032,708	6990,218
			suma	178017	

Modele autoregresyjne jednorównaniowe z opóźnioną zmienną objaśnianą – Przykład, c.d.

Wyniki oszacowania modelu dla $p=2$, tzn. dla modelu:

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \xi_t$$

Po skorzystaniu ze sposobu szacowania parametrów modelu z wieloma zmiennymi objaśniającymi (patrz wykład poprzedni) otrzymujemy model:

$$\hat{y}_t = 71,1 + 0,9059 \cdot y_{t-1} + 0,0187 \cdot y_{t-2}$$

Uwaga! $e(t)^2 = (y_t - \hat{y}_t)^2$

Tabela 3

Modele autoregresyjne jednorównaniowe z opóźnioną zmienną objaśnianą – Przykład, c.d.

Z wartości funkcji $SR(k)$ zawartych w tabeli 2 wynika, że w rozważanym przypadku powinniśmy oszacować model autoregresji z maksymalnym opóźnieniem $p=3$ (dla tego opóźnienia funkcja $SR(k)$ przyjmuje najmniejszą wartość), czyli model postaci:

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \alpha_3 Y_{t-3} + \xi_t$$

Oszacowanie parametrów modelu. Parametry wybranego modelu oszacujemy za pomocą klasycznej metody najmniejszych kwadratów korzystając ze wzoru (6). Dla danych z tabeli 1 odpowiednie macierze mają postać jak na następnym slajdzie.

Modele autoregresyjne jednorównaniowe z opóźnioną zmienną objaśnianą – Przykład, c.d.

362,8	1	345,1	296,2	298,3
356,9	1	362,8	345,1	296,2
428,2	1	356,9	362,8	345,1
378,2	1	428,2	356,9	362,8
376,8	1	378,2	428,2	356,9
373,5	1	376,8	378,2	428,2
418,5	1	373,5	376,8	378,2
474,7	1	418,5	373,5	376,8
516,1	1	474,7	418,5	373,5
552,3	1	516,1	474,7	418,5
634,7	1	552,3	516,1	474,7
702,8	1	634,7	552,3	516,1
708,4	1	702,8	634,7	552,3
635,7	1	708,4	702,8	634,7
754,4	1	635,7	708,4	702,8
809,9	1	754,4	635,7	708,4
752,8	1	809,9	754,4	635,7
679,6	1	752,8	809,9	754,4
783,9	1	679,6	752,8	809,9
817,1	1	783,9	679,6	752,8
802,9	1	817,1	783,9	679,6
801,3	1	802,9	817,1	783,9
911,2	1	801,3	802,9	817,1
960,0	1	911,2	801,3	802,9
895,8	1	960,0	911,2	801,3
779,4	1	895,8	960,0	911,2
909,9	1	779,4	895,8	960,0
1075,2	1	909,9	779,4	895,8
1040,5	1	1075,2	909,9	779,4
877,1	1	1040,5	1075,2	909,9
989,0	1	877,1	1040,5	1075,2
1041,0	1	989,0	877,1	1040,5
949,1	1	1041,0	989,0	877,1

y_t

$X'X = \begin{bmatrix} 33,00 & 22945,70 & 22200,90 & 21510,20 \\ 22945,70 & 17642590,49 & 17048179,77 & 16499275,92 \\ 22200,90 & 17048179,77 & 16499275,92 & 16106987,23 \\ 21510,20 & 16499275,92 & 16106987,23 & 15757505,82 \end{bmatrix}$

$(X'X)^{-1} = \begin{bmatrix} 0,317562449 & -0,000361871 & -0,000008931 & -0,000045461 \\ -0,000361871 & 0,000005874 & -0,000005467 & -0,000000067 \\ -0,000008931 & -0,000005467 & 0,000010976 & -0,000005482 \\ -0,000045461 & -0,000000067 & -0,000005482 & 0,000005800 \end{bmatrix}$

$X'y = \begin{bmatrix} 23549,70 \\ 17933974,25 \\ 17334992,49 \\ 16900081,74 \end{bmatrix}$

i ostatecznie $a = (X'X)^{-1} X'y = \begin{bmatrix} 65,5814 \\ 0,8977 \\ -0,6479 \\ 0,7052 \end{bmatrix}$

Oszacowany model ma zatem postać:

$$y_t = 65,5814 + 0,8977y_{t-1} - 0,6479y_{t-2} + 0,7052y_{t-3} + e_t$$

t	y _t	y(t-1)	y(t-2)	y(t-3)	y(t) [*]	e(t) ²
4	362,8	345,1	296,2	298,3	393,83	962,83
5	356,9	362,8	345,1	296,2	376,56	386,34
6	428,2	356,9	362,8	345,1	394,28	1150,87
7	378,2	428,2	356,9	362,8	474,59	9290,30
8	376,8	378,2	428,2	356,9	379,35	6,48
9	373,5	376,8	378,2	428,2	460,76	7615,04
10	418,5	373,5	376,8	378,2	423,45	24,49
11	474,7	418,5	373,5	376,8	465,00	94,16
12	516,1	474,7	418,5	373,5	483,96	1032,71
13	552,3	516,1	474,7	418,5	516,45	1285,15
14	634,7	552,3	516,1	474,7	561,76	5320,69
15	702,8	634,7	552,3	516,1	641,47	3761,52
16	708,4	702,8	634,7	552,3	674,74	1132,77
17	635,7	708,4	702,8	634,7	693,76	3370,62
18	754,4	635,7	708,4	702,8	672,89	6643,87
19	809,9	754,4	635,7	708,4	830,50	424,30
20	752,8	809,9	754,4	635,7	752,15	0,43
21	679,6	752,8	809,9	754,4	748,64	4766,14
22	783,9	679,6	752,8	809,9	759,06	617,06
23	817,1	783,9	679,6	752,8	859,85	1827,46
24	802,9	817,1	783,9	679,6	770,46	1052,63
25	801,3	802,9	817,1	783,9	809,75	71,41
26	911,2	801,3	802,9	817,1	840,93	4938,29
27	960	911,2	801,3	802,9	930,61	863,95
28	895,8	960	911,2	801,3	902,08	39,47
29	779,4	895,8	960	911,2	890,33	12306,33
30	909,9	779,4	895,8	960	861,85	2308,75
31	1 075,20	909,9	779,4	895,8	1009,14	4363,64
32	1 040,50	1 075,20	909,9	779,4	990,90	2460,59
33	877,1	1 040,50	1 075,20	909,9	944,68	4566,55
34	989	877,1	1 040,50	1 075,20	937,04	2699,45
35	1 041,00	989	877,1	1 040,50	1118,89	6067,29
36	949,1	1 041,00	989	877,1	977,84	826,19
				suma	92277,74	

Modele autoregresyjne jednorównaniowe z opóźnioną zmienną objaśnianą – Przykład, c.d.

Jakość modelu.

Po obliczeniu wartości teoretycznych (kolumna $y(t)^*$ w tabeli 4) i reszt szacujemy (korzystając ze wzoru (9)) wariancję resztową oszacowanego modelu i otrzymujemy:

$$s_e^2(p) = \frac{e'e}{n - p - (p+1)} = \frac{92277,7327}{36 - 3 - (3+1)} = 3181,9908$$

Tabela 4

Modele autoregresyjne jednorównaniowe z opóźnioną zmienną objaśnianą – Przykład, c.d.

Następnie (wykorzystując wzór (8)) szacujemy macierz wariancji i kowariancji estymatorów parametrów modelu:

$$Var(a) = s_e^2(p) \cdot (X'X)^{-1} = \begin{bmatrix} 1010,4808 & -1,1515 & -0,0284 & -0,1447 \\ -1,1515 & \mathbf{0,0187} & -0,0174 & -0,0002 \\ -0,0284 & -0,0174 & \mathbf{0,0349} & -0,0174 \\ -0,1447 & -0,0002 & -0,0174 & \mathbf{0,0185} \end{bmatrix}$$

Na tej podstawie oszacowany model możemy zapisać w postaci (w nawiasach podano średni błąd szacunku danego parametru):

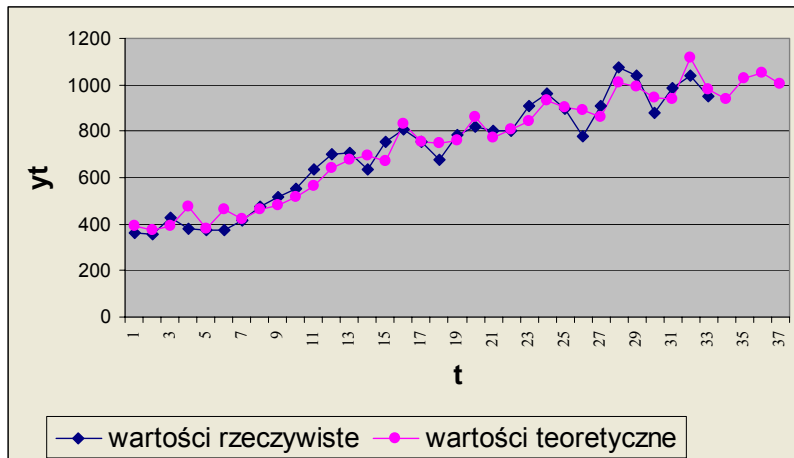
$$\hat{y}_t = 65,5814 + 0,8977 y_{t-1} - 0,6479 y_{t-2} + 0,7052 y_{t-3}$$

(31,7881) (0,1367) (0,1869) (0,1359)

oraz obliczyć wartości następujących miar dopasowania:

$$s_e = 56,4091, \quad V_e = 0,0790, \quad \varphi^2 = 5,7\%, \quad R^2 = 94,3\%$$

Modele autoregresyjne jednorównaniowe z opóźnioną zmienną objaśnianą – Przykład, c.d.



Wykres 1 Wykres dopasowania danych rzeczywistych do oszacowanego modelu

Modele ze zmiennymi binarnymi (zero-jedynkowymi)

- W modelach ekonometrycznych na badaną zmienną często wywierają wpływ takie czynniki, jak miejsce zamieszkania badanej osoby, płeć pracownika, posiadanie własnego mieszkania przez rodzinę o danych dochodach, znajomość języka angielskiego przez pracownika itp. Reprezentujące je zmienne objaśniające wprowadza się do modelu jako tzw. zmienne zero-jedynkowe (dychotomiczne, binarne), przypisując im wartość zero (np. badany nie zna języka obcego), bądź jeden (zna język obcy), zależnie od występującej sytuacji.
- W przypadku występowania *w modelu wśród zmiennych objaśniających zmiennych zero-jedynkowych* postępowanie prognostyczne przebiega tak, jak przy wcześniej opisywanych modelach.

Modele ze zmiennymi binarnymi (zero-jedynkowymi)

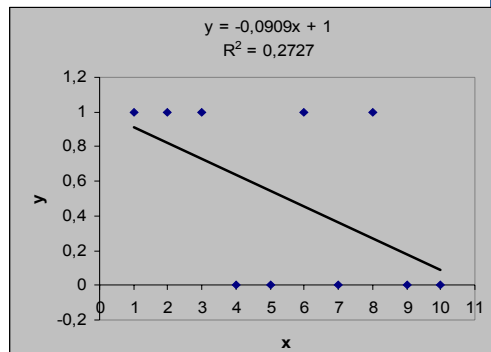
Sytuacja się komplikuje, gdy model ekonometryczny budowany jest w celu wyjaśnienia zjawisk opisywanych przez zmienną jakościową, czyli wtedy, gdy zmienną objaśnianą jest zmienna zero-jedynkowa,

gdyż żaden z poznanych przez nas do tej pory modeli nie daje nam „na wyjściu” tylko dwóch wartości: 0 i 1. Modele opisywane przez nas do tej pory są modelami ciągłymi, tzn. opisują one w sposób ciągły zależność zmiennej objaśnianej od zmiennych objaśniających.

Modele ze zmiennymi binarnymi (zero-jedynkowymi)

Przykład

Załóżmy, że zero-jedynkowa zmienna objaśniana Y zależy od jednej zmiennej objaśniającej X . Wykres hipotetycznej sytuacji dla 10-ciu obserwacji zaprezentowano na wykresie. Zauważmy, że dla oszacowanego modelu liniowego współczynnik determinacji wynosi tylko 0,27, czyli model słabo opisuje zależność między zmiennymi. Ponadto jest modelem ciągłym, a nas interesuje model, który pozwalałby otrzymywać tylko dwie wartości zmiennej objaśnianej, tzn. 0 i 1. Cóż bowiem znaczy wartość np. $y \approx 0.5$ z modelu dla $x=5$?



Wykres 2

Modele ze zmiennymi binarnymi (zero-jedynkowymi) Przekształcenie logitowe

W celu prognozowania wartości zmiennej zero-jedynkowej, stosuje się specjalne przekształcenia modeli (otrzymując tzw. modele probitowe, logitowe i inne).

Przyjmijmy, że interesuje nas model, który objaśnia, czy badana osoba kupiła w ciągu ostatniego roku komputer, czy też nie. Zmienna objaśniana (Y) jest więc zmienną zero-jedynkową przyjmującą wartość jeden, jeżeli dana osoba kupiła komputer, a zero - w przeciwnym przypadku. Ponadto założmy, że jedyną zmienną objaśniającą w tym modelu jest dochód (X) badanej osoby. Model, który opisuje tę sytuację może mieć postać:

$$y_j = \alpha_0 + \alpha_1 x_j + \xi_j \quad (2.1)$$

gdzie j jest numerem badanej osoby.

Modele ze zmiennymi binarnymi (zero-jedynkowymi) Przekształcenie logitowe

Na podstawie obserwacji n osób otrzymamy model:

$$\hat{y}_j = a_0 + a_1 x_j + \xi_j \quad (2.2)$$

gdzie:

$\hat{y}_j$ - realizacja zmiennej losowej Y , przy czym

$$Y = \begin{cases} 1, & \text{jeżeli } j\text{-ta osoba kupiła komputer} \\ 0, & \text{w przeciwnym przypadku} \end{cases}$$

x_j - wysokość dochodów (w zł/osobę) przypadająca na j -tą osobę;

a_0, a_1 - estymatory parametrów α_0, α_1 .

Modele ze zmiennymi binarnymi (zero-jedynkowymi)

Przekształcenie logitowe

Wartość oczekiwana zmiennej objaśnianej może być interpretowana jako

warunkowe prawdopodobieństwo realizacji danego zdarzenia pod warunkiem, że zmienna objaśniająca przyjęła pewną wartość. Wartość $\hat{y}$ możemy traktować jako *warunkowe prawdopodobieństwo realizacji tego zdarzenia*.

Jednak i w tym przypadku możemy mieć problemy. Jak wiadomo prawdopodobieństwo może przyjmować wartości tylko z przedziału $[0, 1]$ natomiast szacując nasz model możemy wykroczyć poza zakres tego przedziału z wartością $\hat{y}$. Na przykład w modelu z wykresu 1 jeżeli przyjęlibyśmy wartość zmiennej $x=20$, to otrzymujemy $=-0.8$.

Modele ze zmiennymi binarnymi (zero-jedynkowymi)

Przekształcenie logitowe

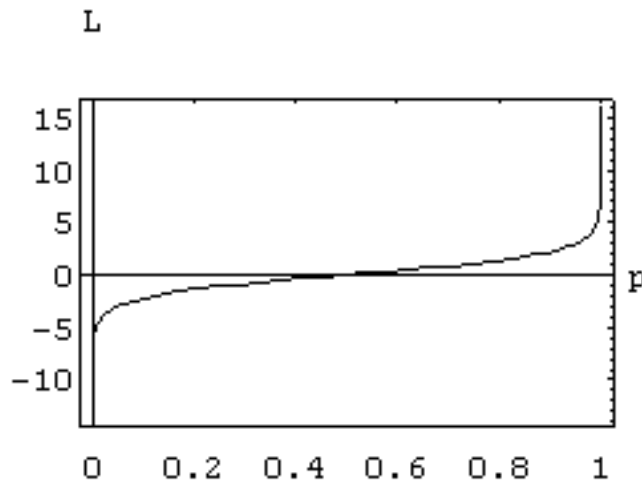
W celu uniknięcia większych od jedności lub mniejszych od zera wartości prawdopodobieństwa dokonuje się monotonicznego przekształcenia

prawdopodobieństwa p z przedziału $[0, 1]$ na przedział $(-\infty, +\infty)$. Jednym z takich przekształceń jest tzw. *transformacja logitowa*. Wykorzystuje ona, do zamiany prawdopodobieństwa p na liczbę z przedziału $(-\infty, +\infty)$, tzw. **logity L** :

$$L = \ln \frac{p}{1-p} \quad (2.3)$$

których wartość dla $p=0$ wynosi $-\infty$, zero dla $p=0.5$ oraz $+\infty$ dla $p=1$.

Modele ze zmiennymi binarnymi (zero-jedynkowymi) Przekształcenie logitowe



Wykres 3 Wykres funkcji logitowej

Modele ze zmiennymi binarnymi (zero-jedynkowymi) Przekształcenie logitowe

W celu oceny parametrów modelu (2.1) odpowiednie wartości L należy podstawić w miejsce zmiennej objaśnianej. Tym samym model przybierze postać:

$$L_j = \beta_0 + \beta_1 x_j + \xi_j \quad (2.4)$$

Aby z modelu (2.4) uzyskać wartość prawdopodobieństwa, trzeba posłużyć się funkcją logitową (2.3) i wyznaczyć p_j . Dokonując ciągu przekształceń wzoru (2.3) otrzymamy:

$$\begin{aligned} e^L = \frac{p}{1-p} &\rightarrow \frac{1}{e^L} = \frac{1-p}{p} = \frac{1}{p} - 1 &\rightarrow \frac{1}{p} = \frac{1}{e^L} + 1 = 1 + e^{-L} &\rightarrow p = \frac{1}{1 + e^{-L}} \end{aligned} \quad \begin{matrix} (2.5) \\ (2.6) \\ (2.7) \\ (2.8) \end{matrix}$$

Modele ze zmiennymi binarnymi (zero-jedynkowymi) Przekształcenie logitowe

Podstawiając $p=p_j$ oraz $L=L_j$ i korzystając z (2.4) otrzymujemy:

$$p_j = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_j)}} \quad (2.9)$$

przy czym pominęliśmy składnik losowy ξ_j , gdyż w klasycznym modelu regresji przyjmowaliśmy założenie, że $E\{\xi_j\}=0$.

Parametry modelu (2.4) szacuje się

uogólnioną metodą najmniejszych kwadratów (metoda najmniejszych kwadratów daje gorsze oszacowania).

Natomiast **prawdopodobieństwa p_j zastępuje się częstościami oszacowanymi na podstawie próby.**

Modele ze zmiennymi binarnymi (zero-jedynkowymi) Przekształcenie logitowe - Przykład

W badaniu budżetów rodzinnych zebrano od 400 rodzin informacje o wysokości dochodu i fakcie zakupu (lub nie) komputera w ciągu roku. W badaniu wyróżniono 8 grup dochodowych. Zebrane informacje zaprezentowano w Tabeli 5. Na podstawie modelu:

$$P_j = \beta_0 + \beta_1 X_j$$

odpowiedzieć na pytania:

1. Jaka jest szansa zakupu komputera przez rodzinę o dochodzie 6 tys. zł?
2. Przy jakim poziomie dochodu rodziny szansa zakupu komputera wynosi 0,9?

Dochód (w tys. zł)	Wartość zmiennej objaśniającej (X)	Liczba badanych gospodarstw domowych (N)	Liczba rodzin kupujących komputer (n)	Częstość (p) $p = \frac{n}{N}$	Logit (L) $L = \ln \frac{p}{1-p}$
0,5-1,5	1	20	2	0,10	-2,20
1,5-2,5	2	40	12	0,30	-0,85
2,5-3,5	3	60	34	0,57	0,27
3,5-4,5	4	70	44	0,63	0,53
4,5-5,5	5	80	54	0,68	0,73
5,5-6,5	6	60	44	0,73	1,01
6,5-7,5	7	50	44	0,88	1,99
7,5-8,5	8	20	18	0,90	2,20

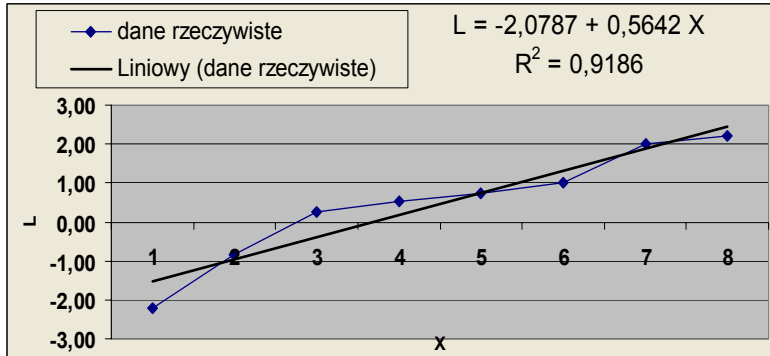
Tabela 5

Modele ze zmiennymi binarnymi (zero-jedynkowymi) Przekształcenie logitowe – Przykład, c.d.

Rozwiązanie (stosujemy metodę najmniejszych kwadratów do szacowania parametrów modelu)

Liniowa funkcja regresji opisująca zależność logitu od wysokości dochodów została przedstawiona na Wykresie 4 i ma postać:

$$L_j = -2.0787 + 0.5642 X_j$$



Modele ze zmiennymi binarnymi (zero-jedynkowymi) Przekształcenie logitowe – Przykład, c.d.

Korzystając z definicji logitu (2.3) wartość prawdopodobieństwa dla liniowej funkcji regresji opisującej logit obliczamy ze wzoru (2.9) (wyniki w tabeli 6):

$$p_j = \frac{1}{1 + e^{-L_j}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_j)}} = \frac{1}{1 + e^{-(-2,0787 + 0,5642 X_j)}}$$

X_j	L_j	p_j	Częstość
1	-1,51	0,1802729	0,10
2	-0,95	0,2788245	0,30
3	-0,39	0,4046565	0,57
4	0,18	0,5444077	0,63
5	0,74	0,6774986	0,68
6	1,31	0,7869269	0,73
7	1,87	0,8665393	0,88
8	2,43	0,9194502	0,90

Tabela 6

Modele ze zmiennymi binarnymi (zero-jedynkowymi) Przekształcenie logitowe – Przykład, c.d.

Możemy obecnie odpowiedzieć na pytania zawarte w treści zadania.

Ad.1

Szansa zakupu komputera przez rodzinę o dochodzie 6 tys. zł wynosi:

$$p_j = \frac{1}{1 + e^{-(2,0787 + 0,5642 \cdot 6)}} \approx 0,79$$

Ad.2

Poziom dochodu rodziny, przy którym szansa zakupu komputera wynosi 0,9 jest równy:

$$-2,0787 + 0,5642X = 2,20 \rightarrow \underline{X=7,58}$$