Interpretacja statystyk modelu dotyczącego otyłości badanych przy użyciu binarnej regresji logistycznej w języku SAS 4GL i narzędziu SAS Enterprise Guide

Autor: Adrian Żelazek

Spis treści

Cel projektu

Celem projektu jest przeanalizowanie poszczególnych statystyk wynikających ze zbudowania modelu binarnej regresji logistycznej, dotyczącego otyłości osób badanych. Należy podkreślić, iż celem niniejszego projektu jest jedynie ćwiczeniowa analiza i interpretacja uzyskanych wyników, a nie zbudowanie najlepszego modelu.

Zapoznanie z danymi

Rezultaty regresji logistycznej

Procedura LOGISTIC

Informacje o modelach		
Zbiór	WORK.SORTTEMPTABLESORTED	
Zmienna objaśniana	nadwaga	nadwaga
Liczba poziomów odpowiedzi	2	
Model	logit binarny	
Metoda optymalizacji	Ocena Fishera	

Liczba obserwacji wczytanych	943
Liczba obserwacji użytych	943

Profil odpowiedzi		
Wartość uporządkowana	nadwaga	Całkowita liczebność
1	0	476
2	1	467

Wymodelowane prawdopodobieństwo wynosi nadwaga=1.

Techniką optymalizacji, czyli poszukiwanie maksimum funkcji wiarygodności, jest Ocena Fishera (udoskonalenie metody Newtona Raphsona).

Liczba obserwacji wczytanych i użytych (943) jest taka sama, ponieważ zbiór nie zawiera braków danych.

Profil odpowiedzi: 0 brak nadwagi oraz 1 nadwaga. Modelowanym zdarzeniem jest obecność nadwagi.

Kodowanie zmiennych jakościowych

Informacje o poziomach klasyfikacji					
Klasa	Wartość	Zmienne projektowe			
sex	Kobieta	1			
	Mezczyz	0			
komp	brak komputera	1			
	komputer	0			
syt_materialna	bardzo dobra	1	0	0	0
	przeciętna	0	1	0	0
	raczej dobra	0	0	1	0
	raczej zła	0	0	0	1
	zła	0	0	0	0
stan_cyw	kawaler, panna	1	0	0	
	rozwiedziony(a	0	1	0	
	wdowiec, wdowa	0	0	1	
	zonaty, mezatk	0	0	0	
fast	często	1			
	rzadko lub prawie nigdy	0			

Informacje o poziomach klasyfikacji obrazują jak zostały zakodowane zmienne jakościowe w modelu. Zmienne projektowe z kolei przedstawiają zmienne sztuczne, które weszły do modelu.

Statystyki Dewiancji i Pearsona

Status zbieżności
Kryterium zbieżności (GCONV=1E-8) spełnione.

Statystyki dewiancji i dobroci dopasowania Pearsona				
Kryterium	Wartość	DF	Wartość/DF	Pr. > chi-kw.
Dewiancja	523.0887	930	0.5625	1.0000
Pearsona	824.3695	930	0.8864	0.9943

Liczba unikatowych profilów: 943

Kryterium zbieżności jest spełnione w omawianym modelu. Jest to pozytywna informacja, ponieważ w przeciwnym razie (gdyby nie było spełnione) wyniki mogłyby być wątpliwe.

W przypadku oceny omawianego modelu nie można wykorzystać statystyk Dewiancji oraz Pearsona. Spowodowane jest to tym, iż próba liczy 943 obserwacje, liczba unikatowych profili również wynosi 943, czyli 1 obserwacja przypada na każdy unikatowy profil. By móc skorzystać z omawianych statystyk przynajmniej 10 obserwacji powinno przypadać na 1 unikatowy profil. Przyczyną tak dużej ilości unikatowych profili są zmienne ciągłe: wiek oraz dochód.

Statystyki dopasowania Akkaike i Schwarza oraz statystyka R-kwadrat

Statystyki dopasowania		
Kryterium	Tylko wyraz wolny	Wyraz wolny i współzmienne
AIC	1309.190	549.089
SC	1314.039	612.127
-2 log L	1307.190	523.089

Statystyki dopasowania AIC i S.C. pozwalają porównać między sobą modele różniące się zestawem zmiennych objaśniających. Im niższa wartość AIC i S.C., tym model jest lepiej dopasowany do danych. Należy używać tych statystyk, gdy porównujemy modele dla tych samych danych, ale różniące się ilością oszacowanych parametrów.

W omawianym projekcie może porównać model z danymi do modelu jedynie z wyrazem wolnym. Tym samym lepszy jest model ze zmiennymi, gdyż wartości AIC oraz SC są o wiele niższe.

Testowanie globalnej hipotezy zerowej

Testowanie globalnej hipotezy zerowej: BETA=0			
Testowanie	Chi-kwadrat	DF	Pr. > chi-kw.
Il. wiarygodn.	784.1010	12	<.0001
Mn. Lagrange'a	647.8598	12	<.0001
Walda	295.8888	12	<.0001

Testowanie globalnej hipotezy zerowej. Formalne wnioskowanie statystyczne. W omawianym przypadku 12 współczynników.

Hipoteza 0 : B1=…=B12=0, gdzie B oznacza Betę

Hipoteza 1: Hipoteza alternatywna

W przypadku omawianego modelu alfa = 0.05. W omawianym przypadku występuje bardzo niskie p-value <0.0001, czyli poniżej 0.05, tym samym odrzucamy H0 na rzecz hipotezy alternatywnej (H1). Po sprawdzeniu istotności wiadome jest, iż chociaż jeden parametr Beta jest różny od zera, co oznacza iż badany model jest lepszy od modelu z wyrazem wolnym i jest istotny statystycznie. Dalsze rozważania i analizy byłyby zbyteczne, gdyby omawiany model nie był lepszy od modelu z jedynie wyrazem wolnym.

Badanie istotności poszczególnych zmiennych

Analiza efektów typu 3

Efekt	DF	Chi-kwadrat Walda	Pr. > chi-kw.
wiek	1	0.2046	0.6510
dochod	1	0.0060	0.9383
sex	1	0.3682	0.5440
komp	1	5.7364	0.0166
syt_materialna	4	4.4318	0.3507
stan_cyw	3	5.7209	0.1260
fast	1	242.8047	<.0001

W analizie efektów typu 3 sprawdzona została istotność poszczególnych zmiennych. DF, czyli liczba stopni swobody, gdzie przykładowo w zmiennej syt_materialna jest ich 4 oznacza, iż zmienna została wprowadzona przy użyciu 4 zmiennych sztucznych.

Na tym etapie testowana jest hipoteza czy wszystkie sztuczne zmienne dla danej zmiennej są równe 0.

Przykładowo dla zmiennej syt_materialna

Hipoteza 0: B5=B6=B7=B8=0, gdzie B oznacza Betę

Hipoteza 1: Hipoteza alternatywna

p-value = 0.3507

alfa = 0.05

p-alue > alfa, tym samym brak podstaw do odrzucenia hipotezy zerowej. Oznacza to, iż sytuacja materialna jest nie istotna statystycznie w omawianym modelu.

Powyższa analiza przeprowadzona również dla pozostałych zmiennych oznacza, że w modelu istotne są jedynie zmienne: komp oraz fast. Nie mniej jednak reszta zmiennych może być zmiennymi zakłócającymi. W przypadku usunięcia jednej zmiennej z modelu przykładowo zmiennej dochód i chcąc zobaczyć jak zmienna fast wpływa na nadwagę to patrząc na to jak zmienia się współczynnik oszacowania zmiennych istnieje możliwość zlokalizowania zmiennych zakłócających. Jeśli różnica

wynosi >10% po usunięciu zmiennej to daną zmienną powinno się przywrócić do modelu, gdyż jest to zmienna zakłócająca do zmiennej fast.

Analiza ocen maksymalnej wiarygodności

Analiza ocen maksymalnej wiarygodności							
Parametr		DF	Ocena	Błąd standardowy	Chi-kwadrat Walda	Pr. > chi-kw.	Exp(oszacowanie)
Intercept		1	-3.6398	0.6666	29.8174	<.0001	0.026
wiek		1	0.00405	0.00895	0.2046	0.6510	1.004
dochod		1	-6.21E-6	0.000080	0.0060	0.9383	1.000
sex	Kobieta	1	0.1639	0.2701	0.3682	0.5440	1.178
komp	brak komputera	1	-0.6692	0.2794	5.7364	0.0166	0.512
syt_materialna	bardzo dobra	1	-0.6559	1.3241	0.2454	0.6204	0.519
syt_materialna	przeciętna	1	0.6895	0.3974	3.0094	0.0828	1.993
syt_materialna	raczej dobra	1	0.4906	0.5077	0.9336	0.3339	1.633
syt_materialna	raczej zła	1	0.4948	0.4383	1.2743	0.2590	1.640
stan_cyw	kawaler, panna	1	0.1156	0.4527	0.0652	0.7984	1.123
stan_cyw	rozwiedziony(a	1	0.7614	0.5556	1.8780	0.1706	2.141
stan_cyw	wdowiec, wdowa	1	-0.8257	0.4907	2.8310	0.0925	0.438
fast	często	1	5.1855	0.3328	242.8047	<.0001	178.654

Analiza ocen maksymalnej wiarygodności zwraca oceny wszystkich parametrów uwzględniając również zmienne sztuczne. W tym miejscu następuje weryfikacja istotności ocen poszczególnych parametrów. Przykładowo w przypadku zmienne syt_materialna wszędzie p-value > 0.05, co oznacza, iż żadna zmienna w obrębie syt_materialna nie jest istotna statystycznie. Jedynie zmienne komp oraz fast są istotne statystycznie i podlegają interpretacji.

Interpretacja zmiennych istotnych statystycznie:

Komp: Szanse na bycie otyłym, a nie nieotyły są dla osoby nie posiadającej komputera o [(0.512-1)*100%]= 48.8% niższe, niż dla osoby posiadającej komputer.

Fast: Szanse na bycie otyłym, a nie nieotyłym są dla osoby jedzącej posiłki typu fast food 178.654 razy większe niż dla osoby jedzącej jedzenia typu fast food rzadko albo prawie nigdy.

Skojarzenie prognozowanych prawdopodobieństw i obserwowanych odpowiedzi

Skojarzenie prognozowanych prawdopodobieństw i obserwowanych odpowiedzi			
Procent zgodnych	92.8	D Somersa	0.857
Procent niezgodnych	7.2	Gamma	0.857
Procent równych	0.0	Tau-a	0.429
Pary	222292	c	0.928

Powyższe statystyki służą do oceny jakości klasyfikacyjnej omawianego modelu.

Model cechuje się wysokim % par zgodnych wynoszącym 92.8%. Co za tym idzie procent par nie zgodnych wynosi zaledwie 7.2%. Statystyki D Sommersa, Gamma, Tau-a oraz c im bliżej znajdują się

wartości 1, tym lepszy jest model i cechuje się większą mocą predykcyjną, przy czym wartość Tau-a standardowo jest najniższa od D Sommersa czy Gamma.

W omawianym przypadku wyniki są bardzo wysokie, a wysokość statystyki c (pole pod krzywą ROC) na poziomie aż 0.928 może wskazywać na przeuczenie modelu. Wartość tej statystyki powinna oscylować w granicach 0.7-0.8.

Oceny parametrów i przedziały ufności Walda wraz z interpretacją wyników

Oceny parametrów i przedziały ufności Walda				
Parametr		Ocena	Przedział ufności 95%	
Intercept		-3.6398	-4.9463	-2.3334
wiek		0.00405	-0.0135	0.0216
dochod		-6.21E-6	-0.00016	0.000151
sex	Kobieta	0.1639	-0.3654	0.6931
komp	brak komputera	-0.6692	-1.2169	-0.1216
syt_materialna	bardzo dobra	-0.6559	-3.2510	1.9392
syt_materialna	przeciętna	0.6895	-0.0895	1.4684
syt_materialna	raczej dobra	0.4906	-0.5045	1.4857
syt_materialna	raczej zła	0.4948	-0.3643	1.3538
stan_cyw	kawaler, panna	0.1156	-0.7717	1.0030
stan_cyw	rozwiedziony(a	0.7614	-0.3275	1.8503
stan_cyw	wdowiec, wdowa	-0.8257	-1.7875	0.1361
fast	często	5.1855	4.5332	5.8377

Jeśli przedział ufności w ocenie parametrów zawiera 0, to oszacowanie jest nie istotne statystycznie i nie podlega interpretacji. Tym samym, w analizowanym modelu, jedynie zmienne: komp oraz fast są istotne statystycznie i podlegają interpretacji.

Interpretacje istotnych statystycznie zmiennych:

Komp:

Fast:

Oceny ilorazów szans i przedziały ufności Walda wraz z interpretacją wyników

Oceny ilorazów szans i przedziały ufności Walda				
Efekt	Jednostka	Ocena	Przedział ufności 95%	
wiek	1.0000	1.004	0.987	1.022
dochod	1.0000	1.000	1.000	1.000
sex Kobieta od Mezczyz	1.0000	1.178	0.694	2.000
komp brak komputera od komputer	1.0000	0.512	0.296	0.886
syt_materialna bardzo dobra od zła	1.0000	0.519	0.039	6.953
syt_materialna przeciętna od zła	1.0000	1.993	0.914	4.342
syt_materialna raczej dobra od zła	1.0000	1.633	0.604	4.418
syt_materialna raczej zła od zła	1.0000	1.640	0.695	3.872
stan_cyw kawaler, panna od zonaty, mezatk	1.0000	1.123	0.462	2.726
stan_cyw rozwiedziony(a od zonaty, mezatk	1.0000	2.141	0.721	6.362
stan_cyw wdowiec, wdowa od zonaty, mezatk	1.0000	0.438	0.167	1.146
fast często od rzadko lub prawie nigdy	1.0000	178.654	93.057	342.987

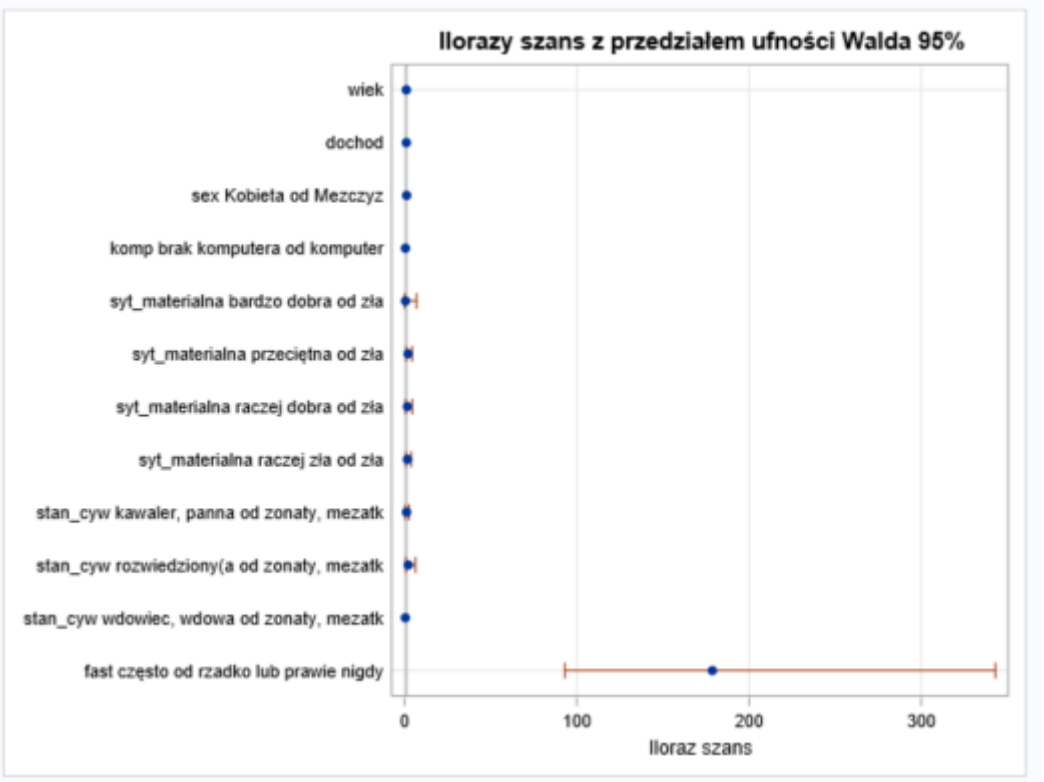
Oceny punktowe dopasowanych ilorazów szans, gdzie jeśli przedział ufności zawiera 1, to oszacowanie jest nie istotne statystycznie i nie podlega interpretacji. Tym samym, w analizowanym modelu, jedynie zmienne: komp oraz fast są istotne statystycznie i podlegają interpretacji.

Interpretacje istotnych statystycznie zmiennych:

Komp: Osoby posiadające komputer mają o [(0.512 -1) *100%] = 48.8% mniejsze szanse na bycie otyłym niż osoby posiadające komputer

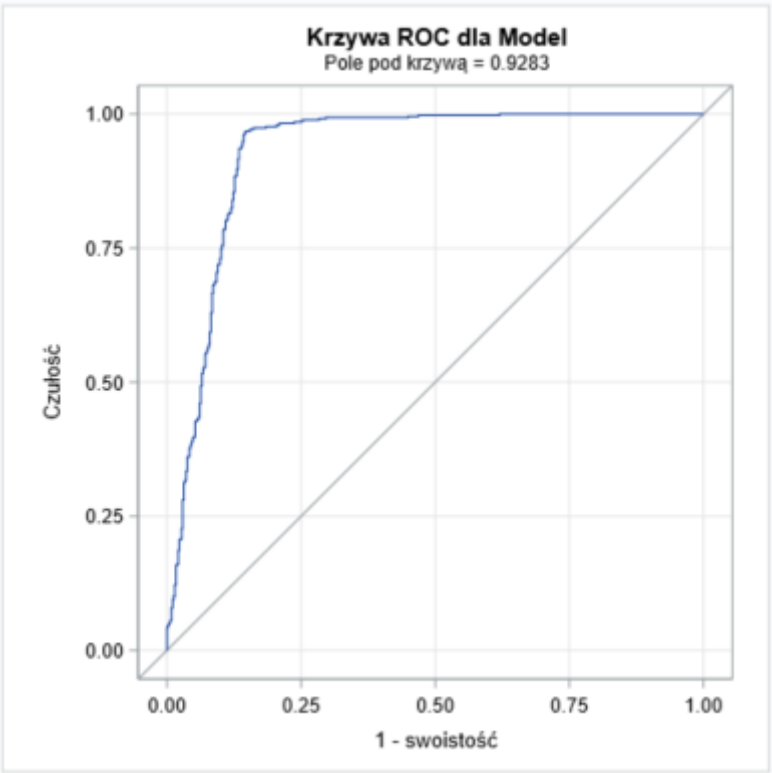
Fast: Osoby często spożywające jedzenie typu fast food mają 178.654 razy większe szanse na bycie otyłym niż osoby spożywające jedzenie typu fast food rzadko lub prawie nigdy.

Wykres ilorazów szans z przedziałem ufności Walda 95%



Wykres obrazuje zmienne wraz z ich przedziałami ilorazu szans. Jak już uprzednio zostało wspomniane, istotne są jedynie te zmienne, których przedział nie zawiera 1 (komp, fast).

Krzywa ROC dla modelu



.....

Test Hosmera i Lemeshowa

Miejsce na test Hosmera i Lemeshowa					
		nadwaga = 1		nadwaga = 0	
Grupa	Suma	Obserwowane	Oczekiwane	Obserwowane	Oczekiwane
1	94	0	1.35	94	92.65
2	94	1	2.30	93	91.70
3	94	2	3.09	92	90.91
4	94	5	4.37	89	89.63
5	94	46	40.30	48	53.70
6	94	79	78.06	15	15.94
7	94	83	80.16	11	13.84
8	94	80	81.93	14	12.07
9	94	84	85.46	10	8.54
10	97	87	89.97	10	7.03

.....

Test zgodności Hosmera i Lemeshowa

Test zgodności Hosmera i Lemeshowa		
Chi-kwadrat	DF	Pr. > chi-kw.
6.7687	8	0.5618

.....