Prognozowanie i symulacje (lab. 11)
Prognozowanie zmiennej dychotomicznej – regresja logistyczna

Wstęp

Przedmiotem przewidywania/prognozowania może być nie tylko przyszła wartość cechy liczbowej (bezrobocia, liczby wypadków czy przewozów transportem kolejowym), ale także sam fakt występowanie pewnego zjawiska. Prognozy takie polegają na modelowaniu prawdopodobieństwa wystąpienia pewnego zdarzenia jako funkcji czynników niezależnych.
W medycynie można na przykład prognozować w ten sposób prawdopodobieństwo wystąpienia powikłań po zabiegu operacyjnym w zależności od czasu trwania operacji, wieku chorego, płci, masy ciała i wartości innych parametrów laboratoryjnych. Modele tego typu są stosowane w ocenie ryzyka kredytowego, gdzie zmienna zależna także ma charakter dychotomiczny (0 – kredyt spłacony, 1 – kredyt niespłacony).
Do prognozowania wartości zmiennej dychotomicznej nie można stosować wcześniej poznanych metod analizy (np. regresji), ponieważ funkcja liniowa, kwadratowa czy wykładnicza przyjmują wartości spoza przedziału (0; 1). Dlatego też w celu modelowania prawdopodobieństwa wystąpienia pewnego zdarzenia wykorzystuje się **funkcję logistyczną**. Funkcja ta, dla wielu zmiennych niezależnych (X_1, …, X_k) i zmiennej zależnej Y, jest następującej postaci:

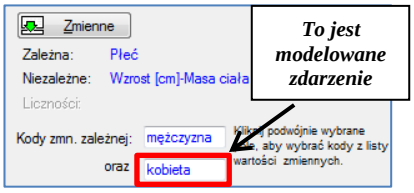
$$P(Y=1) = \frac{e^{b_0+b_1X_1+\ldots+b_kX_k}}{1+e^{b_0+b_1X_1+\ldots+b_kX_k}}$$

Ta „sprytna" funkcja przyjmuje wartości tylko z przedziału 0-1, nadaje się więc do modelowania prawdopodobieństwa.

Przykład 1 (*Dane antropometryczne*) – *ZARZĄDZANIE BEZPIECZEŃSTWEM*

Celem analizy będzie próba określenia płci na podstawie wartości wzrostu i masy ciała. Zagadnienie to może mieć zastosowanie na przykład w kryminalistyce (identyfikacja płci osoby na podstawie danych antropometrycznych), choć przykład ma przede wszystkim charakter poglądowy. Wyobraźmy sobie, że mamy dane o potencjalnych przestępcach dotycząc ich masy ciała i wzrostu (np. z kartoteki szpitalnej), ale zagubieniu (wykasowaniu) uległy dane o płci. Spróbujemy sklasyfikować te osoby, jako **kobiety** lub **mężczyzn** – czyli innymi słowy, będziemy „prognozować" płeć.
Etapy analizy:

1) Po otworzeniu pliku danych za pomocą poleceń ***STATYSTYKA / Zaawansowane modele liniowe i nieliniowe / Estymacja nieliniowa*** a następnie ***Szybka regresja logistyczna*** znajdujemy narzędzie modelowania i prognozowania zmiennej dychotomicznej.
2) Na liści zmiennych wskazujemy ***Płeć*** (jako zmienną zależną) oraz ***Wzrost*** i ***Masę ciała*** jako zmienne niezależne – przechodzimy do kolejnego okna za pomocą przycisku *OK* – po wyborze zmiennych należy zwrócić uwagę na wybór kodów zmiennej zależnej – proszę zapamiętać, że program będzie modelował **prawdopodobieństwo wystąpienia wartości znajdującej się w drugim polu** (a więc w naszym przypadku wyliczone wartości będą oznaczały prawdopodobieństwo sklasyfikowania nieznanej osoby na podstawie danych o wzroście i masie ciała jako kobiety).



3) Za pomocą przycisku *OK* dokonujemy wyliczenia parametrów modelu i dokonujemy interpretacji wyników według następującego schematu:
   - w zakładce ***Podstawowe*** wywołujemy tabelę *Parametry i błędy standardowe* – ważniejsze wyniki zamieszczono obok;
   - zarówno wzrost jak i masa ciała pozwalają w istotny statystycznie sposób prognozować płeć (*p* jest zdecydowanie mniejsze niż 0,05);
   - ***Iloraz szans z jednością***, określa jak zmienia się „szansa", iż nieznana nam osoba jest kobietą gdy wzrost tej osoby będzie większy o 1 cm (a masa ciała większa o 1 kg) – oczywiście, zgodnie z oczekiwaniami, ilorazy szans są mniejsze niż 1, czyli osoby wyższe i cięższe są z mniejszym prawdopodobieństwem płci żeńskiej (a z większym prawdopodobieństwem mogą być sklasyfikowane jako mężczyźni).

N=666	Stała B0	Wzrost [cm]	Masa ciała [kg]
Ocena	63,8	-0,32	-0,11
p	0,0000	0,0000	0,0000
Iloraz szans z.jedn.		0,73	0,89
-95%CL		0,67	0,86
+95%CL		0,78	0,93

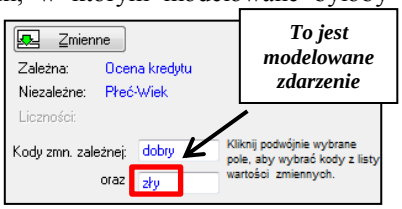
4) Bardzo użytecznym wynikiem jest ocena poprawności klasyfikacji danych na podstawie uzyskanego modelu – w tym celu w zakładce ***Reszty*** klikamy przycisk ***Klasyfikacja przypadków…*** – dowiadujemy się, iż na podstawie posiadanych informacji istnieje możliwość poprawnego sklasyfikowania mężczyzn z prawdopodobieństwem 86,6% zaś kobiet z prawdopodobieństwem 95,0%. Wartości te świadczą o dobrej jakości stworzonego modelu.
5) Aby wyliczyć prawdopodobieństwo, iż dla określonej wartości wzrostu i masy ciała dana osoba jest płci żeńskiej skorzystamy z gotowej formuły dostępnej na wykresie – zakładka ***Więcej*** i przycisk ***Dopasowana funkcja 3W…*** **Aby skopiować wzór klikamy dwa razy powierzchnię na wykresie i kopiujemy wzór z okienka *Z(x, y)*.**
6) W arkuszu danych dodajemy jedną zmienną na końcu arkusza i nazywamy ją *Prawdopodobieństwo sklasyfikowania jako osoby płci żeńskiej*, zmieniamy nazwę zmiennej wzrost na *X* a zmiennej masa ciała na *Y*, po czym wklejamy odpowiednią formułę a następnie wyliczamy wartości tego prawdopodobieństwa – jeżeli wartość prawdopodobieństwa przekracza 0,5 dana osoba jest klasyfikowana jako kobieta w przeciwnym wypadku klasyfikujemy ją jako mężczyznę.

7) Załóżmy, że dysponujemy informacją, iż poszukiwany przestępca miał 177 cm wzrostu i 55 kg wagi. Proszę dodać nowy przypadek w arkuszu danych, wprowadzić te wartości – prawdopodobieństwo sklasyfikowania danej osoby jako kobiety powinno wyświetlić się automatycznie (jeżeli nie, należy nacisnąć klawisz F9 w celu ponownego przeliczenia formuły). Proszę uzupełnić poniższą tabelę dla kolejnych przykładowych kombinacji wzrostu i wagi.

Wzrost	Masa ciała	Prawdopodobieństwo sklasyfikowania jako kobieta	Decyzja prognosty (kobieta lub mężczyzna)
177	55		
166	70		
181	76		

Przykład 2 (*Ryzyko kredytowe*) – *ZARZĄDZANIE RYZYKIEM KREDYTOWYM*

Przedmiotem analizy jest ocena ryzyka kredytowego. Dane są skrajnie uproszczone i zawierają tylko dwa czynniki niezależne: *płeć* i *wiek*. Proszę skonstruować model regresji logistycznej uwzględniający te czynniki, w którym modelowane byłoby prawdopodobieństwo, iż kredyt okaże się zagrożony („zły"). Przypomnijmy, że prognozowane jest prawdopodobieństwo wystąpienia zdarzenia odpowiadającego drugiemu kodowi wskazywanemu na liście kodów – **dlatego też kody te należy określić tak jak na rysunku obok.**



Proszę udzielić odpowiedzi na następujące pytania:

- Czy wpływ wieku i płci na wystąpienie „złego" kredytu jest istotny statystycznie – proszę odczytać wartości *p* dla obu tych zmiennych i ocenić istotność obu tych czynników;
- Proszę ocenić wpływ wieku na zagrożenie kredytu – proszę uzupełnić zdanie: *Iloraz szans z jednością „złego" kredytu dla wieku wynosi …….. Oznacza to, że wraz z wiekiem ryzyko, iż udzielony kredyt nie będzie spłacany w terminie …….. Kredytobiorca starszy o rok to …… o ……% ryzyko niespłacenia kredytu.*
- Jaka jest relacja zagrożenia spłacalności kredytu udzielanego mężczyźnie i kobiecie. Proszę podać i zinterpretować wartość odpowiedniego ilorazu szans: ……
- W analogiczny sposób jak w pkt. 5)-7) w poprzednim przykładzie proszę wyliczyć prawdopodobieństwo niespłacenia kredytu w terminie przez osoby wyszczególnione w poniższej tabeli.

Wiek	Płeć	Prawdopodobieństwo sklasyfikowania kredytu jako „zły"	Decyzja analityka[1)]
20 lat	mężczyzna		
20 lat	kobieta		
45 lat	mężczyzna		

[1)] Przyjmujemy, że bank nie udzieli kredytu jeśli modelowane ryzyko niespłacenia kredytu jest **większe niż 35%**

Przykład 3 (*Efekty rehabilitacji*) – *ZARZĄDZANIE W SŁUŻBIE ZDROWIA*

W bazie danych znajdują się informacje o efektach rehabilitacji pacjentów po udarze mózgu. Końcowa sprawność każdego chorego została oceniona na skali punktowej, a następnie sklasyfikowany jako „wysoka" albo „niska" (zmienna 6). Celem analizy jest prognozowanie końcowej sprawności chorych na podstawie informacji o wyjściowej sprawności chorych (zmienna 4) i ich wieku. W tym celu proszę wykorzystać model **regresji logistycznej**.

Na podstawie uzyskanego modelu proszę oszacować prawdopodobieństwo uzyskania „wysokiej" sprawności po zakończeniu rehabilitacji i podać ostateczną prognozę dla trzech nowych pacjentów:

Sprawność przed rehabilitacją	Wiek	Prawdopodobieństwo „wysokiej" sprawności po rehabilitacji	Decyzja prognosty – przewidywana końcowa sprawność (niska lub wysoka)[1)]
14 pkt	55 lat		
8 pkt	84 lata		
7 pkt	34 lat		

[1)] Jeśli nie ma innych wytycznych przyjmujemy jako próg decyzyjny wartość 0,50 (50%).

UWAGA: Na wykładzie omawiany będzie sposób wyliczania prawdopodobieństwa zdarzeń na podstawie modeli z trzema lub czterema czynnikami. Taki rozszerzony zakres umiejętności obowiązuje na egzaminie!