Analiza danych ankietowych
Sprawozdanie 6

Justyna Niedźwiedzka 229877

12 czerwca 2020

Zadanie 1

W tym zadaniu sprawdzimy, czy wybór kategorii bazowej w uogólnionym modelu logitowym ma duży wpływ na oszacowania odpowiednich prawdopodobieństw oraz czy oszacowania uzyskane z estymacji parametrów uogólnionego modelu logitowego (tzw. estymacji "jednoczesnej") różnią się znacznie od tych uzyskanych z estymacji kilku modeli logitowych dla danych binarnych.

Uogólniony model logitowy, nazywany również uogólnionym modelem logitowym o kategorii bazowej, w terminach prawdopodobieństw jest postaci

$$p_j(\mathbf{x}) = \frac{exp(\beta_j^T \mathbf{x})}{\sum_{i=1}^{J} exp(\beta_i^T \mathbf{x})}, \tag{1}$$

gdzie $\beta_j = (\beta_{j1}, \dots \beta_{jp})^T$ jest j-tym wektorem nieznanych współczynników, związanych z prawdopodobieństwem p_j, $j = 1, \dots, J$. W celu identyfikacji modelu, czyli możliwości estymacji jego parametrów, przyjmuje się, że β_J jest wektorem zer.

Dane w pliku *Gator.csv* zawierają informacje o długości aligatorów (zmienna *Długość*) oraz preferowanym przez nich pożywieniu (zmienna *Pożywienie*, przyjmująca wartości *bezkręgowce, ryby, inne*). Za zmienną objaśnianią przyjmujemy *Pożywienie*, natomiast za zmienną objaśniającą- *Długość*. Na podstawie tych danych oszacujemy parametry uogólnionego modelu logitowego korzystając z estymacji "jednoczesnej" dla danych wielomianowych oraz z estymacji współczynników dwóch modeli logitowych dla danych dwumianowych. Rozpatrzymy następujące przypadki:

(a) kategoria "inne" jest kategorią bazową,

(b) kategoria "ryby" jest kategorią bazową.

(a) Kategoria "inne" jest kategorią bazową.

Uogólniony model logitowy o kategorii bazowej "inne" dla danych wielomianowych jest postaci:

```
> model_mlogit_inne <- mlogit(Pozywienie ~ 0 | Dlugosc, data = gator,
+                             shape = "wide", reflevel = "Inne")
>
```

Tabela 1: Estymacja "jednoczesna" dla danych wielomianowych

parametr	oszacowana wartość
Bezkręgowce	5.69744
Ryby	1.61773
Długość:Bezkręgowce	-2.46545
Długość:Ryby	-0.11011

Modele logitowe odpowiednio 1 oraz 2 dla danych dwumianowych są postaci:

```
> model_bilogit1_inne <- glm(Pozywienie ~ Dlugosc, data = subset(gator,
+     Pozywienie %in% c("Inne", "Bezkręgowce")), family = binomial())
> model_bilogit2_inne <- glm(Pozywienie ~ Dlugosc, data = subset(gator,
+     Pozywienie %in% c("Inne", "Ryby")), family = binomial())
>
```

Tabela 2: Estymacja współczynników dla danych dwumianowych

model	parametr	oszacowana wartość
model 1	Długość	-2.178759
model 2	Długość	-0.1085

(b) Kategoria "ryby" jest kategorią bazową.

Uogólniony model logitowy o kategorii bazowej "ryby" dla danych wielomianowych jest postaci:

```
> model_mlogit_ryby <- mlogit(Pozywienie ~ 0 | Dlugosc, data = gator,
+                             shape = "wide", reflevel = "Ryby")
>
```

Tabela 3: Estymacja "jednoczesna" dla danych wielomianowych

parametr	oszacowana wartość
Bezkręgowce	4.07971
Inne	-1.61773
Długość:Bezkręgowce	-2.35534
Długość:Inne	0.11011

Modele logitowe odpowiednio 3 oraz 4 dla danych dwumianowych są postaci:

```
> model_bilogit1_ryby <- glm(Pozywienie ~ Dlugosc, data = subset(gator,
+     Pozywienie %in% c("Ryby", "Bezkręgowce")), family = binomial())
> model_bilogit2_ryby <- glm(Pozywienie ~ Dlugosc, data = subset(gator,
+     Pozywienie %in% c("Ryby", "Inne")), family = binomial())
>
```

Tabela 4: Estymacja współczynników dla danych dwumianowych

model	parametr	oszacowana wartość
model 3	Długość	2.477518
model 4	Długość	-0.1085

W poniższej tabeli przedstawiono oszacowania prawdopodobieństw wyboru bezkręgowców przez aligatora o długości $2m$. Uwzględniono podział ze względu na kategorię bazową oraz sposób szacowania.

Tabela 5: Oszacowania prawdopodobieństw wyboru bezkręgowców przez aligatora o długości $2m$

	Kategoria "inne"	Kategoria "ryby"
estymacja dla danych wielomianowych	0.299	0.299
estymacja dla danych dwumianowych	0.301	0.272

Widzimy, że wybór kategorii bazowej ma znaczenie w przypadku estymacji prawdopodobieństwa dla danych dwumianowych. Oszacowania uzyskane z estymacji "jednoczesnej" nie różnią się znacznie od tych uzyskanych na podstawie estymacji kilku modeli logitowych dla danych binarnych.

Zadanie 2

W tym zadaniu sprawdzimy, czy przyjęcie zmiennej objaśniającej za zmienną ciągłą czy jakościową ma wpływ na oszacowania prawdopodobieństw w modelu proporcjonalnych szans.

Model proporcjonalnych szans jest postaci

$$log\frac{\gamma_j(\mathbf{x})}{1-\gamma_j(\mathbf{x})} = \theta_j + \beta^T\mathbf{x}, \tag{2}$$

gdzie $\gamma_j(\mathbf{x}) = P(Y \leq j|\mathbf{x})$, $j = 1, \ldots, J-1$. Żeby model miał sens, czyli żeby $p_j \geq 0$, zakładamy $\theta_1 \leq \theta_2 \leq \ldots \leq \theta_{J-1}$.

Dane w pliku *Samopoczucie.csv* zawierają informacje o samopoczuciu 40-tu badanych (zmienna *Samopoczucie*), ich statusie ekonomicznym (zmienna *Status*) oraz liczbie zdarzeń takich jak zmiana pracy, urodzenie dziecka, ślub itp. w ostatnich 3 latach (zmienna *Liczba zdarzeń*). Zmienna *Samopoczucie* przyjmuje 4 wartości w skali od 1 do 4, gdzie 1 oznacza najlepsze, a 4- najgorsze. Zmienna *Status* przyjmuje 2 wartości- 0 i 1, a zmienna *Liczba zdarzeń* 10 wartości w skali od 0 do 9. Jako zmienną zależną przyjmujemy *Samopoczucie*, a zmienne niezależne- *Status* i *Liczba zdarzeń*. Na podstawie tych danych oszacujemy parametry modelu proporcjonalnych szans. Rozpatrzymy następujące przypadki:

(a) zmienna *Liczba zdarzeń* jest predyktorem ciągłym,

(b) zmienna *Liczba zdarzeń* jest predyktorem jakościowym.

(a) *Liczba zdarzeń* jest predyktorem ciągłym.

W poniższej tabeli przedstawiono oszacowane parametry modelu proporcjonalnych szans w przypadku, gdy zmienna *Liczba zdarzeń* przyjmuje ciągłe wartości.

Tabela 6: Parametry modelu proporcjonalnych szans dla predyktora ciągłego

parametr	oszacowana wartość
Status	-1.1112310
Liczba zdarzeń	0.3188613

(b) *Liczba zdarzeń* jest predyktorem jakościowym.

Aby zmienna *Liczba zdarzeń* była predyktorem jakościowym, utworzymy następujące kategorie: 0-4 zdarzeń, 5-9 zdarzeń. W poniższej tabeli przedstawiono oszacowane parametry modelu proporcjonalnych szans dla zmiennej skategoryzowanej.

Tabela 7: Parametry modelu proporcjonalnych szans dla predyktora jakościowego

parametr	oszacowana wartość
Status	-1.065346
Liczba zdarzeń	1.333669

Teraz porównamy oszacowania skumulowanych prawdopodobieństw przy różnych wartościach zmiennej *Status* oraz wybranych wartościach liczby zdarzeń i kategorii liczby zdarzeń.

Tabela 8: Wartości skumulowanych prawdopodobieństw dla $Status = 0$ i liczby zdarzeń od 0 do 4

Typ zmiennej *Liczba zdarzeń*				
ciągła	0.2935834	0.6336943	0.8187148	1.0000000
jakościowa	0.2663148	0.6017586	0.7926061	1.0000000

Tabela 9: Wartości skumulowanych prawdopodobieństw dla $Status = 1$ i liczby zdarzeń od 0 do 4

Typ zmiennej *Liczba zdarzeń*				
ciągła	0.5454794	0.8346781	0.9306580	1.0000000
jakościowa	0.5129834	0.8142921	0.9172877	1.0000000

Możemy zauważyć, że oszacowane prawdopodobieństwa przyjmują mniejszą wartość, jeśli skategoryzujemy zmienną objaśniającą.

Zadanie 3

Modele proporcjonalnego hazardu pozwalają przewidywać czas do wystąpienia jakiegoś zdarzenia, np. śmierci czy awarii urządzenia. Model proporcjonalnego hazardu (inaczej model Coxa) jest postaci

$$log[-log(1-\gamma_j(\mathbf{x}))] = \theta_j + \beta^T\mathbf{x}, \tag{3}$$

gdzie $j = 1, \ldots, J-1$. W tym modelu należy założyć, że $\theta_1 \leq \theta_2 \leq \ldots \leq \theta_{J-1}$.

W tym zadaniu korzystamy z danych z pliku *Zycie.csv*, które zawierają informacje o liczbie osób (zmienna *Liczności*) i długości życia w pięciu kategoriach wiekowych (zmienna *Długość życia*) w zależności od płci i rasy. Zmienna *Płeć* przyjmuje 2 wartości: 0 dla mężczyzn i 1 dla kobiet. Zmienna *Rasa* jest zmienną binarną przyjmującą wartość 0 dla rasy białej oraz 1 dla rasy czarnej. Na podstawie tych danych oszacujemy parametry modelu proporcjonalnego hazardu. Za zmienną zależną przyjmujemy *Długość życia*, a za zmienne niezależne- *Płeć* i *Rasa*.

Zbiór danych *Życie* zawiera 20 rekordów. Aby oszacować parametry modelu, musimy najpierw przekształcić nasz zbiór danych. Za pomocą funkcji *rep* tworzymy nowy zbiór z 400000 rekordami. Teraz nasz zbiór danych zawiera zmienne *Długość*, *Płeć* i *Rasa*. Model proporcjonalnego hazardu ma postać:

```
> model_hazardu <- polr(Dlugosc ~ Plec + Rasa, data = zycie2,
+                       method = c("cloglog"))
>
```

Poniżej przedstawiono oszacowane parametry modelu proporcjonalnego hazardu.

Tabela 10: Parametry modelu proporcjonalnego hazardu

parametr	oszacowana wartość
Płeć	0.657739
Rasa	-0.626438

Oszacujemy jeszcze prawdopodobieństwa przeżycia co najmniej k-tej kategorii ze względu na płeć i rasę.

Tabela 11: Prawdopodobieństwo przeżycia co najmniej k-tej kategorii w zależności od płci i rasy

	0-20	20-40	40-50	50-65	Powyżej 65
biała kobieta	0.9876566	0.9689921	0.9452025	0.8493471	0.0000000
czarna kobieta	0.9770305	0.9427707	0.8999298	0.7367549	0.0000000
biały mężczyzna	0.9763089	0.9410058	0.8969178	0.7296330	0.0000000
czarny mężczyzna	0.9561332	0.8924686	0.8158367	0.5544692	0.0000000

Widzimy, że prawdopodobieństwo przeżycia k-tej kategorii jest największe dla białej kobiety. Dla czarnej kobiety i białego mężczyzny prawdopodobieństwa te są zbliżone, natomiast najmniejsze jest dla czarnego mężczyzny.
Oznaczmy przez ρ_k prawdopodobieństwo przeżycia co najmniej k-tej kategorii dla białej kobiety. Sprawdzimy, że ρ_k^2 jest w przybliżeniu prawdopodobieństwem przeżycia dla czarnej kobiety i białego mężczyzny, a ρ_k^4 dla czarnego mężczyzny.
Dla $k = 1$: $\rho_1 = 0.9876566$, $\rho_1^2 \approx 0.9754656$, $\rho_1^4 \approx 0.9515331$.
Dla $k = 2$: $\rho_2 = 0.9689921$, $\rho_2^2 \approx 0.9389457$, $\rho_2^4 \approx 0.8816190$.
Dla $k = 3$: $\rho_3 = 0.9452025$, $\rho_3^2 \approx 0.8934078$, $\rho_3^4 \approx 0.7981774$.
Dla $k = 4$: $\rho_4 = 0.8493471$, $\rho_4^2 \approx 0.7213905$, $\rho_4^4 \approx 0.5204042$.