Analiza danych jakościowych

Projekt zaliczeniowy

Łukasz Kończyk

1. **Opis danych**

W nocy z 14 na 15 kwietnia 1912 roku doszło do zderzenia statku pasażerskiego Titanic w górą lodową.  Transatlantykiem podróżowało około dwóch tysięcy osób, z czego większość zginęła.

Dane do niniejszego projektu zamieszczone są na stronie <http://www.statsci.org/data/general/titanic.html>. Przedstawiają one dane 1313 pasażerów, których udało się zidentyfikować. W danych zawarto informacje o wieku, płci, klasie, w której podróżowali oraz czy przeżyli katastrofę. Pełna tabela danych dostępna jest w pliku titanic.txt. Do analizy danych będziemy używać następujących zmiennych:

* **PClass**: zmienna określa, którą klasą podróżował podróżny. Dostępne są trzy klasy: pierwsza (najwyższy standard), druga oraz trzecia (najniższy standard).
* **Age**: zmienna określająca wiek. Jest to zmienna ilościowa, jednak zostanie ona dyskretyzowana.
* **Sex**: zmienna określająca płeć: female- kobieta, Male- mężczyzna.
* **Survived**: indykator przeżycia katastrofy przez pasażera.

1. **Sformułowanie problemu badawczego**

Przy pomocy pakietu statystycznego R przeprowadzimy analizę zgromadzonych danych i postaramy się odpowiedzieć na pytania:

* Które zmienne mają wpływ na przeżycie pasażera?
* Które zmienne wpływają na oszacowanie, za pomocą modeli GLM, prawdopodobieństwa przeżycia pasażera?
* Ile wynosi oszacowane, za pomocą modeli GLM, prawdopodobieństwo przeżycia w każdej z grup?

1. **Analiza zależności pomiędzy zmiennymi**

Przystąpimy do badania zależności zmiennej Survived z pozostałymi zmiennymi. Będzie to zmienna objaśniana. Przeglądając dane, widzimy, że wśród danych mamy pasażerów z brakującym atrybutem wieku. Aby uniknąć problemów w dalszej części analizy, usuniemy obserwacje z brakującym wiekiem. Następnie poddamy dyskretyzacji dane dzieląc je na trzy przedziały wiekowe: mniej niż 20 lat, od 20 do 50 lat, więcej niż 50 lat. Będą one zachowane w nowej zmiennej wiek.

Ponieważ frakcja uratowanych z katastrofy jest mniejsza (wynosi 0.4140212), zgodnie z konwencją za sukces będziemy uważać to, że pasażer przeżył katastrofę.

Przejdźmy do analizy tablic kontyngencji.

Przyjrzyjmy się zależności zmiennej Survived od zmiennej PClass.

|  |  |  |  |
| --- | --- | --- | --- |
| PClass | Survived | | frakcja |
| 1 | 0 |
| 1st | 139 | 87 | 0.6150442 |
| 2nd | 96 | 116 | 0.4528302 |
| 3rd | 78 | 240 | 0.2452830 |

Widzimy, że im lepsza klasa rejsu, tym większa frakcja ocalałych. Można więc przypuszczać, że zmienna PClass ma istotny wpływ na zmienną Survived. Przeprowadzimy dokładny test Fishera oraz test chi-kwadrat w celu weryfikacji hipotezy o niezależności zmiennych. P-wartość dokładnego testu Fishera wynosi mniej niż 2.2e-16, co oznacza, że test zdecydowanie odrzuca hipotezę o niezależności zmiennych. Wartość statystyki testu chi-kwadrat wynosi 76.2808, co przy 2 stopniach swobody daje p-wartość mniejszą niż 2.2e-16. Test utwierdza nas w przekonaniu, że należy odrzucić hipotezę o niezależności zmiennych Survived oraz PClass.

Przeanalizujmy zmienną Survived w zależności od zmiennej Sex.

|  |  |  |  |
| --- | --- | --- | --- |
| Sex | Survived | | frakcja |
| 1 | 0 |
| Female | 217 | 71 | 0.7534722 |
| Male | 96 | 372 | 0.2051282 |

Analizując powyższą tabelę, widzimy, że frakcja kobiet, które przeżyły katastrofę, jest znacznie większa niż mężczyzn. Iloraz szans wynosi 11.84331. Na poziomie ufności 95% estymator ilorazu szans zawiera się w przedziale [8.350024; 16.798033]. Estymator jest znacznie większy od 1, należy więc przypuszczać, że zmienne nie są niezależne. W celu weryfikacji tej hipotezy, przeprowadzimy dokładny test Fishera oraz test chi-kwadrat. P-wartość dokładnego testu Fishera wynosi mniej niż 2.2e-16 co oznacza, że test odrzuca hipotezę o niezależności zmiennych. Wartość statystyki testu chi- kwadrat wynosi 218.7079, co przy jednym stopniu swobody daje p-wartość mniejszą niż 2.2e-16. Test potwierdza, że należy odrzucić hipotezę o niezależności zmiennych Survived oraz Sex.

Sprawdzimy jak wygląda zależność zmiennej Survived od zmiennej wiek.

|  |  |  |  |
| --- | --- | --- | --- |
| wiek | Survived | | frakcja |
| 1 | 0 |
| 0-20 | 87 | 85 | 0.5058140 |
| 20-50 | 197 | 317 | 0.3832685 |
| >50 | 29 | 41 | 0.4142857 |

Analizując tabelę, widzimy, że znaczną cześć podróżnych stanowili pasażerowie średniego wieku oraz frakcja ocalałych z katastrofy jest w tej grupie najmniejsza. Przeprowadzimy dokładny test Fishera oraz test chi-kwadrat w celu weryfikacji hipotezy o niezależności zmiennych. P-wartość testu Fishera wynosi 0.01917, co przy poziomie istotności 5% skłania do odrzucenia hipotezy o niezależności zmiennych. Podobny wynik otrzymujemy poprzez test chi-kwadrat. Wartość statystyki testowej wynosi 7.9774, co przy dwóch stopniach swobody daje p-wartość 0.01852. Reasumując, oba przeprowadzane testy odrzucają hipotezę o niezależności zmiennych Survived oraz wiek.

Podsumowując analizę tablic kontyngencji dochodzimy do wniosków:

* Przeżycie katastrofy Titanica mogło zależeć od tego, w której klasie podróżował pasażer.
* Płeć miała wpływ na przeżycie katastrofy
* Wiek podróżującego mógł mieć wpływ na to, czy przeżył on katastrofę.

1. **Specyfikacja i weryfikacja modeli GLM**

Po wstępnej analizie zależności pomiędzy zmiennymi, zauważyliśmy wpływ zmiennych PClass, Sex oraz wiek na zmienną objaśnianą Survived. Wszystkie te zmienne umieścimy w modelu estymującym prawdopodobieństwo przeżycia katastrofy Titanica. Szukamy prawdopodobieństwa przeżycia w zależności od każdej z grup. Użyjemy do tego modeli GLM. Najpierw rozważmy model logitowy postaci:

Z funkcją wiążącą postaci:

Wyestymowane prawdopodobieństwo wyliczymy ze wzoru:

Tabela1. Pierwszy model logitowy.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | z value | Pr(>|z|) |  |
| (Intercept) | 1.15712 | 0.17017 | 6.800 | 1.05e-11 | \*\*\* |
| Sexmale | -2.59971 | 0.19961 | -13.024 | < 2e-16 | \*\*\* |
| wiek.L | -1.14153 | 0.28223 | -4.045 | 5.24e-05 | \*\*\* |
| wiek.Q | -0.01004 | 0.17562 | -0.057 | 0.954 |  |
| PClass.L | 1.61905 | 0.18477 | 8.762 | < 2e-16 | \*\*\* |
| PClass.Q | -0.04895 | 0.16792 | -0.292 | 0.771 |  |

Po dopasowaniu modelu logitowego pełnego, po użyciu funkcji *summary*, widzimy, że poziom zmiennej wiek dla grupy (20,50] jest nieistotny. Oznacza to, że wartość oszacowanego parametru prawdopodobieństwa przeżycia jest taka sama, jak dla grupy (50, Inf). Podobnie mamy dla poziomu zmiennej PClass dla grupy 2nd. Spróbujemy ulepszyć model eliminując czynniki nieistotne.

Na początku, usuwamy całe zmienne, dla których poziomy mają współczynniki nieistotne. Poprzez funkcję *anova* porównujemy modele uproszczone z modelem pełnym. Za każdym razem odrzucana jest hipoteza, że model uproszczony jest lepszy niż pełny.

Od nowa dyskretyzujemy zmienną Age. Tym razem podzielimy ją na dwie grupy: mniej niż 20 lat oraz więcej niż 20 lat i nazwiemy wiek1. Po przeprowadzeniu tej operacji tablica kontyngencji zmiennych Survived oraz wiek1 wygląda następująco:

|  |  |  |  |
| --- | --- | --- | --- |
| wiek1 | Survived | | frakcja |
| 1 | 0 |
| 0-20 | 87 | 85 | 0.5058140 |
| >20 | 226 | 358 | 0.3869863 |

Zweryfikujemy hipotezę o niezależności zmiennych. Wyestymowana wartość ilorazu szans wynosi 1.621343. Na poziomie ufności 95% wartość ta zawiera się w przedziale [1.151529; 2.282837]. Wartość ilorazu szans jest większa od 1, co oznacza, że zmienne mogą nie być niezależne. P-wartość dokładnego testu Fishera wynosi 0.006289, co nasuwa wniosek, aby odrzucić hipotezę o niezależności zmiennych. Statystyka testu chi-kwadrat wynosi 7.251 i przy jednym stopniu swobody dostajemy p-wartość równą 0.007086. Również i ten test nakłania do odrzucenia hipotezy o niezależności zmiennych.

Tabela2. Drugi model logitowy

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | z value | Pr(>|z|) |  |
| (Intercept) | 1.4868 | 0.1647 | 9.030 | < 2e-16 | \*\*\* |
| Sexmale | -2.5887 | 0.1986 | -13.034 | < 2e-16 | \*\*\* |
| wiek1.L | -0.5979 | 0.1637 | -3.653 | 0.000259 | \*\*\* |
| PClass.L | 1.4879 | 0.1749 | 8.506 | < 2e-16 | \*\*\* |
| PClass.Q | -0.1068 | 0.1651 | -0.647 | 0.517806 |  |

W powyższym modelu nadal obserwujemy nieistotność współczynnika, dla drugiej klasy podróżnych. Wartość kryterium Akaike dla tego modelu stosunkowo niewiele wzrosła (z 716.16 na 720.07) w porównaniu do pierwszego rozważanego modelu. W celu pozbycia się z modelu współczynnika nieistotnego, stworzymy nową zmienną klasa, która połączy w sobie dwie najlepsze klasy, czyli 1st oraz 2nd ze zmiennej PClass. Tablica kontyngencji zmiennej Survived oraz zmiennej klasa wygląda następująco:

|  |  |  |  |
| --- | --- | --- | --- |
| klasa | Survived | | frakcja |
| 1 | 0 |
| gorsza | 78 | 240 | 0.2452830 |
| lepsza | 235 | 203 | 0.5365297 |

Zweryfikujemy hipotezę o niezależności zmiennych. Estymator ilorazu szans wynosi 0.2807447 . 95-procentowy przedział ufności [0.2044632; 0.3854854]. Iloraz szans jest znacznie mniejszy od jedynki, a jego wartość wskazuje, że szansa przeżycia katastrofy podróżując gorszą klasą była ponad trzykrotnie mniejsza od podróży lepszą klasą. P-wartość dokładnego testu Fishera wynosi 6.828e-16, co wskazuje, aby odrzucić hipotezę o niezależności zmiennych. Statystyka testu chi-kwadrat wynosi 63.2216 i przy jednym stopniu swobody p-wartość wynosi 1.847e-15, co również skłania do odrzucenia hipotezy o niezależności zmiennych. Spróbujemy dopasować model do nowych zmiennych

Tabela3. Trzeci model logitowy

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | z value | Pr(>|z|) |  |
| (Intercept) | 1.1713 | 0.1543 | 7.593 | 3.12e-14 | \*\*\* |
| Sexmale | -2.5569 | 0.1949 | -13.116 | < 2e-16 | \*\*\* |
| wiek1.L | -0.5008 | 0.1595 | -3.139 | 0.00169 | \*\* |
| klasa.L | 1.1425 | 0.1468 | 7.784 | 7.02e-15 | \*\*\* |

W powyższym modelu, każdy współczynnik jest istotny (czyli różny od 0) na poziomie istotności testu 5%. Wartość kryterium Akaike stosunkowo niewiele wzrosła do 733.47. Jednak ze względu na to, że w tym modelu mamy istotność współczynników, uznajemy go za najlepszy. Porównamy ten model z modelem zerowym. Test ilorazu wiarogodności wykonany funkcją *anova* odrzuca hipotezę, że model zerowy jest lepszy od powyższego modelu.

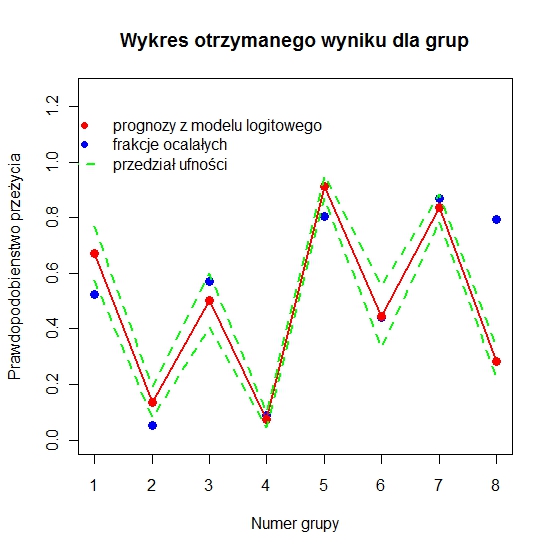
Dla powyższego zestawu zmiennych rozważymy model probitowy. W modelu tym nieznane prawdopodobieństwo przeżycia estymujemy funkcją wiążącą, którą jest dystrybuantą standardowego rozkładu normalnego.

Gdzie jest dystrybuantą rozkładu normalnego o wartości oczekiwanej zero oraz wariancji 1.

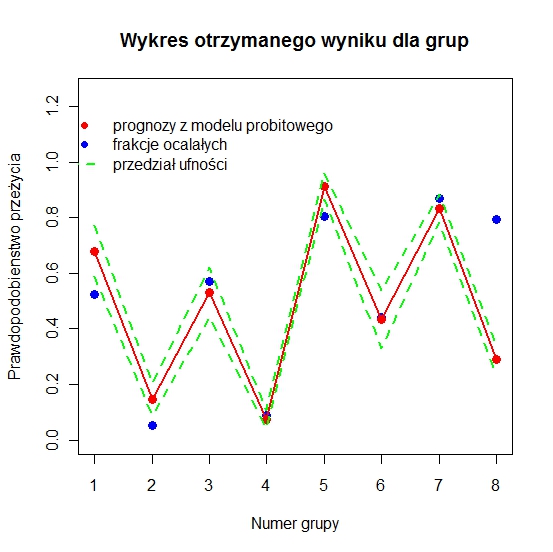
W poniższej tabeli przedstawimy oszacowane prawdopodobieństwo przeżycia katastrofy za pomocą obu wybranych modeli oraz prawdopodobieństwo empiryczne przeżycia dla każdej z zaobserwowanych grup.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Nr gr. | Sex | wiek1 | klasa | Sur | Non Sur | frakcja | logit | probit |
| 1 | female | (0,20] | gorsza | 23 | 21 | 0.52272727 | 0.67207609 | 0.67877489 |
| 2 | male | (0,20] | gorsza | 2 | 36 | 0.05263158 | 0.13713633 | 0.14599318 |
| 3 | female | (20,Inf] | gorsza | 33 | 25 | 0.56896552 | 0.50232297 | 0.52994154 |
| 4 | male | (20,Inf] | gorsza | 13 | 135 | 0.08783784 | 0.07258928 | 0.07452053 |
| 5 | female | (0,20] | lepsza | 45 | 11 | 0.80357143 | 0.91160317 | 0.91231791 |
| 6 | male | (0,20] | lepsza | 15 | 19 | 0.44117647 | 0.44435580 | 0.43530556 |
| 7 | female | (20,Inf] | lepsza | 139 | 21 | 0.86875000 | 0.83549324 | 0.83298160 |
| 8 | male | (20,Inf] | lepsza | 173 | 45 | 0.79357798 | 0.28255956 | 0.29046216 |

Analizując tabelę, widzimy, że oba modele podobnie estymują prawdopodobieństwo przeżycia. W większości grup prawdopodobieństwo empiryczne nie odbiega zbytnio od wartości wyestymowanej. Jedyny niepokój wzbudza grupa 8, gdyż prawdopodobieństwo empiryczne jest znacznie większe od wyestymowanych poprzez modele. Za pomocą wykresu pokażemy przedziały ufności na poziomie 95%, w których powinny się mieścić wyestymowane prawdopodobieństwa.

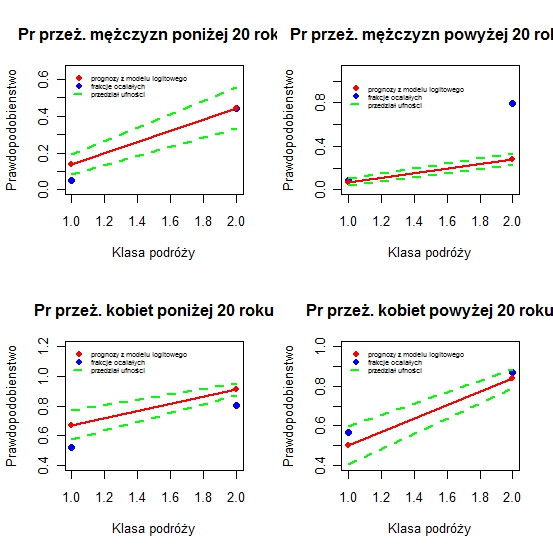


Z powyższego wykresu odczytujemy, że 4 wartości prawdopodobieństwa empirycznego mieszczą się w przedziale ufności prognozy z modelu logitowego. Kolejne trzy obserwacje znajdują się niewiele poniżej przedziału ufności. Jedna wyraźnie odbiega od przedziału ufności.



Dla modelu probitowego otrzymujemy podobną informację. Również 4 frakcje zawierają się w przedziale ufności dla diagnozy oraz 3 znajdują się niewiele poniżej przedziału. Jedna frakcja znacząco odbiega od przedziału ufności.

Poniższe wykresy pomogą ocenić jakość prognozy modelu logitowego. Wykresy dla modelu probitowego znajdują się w dołączonym pliku *projekt.R.*



1. **Interpretacja analiza statystycznej**

Obserwując wykresy, zauważamy, że klasa podróży miała znaczący wpływ na prawdopodobieństwo przeżycia zderzenia z górą lodową. Modele dość dobrze opisują prawdopodobieństwo empiryczne. Jedynie w przypadku mężczyzn powyżej 20 roku życia podróżujących pierwszą lub drugą klasą widzimy, że wyestymowane prawdopodobieństwo jest znacznie zaniżone w porównaniu do frakcji, która przeżyła katastrofę. Zauważmy również, że wyestymowane oraz empiryczne prawdopodobieństwa przeżycia katastrofy przez kobiety, jest wyższe niż mężczyzn. Bardzo ważny był także wiek. Pasażerowie młodsi mieli większą szansę na ocalenie.

1. **Podsumowanie**

Reasumując otrzymane wyniki za pomocą modeli GLM, widzimy, że oba modele podobnie estymują wartość szukanego prawdopodobieństwa, jednak oba odbiegają od danych empirycznych. Wpływ na to może mieć sposób pogrupowania ze względu na wiek. Wybrany model logitowy ma nieznacznie większą wartość kryterium Akaike w porównaniu innych rozważanych modeli, jednak posiada optymalną liczbę współczynników. Uznaliśmy to za kryterium wybrania optymalnego modelu. Wpływ na wynik również mogą mieć brakujące obserwacje wśród pasażerów trzeciej klasy. Od blisko połowy nie można było otrzymać informacji o wieku. Przeprowadziliśmy analizę bez zmiennej określającej wiek, jednak otrzymany model okazał się gorszy od modelu rozważanego w raporcie. Analiza znajduje się w pliku *projekt.R*

Otrzymany wynik zgadza się z rzeczywistością. Titanic był tak skonstruowany, iż na dolnych piętrach znajdowały się kabiny dla podróżujących trzecią klasą. Uderzając w górę lodową podróżujący w tej klasie odczuli zderzenie najmocniej i jako pierwsi. Jednocześnie w klasie pierwszej i drugiej zderzenie było już słabiej odczuwalne. Dodatkowo pasażerowie wyższych klas mieli dostęp do szalup, dzięki czemu mogli opuścić tonący statek. Nie dziwi fakt, że prawdopodobieństwo przeżycia kobiet i młodszych pasażerów jest wyższe. Mężczyźni, zachowując się honorowo, najpierw ratowali kobiety i dzieci.