Uniwersytet Ekonomiczny

Instytut Metod Ilościowych w Naukach Społecznych

Szansa na przetrwanie katastrofy na Titanicu
Analiza modelu logitowego i probitowego

Maciej Laburda
Magdalena Mazur

Analiza Zmiennych Jakościowych

Spis treści

1 Charakterystyka danych oraz wybrane statystyki opisowe **3**
- 1.1 Cel badań . . . 3
- 1.2 Opis danych . . . 3
- 1.3 Wybrane Statystyki Opisowe . . . 5

2 Opis modelu, metodologia oraz wstępne wyniki estymacji **7**
- 2.1 Model logitowy . . . 7
  - 2.1.1 Opis modelu logitowego . . . 7
  - 2.1.2 Wstępne wyniki estymacji logitu . . . 8
- 2.2 Model Probitowy . . . 8
  - 2.2.1 Opis modelu probitowego . . . 8
  - 2.2.2 Wstępne wyniki estymacji probitu . . . 9
- 2.3 Narzędzia i interpretacje . . . 9
  - 2.3.1 Testowanie hipotez . . . 9
  - 2.3.2 Efekty krańcowe . . . 10
  - 2.3.3 Wybrane mierniki dopasowania . . . 11

3 Estymacja modelu i interpretacje wyników **13**
- 3.1 Pierwsze wyniki . . . 13
  - 3.1.1 Interpretacja estymacji β . . . 13
  - 3.1.2 Dopasowanie modeli do danych . . . 14
- 3.2 Redukcja modelu . . . 15
  - 3.2.1 Oceny parametrów β z restrykcjami . . . 15
  - 3.2.2 Test LR . . . 16
- 3.3 Ostateczne modele i zbiór testowy . . . 17
  - 3.3.1 Postać modeli testowych i ich dopasowanie . . . 17
  - 3.3.2 Zestawienie współczynników dopasowania . . . 19
  - 3.3.3 Efekty krańcowe dla modeli testowych . . . 19

Rozdział 1

Charakterystyka danych oraz wybrane statystyki opisowe

1.1 Cel badań

Celem naszego badania jest zbadanie wpływu wybranych zmiennych objaśniających (tj. klasy podróżnej, płci, wieku, liczby członków rodziny na pokładzie, ceny biletu oraz portu źródłowego) na szansę przetrwania katastrofy na Titanicu, która miała miejsce w nocy z 14 na 15 kwietnia 1912 roku za pomocą modelu logitowego. Pozwoli to wyjaśnić, które z danych zmiennych mają istotny wpływ na szansę przeżycia pasażerów.

1.2 Opis danych

Dane pochodzą ze zbioru "Titanic - Machine Learning from Disaster", pochodzącego ze strony: `www.kaggle.com`. Zmienna objaśniana (Survived) jest zmienną binarną i oznacza to, czy dany pasażer przeżył (0 = Nie , 1 = Tak).
W zbiorze występują następujące zmienne objaśniające:

1. **Klasa podróżna** (*Pclass*1, *Pclass*2) - reprezentuje klasę kupionego biletu. Występują 3 klasy biletu, gdzie klasa pierwsza oznacza najwyższy standard podróży. Z racji, że nasze dane reprezentują zmienną jakościową i nie da się na skali ilorazowej przedstawić różnic między pierwszą, drugą i trzecią klasą, to nasze wartości przekodowaliśmy na dwie zmienne logiczne. Pierwsza zmienna *Pclass*1 oznacza, że osoba podróżowała pierwszą klasą, a *Pclass*2 oznacza podróżnika w drugiej klasie. W zbiorze danych występowała 3 wartość odnosząca się do klasy ekonomicznej, której wartości będzie przechwytywać wyraz wolny.

2. **Płeć** (*Sex*) - zmienna binarna, gdzie 0 oznacza kobietę, a 1 mężczyznę.

3. **Wiek** (*Age*) - wiek pasażera wyrażony w latach. Jeśli pasażer ma mniej niż rok, to jego wiek zapisany jest w postaci ułamka dziesiętnego, co oznacza ilość miesięcy.

W przypadku, gdy wiek jest oszacowany, przedstawiany jest formie xx.5, na przykład 43.5. To rozróżnienie pomaga uwzględnić sytuacje, w których dokładny wiek pasażera jest nieznany, a jedynie oszacowany.

4. **Liczba rodzeństwa / małżonków na pokładzie** (*SibSp*) - zbiór danych opisuje relacje rodzinne w następujący sposób

   - Rodzeństwo
     - brat
     - siostra
     - brat przyrodni
     - siostra przyrodnia
   - Małżonek
     - mąż
     - żona

5. **Liczba rodziców / dzieci na pokładzie** (*Parch*) - zbiór danych opisuje relacje rodzinne w następujący sposób:

   - Rodzice
     - ojciec
     - matka
   - Dzieci
     - syn
     - cóka
     - syn przyrodni
     - córka przyrodnia

   Niektóre dzieci podróżowały z opiekunką, wtedy wartość jest równa 0.

6. **Opłata pasażerska** (*Fare*) - cena biletu (podawana w funtach brytyjskich), którą zapłacił pasażer za rejs statkiem.

7. **Port źródłowy** (*EmbarkedS*, *EmbarkedC*) - oznacza port, w którym pasażer wsiadł na pokład statku, wskazując na miejsce rozpoczęcia podróży. Podzieliliśmy tę zmienną na dwie zmienne bianarne z portami w Cherbourg oraz Southampton. W oryginalnym zbiorze danych pojawiał się dodatkowo port w Queenstown, jednak ze względu na brak informacji odnośnie wieku pasażerów wsiadających w Queenstown postanowiliśmy nie uwzględniać ich w regresji.

1.3 Wybrane Statystyki Opisowe

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Survived	0	0	0	0.4045	1	1
Sex	0	0	1	0.6362	1	1
Age	0	20	28	29.64	38	80
SibSp	0	0	0	0.514	1	5
Parch	0	0	0	0.4326	1	6
Fare	0	8.05	15.65	34.57	33	512.33
EmbarkedS	0	1	1	0.7781	1	1
EmbarkedC	0	0	0	0.1826	0	1
Pclass1	0	0	0	0.2584	1	1
Pclass2	0	0	0	0.243	0	1

Tabela 1.1: Tabela statystyk opisowych

Powyższe statystyki opisowe dostarczają istotnych informacji na temat pasażerów na pokładzie Titanica. Warto zauważyć, że większość podróżujących nie przeżyła katastrofy, ponieważ mediana dla zmiennej $Survived$ wynosi 0. Średnia tej zmiennej, równa 0.4045, wskazuje, że około 40.45% pasażerów przeżyło katastrofę.

Analizując klasy podróżnicze, można zauważyć, że większość pasażerów podróżowała drugą klasą lub klasą niższą. Mediana dla zmiennej $Pclass1$ wynosi 0, co sugeruje, że około 25.84% pasażerów podróżowało pierwszą klasą.

Jeśli chodzi o płeć pasażerów, wyniki wskazują, że większość z nich to mężczyźni, gdyż mediana zmiennej Sex wynosi 1, a średnia to 0.6362, co oznacza, że 63.62% pasażerów to mężczyźni.

Analizując wiek pasażerów, warto zauważyć, że najmłodszy pasażer miał około 5 miesięcy, a najstarszy 80 lat. Średni wiek wynosił około 30 lat, a mediana sugeruje, że połowa pasażerów miała mniej niż 28 lat. Kwartyl pierwszy mówi nam o tym, 25% pasażerów było w wieku niższym niż 20 lat, a 75% wyższym. Z wartości kwartyla trzeciego dowiadujemy się o tym, że 75% pasażerów było w wieku niższym niż 38 lat, a 25% wyższym.

Przyjrzymy się teraz sytuacji rodzinnej pasażerów. Większość z nich podróżowała bez rodzeństwa, małżonka, dzieci czy rodziców na pokładzie. Średnia liczba małżonków i rodzeństwa wynosiła 0.514, a średnia liczba dzieci i rodziców wynosiła 0.4326.

W kontekście opłat pasażerskich, najmniejsza opłata wynosiła 0 funtów brytyjskich,co wynikało z różnych okoliczności, jako przykład można podać Thomasa Andrews, który był jednym z konstruktorów statku RMS Titanic. Największa opłata za bilet to 512.33 funtów. Mediana sugeruje, że połowa pasażerów zapłaciła za bilet więcej niż 15.65 funtów.

Jeśli chodzi o port, z którego pasażerowie rozpoczęli rejs, to przeważająca większość zaczęła podróż w Southampton, co stanowiło około 77.81% wszystkich pasażerów na pokładzie. Zdecydowana mniejszośc zaczynała swój rejs w porcie w Cherbourg. Było to około 18.26%.

Rozdział 2

Opis modelu, metodologia oraz wstępne wyniki estymacji

2.1 Model logitowy

2.1.1 Opis modelu logitowego

Modele logitowe są modelami regresyjnymi, opisującymi relację między zmienną objaśnianą, która ma charakter dychotomiczny (przyjmuje dwie wartości; jedną z nich nazywamy sukcesem). W tym modelu interesuje nas regresja między prawdopodobieństwem sukcesu wyrażonym w skali logitowej, a zmiennymi objaśniającymi.[1]
Rozpatrywaną w projekcie zmienną objaśnianą jest szansa przeżycia katastrofy na statku RMS Titanic.

Dla uproszczenia notacji zastosujemy zapis macierzowy:

$$X_t\beta = \beta_0 + \beta_1 Pclass1_t + \beta_2 Pclass2_t + \beta_3 Sex_t + ... + \beta_8 EmbarkedS_t + \beta_9 EmbarkedC_t$$

W modelach logitowych korzysta się z rozkładu logistycznego, a dokładniej mówiąc z dystrybuanty jego zestandaryzowanej wersji z wartością oczekiwaną równą zero i z wariancją równą $\frac{\pi^2}{3}$. W konsekwencji równanie na prawdopodobieństwo sukcesu będzie miało następującą postać:

$$p_t = \frac{e^{X_t\beta}}{1 + e^{X_t\beta}}$$

Gdzie p_t oznacza prawdpodobieństwo, że zmienna $Survived_t$ przyjmie wartość 1.

Do estymacji parametrów wykorzystana zostanie funkcjonalność języka R, która umożliwia przeprowadzenie regresji logitowej przy użyciu funkcji `glm`. Umożliwia ona estymację

parametrów dla wielu rodzajów modeli liniowych. W szczególności, parametr `family` pozwala na określenie rodzaju regresji, z którą mamy do czynienia. W tym wypadku zostanie wykorzystana rodzina `binomial`, która mówi nam, że zmienna objaśniana ma rozkład dwumianowy.

2.1.2 Wstępne wyniki estymacji logitu

Po podzieleniu zbioru danych na zbiór uczący i testowy, przeprowadziliśmy regresję z pomocą języka R, otrzymując następujące wyniki:

Zmienna	Oszacowanie	Błąd standardowy	Iloraz t	p-value
(Intercept)	1.214	0.583	2.084	0.037
Sex	-2.638	0.223	11.829	0.000
Age	-0.043	0.008	5.204	0.000
SibSp	-0.363	0.129	2.807	0.005
Parch	-0.060	0.124	0.487	0.626
Fare	0.001	0.003	0.559	0.576
EmbarkedS	0.421	0.556	0.756	0.450
EmbarkedC	0.823	0.600	1.372	0.170
Pclass1	2.395	0.343	6.976	0.000
Pclass2	1.206	0.250	4.827	0.000

Tabela 2.1: Wynik działania funkcji `glm` na zbiorze testowym

Na pierwszy rzut oka największy wpływ ma zmienna oznaczająca płeć. Wyższa wartość tej zmiennej odpowiadająca płci męskiej powoduje znaczące zmniejszenie się prawdpodobieństwa przetrwania katastrofy. Innymi znaczącymi zmiennymi na podstawie ilorazu t są wiek pasażera, klasa podróżna oraz ilość rodzeństwa i małżonków.

2.2 Model Probitowy

2.2.1 Opis modelu probitowego

Model probitowy to jedna z form modeli regresyjnych używanych do analizy zmiennych jakościowych, przyjmujących wartości zero-jedynkowe. Jest nieco bardziej skomplikowaną formą modelu logitowego, jednak jego założenie i sposób testowania hipotez jest identyczny. Główną róznicą między modelem logitowym jest jego postać funkcji prawdopodobieństwa. W przypadku logitu stosowaliśmy standaryzowaną regresję logistyczną, w tej wersji do obliczenia prawdopodobieństwa p_t użyjemy dystrybuanty standaryzowanego rozkładu normalnego:

$$E(Y|X_t) = Pr(Y = 1|X_t) = \Phi(X_t\beta)$$

Zatem prawdopodobieństwo p_t otrzymamy za pomocą:

$$p_t = \Phi(X_t\beta) = Pr(U \leq X_t\beta), U \sim N(0,1)$$

Obliczenie wektora ocen parametrów β będzie możliwe za pomocą biblioteki `glm2` w języku R. Estymacja będzie przebiegać analogicznie jak w przypadku modelu logitowego, stosujemy do tego funkcję `glm`, ze zdefiniowanym argumentem family na "probit".

2.2.2 Wstępne wyniki estymacji probitu

Po zastosowaniu funkcji glm dla probitu na naszym zbiorze uczącym otrzymujemy nastepujące wyniki wstępnej estymacji:

Zmienna	Oszacowanie	Błąd standardowy	Iloraz t	p-value
(Intercept)	0.736	0.337	2.182	0.029
Sex	-1.557	0.124	12.512	0.000
Age	-0.024	0.005	5.180	0.000
SibSp	-0.210	0.073	2.862	0.004
Parch	-0.048	0.073	0.663	0.507
Fare	0.001	0.002	0.633	0.527
EmbarkedS	0.239	0.318	0.752	0.452
EmbarkedC	0.482	0.343	1.405	0.160
Pclass1	1.340	0.194	6.896	0.000
Pclass2	0.665	0.141	4.711	0.000

Tabela 2.2: Wynik działania funkcji `glm` na zbiorze testowym

Istotność parametrów kształtuje się bardzo podobnie jak w przypadku logitu. Ponownie mamy wysoką istotność dla zmiennej płci i dla pierwszych dwóch klas poróży. Wiek podróżnika w tym modelu również jest destymulantą i możemy wnioskować, że starsze osoby miały mniejsze szanse na przeżycie katastrofy.

2.3 Narzędzia i interpretacje

2.3.1 Testowanie hipotez

W naszym modelu będziemy testować istotność poszczególnych elementów wektora ocen parametrów przy użyciu statystyki testu T-Studenta. Naszą ocenę punktową wektora beta będziemy dzielić przez poszczególne błędy szacunku oraz przy pomocy trzech konwencjonalnych wartości $\alpha = (0.1; 0.05; 0.01)$ ocenimy istotnośc naszych parametrów.

Drugim elementem testowania będzie sprawdzanie hipotez odnoszących się do łącznej

redukcji zmiennych w modelu. Do tego posłuży nam odpowiednik testu F w klasycznym modelu normalnej regresji liniowej, a dokładniej LR test (Likelihood Ratio). Będziemy poddawali restrykcjom pewną grupę parametrów zawartych w wektorze $\beta^{(0)}$ i za pomocą logarytmów ilorazu wiarygodności modelu z restrykcjami i modelu wyjściowego będziemy badać łączną istotność parametrów statystyką LR.

Układ hipotez prezentuje się nastepująco:

$$H_0 : \beta^{(0)} = \mathbf{0}$$
$$H_1 : \beta^{(0)} \neq \mathbf{0}$$

Wartość statystyki testowej obliczymy za pomocą wzoru:

$$LR = -2(ln(L_r(\hat{\beta_{MNW}}) - ln(L_y(\hat{\beta_{MNW}}))$$

Gdzie $ln(L_r(\hat{\beta_{MNW}})$ i $ln(L_y(\hat{\beta_{MNW}})$ oznaczają kolejno logarytmy funkcji wiarygodności dla modelu z restrykacjami i dla modelu bez restrykcji.

Decyzję podejmiemy na podstawie prawostronnego testu χ^2 o v stopniach swobody, przy v = (liczba parametrów modelu wyjściowego - liczba parametrów modelu zredukowanego).

$$LR \sim \chi^2_v | H_0$$

Jeżeli statystyka LR będzie większa od kwantyla rozkladu χ^2_α, to z perpektywy H_0 zajdzie zjawisko nieprawdopodobne i przyjmiemy prawdziwość hipotezy alternatywnej.

2.3.2 Efekty krańcowe

Efekty krańcowe informują nas o tym, o ile zmienia się prawdopodobieństwo, że zmienna objaśniana przyjmuje wartość 1 przy wzroście zmiennej objaśniającej o jedną jednostkę. Do obliczenia efektów krańcowych w przypadku obu modeli będziemy wyznaczali pochodną z dystrybuanty $F(X_t\beta)$ względem zmienenj x_{th}.

Przykładowy wzór na efekt krańcowy dla zmiennej Age_t:

$$\frac{\partial p_t}{\partial Age_t} = \frac{\partial F(X_t\beta)}{\partial Age_t} = f(X_t\beta)\beta_4$$

Gdzie $f(\cdot)$ jest funkcją gęstości rozkładu, a β_4 oceną parametru zmiennej Age_t.

Stosując przekształcenia analityczne jesteśmy w stanie wyznaczyć wzór na efekty krańcowe dla logitu:

$$\frac{\partial p_t}{\partial Age_{it}} = \beta_4 \frac{e^{-X_t\beta}}{(e^{-X_t\beta} + 1)^2}$$

W przypadku modelu probitowego nie da się zastosować podobnych przekształceń, co oznacza, że należy obliczać efekty krańcowe z funkcji gęstości $f(X_t\beta)$ dla standaryzowanego rozkładu normalnego.

Interpretacje efektów krańcowych dla zmiennych ilościowych, czyli Age_t, $Sibsp_t$, $Parch_t$, $Fare_t$, będą interpretowane we wczesniej wspomniany sposób: "Wraz ze wzrostem zmiennej o 1 jednostkę prawdopodobieństwo sukcesu wzrośnie o $f(X_t\beta)\beta_h$ jednostek".
W przypadku zmiennych przyjmujących wartości logiczne interpretacja będzie następująca: "Jeżeli wystąpiło dane zjawisko to prawdopodobieństwo sukcesu wzrośnie o $f(X_t\beta)\beta_h$."

2.3.3 Wybrane mierniki dopasowania

Pierwszym wykorzystanym miernikiem dopasowania modelu do danych będzie współczynnik determinacji Efron'a wyrażający się wzorem:

$$R^2_{Efrona} = 1 - \frac{\sum_{t=1}^{T}(y_t - \hat{p}_t)^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2} = 1 - \frac{\sum_{t=1}^{T}(y_t - \hat{p}_t)^2}{T \cdot \bar{y} \cdot (1 - \bar{y})}$$

Jako drugi miernik determinacji zastosujemy R^2_{MZ}, czyli współczynnik determinacji Zavoin'a i McKelvey'a w wersji dla logitu i probitu:

$$R^2_{MZ} = \frac{\sum_{t=1}^{T}(\hat{z}_t - \bar{\hat{z}})^2}{T + \sum_{t=1}^{T}(\hat{z}_t - \bar{\hat{z}})^2}$$

W wersji dla logitu w mianowniku powinnismy uwzględnić wariancję standaryzowanego rozkładu logistycznego, więć zamiast T będzie $T \cdot \frac{\pi^2}{3}$.

Trzecim miernikiem będzie tabela trafności i zliczeniowe R^2.

Do stworzenia tabeli trafności będziemy potrzebować rzeczywistych wartości y oraz tych oszacowanych $\hat{y}$ za pomocą naszego modelu. Do odpowiednich miejsc w tabeli wpisujemy ilości rzeczywistych i prognozowanych sukcesów (TP) oraz rzeczywiste porażki i prognozowane sukcesy (FP). Podobnie można postąpić w przypadku prognozowanych porażek. Przy prognozowanym $\hat{y}$ będziemy musieli arbitralnie ustalić progowe p^*, gdzie $p_t > p^*$ oznaczać będzie sukces, a $p_t \leq p^*$ porażkę.

		Klasa rzeczywista	
		pozytywna	**negatywna**
Klasa predykowana	**pozytywna**	prawdziwie pozytywna (TP)	fałszywie pozytywna (FP)
	negatywna	fałszywie negatywna (FN)	prawdziwie negatywna (TN)

Rysunek 2.1: Tabela trafności. Źródło: Wikipedia[2]

Na podstawie tabeli trafności jesteśmy w stanie sporządzić mierniki cząstkowe i zliczeniowe R^2, za pomocą których będziemy w stanie obliczyć trafnosci prognoz:

1. Miernik cząstkowy dla trafności prognoz $y_t = 1$:
$$\frac{TP}{TP + FN} \cdot 100\%$$
2. Miernik cząstkowy dla trafności prognoz $y_t = 0$:
$$\frac{TN}{FP + TN} \cdot 100\%$$
3. Zliczeniowe R^2 (count R^2) będące średnią ważoną mierników cząstkowych dla $y_t = 1$ i $y_t = 0$:
$$countR^2 = \frac{TP + TN}{T} \cdot 100\%$$
4. Odsetek nietrafionych prognoz dla $y_t = 1$ i $y_t = 0$:
$$\frac{FN + FP}{T} \cdot 100\%$$

Interpretacja każdego z R^2 będzie mówić jaki odsetek zmienności udało się wyjaśnić za pomocą oszacowanego modelu. W przypadku mierników cząstkowych będzie oznaczać to, jaki odesetek sukcesów i porażek (osobno) udało nam się wyjaśnić za pomocą naszego modelu). Miernik nietrafionych prognoz będzie oznaczać, jaki odsetek obserwacji nie udało się wyjasnić przez nasz model.

Rozdział 3

Estymacja modelu i interpretacje wyników

3.1 Pierwsze wyniki

3.1.1 Interpretacja estymacji β

Oszacowanie parametrów w modelach odbywa się za pomocą funkcji `glm` z pakietu "glm2"do języka R. Funkcja `glm` wyznacza wektor $\hat{\beta}$ za pomocą maksymalizacji funkcji wiarygodności. Otrzymane w ten sposób estymatory $\hat{\beta}_{MNW}$ mają własności asymptotyczne i wraz ze wzrostem liczebności naszych obserwacji stają się bardziej efektywne. Porównania wyników dla logitu i probitu są przedstawione w poniższej tabeli.

Zmienna	Logit	Probit	Logit - p-value	Probit - p-value
(Intercept)	1.214	0.736	0.037	0.029
Sex	-2.638	-1.557	0.000	0.000
Age	-0.043	-0.024	0.000	0.000
SibSp	-0.363	-0.210	0.005	0.004
Parch	-0.060	-0.048	0.626	0.507
Fare	0.001	0.001	0.576	0.527
EmbarkedS	0.421	0.239	0.450	0.452
EmbarkedC	0.823	0.482	0.170	0.160
Pclass1	2.395	1.340	0.000	0.000
Pclass2	1.206	0.665	0.000	0.000

Tabela 3.1: Porównanie parametrów i p-value w modelach

Oszacowanie obu modeli jest zgodne co do kierunku zmian wszystkich zmiennych. Posiadają jednak inne siły, co w różny sposob może wpłynąć na wartości oczekiwane w

rozkładach prawdopodobieństwa. Wartość jeden w zmiennej *Sex*, oznaczajaca płeć męską pasażera, będzie wpływać niekorzystnie na przeżycie katstrofy Titanica. Wartośc ujemna w zmiennych *Age* i *SibSb* informują, że wraz z większa ilością lat i liczbą członków rodziny na pokładzie nasze szanse na przeżycie katastrofy maleją. Pokrywa się to z kontekstem hisotrycznym, ponieważ wiemy, że w pierwszej kolejnosci ratowane były kobiety i dzieci. Duże rodziny mogły mieć problem z pomieszczeniem się na pokładzie łodzi ratunkowych i z przedostaniem się przez liczny tłum na statku. Dodatkowo zauważamy, że w obu modelach znacznie rosły szanse przeżycia, gdy ktoś był pasażerem pierwszej lub drugiej klasy.

Za pomocą p-value możemy wnioskowac o istotności poszczególnych parametrów. Jesteśmy w stanie stwierdzić, że zmienne *Sex*, *Age*, *SibSp*, $Pclass1$ i $Pcalss2$ mają istotny wpływ na zmienną objaśnianą przy każdym konwencjonalnym poziomie istotności α. Pozostałe parametry, poza wyrazem wolnym, który musi pozostać w modelu nie zależnie od jego istotności, będą testowane pod kątem ich łącznej istotności zanim zostaną zredukowane.

3.1.2 Dopasowanie modeli do danych

Pierwszym krokiem do wyznaczenia wartości zero-jedynkowych oznaczających sukces i porażkę w modelach zmiennych jakościowcyh będzie przeliczenie prawdopodobieństwa sukcesu na podstawie dystrybuant odpowiednich rozkładów. Kombinacja liniowa parametrów i zmiennych tworzą wartosć oczekiwaną do wyznaczenia gęstości prawdopodobieństwa dla modeli. W logicie użyjemy do tego dystrybuanty standaryzowanego rozkładu logistycznego, a w przypadku probitu standaryzowanego normalnego. W ten sposób otrzymamy wartości z przedziału od zera do jeden, które będą oznaczać szanse każdego poszczególnego pasażera na przeżycie katastrofy.

Następnym etapem jest ustalenie progowego p^*, za pomocą którego będziemy przypisywać wartość $Y = 1$ dla $p > p^*$ i wartośc $Y = 0$ dla $p \leq p^*$. W naszych obliczeniach zastosowaliśmy średnią wartosc zmiennej objaśnianej równą:

$$p = 0.4045$$

Po ustaleniu p^* i obliczeniu prognozowanych $\hat{Y}$ możemy przejśc do konstruowania tabeli trafności, a na jej podstawie zostanie obliczona miara dopasowania $countR^2$.

	Logit			Probit		
Prognozy	$y_t = 1$	$y_t = 0$	Suma	$y_t = 1$	$y_t = 0$	Suma
$\hat{y}_t = 1$	221	81	302	222	82	304
$\hat{y}_t = 0$	67	343	410	66	342	408
Suma	288	424	712	288	424	712

Tabela 3.2: Tabela trafności logitu i probitu

Za pomocą tabeli trafności jesteśmy w stanie zauważyć, że oba modele podobnie poradziły sobie z przewidzeniem wartości Y. Oba modele przewidziały kolejno 221 i 222 trafnych prognoz dla $Y = 1$. Dosyć dobrze były też w stanie przewidzieć wartości $Y = 0$ w ilościach 343 i 342. Do rozstrzygnięcia, który z nich jest lepiej dopasowany do danych będziemy potrzebować większej ilości mierników. Przedstawimy je w kolejnej tabeli:

Miara dopasowania	Logit	Probit
$countR^2$	0.7921	0.7921
R^2_{Efrona}	0.4164	0.4138
R^2_{MZ}	0.0292	0.0891

Tabela 3.3: Mierniki dopasowania

Niestety na podstawie samych mierników nie jesteśmy w stanie stwierdzić, który z wyżej zaprezentowanych modeli jest lepiej dopasowany. Oba mierniki $countR^2$ wskazują na to, że udało nam się wyjasnić około 80% zmienności modelu. Pozostałe dwie miary wskazują na lepsze dopasowanie albo probitu albo logitu przez co nie możemy wskazać lepszego modelu. W tym momencie pomocna może się okazać analiza ex-post, którą zaprezentujemy w kolejnych etapach.

3.2 Redukcja modelu

3.2.1 Oceny parametrów β z restrykcjami

Sprawdzamy, czy można przeprowadzić łączną redukcję zmiennych, które podejrzewamy o brak istotności. W celu wyznaczenia nowego oszacowania parametru β z restrykcjami dla zmiennych $Parch$, $Fare$, $EmbarkedS$ i $EmbarkedC$. Tworzymy nowy zbiór danych i wywołujemy funkcję `glm` i otrzymujemy nowe oszacowanie:

Zmienna	**Logit**	**Probit**	**Logit - p-value**	**Probit - p-value**
(Intercept)	1.680	1.00	0.000	0.000
Sex	-2.619	-1.542	0.000	0.000
Age	-0.045	-0.025	0.000	0.000
SibSp	-0.379	-0.224	0.002	0.001
Pclass1	2.645	1.492	0.000	0.000
Pclass2	1.239	0.681	0.000	0.000

Tabela 3.4: Porównanie parametrów i p-value w modelach

Otrzymalismy teraz ocene parametru β, w którym wszystkei zmienne są istotne, jednak aby stwierdzić o zasadności redukcji potrzebujemy przeprowadzić test LR.

3.2.2 Test LR

Na początek zastanówmy się nad układem hipotez do testu LR. Hipoteza zerowa będzie wskazywać, że redukcja modelu jest zasadna i model zredukowany jest tak samo dobry, jak model bez restrykcji. W hipotezie alternatywnej, która jest zaprzeczeniem zerowej otrzymamy, że przynajmniej jedna z redukowanych zmiennych ma istotny wpływ na wyjaśnienie zmienności modelu. W ten sposób otrzymujemy następujący układ hipotez:

$$H_0 : Parch = Fare = EmbarkedS = EmbarekdC = 0$$

$$H_1 : Parch \neq 0 \vee Fare \neq 0 \vee EmbarkedS \neq 0 \vee EmbarkedC \neq 0$$

Do kolejnego etapu testu potrzebujemy wyznaczyć funkcje wiarygodności dla modeli z restrykcjami i bez restrykcji. Pakiet `glm2` nie zwraca nam bezpośrednio takich wartości, jednak możemy obliczyć to sami korzystając z kryterium AIC, dostepnego w funkcji `glm` pod zmienna "aic". Wartość ta jest obliczna na podstawie wzoru:

$$AIC = 2k - 2l_y(\theta)$$

Gdzie k to liczba parametrów w modelu, a $l_y(\theta)$ to logarytm funkcji wiarygodności.

Po odpowiednich przekształceniach wzoru, jesteśmy w stanie wyznaczyć wartości logarytmów funkcji wiarygodności dla obu modeli. Ich wyniki prezentują się następująco:

- Logit :
  - Bez restrykcji - $l_y(\hat{\beta_{MNW}}) = -316.171$
  - Z restrykcjami - $l_r(\hat{\beta_{MNW}}) = -318.020$
- Probit:
  - Bez restrykcji - $l_y(\hat{\beta_{MNW}}) = -316.924$
  - Z restrykcjami - $l_r(\hat{\beta_{MNW}}) = -319.000$

Widzmy zatem, że redukując modele maleje wartość funkcji wiarygodności. Jest to zasadne, ponieważ wraz ze spadkiem ilości zmiennych w modelu tracimy pewną część informacji. Musimy się teraz zastanowić, czy ta utrata z perspektywy naszych modeli jest znacząca. W tym momencie wyznaczymy wartość statystyki testowej LR.

$$LR = -2(l_r(\hat{\beta_{MNW}}) - l_y(\hat{\beta_{MNW}})$$

Dla logitu i probitu wartości statystyk testowych prezentują się nastepująco:

$$logitLR = 3.6976$$

$$probitLR = 4.1518$$

Przy prawdziwości hipotezy zerowej statystyka LR ma rozkład χ^2 o v stopniach swobody, gdzie v oznacza ilość zmiennych poddawanych restrykcjom. W środowisku R, do przeprowadzenia testu użyjemy wartości p-value, uzuskanej za pomocą funkcji `pchisq` zwracającej wartość rozkładu skumulowanego. Do przeprowadzenia testu potrzebujemy obliczyć wartości z prawego ogona, co otrzymamy definiując argument `lower.tail` na wartośc `TRUE`. Jako kwantyl podajemy statystykę testową LR.

$$p - value_{logit} = 0.5515$$

$$p - value_{probit} = 0.6142$$

Ponieważ p-value w obu modelach jest większe od każdego konwencjonalnego poziomu istotności α nie mamy podstaw do odrzucenie hipotezy zerowej, mówiącej, że redukcja modelu jest zasadna. W ten sposób otrzymujemy modele, w których wszystkie zmienne są istotne.

3.3 Ostateczne modele i zbiór testowy

3.3.1 Postać modeli testowych i ich dopasowanie

Za pomocą redukcji otrzymaliśmy model logitowy oraz probitowy, w którym każda ze zmiennych ma istotny wpływ na zmienną objaśnianą. Za pomocą tych modeli przystąpimy do analizy ex post, jednak najpierw przypomnijmy ostateczną postać modelu.

$$X_t\beta = \beta_0 + \beta_1 Pclass1_t + \beta_2 Pclass2_t + \beta_3 Sex_t + \beta_4 Age_t + \beta_5 SibSp_t$$

Ze zbioru testowego wyodrębnimy macierz zawierającą tylko interesujące nas zmienne. Nastepnie przemnożymy ją przez oszacowanie parametrów β dla modelu logitowego i probitowego. W ten sposób będziemy mogli wyznaczyć prawdopodobieństwo w zbiorze testowym. Postepując podobnie jak przy zbiorach uczących wyznaczymy tabele trafności i mierniki dopasowania. W tym momencie będziemy w stanie stwierdzić, który z modeli lepiej poradził sobie przy zbiorze testowym. Granicznym progiem p^* pozostanie średnią wartość zmiennej objaśnianej ze zbioru uczącego.

	Logit			Probit		
Prognozy	$y_t = 1$	$y_t = 0$	Suma	$y_t = 1$	$y_t = 0$	Suma
$\hat{y}_t = 1$	123	31	154	123	32	155
$\hat{y}_t = 0$	4	173	177	4	172	175
Suma	127	204	331	127	204	331

Tabela 3.5: Tabela trafności zbioru testowego

Zanim przejdziemy do wniosków, który z modeli jest lepszy zobaczmy jak wygląda porównanie pozostałych miar dopasowania.

Miara dopasowania	Logit	Probit
$countR^2$	0.8943	0.8912
R^2_{Efrona}	0.6577	0.6618
R^2_{MZ}	0.0277	0.0844

Tabela 3.6: Mierniki dopasowania dla analizy ex post

W tym momencie jesteśmy w stanie stwierdzić, że model logitowy poradził sobie lepiej z przewidzeniem sukcesu w modelu testowym. Przewaga jest nieznaczna, ponieważ to kwestia jednej obserwacji i mimo minimalnie lepszego dopasowania dla współczynników R^2_{Efrona} i R^2_{MZ}, to model logitowy poradził sobie lepiej z przewidzeniem zjawiska.

Zastanówmy się jeszcze jak zmienia się wyniki w przypadku ustalenia innego p^*, czy w tej sytuacji któryś z modeli okaże się sprawniejszy? Ustalmy nową wartość prawdopodobieństwa progowego na 0.5 i ponownie dokonajmy analizy na podstawie tabeli trafności.

	Logit			Probit		
Prognozy	$y_t = 1$	$y_t = 0$	Suma	$y_t = 1$	$y_t = 0$	Suma
$\hat{y}_t = 1$	117	16	123	119	16	125
$\hat{y}_t = 0$	10	188	198	8	188	196
Suma	127	204	331	127	204	331

Tabela 3.7: Porównanie tabeli trafności przy innej wartości p^*

Przy takiej tabeli trafności zliczeniowe R^2 prezentuje się nastepująco:

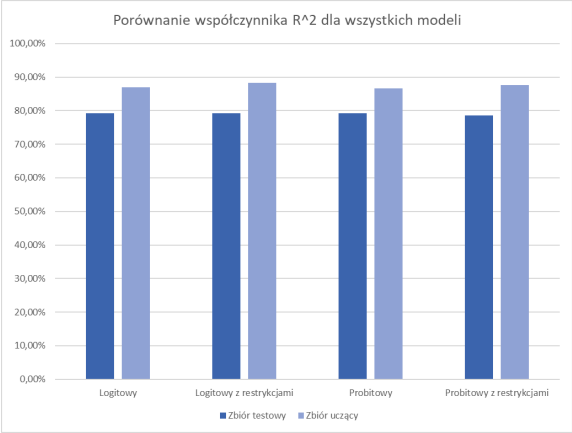
$$countR^2_{logit} = 0.9215$$

$$countR^2_{probit} = 0.9275$$

Manipulując prawdopodobieństwem progowym udało się otrzymać jeszcze bardziej trafne prognozy dla obu modeli. Mimo, że wczesniej model logitowy miał przewagę jednej lepszej prognozy, to w tym momencie lepszy okazał się probit. Są to jednak zbyt małe różnice, by móc ostatecznie stwierdzić, który z nich jest lepszy. Oba z modeli były w stanie wyjaśnić bardzo duży poziom zmienności Y (w okolicach 90% w zależności od p^*, przez co zarówno model logitowy, jak i probitowy jest wystarczająco skuteczny do analizowania katastrofy Titanica.

3.3.2 Zestawienie współczynników dopasowania

W celu lepszego zwizualizowania efektywności modeli, w tabeli 3.1 zaprezentowano wartości współczynników R^2 dla wszystkich ośmiu przetestowanych przypadków. Z wykresu można wywnioskować, że współczynniki dopasowania we wszystkich przypadkach były bardzo zbliżone i żaden model nie miał wyraźnej przewagi nad innymi.



Rysunek 3.1: Porównanie współczynników dopasowania dla wszystkich modeli

3.3.3 Efekty krańcowe dla modeli testowych

Dodatkowo, obliczono średnie efekty krańcowe dla poszczególnych zmiennych. Najbardziej interesujące są wyniki dotyczące modeli, w których weryfikację danych przeprowadzono na zbiorze testowym. Poniżej zostały zaprezentowane wyliczone efekty krańcowe kolejno dla: modelu logitowego bez restrykcji, modelu logitowego z restrykcjami, modelu probitowego bez restrykcji, modelu probitowego z restrykcjami.

Zmienna	Logitowy	Logitowy z restrykcjami	Probitowy	Probitowy z restrykcjami
Sex	-0.3988	-0.4002	-0.4100	-0.4091
Age	-0.0065	-0.0066	-0.0064	-0.0067
SibSp	-0.0549	-0.0551	-0.0553	-0.0595
Parch	-0.009	-	-0.0127	-
Fare	0.0002	-	0.0002	-
EmbarkedS	0.0636	-	0.0630	-
EmbarkedC	0.1245	-	0.1270	-
Pclass1	0.3621	-0.0092	0.3529	0.3957
Pclass2	0.1823	0.0002	0.1752	0.1805

Tabela 3.8: Średnie efekty krańcowe dla zbioru testowego

Patrząc na wyniki ze zbioru testowego można zauważyć, że prawdopodobieństwo przeżycia katastrofy na Titanicu maleje o około 0.4 jeśli pasażer jest mężczyzną. Istotny wpływ na zmianę prawdopodobieństwa sukcesu ma również zmienna *Pclass*. Widzimy, że szanse na przeżycie są większe o około 0.4 i 0.2 odpowiednio dla klasy pierwszej oraz drugiej w modelu logitowym bez restrykcji oraz probitowym. Dla modelu logitowego z restrykcjami klasa podróżnicza nie zmienia znacząco prawdpodobieństwa sukcesu. Z tabeli można również wywnioskować, że w przypadku dwóch pasażerów, z czego jeden z nich jest o 10 lat starszy od drugiego, szansa na przeżycie starszego jest mniejsza o 0.065.

Dodatkowo, w celu zwizualizowania wszystkich obliczonych danych, w tabeli 3.9 przedstawiono efekty krańcowe dla obliczone dla zbioru uczącego:

Zmienna	Logitowy	Logitowy z restrykcjami	Probitowy	Probitowy z restrykcjami
Sex	-0.3748	-0.3775	-0.3890	-0.3883
Age	-0.0062	-0.0062	-0.0061	-0.0063
SibSp	-0.0516	-0.0519	-0.0524	-0.0565
Parch	-0.0086	-	-0.0121	-
Fare	0.0002	-	0.0002	-
EmbarkedS	0.0598	-	0.0598	-
EmbarkedC	0.1170	-	0.1205	-
Pclass1	0.3403	-0.0086	0.3348	0.3756
Pclass2	0.1713	0.0002	0.1662	0.1714

Tabela 3.9: Średnie efekty krańcowe dla różnych modeli dla zbioru uczącego.

Wyniki dla zbioru uczącego nie różnią się zbytnio od tych dla zbioru testowego.