1) Wstęp

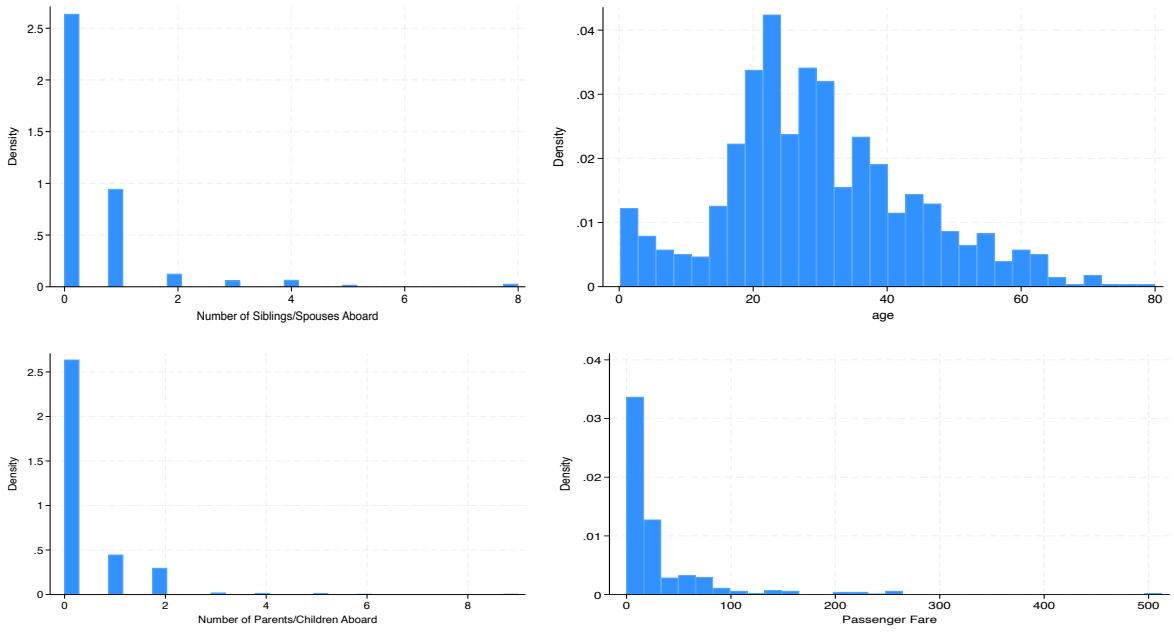
W 1912 roku zdarzyła się najbardziej katastrofalny wypadek morski w historii – zatonięcie Titanica, w którym zginęło około 1500 osób. Raport jest stworzony z zbioru danych z Kaggle. Analiza została przeprowadzona podczas zajęć mikroekonometrii za pomocą programu Stata na pliku, który jest dodatkowo załączony.

Problem

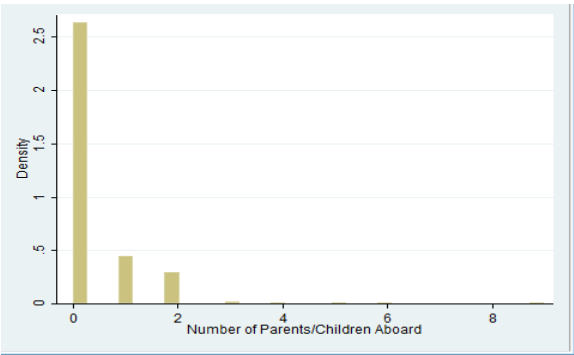
Celem tego badania jest stworzenie modeli predykcyjnych w celu prognozowania przeżywalności pasażerów Titanica. Wykorzystując do tego regresję logistyczną uwzględniając zmienne, które są według badanego kluczowe do przedstawienia tego modelu oraz które są zgodne z teorią ekonometrii.

2) Opis zmiennych

1. survival : Przeżycie pasażera (0 = No; 1 = Yes); (No– 61,80%, Yes – 38,20%)
2. pclass : Klasa pasażera (1 = 1st; 2 = 2nd; 3 = 3rd) (1rd=25%; 2nd=21%, 3rd=54%)
3. sex : Płeć (Male=0, Female=1); (Male – 64,40%; Female – 35,60%)
4. age : Wiek pasażera w latach
5. sibsp : Liczba rodzeństwa i małżonków podróżujących razem
6. parch : Liczba rodziców i dzieci podróżujących razem.
7. fare : Opłata pasażera
8. Embarked : Port zaokrętowania (C = Cherbourg; Q = Queenstown; S = Southampton) (Cherbourg=21%; Queenstown=9%; Southampton=70%)



Binaryzacja zmiennej parch na parch_bin (0=samodzielnie, 1= z osobą/osobami)



parch_bin	Freq.	Percent	Cum.
0	1,002	76.49	76.49
1	308	23.51	100.00
Total	1,310	100.00	

3) Charakterystyka spodziewanych zależności hipotez

Hipoteza 1: zmienna pclass (+)
Pasażerowie pierwszej klasy mieli lepszy dostęp do łodzi ratunkowych i mniejszy tłok w swojej części statku, co zwiększało szanse na przeżycie.

Hipoteza 2: zmienna sex (+)
Historyczne relacje mówią „kobiety i dzieci pierwsze", więc mężczyźni byli narażeni na niższe szanse przeżycia.

Hipoteza 3: zmienna age (-)
W sytuacjach ratunkowych preferowano dzieci i młodsze osoby, a starsi mogli mieć trudności w szybkim dotarciu do łodzi.

Hipoteza 4: zmienna fare (+)
Wyższe opłaty były związane z lepszą klasą kabiny, co korelowało z lepszym dostępem do ratunku.

Hipoteza 5: zmienna sibsp(-)
Mała liczba towarzyszy mogła ułatwić szybkie działanie i zwiększyć szanse przeżycia.

Hipoteza 6: zmienna parch (-)
Dorosły pasażer mógł najpierw pomagać dzieciom lub starszym członkom rodziny, co spowalniało jego ewakuację.

Hipoteza 7: zmienna embarked (?)
Ze względu na brak jednoznacznych przesłanek historycznych i różnic w warunkach ewakuacji między portami, trudno przewidzieć kierunek wpływu zmiennej Embarked na przeżycie pasażera. Pozostałe zmienne są problematyczne do postawienia hipotezy.

4) Model wstępny oraz ostateczny

W tabeli poniżej przedstawiono wyniki modelu wstępnego i ostatecznego. Model wstępny uwzględniał wszystkie dostępne zmienne, natomiast model ostateczny zawiera jedynie te, które okazały się istotne statystycznie lub miały uzasadnienie teoretyczne.

Model wstepny

survived	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
pclass						
2	-1.097727	.2415174	-4.55	0.000	-1.571093	-.6243618
3	-2.039685	.2779234	-7.34	0.000	-2.584405	-1.494965
1.sex	-2.586742	.1760145	-14.70	0.000	-2.931724	-2.24176
age	-.0664068	.0229502	-2.89	0.004	-.1113883	-.0214254
c.age#c.age	.0004614	.0003155	1.46	0.144	-.000157	.0010798
sibsp	-.447166	.111914	-4.00	0.000	-.6665135	-.2278186
1.parch_bin	.3942441	.2090851	1.89	0.059	-.0155552	.8040434
fare	-.0002376	.0017132	-0.14	0.890	-.0035955	.0031203
embarked						
2	-1.35585	.43867	-3.09	0.002	-2.215627	-.4960722
3	-.6416106	.2148795	-2.99	0.003	-1.062767	-.2204545
_cons	4.530771	.5373268	8.43	0.000	3.47763	5.583912

Model ostateczny

survived	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
pclass						
2	-1.084715	.2176364	-4.98	0.000	-1.511274	-.6581555
3	-2.024979	.2438069	-8.31	0.000	-2.502832	-1.547127
1.sex	-2.586127	.1755068	-14.74	0.000	-2.930114	-2.24214
age	-.0665256	.0229045	-2.90	0.004	-.1114176	-.0216336
c.age#c.age	.0004624	.0003145	1.47	0.141	-.000154	.0010787
sibsp	-.4485436	.1112495	-4.03	0.000	-.6665886	-.2304985
1.parch_bin	.3883182	.2028811	1.91	0.056	-.0093215	.7859579
embarked						
2	-1.352773	.43751	-3.09	0.002	-2.210277	-.4952693
3	-.6380704	.210951	-3.02	0.002	-1.051527	-.224614
_cons	4.513611	.5149959	8.76	0.000	3.504237	5.522984

5) Zapis modelu ostatecznego w formie równania

log(P(y=1))/(1-P(y=1))=4,53 − 1,08·(pclass=2) − 2.02·(pclass=3) − 2,59·(sex=0) + −0,07·age + 0.001·age2 −0,45·sibsp + 0.39·(parch_bin=1) − 1,36·(embarked=2) + −0,64·(embarked=3)

6) Interpretacja modelu ostatecznego

pclass (klasa): Hipoteza potwierdzona. Im niższa klasa (wyższy numer), tym mniejsze szanse na przeżycie (współczynniki silnie ujemne i istotne).

sex (płeć): Hipoteza potwierdzona. Mężczyźni mieli radykalnie niższe szanse na przeżycie niż kobiety (współczynnik -2,59, bardzo istotny).

age (wiek): Hipoteza częściowo potwierdzona. Wiek ma istotny, negatywny wpływ na szanse przeżycia. Efekt kwadratowy wieku jest nieistotny.

sibsp (rodzeństwo/małżonek): Hipoteza potwierdzona. Większa liczba rodzeństwa lub małżonka istotnie obniżała szanse na przeżycie.

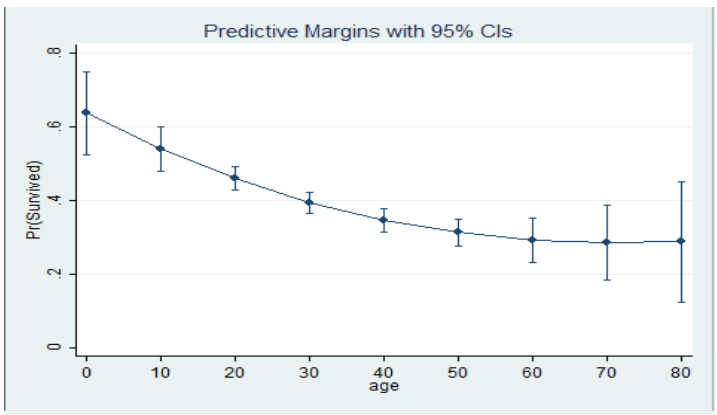
parch_bin (rodzice/dzieci): Hipoteza odrzucona. Obecność rodziców/dzieci miała dodatni (choć słabo istotny) wpływ na szanse przeżycia, co jest przeciwne do założenia.

embarked (port): Wpływ istotny. Pasażerowie, którzy weszli w portach 2 i 3, mieli istotnie niższe szanse na przeżycie niż ci z portu referencyjnego.

Efekty krańcowe

Płeć (male): Bycie mężczyzną obniża prawdopodobieństwo przeżycia o ok. 49 punktów procentowych. To najsilniejszy efekt w modelu.
Podróżowanie klasą 2. zamiast 1. obniża prawdopodobieństwo przeżycia o ok. 19 p.p.
Podróżowanie klasą 3. zamiast 1. obniża prawdopodobieństwo przeżycia o ok. 33 p.p.
Wiek (age): Każdy dodatkowy rok życia obniża prawdopodobieństwo przeżycia o ok. 0,56 punktu procentowego.
Port (embarked): Wejście na statek w Queenstown lub Southampton (w porównaniu do Cherbourga) wiązało się z niższym prawdopodobieństwem przeżycia (o ok. 20 i 10 p.p. odpowiednio).
Pozostałe zmienne (takie jak fare lub parch_bin) są nieistotne lub słabo istotne, więc ich efektów nie interpretuje się jako mających realny wpływ.

Nieliniowość – graficzna analiza



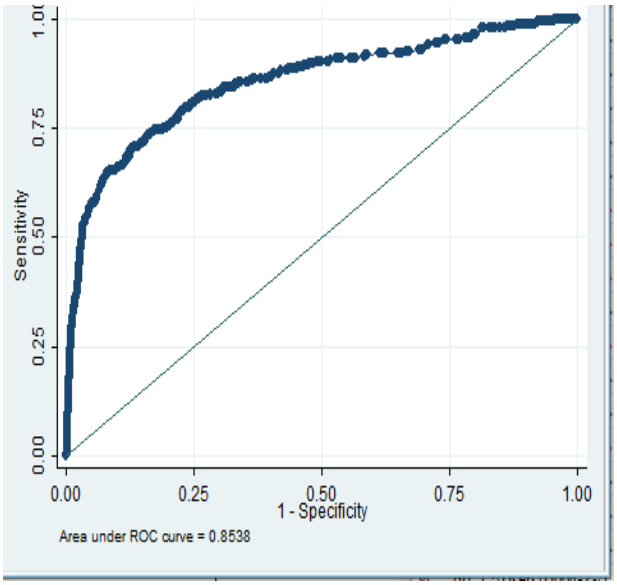
7) Ocena dopasowania modelu ostatecznego

Pseudo R^2 wynosi 0,329, co oznacza, że model logitowy wyjaśnia ok. 32,8 % funkcji wiarygodności do przeżycia. Jest to dobry wynik.

Classified	True D	True ~D	Total
+	329	136	465
-	96	483	579
Total	425	619	1044

Classified + if predicted Pr(D) >= .38
True D defined as survived != 0

Sensitivity	Pr(+| D)	77.41%
Specificity	Pr(-|~D)	78.03%
Positive predictive value	Pr(D| +)	70.75%
Negative predictive value	Pr(~D| -)	83.42%
False + rate for true ~D	Pr(+|~D)	21.97%
False - rate for true D	Pr(-| D)	22.59%
False + rate for classified +	Pr(~D| +)	29.25%
False - rate for classified -	Pr(D| -)	16.58%
Correctly classified		77.78%



Model poprawnie klasyfikuje 77.8% przypadków, co wskazuje na jego dobrą ogólną skuteczność. Jego czułość na poziomie 77.4% oznacza, że dobrze identyfikuje osoby, które przeżyły, zaś swoistość 78.0% potwierdza, że równie trafnie rozpoznaje ofiary katastrofy.

Krzywa ROC ilustruje ogólną zdolność modelu do odróżniania obu grup. Pole pod tą krzywą (AUC) wynoszące około 0.85 potwierdza bardzo dobrą jakość modelu w rozdzielaniu osób, które przeżyły, od tych które nie przeżyły.

8) Wnioski

Dalsza eksploracja zbioru danych oraz wprowadzenie dodatkowych zmiennych mogłoby być interesujące dla pogłębienia analizy. Dodanie większej liczby cech może potencjalnie poprawić dokładność i szczegółowość uzyskanych wyników.

Eksperymentowanie rożnymi zmiennymi może umożliwić głębsze zrozumienie danych i do odkrycia nowych wzorców lub zależności.

9) Literatura

Aamir, M. (2023). Titanic survival prediction , ResearchGate

https://www.researchgate.net/publication/383295501_TITANIC_SURVIVAL_PREDICTION