MIĘDZY WIERNOŚCIĄ A POKUSĄ: MODELE LOGITOWY I PROBITOWY W PRZEWIDYWANIU ZDRAD MAŁŻEŃSKICH

Modelowanie parametryczne

Wydział Zarządzania
Informatyka i Ekonometria
Analiza danych

K. M.
Szymańska Natalia

B. K.
N. Ł.

Wstęp

Zdrady małżeńskie są zjawiskiem, które od wieków zarówno fascynuje, jak i niepokoi ludzi na całym świecie. W kontekście związków partnerskich, wierność jest uważana za fundament zaufania, lojalności i intymności. Jednak w obliczu rozmaitych pokus, emocji i napięć, wiele osób znajduje się w sytuacjach, które stawiają ich wierność na próbę.

Praca powstała w wyniku współpracy M. K., Natalii Szymańskiej, K. B. oraz Ł. N.. Współpraca opierała się na wspólnych spotkaniach na platformie Discord oraz dyskutowaniu na każdym etapie prac, przez co trudno jest wyszczególnić konkretny podział zadań.

Opis problemu badawczego i celu badania

Celem niniejszej pracy jest przeprowadzenie analizy porównawczej modelu logitowego i probitowego w kontekście przewidywania zdrad małżeńskich. Poprzez zastosowanie tych dwóch podejść, chcemy zbadać, który z modeli lepiej radzi sobie z prognozowaniem zdrad, a także zrozumieć, jakie czynniki mogą mieć wpływ na wystąpienie zdrady małżeńskiej.

W ramach tej pracy skupimy się na wykorzystaniu danych związanych ze zdradami małżeńskimi, dostępnych na platformie Kaggle[1]. Przeanalizujemy zbiór danych pod kątem odpowiednich zmiennych, takich jak wiek, staż małżeństwa, poziom zadowolenia z małżeństwa, a także inne czynniki, które mogą mieć wpływ na zdradę.

Przygotowanie danych i prezentacja zbioru danych

W celu przeprowadzenia analizy dotyczącej zdrad małżeńskich, skorzystaliśmy z zestawu danych o nazwie "Fair's Affairs", które są wynikiem przekrojowego badania przeprowadzonego przez Psychology Today w 1969 roku. Zbiór danych zawiera informacje na temat 601 badanych osób, które

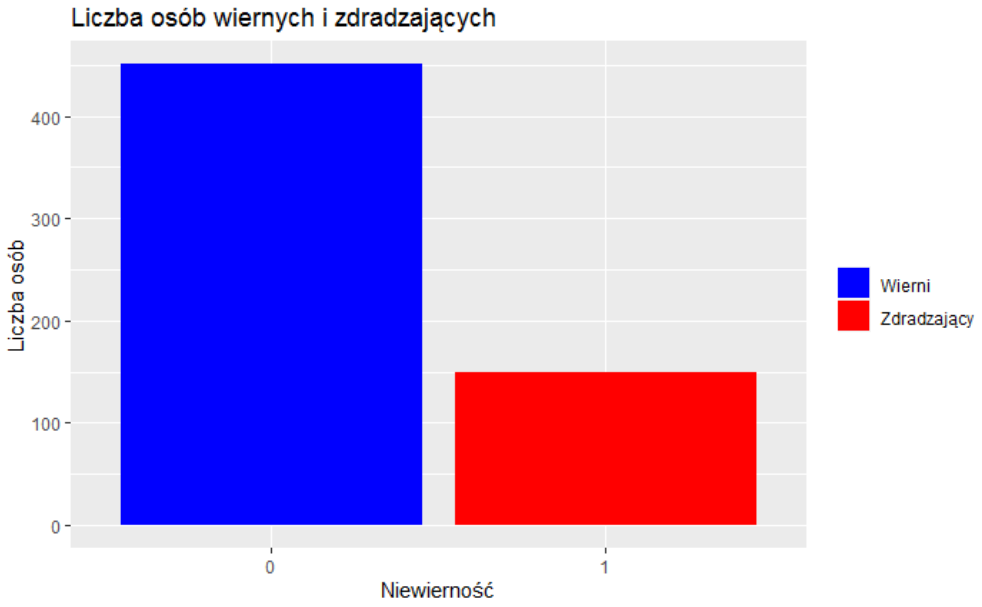
[1] Link: https://www.kaggle.com/datasets/utkarshx27/fairs-extramarital-affairs-data

zostały opisane za pomocą 9 zmiennych związanych z zdradami małżeńskimi, co umożliwia nam dogłębne zbadanie tego zjawiska.

Poniżej przedstawione są zmienne, które zostały uwzględnione w analizie oraz krótkie opisy dotyczące ich istoty, znaczenia oraz ich statystyki opisowe. W przypadku zmiennych „affairs", „gender" oraz „children" zdecydowano się skupić na obliczaniu liczebności i proporcji dla tych zmiennych, ponieważ inne metody analizy statystycznej nie mają sensu ze względu na ich naturę. Zmienne te są dychotomiczne dlatego nie ma możliwości przeprowadzenia bardziej zaawansowanych analiz, takich jak średnie, odchylenia standardowe czy testy statystyczne oparte na założeniach o rozkładach danych. W przypadku tych zmiennych, istotne są przede wszystkim dwie rzeczy: określenie proporcji, czyli udziału jednej kategorii w stosunku do drugiej oraz obliczenie liczebności, czyli ilości respondentów należących do danej kategorii. Takie podejście pozwoli nam zrozumieć, jak często występują poszczególne zjawiska w populacji i czy występują istotne różnice między grupami. Umieszczono również wykresy w celu lepszego zilustrowania wartości danej zmiennej.

AFFAIRS (CZĘSTOTLIWOŚĆ ROMANSÓW POZAMAŁŻEŃSKICH W CIĄGU OSTATNIEGO ROKU)

W kontekście naszego badania, zmienna affairs jest zmienną zależną i określa, czy respondent był zaangażowany w romans pozamałżeński w ciągu ostatniego roku. Zmienna została zakodowana jako 1, w przypadku wystąpienia romansu oraz jako 0, w przypadku jego braku.



Wykres 1. Źródło: opracowanie własne w programie RStudio.

Statystyka	Wartość
% wystąpienie romansu	24,96
% braku romansu	75,04
Liczebność osób niewiernych	150
Liczebność osób wiernych	451

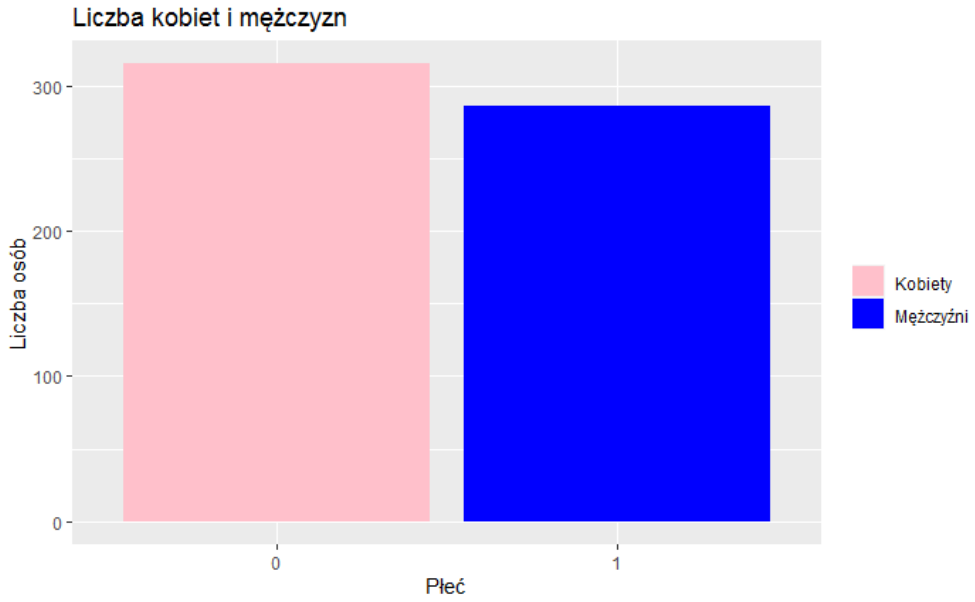
Tabela 1. Źródło: opracowanie własne w programie Excel.

Wykres 1 oraz tabela 1 przedstawiają stosunek występowania romansu w badanej populacji do braku występowania romansu. Liczba osób niewiernych wynosi 150, a liczba osób wienych 451, co oznacza, że według analizy statystycznej, 24,96% respondentów przyznało się do zaangażowania w romans pozamałżeński w ciągu ostatniego roku. Oznacza to, że około jedna czwarta badanej populacji doświadczyła romansu w tym okresie. Natomiast 75,04% respondentów deklarowało wierność w ciągu ostatniego roku, przez co grupa ta stanowi większość badanej populacji.

Analiza tych wartości pozwala nam zrozumieć, że romanse pozamałżeńskie występują w badanej populacji, są one jednak mniej powszechne niż brak romansu. Proporcjonalnie, mniej niż jedna czwarta respondentów doświadczyła takiego zdarzenia w ciągu ostatniego roku.

GENDER (PŁEĆ)

Ta zmienna określa płeć respondentów. Przyporządkowano wartość 1 dla mężczyzn i 0 dla kobiet.



Wykres 2. Źródło: opracowanie własne w programie RStudio.

Statystyka	Wartość
% Mężczyzn	47,59
% Kobiet	52,41
Liczebność mężczyzn	286
Liczebność kobiet	315

Tabela 2. . Źródło: opracowanie własne w programie Excel.

Wykres 2 przedstawia liczebność mężczyzn i kobiet w danej populacji. Liczba mężczyzn wynosi 286, natomiast liczba kobiet wynosi 315. Zatem w badaniu wzięła udział nieco większa liczba kobiet, niż mężczyzn.

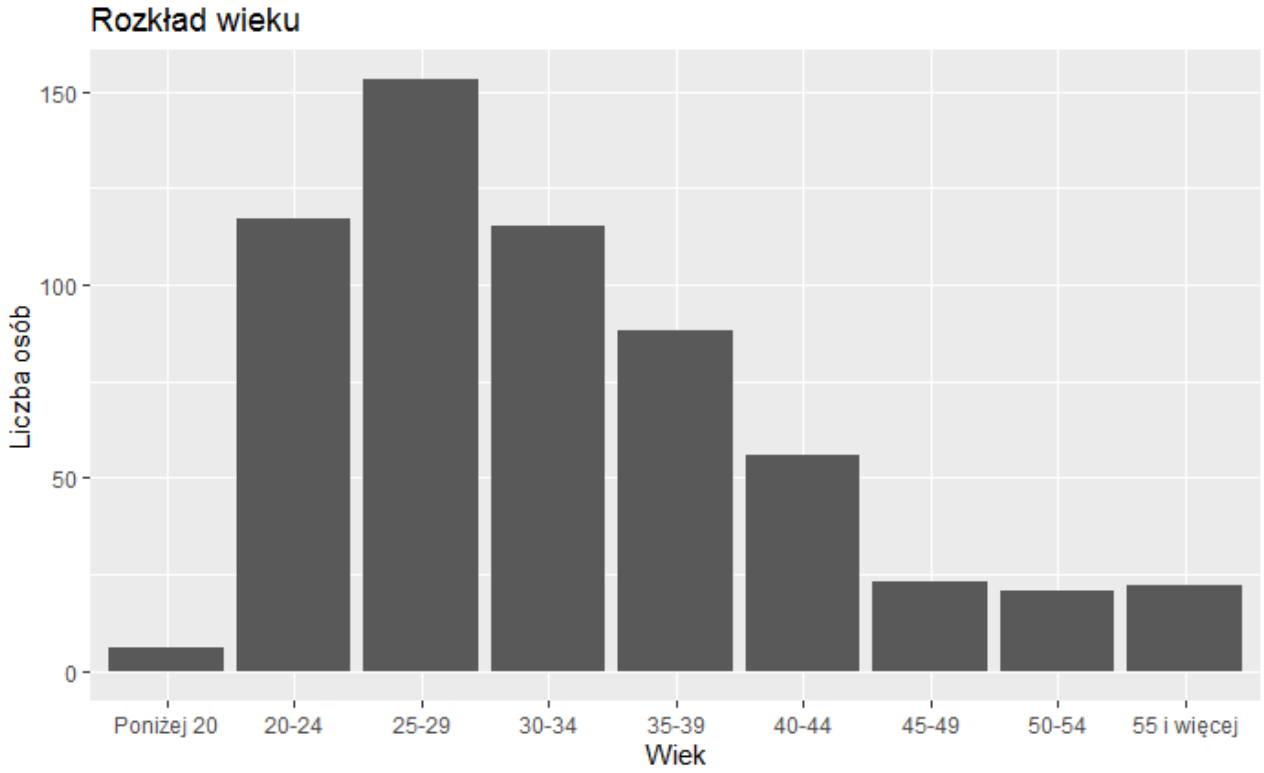
Tabela 2 zawiera dodatkowe informacje na temat udziału procentowego mężczyzn i kobiet w populacji. Procent mężczyzn wynosi 47,59%, a procent kobiet wynosi 52,41%.

AGE (WIEK)

Ta zmienna odnosi się do wieku respondentów, wyrażonego w latach. Została ona zakodowana w następujący sposób: 17.5: poniżej 20 lat, 22: 20-24 lata, 27: 25-29 lat, 32: 30-34 lata, 37: 35-39 lat, 42: 40-44 lata, 47: 45-49 lat, 52: 50-54 lata oraz 57: 55 lat i więcej.

Obserwacja	Wartość w populacji [%] zaokrąglona do dwóch miejsc po przecinku	Liczba w populacji
17.5	1,00	6
22	19,47	117
27	25,46	153
32	19,13	115
37	14,64	88
42	9,32	56
47	3,83	23
52	3,66	21
57	3,66	22

Tabela 3. Źródło: opracowanie własne w programie Excel.

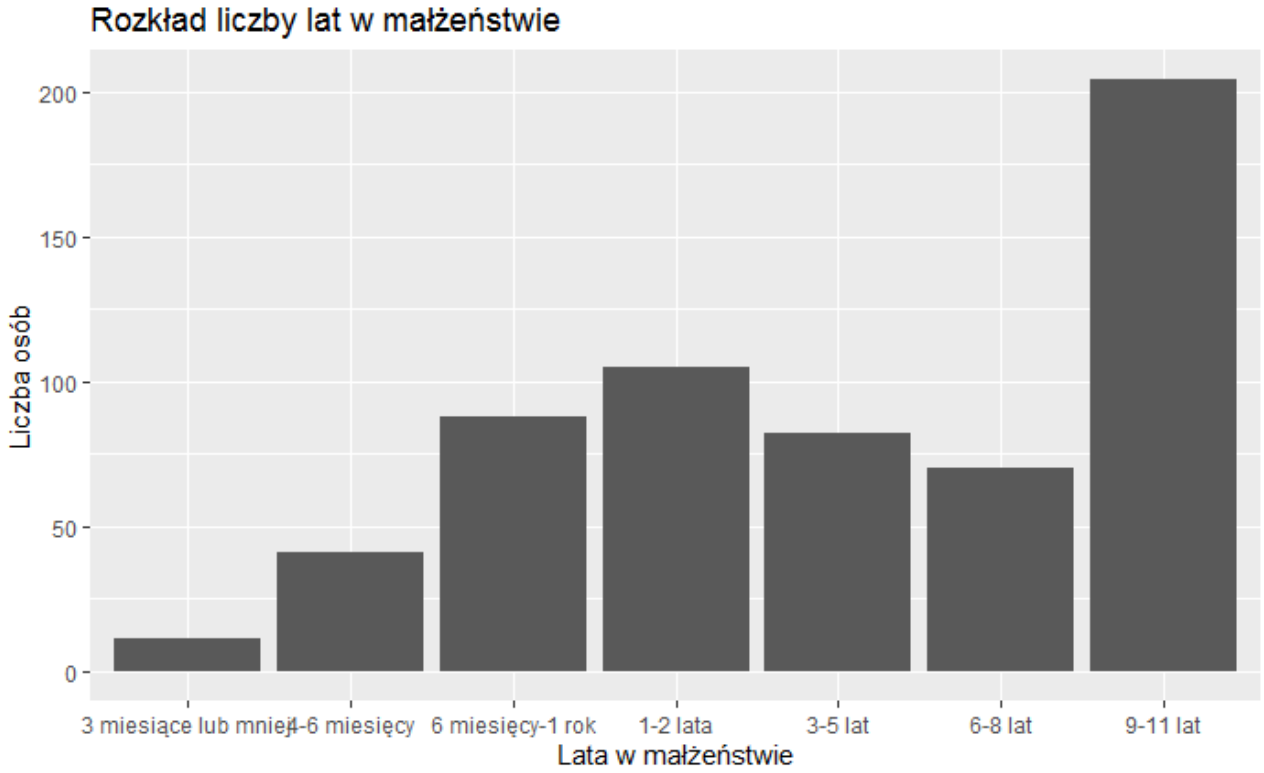


Wykres 3. . Źródło: opracowanie własne w programie RStudio

Na podstawie wykresu 3 i tabeli 3 można zauważyć jaki jest rozkład pomiędzy grupami wiekowymi w badanej populacji. Największa liczba obserwacji przypada na grupę wiekową "25-29 lat" (27). Najmniejsza liczba obserwacji przypada na grupę wiekową osób poniżej 20 lat, stanowi ona zaledwie 1% badanej populacji.

YEARSMARRIED (LICZBA LAT MAŁŻEŃSTWA)

Zmienna yearsmarried reprezentuje staż małżeństwa. Kodowanie przyporządkowane różnym przedziałom czasowym jest opisane w następujący sposób: 0.125: 3 miesiące lub krócej, 0.417: 4–6 miesięcy, 0.75: 6 miesięcy–1 rok, 1.5: 1–2 lata, 4:3–5 lat, 7: 6–8 lat, 10: 9–11 lat oraz 15: 12 lub więcej lat.



Wykres 4. . Źródło: opracowanie własne w programie RStudio

Obserwacja	Wartość w populacji [%] zaokrąglona do dwóch miejsc po przecinku	Liczba w populacji
0.13	1,83	11
0.42	1,66	10
0.75	5,16	31
1.5	14,64	88
4	17,47	105
7	13,64	82
10	11,65	70
15	33,94	204

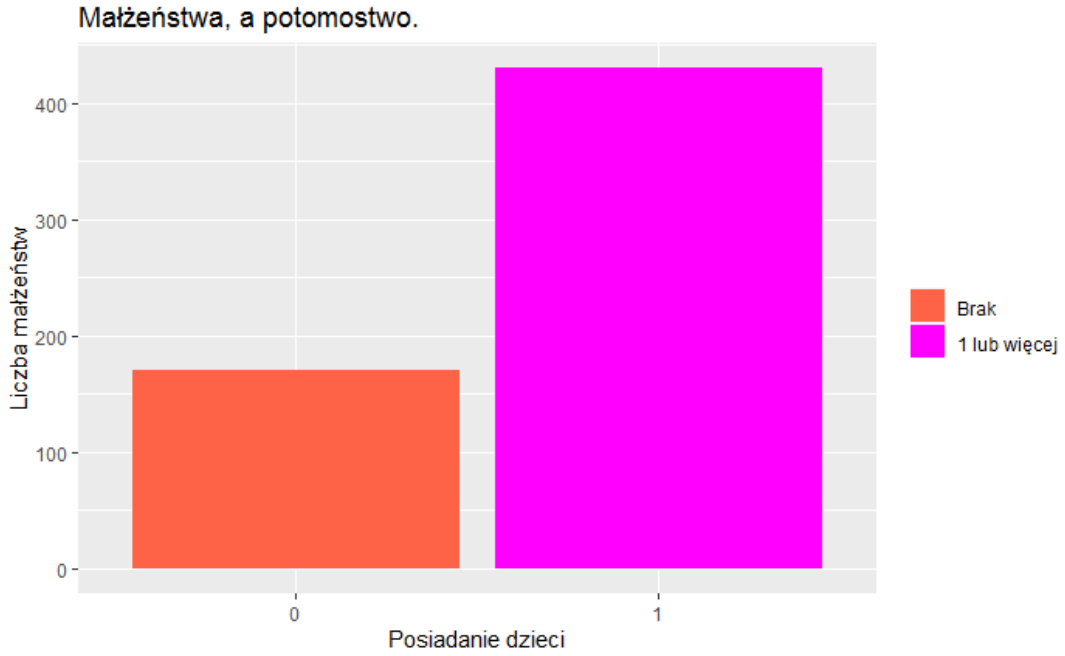
Tabela 4. Źródło: opracowanie własne w programie Excel

Na podstawie przedstawionego wykresu 4 i tabeli 4 można zidentyfikować różne przedziały czasowe reprezentujące staż małżeństwa w analizowanej populacji. Analiza wskazuje na zróżnicowanie w stażu małżeństwa w badanej populacji. Największą liczbę obserwacji obserwuje się w przedziale "12 lub więcej lat" (15), gdzie wartość wynosi 204. To sugeruje, że w badanej populacji istnieje duża liczba osób będących w długotrwałych małżeństwach.

Przedziały czasowe "3 i mniej miesięcy (0.13)", "4-6 miesięcy" (0.42) i "6 miesięcy-1 rok" (0.75) mają najmniejsze liczby obserwacji, wynoszące odpowiednio 11, 10, 31, co świadczy o niskim udziale małżeństw z niskim stażem w badaniu.

CHILDREN (CZY SĄ DZIECI W MAŁŻEŃSTWIE)

Zmienna „children” określa, czy w danym małżeństwie są dzieci. Przyporządkowano wartość 1, jeśli w małżeństwie są dzieci, oraz wartość 0, jeśli nie ma dzieci.



Wykres 5. Źródło: opracowanie własne w programie RStudio

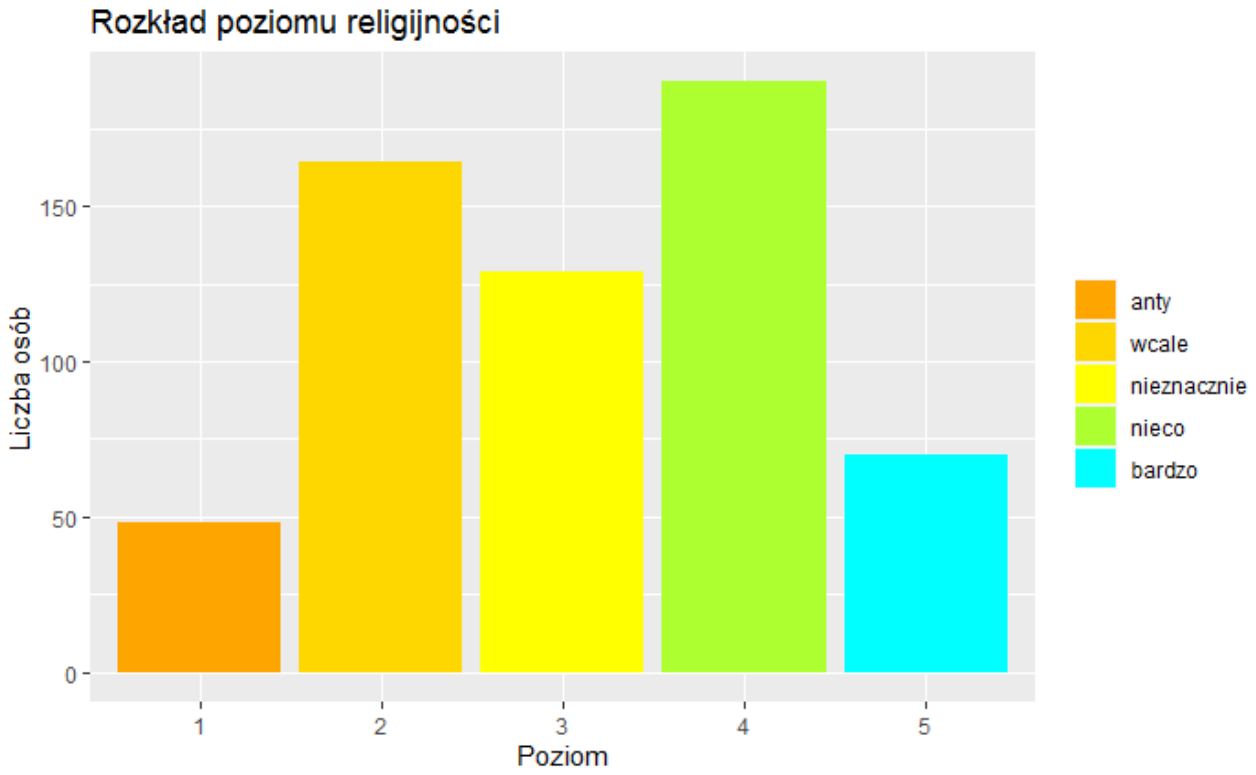
Statystyka	Wartość
% małżeństw z przynajmniej jednym dzieckiem	28,45
% małżeństw bez dzieci	71,55
Liczba małżeństw z przynajmniej jednym dzieckiem	171
Liczba małżeństw bez dzieci	430

Tabela 5. Źródło: opracowanie własne w programie Excel

Wykres 5 przedstawia dwie grupy małżeństw: te z przynajmniej jednym dzieckiem i te bez dzieci. Liczba małżeństw z przynajmniej jednym dzieckiem wynosi 430 (71,55%), co sugeruje, że w badanej populacji większość jednostek posiada chociaż jedno dziecko. Z kolei liczba małżeństw bez dzieci wynosi 171 (26,45%).

RELIGIOUSNESS (POZIOM RELIGIJNOŚCI)

Ta zmienna odzwierciedla stopień religijności respondentów. Wartości kodowania są zakodowane od 1 do 5, gdzie 1 oznacza „antyreligijny”, a 5 oznacza „bardzo religijny”.



Wykres 6. Źródło: opracowanie własne w programie RStudio

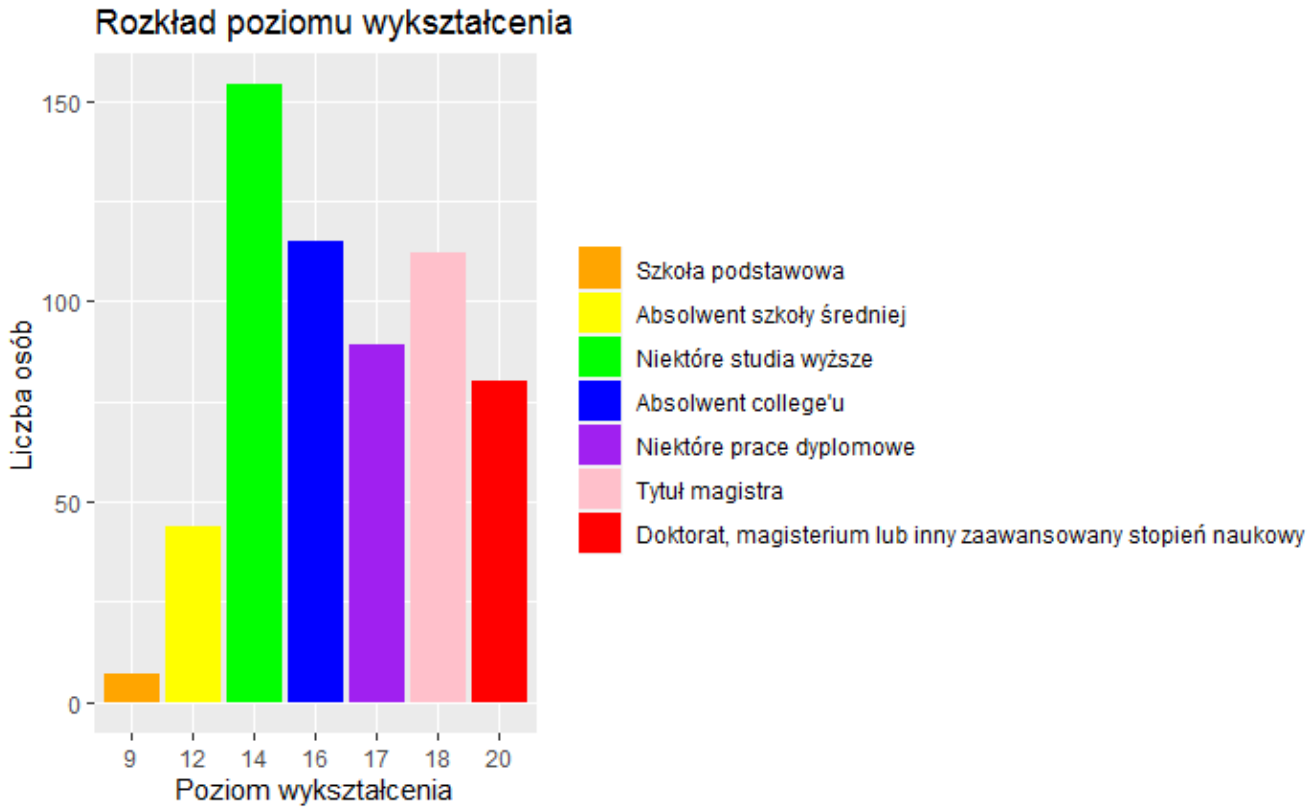
Obserwacja	Wartość w populacji [%] zaokrąglona do dwóch miejsc po przecinku	Liczba w populacji
1	7,99	48
2	27,29	164
3	21,46	129
4	31,61	190
5	11,65	70

Tabela 6. Źródło: opracowanie własne w programie Excel

Zdecydowana większa liczba respondentów deklaruje się jako osoby religijne. Odsetek osób, które deklarują się jako osoby antyreligijne wynosi 7.88%, co stanowi nieznaczny odsetek w badanej populacji. Osoby, które deklarują się jako osoby bardzo religijne stanowią 11.65% w badanej populacji.

EDUCATION (POZIOM WYKSZTAŁCENIA)

Zmienna „education” opisuje poziom wykształcenia respondentów. Kodowanie przypisane różnym poziomom edukacji: 9: ukończona szkoła podstawowa, 12: absolwent szkoły średniej, 14: niektóre studia wyższe, 16: absolwent college’u, 17: niektóre prace magisterskie, 18: tytuł magistra, 20: doktorant, magisterium lub inny zaawansowany tytuł naukowy.



Wykres 7. Źródło: opracowanie własne w programie RStudio

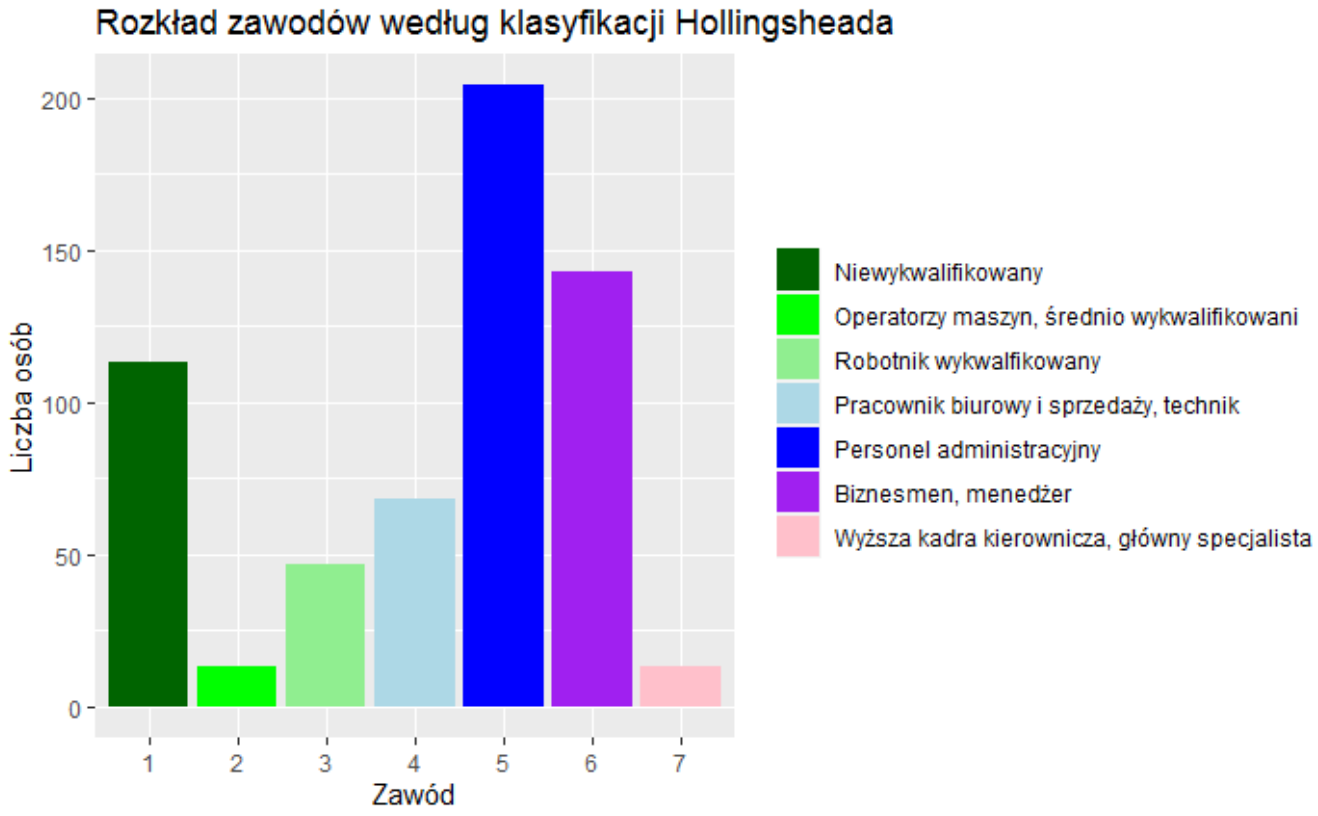
Obserwacja	Wartość w populacji [%] zaokrąglona do dwóch miejsc po przecinku	Liczba w populacji
9	1,16	7
12	7,32	44
14	25,62	154
16	19,13	115
17	14,81	89
18	18,64	112
20	13,31	80

Tabela 7. Źródło: opracowanie własne w programie Excel

Analizując procentowe rozkłady przedstawione w tabeli 7 oraz na wykresie 7, można zauważyć, że największa grupa respondentów (25,62%) stanowią niektóre studia wyższe (wartość kodowania 14). Najmniejszą grupę respondentów stanowią ci, którzy ukończyli szkołę podstawową (wartość kodowania 9) stanowiący 1,16% badanej populacji.

OCCUPATION (ZAWÓD)

Zmienna „occupation" odnosi się do zawodu respondentów i jest zakodowana zgodnie z klasyfikacją Hollingsheada, przy czym większe wartości kodów oznaczają zawody bardziej prestiżowe i społecznie wysoko cenione. Przykładowo, wartość kodu 1 oznacza zawody o niższym statucie społecznym, podczas gdy wyższe wartości kodów wskazują na zawody o wyższym statucie społecznym. Dzięki temu kodowaniu możliwe jest porównywanie i analizowanie różnych zawodów pod względem prestiżu i pozycji społecznej,



Wykres 8. Źródło: opracowanie własne w programie RStudio.

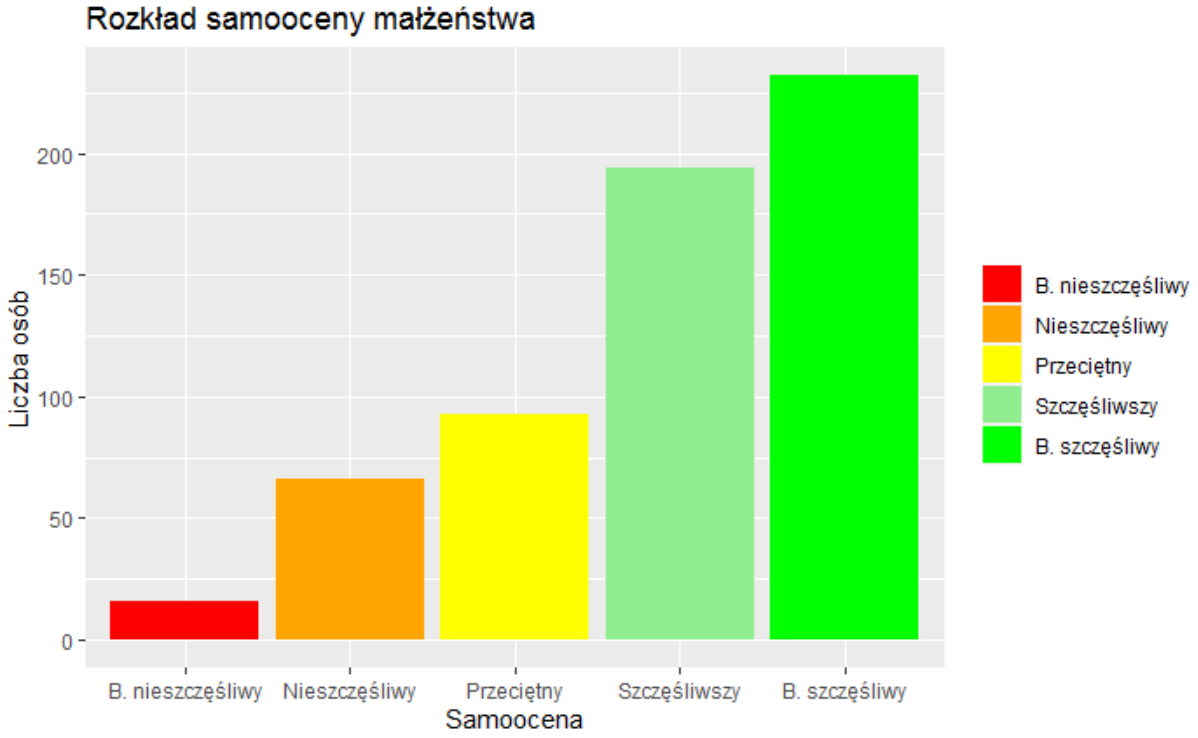
Obserwacja	Wartość w populacji [%] zaokrąglona do dwóch miejsc po przecinku	Liczba w populacji
1	18,80	113
2	2,16	13
3	7,82	47
4	11,31	68
5	33,94	204
6	23,79	143
7	2,16	13

Tabela 8. Źródło: opracowanie własne w programie Excel

Wykres 8 przedstawia liczbę obserwacji dla poszczególnych kodów zawodów, a tabela 8 prezentuje dodatkowo procentowy udział tych kodów w populacji. Na podstawie tych danych można zauważyć, że zawód o kodzie 5 ma największą liczbę obserwacji sięgająco 33%. Kolejną najliczniejszą grupą stanowią osoby z zawodem o kodzie 6 z liczbą obserwacji wynoszącą 143 (23,79%), co oznacza to, że w badanej populacji zawody o wyższym prestiżu i statusie społecznym są najliczniejsze. W badaniu wzięło udział 113 niewykwalikowanych osób, co stanowi 18,8% populacji.

RATING (SAMOOCENA MAŁŻEŃSTWA)

Ta zmienna opisuje ocenę respondentów dotyczącą ich własnego małżeństwa. Kodowanie wartości obejmuje skalę od 1 do 5, gdzie 1 oznacza "bardzo nieszczęśliwe", a 5 oznacza "bardzo szczęśliwe".



Wykres 9. Źródło: opracowanie własne w programie RStudio

Obserwacja	Wartość w populacji [%] zaokrąglona do dwóch miejsc po przecinku	Liczba w populacji
1	2,66	16
2	10,98	66
3	15,47	93
4	32,28	194
5	38,60	232

Tabela 9. Źródło: opracowanie własne w programie Excel

Tabela numer 9 przedstawia liczbę obserwacji samooceny małżeństwa wystawianych przez respondentów. Najmniejszy odsetek respondentów (2,66%) ocenił swoje małżeństwo jako "bardzo nieszczęśliwe" (ocena 1), podczas gdy największy odsetek (38,60%) określił je jako "bardzo szczęśliwe" (ocena 5). Wykres 9 dodatkowo wizualizuje te dane, przedstawiając liczbę obserwacji dla każdej z ocen (1-5).

INDEKS (INDEKS)

Zmienna „indeks" jest oznaczeniem indeksu. Z uwagi na brak istotności tej zmiennej w kontekście naszej analizy, postanawiamy z niej zrezygnować. Wynika to z faktu, że nie dostarcza ona istotnych informacji ani wartości w naszym badaniu. Skoncentrujemy się na pozostałych zmiennych, które są bardziej istotne i znaczące dla naszej analizy.

WSPÓŁCZYNNIK TAU KENDALLA (τ-KENDALLA)

Siłę współzależności zmiennych można wyrazić liczbowo na różne sposoby. Najbardziej popularnym jest współczynnik korelacji Pearsona. Nie możemy jednak go użyć ze względu na charakter zmiennych. W celu sprawdzenia, czy w zbiorze nie występują nadmiernie skorelowane cechy wykorzystano współczynnik (τ-Kendalla). Opiera się on na różnicy pomiędzy prawdopodobieństwem tego, że dwie zmienne układają się w tym samym porządku, a prawdopodobieństwem, tego że ich uporządkowanie się różni. Jest on wykorzystywany w celu określenia zgodności uporządkowań dla zmiennych określonych co najmniej na skali porządkowej. Współczynnik wynoszący co do modułu wartość większą od 0,7 świadczy o nadmiernym skorelowaniu cech. Sugeruje się usunięcie zmiennych, których współczynnik tau-Kendalla przekracza 0,7, w celu uniknięcia utraty informacji. Nadmierna

korelacja między zmiennymi może sugerować, że zawierają one podobne lub redundantne informacje, co może wprowadzać zakłócenia lub nadmierną wagę w analizie. Usunięcie tych zmiennych daje zmniejszoną , można zmniejszyć złożoność modelu i potencjalnie poprawia interpretowalność wyników.

	affairs	age	yearsmarried	children	religiousness	education	occupation	rating
affairs	1.0000	0.0788	0.1322	0.1336	-0.1169	0.0289	0.0284	-0.2229
age	0.0788	1.0000	0.7124	0.4333	0.1591	0.1600	0.1625	-0.1884
yearsmarried	0.1322	0.7124	1.0000	0.5312	0.1749	0.0544	0.0682	-0.2061
children	0.1336	0.4333	0.5312	1.0000	0.1186	0.0029	-0.0325	-0.1968
religiousness	-0.1169	0.1591	0.1749	0.1186	1.0000	-0.0342	-0.0173	0.0217
education	0.0289	0.1600	0.0544	0.0029	-0.0342	1.0000	0.4858	0.0454
occupation	0.0284	0.1625	0.0682	-0.0325	-0.0173	0.4858	1.0000	-0.0092
rating	-0.2229	-0.1884	-0.2061	-0.1968	0.0217	0.0454	-0.0092	1.0000

Źródło: opracowanie własne w programie RStudio

Jak wynika z obliczonego wskaźnika t-Kendalla staż małżeństwa jest nadmiernie skorelowany ze zmienną age. Zazwyczaj, im dłużej trwa małżeństwo, tym starsi są małżonkowie. Do dalszego badania zdecydowano się zachować zmienną yearsmarried. Dłuższy staż małżeństwa może oznaczać większą stabilność i więź emocjonalną między partnerami. Jednak długotrwałe małżeństwo nie jest gwarancją wierności. Wiele czynników związanych ze stażem małżeństwa takie jak monotonia, czy potrzeba nowości może mieć wpływ na decyzję osoby o zdradzie. Dodatkowo, zmienna yearsmarried charakteryzuje się niewiele większą korelacją ze zmienną objaśnianą affairs, czyli ma większy wpływ na nią niż zmienna age.
Ostatecznie do budowy modeli została wybrana zmienna objaśniana affairs, oraz potencjalne predyktory: gender, yearsmarried, children, religiousness, education, occupation i rating.

Zbiór danych został podzielony na zbiór uczący (tzw. treningowy) oraz zbiór testowy. Później dane uczące zostały wykorzystane do zbudowania modelu, a zbiór testowy do jego oceny. Dokonano losowego podziału w proporcji odpowiednio 70% i 30%.

Opis etapów budowy modeli logitowego i probitowego

Poziom istotności we wszystkich testach został przyjęty na poziomie $\alpha = 0{,}05$.

MODEL LOGITOWY

Na samym początku zbudowano model logitowy – logit0, gdzie do zmiennych objaśnianych wzięto wszystkie potencjalne predyktory:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.7447	1.9070	-0.9149	0.3602
gender1	-0.0395	0.3152	-0.1252	0.9004
yearsmarried0.42	0.2566	1.6302	0.1574	0.8749
yearsmarried0.75	-0.5300	1.5707	-0.3375	0.7358
yearsmarried1.5	-0.2378	1.2331	-0.1928	0.8471
yearsmarried4	0.6917	1.2053	0.5739	0.5661
yearsmarried7	0.3221	1.2253	0.2629	0.7926
yearsmarried10	0.8871	1.2238	0.7249	0.4685
yearsmarried15	0.6264	1.2040	0.5203	0.6028
children1	0.6372	0.4078	1.5623	0.1182
religiousness2	-1.0616	0.4819	-2.2030	0.0276
religiousness3	-0.6166	0.4847	-1.2723	0.2033
religiousness4	-1.2730	0.4744	-2.6835	0.0073
religiousness5	-1.0776	0.5648	-1.9079	0.0564
education12	0.8525	1.3854	0.6154	0.5383
education14	0.8318	1.3456	0.6182	0.5365
education16	0.9603	1.3731	0.6993	0.4843
education17	1.3951	1.3636	1.0231	0.3063
education18	1.1544	1.3637	0.8465	0.3973
education20	0.4355	1.4116	0.3085	0.7577
occupation2	1.1672	0.9647	1.2100	0.2263
occupation3	0.8245	0.6067	1.3590	0.1741
occupation4	1.3663	0.5536	2.4678	0.0136
occupation5	0.8988	0.4551	1.9749	0.0483
occupation6	1.3084	0.5478	2.3884	0.0169
occupation7	1.4156	0.8625	1.6413	0.1007
rating2	-0.1819	0.7985	-0.2278	0.8198
rating3	-1.1708	0.8081	-1.4487	0.1474
rating4	-1.5979	0.7827	-2.0415	0.0412
rating5	-1.8538	0.7905	-2.3451	0.0190

Źródło: opracowanie własne w programie RStudio

Test lokalny o istotności parametrów wykazał, iż zmienne które nie zostały podkreślone na czerwono nie są istotnie statystyczne. Oznacza to, że nie wpływają na zmienną objaśnianą. Podjęto decyzję o wykluczaniu po kolei nieistotnych predyktorów tak, aby zbudować model, w którym p-value przy większości zmiennych objaśniających będzie mniejsze od poziomu istotności. Pozwoli to na odrzucenie hipotezy zerowej przy każdym estymowanym parametrze strukturalnym, przez co każdy z parametrów będzie istotny statystycznie.

W modelu logit1 nie uwzględniono wśród zmiennych objaśniających zmiennej gender. Otrzymane wyniki nie wniosły za wiele zmian, ponieważ w dalszym ciągu większość parametrów nie była istotna statystycznie. Z tego właśnie

powodu, nie zdecydowano się przedstawić modelu logit1 w niniejszym raporcie.

Zmienne dla modelu logit2 to: children, religiousness, education, occupation I rating. Po zrezygnowaniu ze zmiennej opisującej staż małżeństwa, zmienna children jest istotna statystycznie. Zmienna religiousness również wypadła lepiej niż w modelu logit0, ponieważ tylko dla religiousness3 nie mamy podstaw do odrzucenia hipotezy zerowej. Zmienna education na żadnym poziomie nie okazała się istotna, zatem podjęto decyzję o zbudowaniu kolejnego modelu bez tej zmiennej.

Wyniki estymacji modelu logitowego logit3 o zmiennych objaśniających: children, religiousness, occupation i rating:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.4738	0.9195	-0.5152	0.6064
children1	0.9543	0.3364	2.8370	0.0046
religiousness2	-1.0460	0.4623	-2.2623	0.0237
religiousness3	-0.5677	0.4682	-1.2125	0.2253
religiousness4	-1.2838	0.4532	-2.8325	0.0046
religiousness5	-1.1503	0.5345	-2.1519	0.0314
occupation2	0.7570	0.8937	0.8471	0.3969
occupation3	0.7395	0.5710	1.2952	0.1953
occupation4	1.2813	0.5180	2.4733	0.0134
occupation5	1.0120	0.4080	2.4803	0.0131
occupation6	1.2804	0.4347	2.9455	0.0032
occupation7	1.4513	0.7814	1.8573	0.0633
rating2	-0.3686	0.7664	-0.4809	0.6306
rating3	-1.2906	0.7719	-1.6721	0.0945
rating4	-1.6686	0.7431	-2.2454	0.0247
rating5	-1.9358	0.7472	-2.5907	0.0096

Źródło: opracowanie własne w programie RStudio

Z uwagi na to, że zmienne podkreślone na czerwono dalej nie są istotne statystycznie, została podjęta decyzja o rekategoryzacji zmiennych religiousness, occupation i rating w celu sprawdzenia czy rekategoryzacja wpłynie na poprawienie istotności parametów w modelu. Zmienna opisująca poziom religijności została skategoryzowana w następujący sposób: 1 – ateista (anty), 2 – niepraktykujący (wcale + nieznacznie) oraz 3 – praktykujący (nieco + bardzo). Nowe kodowanie zmiennej occupation prezentuje się następująco: 1 – niski prestiż (dawne 1,2), 2 – średni prestiż (dawne 3,4,5) i 3 – wysoki prestiż (dawne 6 i 7). Zmienna opisująca zadowolenie ze związku małżeńskiego zyskała kodowanie: 1 – niezadowoleni (dawne 1,2), 2 – średnio zadowoleni (dawne 3) oraz 3 – zadowoleni (dawne 4,5). Na nowych danych zbudowano kolejny model logitowy – logit4. Wyniki estymacji parametrów strukturalnych logit4:

```
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.8338     0.6326 -1.3180   0.1875
children1         1.0156     0.3237  3.1378   0.0017
religiousness2   -0.7697     0.4183 -1.8402   0.0657
religiousness3   -1.1676     0.4273 -2.7325   0.0063
occupation2       0.9213     0.3647  2.5261   0.0115
occupation3       1.1646     0.3985  2.9225   0.0035
rating2          -0.9475     0.4085 -2.3192   0.0204
rating3          -1.4224     0.3191 -4.4580   0.0000
```

Źródło: opracowanie własne w programie RStudio

Dla modelu logitowego logit4 testy statystyczne dla poszczególnych parametrów przy zmiennych children1, religiousness3, occupation i rating wykazały, że każdy z tych parametrów jest istotny statystycznie. Jedynie jeden poziom zmiennej religiousness oraz wyraz wolny nie jest istotny statystycznie. Zdecydowano jednak nie wyrzucać wyrazu wolnego ze względu na to, że będzie odzwierciedlał szansę na niedotrzymanie wierności dla grupy referencyjnej. Poziom religijności ma znaczenie teoretyczne i może on wpływać na postawy, czy wartości związane z tematem badania - wiernością małżeńską.

Wykonano również testy LR oraz Walda. Obydwa testy służą do porównania dwóch modeli: Modelu 1, który obejmuje zmienne niezależne (children, religiousness, occupation, rating), oraz Modelu 2, który jest modelem uproszczonym i zawiera tylko wyraz wolny.

```
Likelihood ratio test

Model 1: affairs ~ children + religiousness + occupation + rating
Model 2: affairs ~ 1
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   8 -203.78
2   1 -228.44 -7 49.316  1.967e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Wald test

Model 1: affairs ~ children + religiousness + occupation + rating
Model 2: affairs ~ 1
  Res.Df Df      F    Pr(>F)
1    413
2    420 -7 6.0588 9.599e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Źródło: opracowanie własne w programie RStudio

P-value dla obydwu testów jest zdecydowanie mniejsze od przyjętego poziomu istotności $\alpha = 0.05$, zatem słusznym jest odrzucenie hipotezy zerowej na rzecz hipotezy alternatywnej. Oznacza to, że obydwa testy potwierdzają, że Model 1. jest statystycznie istotnie lepszy od Modelu 2.

Z pomocą powyższych rozważań i wykluczania nieistotnych statystycznie zmiennych doszliśmy do modelu logitowego, w którym korzystamy ze zmiennych children, religiousness, occupation i rating.

Postać modelu logitowego:

$$logit(p) = -0{,}8338 + 1{,}0156 \cdot children1 - 0{,}7697 \cdot religiousness2 - 1{,}1676 \cdot religiousness3 + 0.9213 \cdot occupation2 + 1.1646 \cdot occupation3 - 0.9475 \cdot rating2 - 1{,}4224 \cdot rating3$$

Grupą referencyjną w tym modelu są osoby w związku małżeńskim które nie mają dzieci, są ateistami, ich zawód jest nisko prestiżowy oraz nie są szczęśliwi ze swojej relacji ze współmałżonkiem.

MODEL PROBITOWY

Do modelowania cechy dychotomicznej wykorzystuje się również model probitowy, gdzie funkcją wiążącą jest funkcja probit. Zdecydowano się zbudować model o tych samych zmiennych niezależnych co w modelu logitowym:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.4203	0.3681	-1.1418	0.2535
children1	0.5540	0.1766	3.1367	0.0017
religiousness2	-0.4549	0.2510	-1.8123	0.0699
religiousness3	-0.6644	0.2544	-2.6113	0.0090
occupation2	0.4921	0.1993	2.4686	0.0136
occupation3	0.6308	0.2207	2.8584	0.0043
rating2	-0.5861	0.2450	-2.3924	0.0167
rating3	-0.8429	0.1917	-4.3974	0.0000

Źródło: opracowanie własne w programie RStudio

Dla modelu probitowego, testy statystyczne dla poszczególnych parametrów beta pokazały, że parametry przy zmiennych opisujących: posiadanie dzieci, wierzący (praktykujący), poziom prestiżu zawodu oraz poziom zadowolenia z małżeństwa są istotne statystycznie. Tak samo jak w przypadku modelu logitowego nie decydujemy się na usunięcie nieistotnych statystycznie zmiennych: wyrazu wolnego oraz religii na poziomie słabo wierzącym (niepraktykujący).

Model probitowy:

$$g(\mu) = -0{,}4203 + 0.5540 \cdot children1 - 0{,}4549 \cdot religiousness2 - 0{,}6644 \cdot religiousness3 + 0{,}4921 \cdot occupation2 + 0{,}6308 \cdot occupation3 - 0{,}5861 \cdot rating2 - 0{,}8429 \cdot rating3$$

Ocena modeli

W celu sprawdzenia, który z modeli - logitowy czy probitowy lepiej opisuje przewidywanie zdrad małżeńskich zrobiono porównanie dobroci dopasowania modeli logit4 i probit1:

	kryterium_AIC <dbl>	McFadden <dbl>	Cragg_Uhler <dbl>
logitowy	423.5638	0.1079408	0.1669340
probitowy	424.7177	0.1054153	0.1632475

Źródło: opracowanie własne w programie RStudio

Według kryterium informacyjnego oraz według miar pseudo R^2 lepszym modelem jest model logitowy. Dla tego modelu kryterium AIC przyjęło najniższą wartość, a miary pseudo R^2 największe. Warto zauważyć, że pod kątem dopasowania jest on jedynie nieco lepszym wyborem.

Następnie dokonano porównania jakości predykcji modeli logitowego i probitowego. Tablice trafności zostały zbudowane dla punktu odcięcia p*, które równe jest proporcji z próby uczącej.

```
Tablica trafności dla modelu logitowego - próba ucząca
            przewidywane
obserwowane   0   1
          0 215 108
          1  31  67
Tablica trafności dla modelu probitowego - próba ucząca
            przewidywane
obserwowane   0   1
          0 213 110
          1  30  68
Tablica trafności dla modelu logitowego - próba testowa
            przewidywane
obserwowane  0  1
          0 81 47
          1 20 32
Tablica trafności dla modelu probitowego - próba testowa
            przewidywane
obserwowane  0  1
          0 81 47
          1 20 32
```

Źródło: opracowanie własne w programie RStudio

Na podstawie tablic trafności zostały wyliczone miary jakości predykcji.

Pozwalają one ocenić, jak dobrze modele radzą sobie z przewidywaniem zmiennych odpowiedzi, czyli z przewidywaniem niedotrzymania wierności w związku małżeńskim.

	ACC	ER	SENS	SPEC	PPV	NPV
	Zbiór danych uczących					
Logit	0.6698	0.3302	0.6837	0.6656	0.3829	0.8740
Probit	0.6675	0.3325	0.6939	0.6594	0.3820	0.8765
	Zbiór danych testowych					
Logit	0.6278	0.3722	0.6154	0.6328	0.4051	0.802
Probit	0.6278	0.3722	0.6154	0.6328	0.4051	0.802

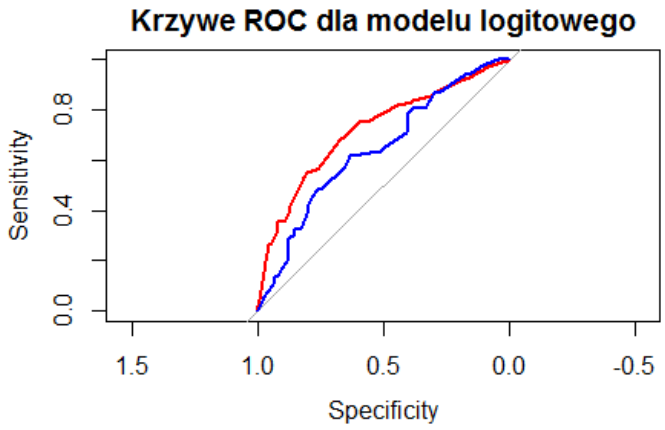
Tabela10: Źródło: opracowanie własne w programie RStudio

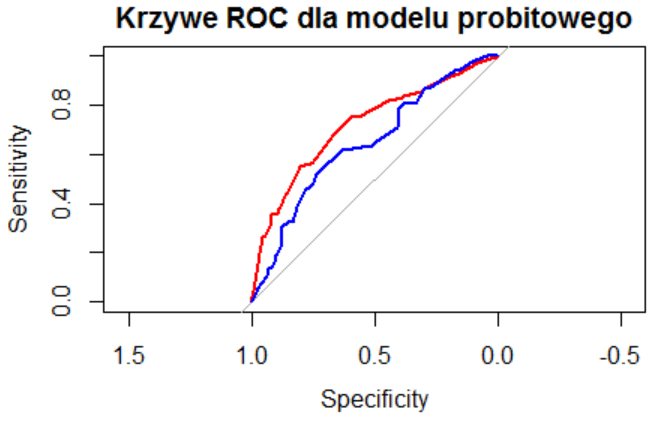
Modele logitowe i probitowe osiągają podobne wyniki dla miar jakości predykcji na zbiorze uczącym oraz na zbiorze danych testowych. Oba modele mają podobną dokładność, cechują się podobną wrażliwością, specyficznością, oraz dodatnią i ujemną wartością predykcyjną. Następnie, porównując miary na zbiorze uczącym i testowym, można stwierdzić, że modele zachowują swoją skuteczność przy zmianie zbioru danych.

Dokładność (ACC) to stosunek liczby poprawnie zaklasyfikowanych przypadków do liczby wszystkich przypadków. W przypadku zbioru uczącego wynosi około 67%, zaś na zbiorze testowym około 63%. Oznacza to, że modele są w stanie poprawnie przewidzieć około 67% (lub 63%) przypadków. Modele mają około 33% (zbiór uczący) lub 37% (zbiór testowy) błędu klasyfikacji. Patrząc na czułość możemy powiedzieć, że modele mogą poprawnie zidentyfikować około 68%(logit), 69%(probit) lub 62% osób, które popełniły zdradę. Jednocześnie modele mogą poprawnie zidentyfikować około 67%(logit) 66%(probit) lub około 63% osób które nie popełniły zdrady. W przypadku zbioru uczącego dodatnia wartość predykcyjna wynosi około 38% dla obydwu modeli, zaś dla zbioru testowego około 41%. Oznacza to, że około 38% lub 41% osób zakwalifikowanych faktycznie popełniło zdradę. Nie jest to jendak zbyt dobry wynik. Gdyby wierzyć tylko i wyłącznie samemu modelowi, mogłoby dojść do rozpadu wielu małżeństw. Z drugiej strony, jeżeli oddalibyśmy do użytku model terapeutom do spraw problemów małżeńskich to taki wynik jest jak najbardziej satysfakcjonujący, z uwagi na to, że mogą oni zrozumieć jakie potencjalne cechy mogą wpływać na zdradę w małżeństwie. Dobrze o naszych modelach świadczy ujemna wartość predykcyjna. Na zbiorze uczącym wynosi ona około 87%(logit) 88%(probit), na zbiorze danych testowych wynosi około 80%. Świadczy to, że około 87% lub 80% osób

zakwalifikowanych jako niezdradzające faktycznie nie popełniło zdrady. Model logitowy jak i probitowy dobrze radzą sobie z wykrywaniem małżonków wiernych. Podsumowując analizę miar jakości predykcji, stworzony model logitowy i probitowy nie charakteryzują się bardzo wysoką skutecznością w przewidywaniu niewierności, co oczekuje się od wybranego modelu. Gdyby model miał być użyteczny w celach komercyjnych należałoby go ulepszyć w celu uzyskania lepszych wyników predykcji.

Następnie sporządzono wykresy krzywej ROC, które prezentują jakość predykcji modelu dla wszystkich możliwych punktów ocięcia p*. Dla każdego modelu przedstawiono dwa zestawy danych: zbiór uczący i zbiór testowy. Wykresy krzywej ROC dla zbioru uczącego zostały oznaczone czerwoną linią, zaś dla zbioru testowego niebieską. Wykresy pozwolą zaobserwować, jak dobrze stworzone modele radzą sobie w rozróżnianiu wierności od niewierności na podstawie dostępnych zmiennych. Zaprezentowanie krzywej ROC na dwóch zbiorach danych pozwoli również na stwierdzenie czy modele zachowują skuteczność na nowych danych (zbiór testowy) oraz czy są w stanie rozróżnić wierność od niewierności w zadowalający sposób.





Źródło: opracowanie własne w programie RStudio

Jednoznacznie widzimy, iż modele z danych uczących są lepiej dopasowane niż z danych testowych. Wynika to z faktu, iż czerwona linia ma bardziej pełny kształt i jest bliżej punktu (0,1). Kształt krzywych dla obydwu modeli niestety można zaklasyfikować bardziej jako półokrąg niż prostokąt, więc możemy się domyślić, że modele te nie są „idealnie" dopasowane do danych. Z drugiej strony, krzywe te nie opierają się o przekątną linię referencyjną. Na podstawie wykresu możemy stwierdzić, iż modele te mają średnie dopasowanie. Aby móc to szczegółowo zweryfikować, należy obliczyć miarę AUC dla obydwu modeli.

Celem przeprowadzenia jeszcze bardziej szczegółowej analizy skuteczności modeli w przewidywaniu niewierności, obliczono miarę AUC. Jest to miara, która wyraża pole powierzchni pod krzywą ROC.

Model logitowy		Model probitowy	
Z. d. uczący	Z. d. testowy	Z. d. uczący	Z. d. testowy
0.72	0.6411	0.7191	0.6448

Tabela 11: Źródło: opracowanie własne w programie RStudio

Analizując wartości pola pod wykresem ROC model logitowy na zbiorze danych uczących osiągnął niewiele wyższą wartość niż model probitowy. Na zbiorze danych testowych zaobserwowano odwrotną zależność. Nie są to jednak znaczące różnice, ponieważ dokonując przybliżenia do dwóch miejsc po przecinku w modelu probitowym otrzymano by te same wartości pól. Zatem model logitowy i probitowy jest tak samo skuteczny w przewidywaniu niewierności. Porównując AUC między zbiorem danych uczących, a zbiorem danych testowych zaobserwowano spadek wartości dla obydwu modeli. Oznacza to lekki spadek zdolności predykcyjnej w rozpoznawaniu niewierności na zbiorze testowym. Może to wynikać z różnych przyczyn, na przykład z przeuczenia zbioru uczącego. Model przeuczony przeważnie zawsze dopasowuje się gorzej do nowych danych.

W oparciu o powyższe obserwacje i wyniki do dalszej analizy został wybrany model logitowy. Jego lepsze wyniki w kryterium informacyjnym AIC, miarach pseudo R^2, miarach jakości predykcji oraz pole pod krzywą ROC potwierdzają o jego (minimalnie) lepszej skuteczności do badanego zagadnienia. Dodatkowo należy zaznaczyć, że model logitowy ma o wiele ciekawszą interpretację, co też przemawia za jego wyborem.

Interpretacja modelu wynikowego

Postać modelu wynikowego (logitowego – logit4):

$$logit(p) = -0{,}8338 + 1{,}0156 \cdot children1 - 0{,}7697 \cdot religiousness2 - 1{,}1676 \cdot religiousness3 + 0.9213 \cdot occupation2 + 1.1646 \cdot occupation3 - 0.9475 \cdot rating2 - 1{,}4224 \cdot rating3$$

Celem dokonania interpretacji obliczono ilorazy szans dla poszczególnych zmiennych:

(Intercept)	children1	religiousness2	religiousness3	occupation2
0.434	2.761	0.463	0.311	2.512
occupation3	rating2	rating3		
3.205	0.388	0.241		

Źródło: opracowanie własne w programie RStudio

Szansa na niewierność wśród grupy referencyjnej (osoby w związku małżeńskim którzy nie mają dzieci, są ateistami, ich zawód jest nisko prestiżowy oraz nie są szczęśliwi ze swojej relacji ze współmałżonkiem) wynosiła około 0,434.

Szansa na niewierność wśród osób mających dzieci wzrasta około 2,761 razy niż u małżeństw bez dzieci, ceteris paribus, to znaczy na tym samym poziomie religijności, w tym samym zawodzie i na tym samym poziomie zadowolenia z małżeństwa.

Dla zmiennych religiousness2 oraz religiousness3 ilorazy szans wynoszą odpowiednio około 0,463 i 0,311. Oznacza to, że im osoba bardziej wierzy i praktykuje są mniejsze szanse na to, że pokusi się o zdradę, ceteris paribus.

Osoby związane z zawodami o średnim i wysokim prestiżu mają odpowiednio 2.512 i 3.205 razy większe szanse na zdradę w porównaniu do osób z innymi zawodami, przy zachowaniu stałości pozostałych zmiennych.

Szansa na niewierność u osób średnio i zadowolonych z relacji w małżeństwie jest odpowiednio 0,388 oraz 0,241 większa niż u osób niezadowolonych z małżeństwa, ceteris paribus. Oznacza to, że im osoba jest bardziej zadowolona z kondycji swojej relacji tym są mniejsze szanse na zdradę, przy zachowaniu stałości pozostałych zmiennych.

Zakończenie

Zdrady małżeńskie od wieków intrygują i niepokoją ludzi na całym świecie. Wierność jest podstawowym fundamentem zaufania. Z drugiej strony, świat jest pełen pokus, przez co wierność jest często wystawiana na próbę. W naszej pracy próbowaliśmy zgłębić tajemnicę zdrad i odkryć zmienne które wpływają na ich występowanie. Wykorzystaliśmy w tym celu dwa modele: model logitowy i probitowy. Ostatecznie po porównaniu ich dopasowania oraz ich jakości predykcji postanowiliśmy wybrać model logitowy jako ten, który pozwoli zgłębić czynniki wpływające na szansę zdrad w związku. Przeszukaliśmy dane, a następnie zbadaliśmy zmienne pod kątem ich istotności, co doprowadziło nas do odkrycia małego ułamka kulis zdrad małżeńskich. Nasze wyniki są fascynujące ale jednocześnie niejednoznaczne, ponieważ model miał mniejsze pole pod wykresem ROC na zbiorze danych testowych niż uczących. Nie można zapomnieć, że relacje międzyludzkie są złożone i pełne niewiadomych, przez co można śmiało powiedzieć, iż nasz model stanowi jedynie część całej układanki. W celu ulepszenia modelu należałoby spotkać się ze specjalistami w dziedzinie terapii małżeńskich i omówić z nimi dokładny mechanizm zdrad. Niemniej jednak uzyskane rezultaty oraz dość zadowalające miary jakości predykcji wskazują na potencjał wykorzystania naszego modelu w przyszłości.