Sprawozdanie 3

Analiza Danych Ankietowych

Mateusz Machaj (262288), Adam Kawałko (262329)

26.05.2023

Wstęp

W naszym raporcie zaprezentujemy funkcjonalność pakietu *R* związaną z analizami dla badań wykonywanych kilkukrotnie, modelami log-liniowymi oraz zaawansowanymi przypadkami badania niezależności. Odpowiemy na problemy zawarte w zadaniach zestawu trzeciego.

Będziemy w praktyczny sposób stosować funkcje zarówno dostępne w bibliotekach *R*, jak i samodzielnie zaimplementowane. Wykonamy też badania, pozwalające na lepsze zrozumienie działania testów. W zadaniach, które dotyczą otrzymanych danych zakładamy, że zostały przeprowadzone badania na dwustu pracownikach pewnej, dużej korporacji. Zadawano im konkretne pytania, na które musieli podać jedną z dostępnych odpowiedzi. Sprawdzano między innymi wiek, wykształcenie czy dział zatrudnienia. W przykładach kodu tabela z tymi danymi jest przechowywana w zmiennej `personel`. Kilka zaś jej wierszy przedstawiamy poglądowo w tab. 1

Używamy w analizach następującej konwencji nazewniczej kolumn:

- **D** – Dział zatrudnienia pracownika (Z – zaopatrzenia, P – produkcyjny, S – sprzedaży/marketingu, O – obsługi kadrowo-płacowej).
- **S** – Stanowisko (1 jeśli kierownicze, w przeciwnym razie 0).
- **A1** – Zadowolenie z atmosfery w pierwszym badanym okresie (w skali od -2 do 2).
- **A2** – Zadowolenie z atmosfery w drugim badanym okresie (w skali od -2 do 2).
- **W1** – Zadowolenia z wynagrodzenia w pierwszym badanym okresie (w skali od -2 do 2 bez 0).
- **W2** – Zadowolenia z wynagrodzenia w drugim badanym okresie (w skali od -2 do 2 bez 0).
- **P** – Płeć (K, M).
- **Wiek** – Przedział wiekowy pracownika (skala od 1 do 4 obejmuje w latach grupy < 26, $26-35$, $36-50$ oraz > 50).
- **Wyk** – Rodzaj wykształcenia (zawodowe - 1, średnie - 2, wyższe 3).

Jeżeli chodzi o poziom istotności, na którym wykonywano testy, jeśli nie podano inaczej, jest to zawsze $\alpha = 0.05$.

Tabela 1: Fragment tabeli analizowanych danych dotyczących personelu.

D	S	A1	A2	W1	W2	P	Wiek	Wyk
O	0	1	1	-2	-2	M	4	2
O	0	0	0	-2	-2	M	4	2
O	0	1	1	2	2	M	4	2
O	0	-1	0	-2	-2	K	4	2
O	1	1	1	2	2	K	4	3
O	1	0	0	1	2	K	4	3

Tabela 2: Odpowiedzi ankietowanych w podziale na zadowolenie z atmosfery w pierwszym okresie (wiersze) oraz drugim okresie (kolumny).

	-2	-1	0	1	2
-2	10	2	1	1	0
-1	0	15	1	1	0
0	1	1	32	6	0
1	0	0	1	96	3
2	1	1	0	1	26

Tabela 3: Odpowiedzi ankietowanych w podziale na zadowolenie z wynagrodzenia w pierwszym okresie (wiersze) oraz drugim okresie (kolumny).

	-2	-1	1	2
-2	74	0	0	0
-1	0	19	1	0
1	0	0	1	1
2	0	0	0	104

Część I

Zadanie 1

W pierwszym zadaniu zweryfikujemy hipotezę, że atmosfera w miejscu pracy w pierwszym badanym okresie oraz po roku od pierwszego badania odpowiada modelowi symetrii.

Najpierw próbujemy wykonać test McNemary. Używamy do tego takiego kodu, jak poniżej.

```
ftable(personel$A1, personel$A2) %>% as.table() %>% mcnemar.test()
```

Okazuje się, że p-wartość nie istnieje. Dzieje się tak, ponieważ dla naszych danych w macierzy istnieją zera na odpowiadających sobie miejscach. Można to łatwo zaobserwować w tab. 2.

Mimo to, możemy ciagle zastosować test oparty na ilorazie wiarygodności:

```
grps <- personel %>% select(c(A1, A2)) %>% group_by(A1, A2, .drop = F) %>% count()
symmetry <- glm(n ~ Symm(A1, A2), data = grps, family = poisson)
p <- 1 - pchisq(symmetry$deviance, symmetry$df.residual)
```

Otrzymujemy p-wartość 0.206. Jest ona większa od poziomu $\alpha = 0.05$, zatem przyjmujemy hipotezę zerową. Uznajemy, że zadowolenie z atmosfery nie zmieniło się wraz z okresem.

Zadanie 2

W kolejnym zadaniu zarówno problem jak i rozwiązanie są analogiczne. Nie przyglądamy się natomiast zadowoleniu z atmosfery a wynagrodzenia w dwóch badanych okresach.

Okazuje się, iż z tego samego powodu co poprzednio (wystarczy spojrzeć na tab. 3), test McNemary nie może zwrócić żadnej p-wartości. Z problemem radzimy sobie dzięki testowi opartemu na ilorazie wiarygodności i znów przyjmujemy H_0, z p-wartością na poziomie 0.837. Nie ma powodu aby sądzić, iż zadowolenie z wynagrodzenia zmieniło się w czasie.

Zadanie 3

W tym zadaniu ponownie podejdziemy do badania zadowolenia z wynagrodzenia w czasie, ale skonsolidujemy oba poziomy negatywnych odpowiedzi do jednej i to samo z pozytywnymi. Rezultat przedstawiony jest

Tabela 4: Odpowiedzi ankietowanych skonsolidowane do negatywnych i pozytywnych w podziale na zadowolenie z wynagrodzenia w pierwszym okresie (wiersze) oraz drugim okresie (kolumny).

	-1	1
-1	93	1
1	0	106

bardzo przejrzyście w tab. 4.

Jak można się domyślać, tym razem p-wartość w teście McNemary istnieje. Wynosi ona 1, a zatem nie odrzucamy hipotezy zerowej. Teraz też tabla ma z resztą wymagany dla testu wymiar 2x2. Poziomy krytyczne uzyskane w testach Z i Z_0 to odpowiednio 0.316 oraz 0.317. Nie ma zatem w ogóle podstaw do sądzenia, że zaszły znaczne zmiany w nastrojach odnośnie wyngrodzenia.

Zadanie 4

To zadanie zaczniemy od prezentacji zaimplementowanych testów Z oraz Z_0 symetrii.

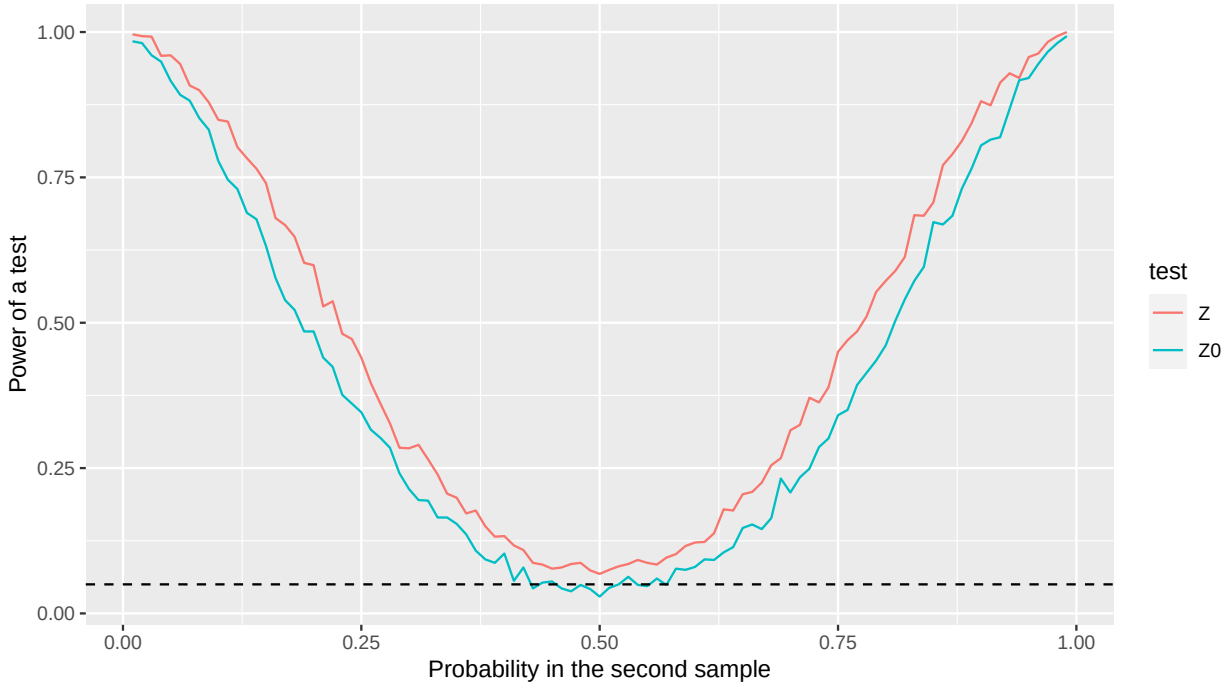
```
sym.z.test <- function(x) {
  P <- (x / sum(x)) %>% as.table()
  D <- (rowSums(P)[1] - colSums(P)[1]) %>% as.numeric()
  # sigma
  sg <- sqrt((
    rowSums(P)[1] * (1 - rowSums(P)[1])
    +  colSums(P)[1] * (1 - colSums(P)[1])
    - 2 * (P[1, 1] * P[2, 2] - P[1, 2] * P[2, 1])
  ) / sum(x)) %>% as.numeric()
  # stat
  z <- D / sg
  # p-value
  p <- 2 * (1 - pnorm(abs(z)))
  return(p)
}
```

```
sym.z0.test <- function(x) {
  # stat
  z0 <- (x[1, 2] - x[2, 1]) / sqrt((x[1, 2] + x[2, 1]))
  # p-value
  p <- 2 * (1 - pnorm(abs(z0)))
  return(p)
}
```

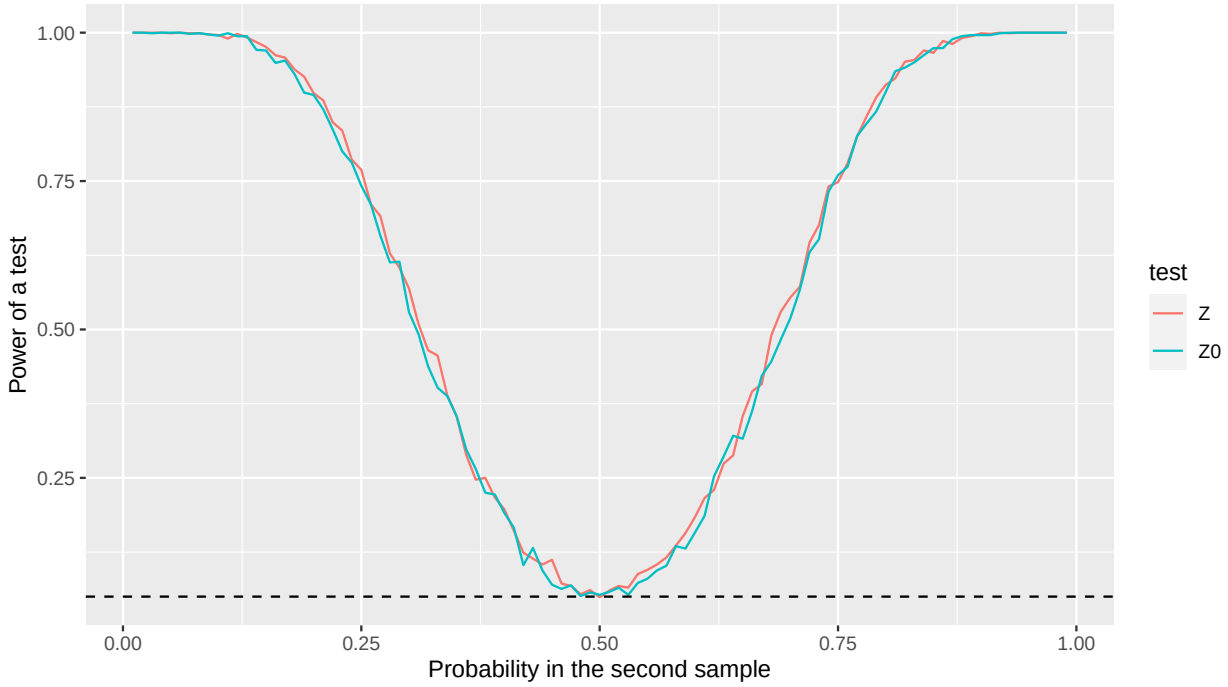
Obie funkcje bazują na wzorach podanych na wykładzie.

Korzystając z p-wartości wyznaczanych przez te funkcje, będziemy teraz badać moce obu testów. W symulacjach używać będziemy dwóch prób z rozkładu dwupunktowego o różnych rozmiarach – $n \in \{20, 50, 100, 1000\}$. Pierwsza z prób będzie miała prawdopodobieństwo sukcesu równe $p_1 = 0.5$, a druga zmienne $p_2 \in (0, 1)$. Przy liczbie powtórzeń Monte Carlo $M = 1000$ będziemy sprawdzać, ile odrzuceń fałszywych hipotez otrzymujemy.

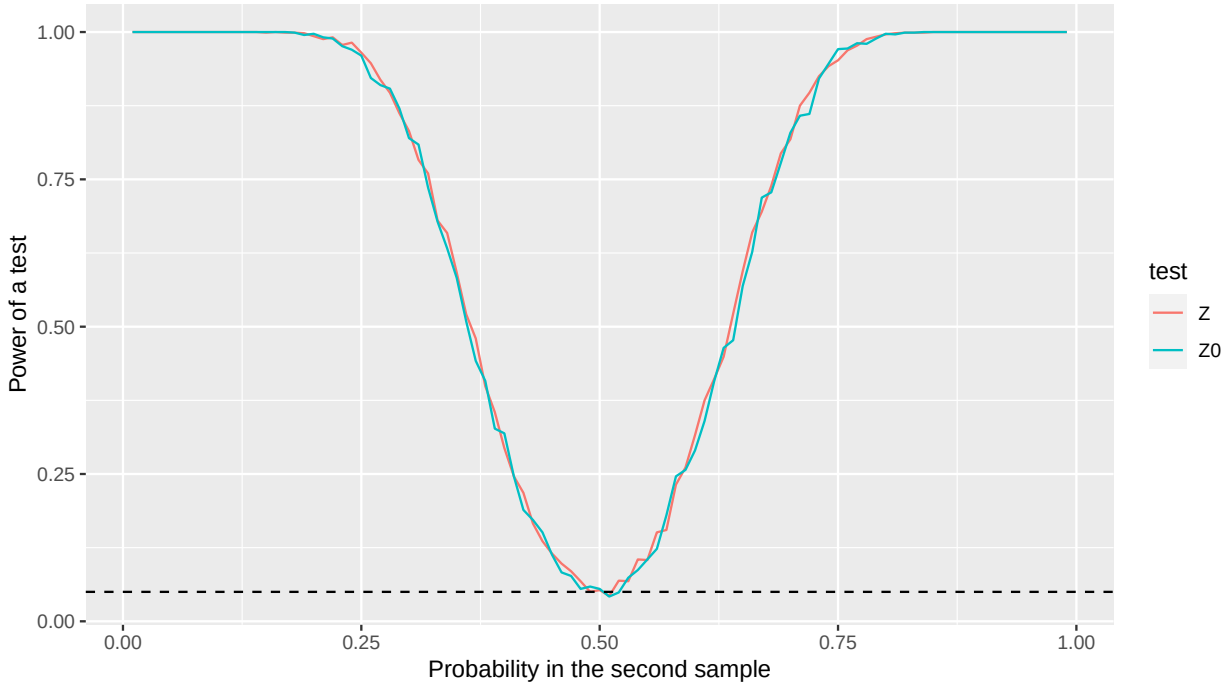
Analizując wyniki w formie wykresów w porównaniu z poziomem istotności zauważamy, że dla małych n (rys. 1) test Z cechuje się znacznie większą od testu Z_0 mocą dla całego zakresu p_2. Różnica pomiędzy wartościami jest też generalnie podobna. Można jedynie zauważać podobne do siebie wyniki w okolicach bardzo bliskich brzegom przedziału (0, 1). Tam p-wartości zbliżają się do jedynki, pozostając na podobnym poziomie. Warto też zauważyć, że gdy hipoteza zerowa jest prawdziwa, test Z ma wyraźnie większą moc, niż poziom istotności. Nie jest to prawdą dla testu Z_0.



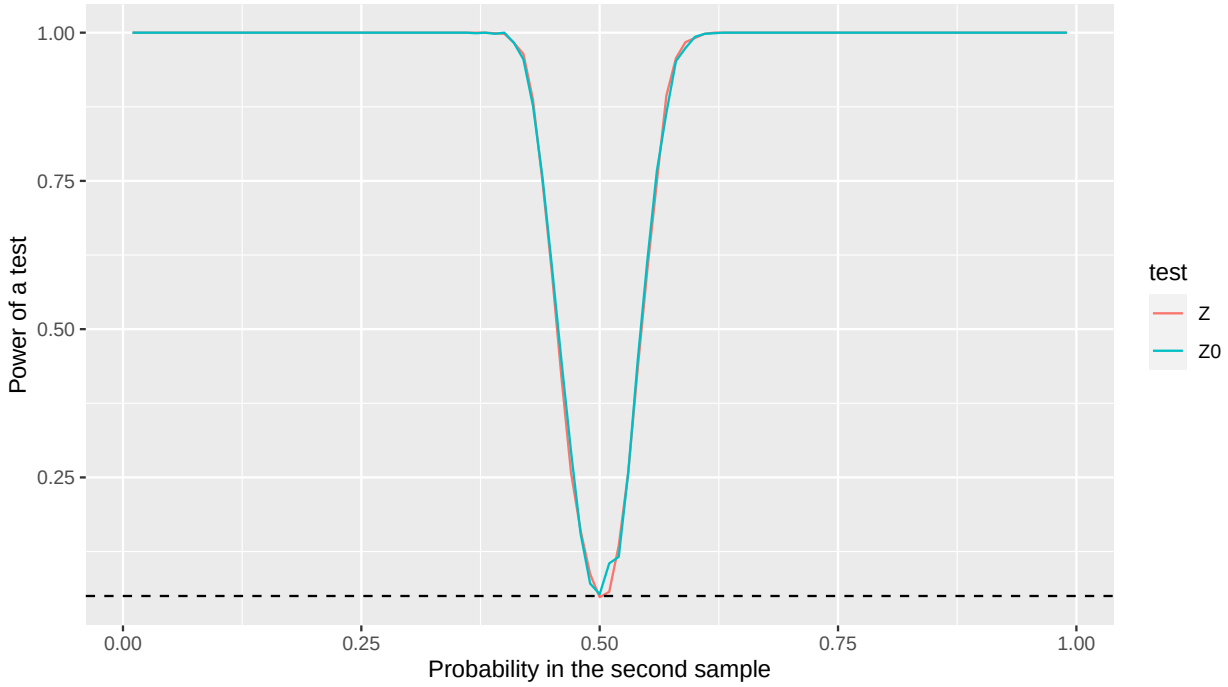
Rysunek 1: Moce testów Z i Z_0 dla różnych prawdopodobieństw p_2 oraz stałego $p_1 = 0.5$, na poziomie $\alpha = 0.05$. Rozmiar prób wynosi tu $n = 20$.



Rysunek 2: Moce testów Z i Z_0 dla różnych prawdopodobieństw p_2 oraz stałego $p_1 = 0.5$, na poziomie $\alpha = 0.05$. Rozmiar prób wynosi tu $n = 50$.



Rysunek 3: Moce testów Z i Z_0 dla różnych prawdopodobieństw p_2 oraz stałego $p_1 = 0.5$, na poziomie $\alpha = 0.05$. Rozmiar prób wynosi tu $n = 100$.



Rysunek 4: Moce testów Z i Z_0 dla różnych prawdopodobieństw p_2 oraz stałego $p_1 = 0.5$, na poziomie $\alpha = 0.05$. Rozmiar prób wynosi tu $n = 1000$.

Tabela 5: Zestawienie predykcji i empirycznych wartości w grupach dla modelu log-liniowego [1 3] i zmiennych 1 - S, 2 - W1, 3 - Wyk.

S	W1	Wyk	n	fitted_n	abs_error
0	-2	1	19	8.866	10.134
0	-2	2	40	30.275	9.725
0	-2	3	5	4.109	0.891
0	-1	1	3	8.866	5.866
0	-1	2	15	30.275	15.275
0	-1	3	0	4.109	4.109
0	1	1	0	8.866	8.866
0	1	2	0	30.275	30.275
0	1	3	0	4.109	4.109
0	2	1	18	8.866	9.134
0	2	2	68	30.275	37.725
0	2	3	5	4.109	0.891
1	-2	1	1	1.384	0.384
1	-2	2	5	4.725	0.275
1	-2	3	4	0.641	3.359
1	-1	1	0	1.384	1.384
1	-1	2	2	4.725	2.725
1	-1	3	0	0.641	0.641
1	1	1	0	1.384	1.384
1	1	2	0	4.725	4.725
1	1	3	2	0.641	1.359
1	2	1	0	1.384	1.384
1	2	2	10	4.725	5.275
1	2	3	3	0.641	2.359

Patrząc na dalsze ilustracje – rys. 2 - 4 obserwujemy, że różnica w mocach pomiędzy omawianymi testami zanika wraz ze wzrostem n. Dla $n = 50$, czyli na rys. 2, jest już właściwie w ogóle nie widoczna. P-wartości na tych trzech wykresach w przypadku $p_2 = p_1$ zdają się pozostawać dokładnie na poziomie α już w obu przypadkach. Widziomy też, że im większe n, tym szerszy zakres p_2 może cieszyć się mocą niemalże równą 1. Podczas, gdy dla $n = 100$ i obu testów, przy $|p_2 - p_1| = 0.1$, $\beta_\varphi(p_2) < 0.5$ (rys. 3). Gdy mamy do czynienia z dziesięciokrotnie większym n (rys. 4), o teij samej funkcj mocy we wspomnianych punktach powiemy, że $\beta_\varphi(p_2) \approx 1$.

Część II

Zadanie 5

Rozpatrzymy teraz modele log-liniowe dla różnych zmiennych. W tym zadaniu przyjmiemy za zmienną 1 zajmowane stanowisko (S), za 2 – zadowolenie z wynagrodzenia w pierwszym okresie ($W1$), a przez 3 rozumiemy wykształcenie (Wyk). Będziemy rozważać kolejne modele, podając ich interpretację oraz przewidywania.

Model [1 3]

Zakładamy, że stanowisko i wykształcenie mają dowolne rozkłady, a zadowolenie z wynagrodzenia rozkład równomierny. Dodatkowo, wszystkie te zmienne są niezależne.

Mając liczności w grupach w tabeli `personel.s_w1_wyk`, wywołujemy funkcję `glm` z następującymi argumentami:

Tabela 6: Zestawienie predykcji i empirycznych wartości w grupach dla modelu log-liniowego [13] i zmiennych 1 - S, 2 - W1, 3 - Wyk.

S	W1	Wyk	n	fitted_n	abs_error
0	-2	1	19	10.00	9.00
0	-2	2	40	30.75	9.25
0	-2	3	5	2.50	2.50
0	-1	1	3	10.00	7.00
0	-1	2	15	30.75	15.75
0	-1	3	0	2.50	2.50
0	1	1	0	10.00	10.00
0	1	2	0	30.75	30.75
0	1	3	0	2.50	2.50
0	2	1	18	10.00	8.00
0	2	2	68	30.75	37.25
0	2	3	5	2.50	2.50
1	-2	1	1	0.25	0.75
1	-2	2	5	4.25	0.75
1	-2	3	4	2.25	1.75
1	-1	1	0	0.25	0.25
1	-1	2	2	4.25	2.25
1	-1	3	0	2.25	2.25
1	1	1	0	0.25	0.25
1	1	2	0	4.25	4.25
1	1	3	2	2.25	0.25
1	2	1	0	0.25	0.25
1	2	2	10	4.25	5.75
1	2	3	3	2.25	0.75

```
glm(n ~ S + Wyk, data = personel.s_w1_wyk, family = poisson)
```

Po wynikach z tab. 5 widzimy, iż niektóre z kategorii predykowane są z dość małym błędem względem zebranych danych. Niestety w ogólnym sensie, wartości dopasowane w ogóle nie przypominają empirycznych.

Model [13]

Zakładamy, że stanowisko i wykształcenie mają dowolne rozkłady, a zadowolenie z wynagrodzenia rozkład równomierny. Dodatkowo, zadowolenie z wynagrodzenia jest niezależne od stanowiska i od wykształcenia, ale samo stanowisko i wykształecenie nie są niezależne.

Wywołujemy tym razem

```
glm(n ~ S + Wyk + S * Wyk, data = personel.s_w1_wyk, family = poisson)
```

Dopasowanie z tab. 6 ciągle nie satysfakcjonuje. Taki model nie wydaje się być dobrym kandydatem.

Model [1 2 3]

Zakładamy, że stanowisko, zadowolenie z wynagrodzenia oraz wykształcenie, mają dowolne rozkłady. Wszystkie te zmienne są zaś wzajemnie niezależne.

```
glm(n ~ S + W1 + Wyk, data = personel.s_w1_wyk, family = poisson)
```

Tym razem w tab. 7 odczytujemy już poprawę zachowania. Nie ma już spotykanych wcześniej predykcji na

Tabela 7: Zestawienie predykcji i empirycznych wartości w grupach dla modelu log-liniowego [1 2 3] i zmiennych 1 - S, 2 - W1, 3 - Wyk.

S	W1	Wyk	n	fitted_n	abs_error
0	-2	1	19	13.122	5.878
0	-2	2	40	44.807	4.807
0	-2	3	5	6.081	1.081
0	-1	1	3	3.547	0.547
0	-1	2	15	12.110	2.890
0	-1	3	0	1.644	1.644
0	1	1	0	0.355	0.355
0	1	2	0	1.211	1.211
0	1	3	0	0.164	0.164
0	2	1	18	18.442	0.442
0	2	2	68	62.972	5.028
0	2	3	5	8.546	3.546
1	-2	1	1	2.048	1.048
1	-2	2	5	6.993	1.993
1	-2	3	4	0.949	3.051
1	-1	1	0	0.554	0.554
1	-1	2	2	1.890	0.110
1	-1	3	0	0.256	0.256
1	1	1	0	0.055	0.055
1	1	2	0	0.189	0.189
1	1	3	2	0.026	1.974
1	2	1	0	2.878	2.878
1	2	2	10	9.828	0.172
1	2	3	3	1.334	1.666

Tabela 8: Zestawienie predykcji i empirycznych wartości w grupach dla modelu log-liniowego [12 3] i zmiennych 1 - S, 2 - W1, 3 - Wyk.

S	W1	Wyk	n	fitted_n	abs_error
0	-2	1	19	13.12	5.880
0	-2	2	40	44.80	4.800
0	-2	3	5	6.08	1.080
0	-1	1	3	3.69	0.690
0	-1	2	15	12.60	2.400
0	-1	3	0	1.71	1.710
0	1	1	0	0.00	0.000
0	1	2	0	0.00	0.000
0	1	3	0	0.00	0.000
0	2	1	18	18.66	0.655
0	2	2	68	63.70	4.300
0	2	3	5	8.64	3.645
1	-2	1	1	2.05	1.050
1	-2	2	5	7.00	2.000
1	-2	3	4	0.95	3.050
1	-1	1	0	0.41	0.410
1	-1	2	2	1.40	0.600
1	-1	3	0	0.19	0.190
1	1	1	0	0.41	0.410
1	1	2	0	1.40	1.400
1	1	3	2	0.19	1.810
1	2	1	0	2.66	2.665
1	2	2	10	9.10	0.900
1	2	3	3	1.24	1.765

poziomi 30 podczas, gdy wartości w oryginalnych danych były równe 0. Przy liczbie 200 ankietowanych i 24 grupach danych był to istotnie nieporządany rezultat. Tu błąd bezwzględny w ogóle nie przekracza poziomu 6. Wygląda na to, iż żadna ze zmiennych nie ma równomiernego rozkładu.

Model [12 3]

Zakładamy, że stanowisko, zadowolenie z wynagrodzenia oraz wykształcenie, mają dowolne rozkłady. Wykształcenie jest niezależne od stanowiska i zadowolenia z wynagrodznia, ale samo stanowisko oraz zadowolenie z wynagrodznia nie są niezależne.

```
glm(n ~ S + W1 + Wyk + S * W1, data = personel.s_w1_wyk, family = poisson)
```

Dla kolejnego modelu nie obserwujemy istotnej poprawy, chodziaż po dodaniu jednej zależności można zaobserwować nieco lepsze wyniki w tab. 8.

Model [12 13]

Zakładamy, że stanowisko, zadowolenie z wynagrodzenia oraz wykształcenie, mają dowolne rozkłady. Do tego, przy ustalonym stanowisku, wykształcenie i zadowolenie z wynagrodzenia są niezależne (*W1* i *Wyk* są warunkowo niezależne).

```
glm(n ~ S + W1 + Wyk + S * W1 + S * Wyk, data = personel.s_w1_wyk, family = poisson)
```

Dla kolejnego przykładu – z tab. 9 – można powiedzieć, że wyniki są znów lepsze, choćby dlatego, że maksymalny błąd się zmniesza.

Tabela 9: Zestawienie predykcji i empirycznych wartości w grupach dla modelu log-liniowego [12 13] i zmiennych 1 - S, 2 - W1, 3 - Wyk.

S	W1	Wyk	n	fitted_n	abs_error
0	-2	1	19	14.798	4.202
0	-2	2	40	45.503	5.503
0	-2	3	5	3.699	1.301
0	-1	1	3	4.162	1.162
0	-1	2	15	12.798	2.202
0	-1	3	0	1.040	1.040
0	1	1	0	0.000	0.000
0	1	2	0	0.000	0.000
0	1	3	0	0.000	0.000
0	2	1	18	21.040	3.040
0	2	2	68	64.699	3.301
0	2	3	5	5.260	0.260
1	-2	1	1	0.370	0.630
1	-2	2	5	6.296	1.296
1	-2	3	4	3.333	0.667
1	-1	1	0	0.074	0.074
1	-1	2	2	1.259	0.741
1	-1	3	0	0.667	0.667
1	1	1	0	0.074	0.074
1	1	2	0	1.259	1.259
1	1	3	2	0.667	1.333
1	2	1	0	0.481	0.481
1	2	2	10	8.185	1.815
1	2	3	3	4.333	1.333

Tabela 10: Zestawienie predykcji i empirycznych wartości w grupach dla modelu log-liniowego [1 23] i zmiennych 1 - S, 2 - W1, 3 - Wyk.

S	W1	Wyk	n	fitted_n	abs_error
0	-2	1	19	17.300	1.700
0	-2	2	40	38.925	1.075
0	-2	3	5	7.785	2.785
0	-1	1	3	2.595	0.405
0	-1	2	15	14.705	0.295
0	-1	3	0	0.000	0.000
0	1	1	0	0.000	0.000
0	1	2	0	0.000	0.000
0	1	3	0	1.730	1.730
0	2	1	18	15.570	2.430
0	2	2	68	67.470	0.530
0	2	3	5	6.920	1.920
1	-2	1	1	2.700	1.700
1	-2	2	5	6.075	1.075
1	-2	3	4	1.215	2.785
1	-1	1	0	0.405	0.405
1	-1	2	2	2.295	0.295
1	-1	3	0	0.000	0.000
1	1	1	0	0.000	0.000
1	1	2	0	0.000	0.000
1	1	3	2	0.270	1.730
1	2	1	0	2.430	2.430
1	2	2	10	10.530	0.530
1	2	3	3	1.080	1.920

Tabela 11: Zestawienie predykcji i empirycznych wartości w grupach dla modelu log-liniowego [1 3] i zmiennych 1 - S, 2 - P, 3 - Wyk.

S	P	Wyk	n	fitted_n	abs_error
0	K	1	1	17.73	16.733
0	K	2	54	60.55	6.550
0	K	3	8	8.22	0.217
0	M	1	39	17.73	21.267
0	M	2	69	60.55	8.450
0	M	3	2	8.22	6.217
1	K	1	0	2.77	2.768
1	K	2	4	9.45	5.450
1	K	3	4	1.28	2.717
1	M	1	1	2.77	1.768
1	M	2	13	9.45	3.550
1	M	3	5	1.28	3.717

Model [1 23]

Zakładamy, że stanowisko, zadowolenie z wynagrodzenia oraz wykształcenie, mają dowolne rozkłady. Stanowisko jest niezależne od wykształcenia i zadowolenia z wynagrodznia, ale samo wykształcenie oraz zadowolenie z wynagrodznia nie są niezależne.

```
glm(n ~ S + W1 + Wyk + W1 * Wyk, data = personel.s_w1_wyk, family = poisson)
```

Ostatecznie, dochodzimy do wniosku, że z zaproponowanej listy warto wybrać ostatni model. Choć obejmuje mniej zależności pomiędzy zmiennymi daje najciekawsze rezultaty (tab. 10). Mówimy to w takim sensie, że błąd w żadnej grupie nie przekacza tu już nawet liczby 3 ankietowanch. Pokazuje to, że model należy dobierać z uwzględnieniem analiz niezależności, a większa liczba powiązań w modelu wcale nie musi oznaczać lepszego dopasowania. Jeżeli oceniamy model w całości, a nie patrząc na poszczególne grupy (`deviance`), można dojść zaś do wniosku, że model [12 13] będzie najodpowiedniejszy. Oba prezentują dość podobne, dobre dopasowania.

Zadanie 6

W tym zadaniu powtórzymy procedury z poprzedniego zadania. Będziemy jednak mówić o innych zmiennych: 1 – zajmowane stanowisko (S), 2 – płeć (P), 3 – wykształcenie (Wyk).

Model [1 3]

Zakładamy, że stanowisko i wykształcenie mają dowolne rozkłady, a płeć rozkład równomierny. Dodatkowo, wszystkie te zmienne są niezależne.

Tutaj posłużymy się analogicznym kodem, ale już z tabelą danych `personel.s_p_wyk`, uwzględniającą inne zmienne.

```
glm(n ~ S + Wyk, data = personel.s_p_wyk, family = poisson)
```

Pamiętając podobny model i zmienne z poprzedniego zadania nie spodziewamy się świetnych rezultatów modelu [1 3]. Faktycznie, w tab. 11 zauważamy znów ogromne rozbieżności.

Model [13]

Zakładamy, że stanowisko i wykształcenie mają dowolne rozkłady, a płeć rozkład równomierny. Dodatkowo, płeć jest niezależne od stanowiska i od wykształcenia, ale samo stanowisko i wykształecenie nie są niezależne.

Tabela 12: Zestawienie predykcji i empirycznych wartości w grupach dla modelu log-liniowego [13] i zmiennych 1 - S, 2 - P, 3 - Wyk.

S	P	Wyk	n	fitted_n	abs_error
0	K	1	1	20.0	19.0
0	K	2	54	61.5	7.5
0	K	3	8	5.0	3.0
0	M	1	39	20.0	19.0
0	M	2	69	61.5	7.5
0	M	3	2	5.0	3.0
1	K	1	0	0.5	0.5
1	K	2	4	8.5	4.5
1	K	3	4	4.5	0.5
1	M	1	1	0.5	0.5
1	M	2	13	8.5	4.5
1	M	3	5	4.5	0.5

Tabela 13: Zestawienie predykcji i empirycznych wartości w grupach dla modelu log-liniowego [1 2 3] i zmiennych 1 - S, 2 - P, 3 - Wyk.

S	P	Wyk	n	fitted_n	abs_error
0	K	1	1	12.590	11.590
0	K	2	54	42.991	11.009
0	K	3	8	5.834	2.166
0	M	1	39	22.875	16.125
0	M	2	69	78.109	9.109
0	M	3	2	10.601	8.601
1	K	1	0	1.965	1.965
1	K	2	4	6.710	2.710
1	K	3	4	0.911	3.089
1	M	1	1	3.570	2.570
1	M	2	13	12.191	0.809
1	M	3	5	1.654	3.346

```
glm(n ~ S + Wyk + S * Wyk, data = personel.s_p_wyk, family = poisson)
```

Tab. 6 sugeruje zaś, że ten pomysł nie jest perfekcyjnym rozwiązaniem. W stosunku do modelu [1 3], błędy w niektórych grupach zmalały, ale w innych zauważalnie wzrosły.

Model [1 2 3]

Zakładamy, że stanowisko, płeć oraz wykształcenie, mają dowolne rozkłady. Wszystkie te zmienne są zaś wzajemnie niezależne.

```
glm(n ~ S + P + Wyk, data = personel.s_p_wyk, family = poisson)
```

Wprowadzenie trzech niezależnych zmiennych do modelu powoduje w tab. 7 spore zamieszanie. Błędy są w znacznej mierze jeszcze większe, niż wcześniej.

Model [12 3]

Zakładamy, że stanowisko, płeć oraz wykształcenie, mają dowolne rozkłady. Wykształcenie jest niezależne od stanowiska i płci, ale samo stanowisko oraz płeć nie są niezależne.

Tabela 14: Zestawienie predykcji i empirycznych wartości w grupach dla modelu log-liniowego [12 3] i zmiennych 1 - S, 2 - P, 3 - Wyk.

S	P	Wyk	n	fitted_n	abs_error
0	K	1	1	12.91	11.91
0	K	2	54	44.10	9.90
0	K	3	8	5.99	2.02
0	M	1	39	22.55	16.45
0	M	2	69	77.00	8.00
0	M	3	2	10.45	8.45
1	K	1	0	1.64	1.64
1	K	2	4	5.60	1.60
1	K	3	4	0.76	3.24
1	M	1	1	3.90	2.90
1	M	2	13	13.30	0.30
1	M	3	5	1.80	3.19

Tabela 15: Zestawienie predykcji i empirycznych wartości w grupach dla modelu log-liniowego [12 13] i zmiennych 1 - S, 2 - P, 3 - Wyk.

S	P	Wyk	n	fitted_n	abs_error
0	K	1	1	14.566	13.566
0	K	2	54	44.792	9.208
0	K	3	8	3.642	4.358
0	M	1	39	25.434	13.566
0	M	2	69	78.208	9.208
0	M	3	2	6.358	4.358
1	K	1	0	0.296	0.296
1	K	2	4	5.037	1.037
1	K	3	4	2.667	1.333
1	M	1	1	0.704	0.296
1	M	2	13	11.963	1.037
1	M	3	5	6.333	1.333

```
glm(n ~ S + P + Wyk + S * P, data = personel.s_p_wyk, family = poisson)
```

Rozpoznajemy w tab. 8 dla takiego modelu dalej duże błędy, największy to ponad 16, przy empirycznej liczności 39.

Model [12 13]

Zakładamy, że stanowisko, płeć oraz wykształcenie, mają dowolne rozkłady. Do tego, przy ustalonym stanowisku, wykształcenie i płeć są niezależne (P i Wyk są warunkowo niezależne).

```
glm(n ~ S + P + Wyk + S * P + S * Wyk, data = personel.s_p_wyk, family = poisson)
```

Taki model, jak widzimy na tab. 9, nie wnosi nic specjalnie dobrego. Prezentuje równie duże odchyły, jak wcześniej.

Model [1 23]

Zakładamy, że stanowisko, zadowolenie z wynagrodzenia oraz wykształcenie, mają dowolne rozkłady. Stanowisko jest niezależne od wykształcenia i zadowolenia z wynagrodznia, ale samo wykształcenie oraz

Tabela 16: Zestawienie predykcji i empirycznych wartości w grupach dla modelu log-liniowego [1 23] i zmiennych 1 - S, 2 - P, 3 - Wyk.

S	P	Wyk	n	fitted_n	abs_error
0	K	1	1	0.865	0.135
0	K	2	54	50.170	3.830
0	K	3	8	10.380	2.380
0	M	1	39	34.600	4.400
0	M	2	69	70.930	1.930
0	M	3	2	6.055	4.055
1	K	1	0	0.135	0.135
1	K	2	4	7.830	3.830
1	K	3	4	1.620	2.380
1	M	1	1	5.400	4.400
1	M	2	13	11.070	1.930
1	M	3	5	0.945	4.055

Tabela 17: Porównanie prawdopodobieństw empirycznych i predykowanych znalezienia się w grupie osób zdecydowanie zadowolonych ze swojego wynagrodzenia, mając stanowisko kierownicze.

	Prawdopodobieństwo
Empiryczna wartość	0.481
Predykcja modelu [13 23]	0.507
Predykcja modelu [123]	0.481

zadowolenie z wynagrodznia nie są niezależne.

```
glm(n ~ S + P + Wyk + P * Wyk, data = personel.s_p_wyk, family = poisson)
```

W widoczny sposób dla tych zmiennych zwycięża model [1 23], dając – tak jak w tab. 10 – rezultaty o najmniejszych błędach. Wnioski są podobne jak poprzednio – warto pochylić się nad zależnościami, zanim zaczniemy dobierać model.

Część III

Zadanie 7

W tym zadaniu, weźmiemy znów zmienne 1 – stanowisko (S), 2 – zadowolenie z wynagrodzenia w pierwszym okresie ($W1$) oraz 3 – wykształcenie (Wyk). Szacować będziemy prawdopodobieństwa różnych zdarzeń dwoma modelami, [13 23] oraz [123], a wyniki porównamy do empirycznych liczności.

W pierwszym zakładamy, że stanowisko, zadowolenie z wynagrodzenia oraz wykształcenie, mają dowolne rozkłady, a zajmowane stanowisko i zadowolenie z wynagrodzenia są warunkowo niezależne. Użyjemy kodu

```
glm(n ~ S + W1 + Wyk + S * Wyk + W1 * Wyk, data = personel.s_p_wyk, family = poisson)
```

W przypadku drugiego modelu za to wszystkie trzy rozpatrywane zmienne nie są niezależne i żadna z par nie jest warunkowo niezależna. Tu funkcja będzie wywołana następująco:

```
glm(n ~ (S + W1 + Wyk) ^ 2, data = personel.s_p_wyk, family = poisson)
```

Trzy problemy przedstawimy w punktach.

Tabela 18: Porównanie prawdopodobieństw empirycznych i predykowanych posiadania stanowiska kirowniczego, mając wykształcenie zawodowe.

	Prawdopodobieństwo
Empiryczna wartość	0.024
Predykcja modelu [13 23]	0.024
Predykcja modelu [123]	0.024

Tabela 19: Porównanie prawdopodobieństw empirycznych i predykowanych nie posiadania stanowiska kierowniczego, mając wyższe wykształcenie.

	Prawdopodobieństwo
Empiryczna wartość	0.526
Predykcja modelu [13 23]	0.526
Predykcja modelu [123]	0.526

Punkt (a)

Chcemy oszacować prawdopodobieństwo, że osoba pracująca na stanowisku kierowniczym jest zdecydowanie zadowolona ze swojego wynagrodzenia. Okazuje się, że model [123] dopasowuje się do danych treningowych lepiej, niż model [13 23]. Należałoby natomiast zapytać, jak sytuacja ma się w przypadku ewentualnego zbioru testowego, gdyż może to być objaw *overfittingu*. Dokładne wyniki prezentujemy w tab. 17.

Punkt (b)

Szacujemy prawdopodobieństwo, iż osoba z wykształceniem zawodowym pracuje na stanowisku kierowniczym. Z tab. 18 czytamy, iż te prawdopodobieństwa są niezwyke małe. Z zapisywaną zaś precyzją, dla obu modeli i liczb z ankiet, prawdopodobieństwa są takie same.

Punkt (c)

Tym razem patrzymy na sytuację, że osoba z wyższym wykształceniem nie pracuje na stanowisku kierowniczym. Tab. 19 pokazuje znów zgodne wersje z dokładnością do trzeciego miejsca po przecinku. Oba modele idelanie dopasowują się do treningowych danych.

Zadanie 8

Tutaj wrócimy do zmiennyc 1 – stanowisko (S), 2 – płeć (P) oraz 3 – wykształcenie (Wyk). Rozważymy model [13 23]. Zobaczymy też jak przewidywania w punkcie (b) mają się do analogicznego szacowania w punkcie (b) zadania 7.

W opisywanym modelu – analogicznie do poprednio rozważanych przypadków – zakładamy, że stanowisko i płeć są warunkowo niezależne.

Problemy znów wypisujemy jako osobne punkty.

Punkt (a)

Oceniamy prawdopodobieństwo, że osoba pracująca na stanowisku kierowniczym jest kobietą. Wyniki takich przewidywań w oparciu o model log-liniowy, z tab. 20, dość mocno odbiegają jednak bezpośredniego wyniku z danych.

Tabela 20: Porównanie prawdopodobieństw empirycznych i predykowanych posiadania płci żeńskiej, pracując na stanowisku kierowniczym.

	Prawdopodobieństwo
Empiryczna wartość	0.296
Predykcja modelu [13 23]	0.472

Tabela 21: Porównanie prawdopodobieństw empirycznych i predykowanych pracy na stanowisku kierowniczym, mając wykształcenie zawodowe.

	Prawdopodobieństwo
Empiryczna wartość	0.024
Predykcja modelu [13 23]	0.024

Punkt (b)

Szukamy prawdopodobieństwa, że osoba z wykształceniem zawodowym pracuje na stanowisku kierowniczym. Zauważmy, iż jest to taki sam problem jak w zadaniu poprzednim, choć modele log-liniowe się różnią. Oszacowane prawdopodobieństwa są zaś w obu przypadkach z zapisywaną dokładnością identyczne. Widimy to porównując tab. 18 oraz tab. 21.

Punkt (c)

Patrzymy teraz jeszcze na prawdopodobieństwo, że osoba z wykształceniem wyższym jest mężczyzną. Korespondujące prawdopodobieństwa z tab. 22 sugerują, że wśród pracowników tej firmy z wykształceniem wyższym, lekko ponad trzecia część to mężczyźni.

Część III

Zadanie 9

W tym zadaniu, z użyciem funkcji `anova` zweryfikujemy szereg hipotez o niezależności. Pisząc skrótowo symbole modeli, będziemy oczywiście trzymać się konwencji kolejności zmiennych, w której zostały one na początku podane.

Punkt (a)

Zaczniemy od hipotezy, iż zmienne losowe S, $W1$ i Wyk są wzajemnie niezależne. Testujemy H_0 : [1 2 3] przeciwko dwom alternatywom – [1 23] oraz [123]. Modele dopasowujemy więc następującymi liniami kodu:

```
mod9a.0 <- glm(n ~ S + W1 + Wyk, data = personel.s_w1_wyk, family = poisson)
mod9a.1 <- glm(n ~  S + W1 * Wyk, data = personel.s_w1_wyk, family = poisson)
mod9a.2 <- glm(n ~ (S + W1 + Wyk) ** 2, data = personel.s_w1_wyk, family = poisson)
```

Okazuje się, że otrzymujemy odpowiednio do podanej kolejności p-wartości 0.006 oraz 2.773×10^{-5}. Obie są mniejsze od poziomu $\alpha = 0.05$, zatem hipotezę zerową odrzucamy.

Tabela 22: Porównanie prawdopodobieństw empirycznych i predykowanych posiadania płci męskiej, mając wyższe wykształcenie.

	Prawdopodobieństwo
Empiryczna wartość	0.368
Predykcja modelu [13 23]	0.368

Punkt (b)

Jeżeli chodzi o to, czy zmienna losowa *W1* jest niezależna od pary zmiennych *S1* i *Wyk*, to przeciwko musimy hipotezie [2 13] przetestować przykładowo $H_{1,a}$: [12 13] oraz $H_{1,a}$: [123].

```
mod9b.0 <- glm(n ~ W1 + S * Wyk, data = personel.s_w1_wyk, family = poisson)
mod9b.1 <- glm(n ~ S * W1 + S * Wyk, data = personel.s_w1_wyk, family = poisson)
mod9b.2 <- glm(n ~ (S + W1 + Wyk) ** 2, data = personel.s_w1_wyk, family = poisson)
```

Wynikowe wartości poziomu krytycznego – odpowiednio 0.04 i 0.01 – są podstawą do odrzucenia przedstawionej hipotezy zerowej.

Punkt (c)

Dalej, zastanowimy się nad niezależnością zmiennej losowej *W1* od zmiennej *S*, przy ustalonej wartości zmiennej *Wyk*. Mówimy tutaj o hipotezie zerowej postaci [13 23]. W związku z tym, iż z odgórnego założenia o nadmodelach, jako testowanych obiektach, pozostaniemy tylko przy alternatywie [123]. Tutaj już duża p-wartość (0.35) nie daje podstaw do odrzucenia hipotezy, że zmienne są modelowane na sposób [13 23].

```
mod9c.0 <- glm(n ~ S * Wyk + W1 * Wyk, data = personel.s_w1_wyk, family = poisson)
mod9c.1 <- glm(n ~ (S + W1 + Wyk) ^ 2, data = personel.s_w1_wyk, family = poisson)
```

Punkt (e)

Jako ostatnią, weźmiemy hipotezę o niezależności zmiennej losowej S od zmiennej P, przy ustalonej wartości zmiennej *Wyk*. Tworzymy więc modele:

```
mod9d.0 <- glm(n ~ S * Wyk + P * Wyk, data = personel.s_p_wyk, family = poisson)
mod9d.1 <- glm(n ~ (S + P + Wyk) ^ 2, data = personel.s_p_wyk, family = poisson)
```

Hipotezą zerową jest (z kolejnością *S*, *P*, *Wyk*) [13 23], a alternatywną [123]. Otrzymana p-wartość jest równa 0.024, co oznacza, iż odrzucamy hipotezę zerową, na korzyść stwierdzenia, iż wszystkie zmienne są parami zależne.

Część IV

Zadanie 10

Dwa zadania kończące listę obowiązkowego zakresu raportu, polegają na wyborze modelu dla zestawu zmiennych, w oparciu o testy, kryterium AIC oraz kryterium BIC.

W pierwszym przypadku zmiennymi będą *A1*, *W1* oraz *P*.

%%%%%%%%%%%%%

Zadanie 11

W drugim z zadań, rozpatrujemy zmienne *D*, *A1* oraz *P*.

%%%%%%%%%%%%%

Podsumowanie

Reasumując, udało nam się przyjrzeć działaniu modeli symetrii, stosując je do analizy zmian charakteru zjawisk w czasie. Dwa z tych testów wnikliwie zbadaliśmy pod kątem ich mocy. Poza tym, poruszyliśmy wiele aspektów związanych z modelami log-liniowymi. Obserwowaliśmy oraz porównywaliśmy rezultaty dopasowań w różnych przypadkach, a same przypadki modeli za każdym razem interpretowaliśmy. Potem, stosowaliśmy

takie modele do szacowania prawdopodobieństw, a także niezależności ??????????????????????????????????. Na końcu poruszyliśmy też kwestię kryteriów wyboru modelu.

Zestaw narzędzi, którego używaliśmy, jest kolejna część przybornika pełnego technik, umożliwiających coraz lepszą analizę danych ankietowych. Forma zaś pracy z tymi narzędziami, pozwoliła nam je nie tylko zaprezentować, ale także doskonale zrozumieć.