Sprawozdanie 3 ADA

Krzysztof Radomski 275968

22 stycznia 2025

Spis treści

1 Zadanie 1

Paradoks Simpsona opisuje sytuację, w której trend obserwowany w kilku grupach danych znika lub odwraca się po połączeniu tych grup w jedną całość. Matematycznie, jeśli A, B i C są zmiennymi losowymi, paradoks zachodzi, gdy:

$$P(A \mid B) < P(A \mid \neg B) \quad \text{oraz} \quad P(A \mid B, C) > P(A \mid \neg B, C) \quad \text{i} \quad P(A \mid B, \neg C) > P(A \mid \neg B, \neg C).$$

Dla każdego z podpunktów sprawdzano, czy zachodzi paradoks Simpsona dla wybranych zmiennych. Wyniki wskazują, że w żadnym przypadku paradoks nie wystąpił.

Test 1: A = female, B = nonwhite, C = married

Podział zmiennych:

- **A:** female (płeć żeńska, zmienna binarna: 1 - tak, 0 - nie),
- **B:** nonwhite (osoba niebiałoskóra, zmienna binarna: 1 - tak, 0 - nie),
- **C:** married (osoba zamężna, zmienna binarna: 1 - tak, 0 - nie).

Wynik: Paradoks Simpsona nie wystąpił.

```
#Test 1 A:female=1 B:nonwhite=1 C:married=1
P.A.pod.war.B<-sum(dane$female==1 & dane$nonwhite==1)/sum(dane$nonwhite==1)
P.A.pod.war.B_<-sum(dane$female==1 & dane$nonwhite==0)/sum(dane$nonwhite==0)
P.A.pod.war.B.C<-sum(dane$female==1 & dane$nonwhite==1 & dane$married==1)/
  sum(dane$nonwhite==1 & dane$married==1)
P.A.pod.war.B_.C<-sum(dane$female==1 & dane$nonwhite==0 & dane$married==1)/
  sum(dane$nonwhite==0 & dane$married==1)
P.A.pod.war.B.C_<-sum(dane$female==1 & dane$nonwhite==1 & dane$married==0)/
  sum(dane$nonwhite==1 & dane$married==0)
P.A.pod.war.B_.C_<-sum(dane$female==1 & dane$nonwhite==0 & dane$married==0)/
  sum(dane$nonwhite==0 & dane$married==0)

(P.A.pod.war.B<P.A.pod.war.B_) ==
  (P.A.pod.war.B.C>P.A.pod.war.B_.C & P.A.pod.war.B.C_>P.A.pod.war.B_.C_
)
```

[1] FALSE

Test 2: A = experCat, B = wageCat, C = female

Podział zmiennych:

- **A:** experCat (kategoria doświadczenia zawodowego: Low, Medium, High),
- **B:** wageCat (kategoria zarobków: 1, 2, 3, 4),
- **C:** female (płeć żeńska, zmienna binarna: 1 - tak, 0 - nie).

Wynik: Paradoks Simpsona nie wystąpił.

```
# Test 2: A = experCat, B = wageCat, C = female
P.A.pod.war.B.2 <- sum(dane$experCat == "High" & dane$wageCat == 4) /
  sum(dane$wageCat == 4)
P.A.pod.war.B_.2 <- sum(dane$experCat == "High" & dane$wageCat != 4) /
  sum(dane$wageCat != 4)
P.A.pod.war.B.C.2 <- sum(dane$experCat == "High" & dane$wageCat == 4 &
                           dane$female == 1) /
  sum(dane$wageCat == 4 & dane$female == 1)
P.A.pod.war.B_.C.2 <- sum(dane$experCat == "High" & dane$wageCat != 4 &
                            dane$female == 1) /
  sum(dane$wageCat != 4 & dane$female == 1)
P.A.pod.war.B.C_.2 <- sum(dane$experCat == "High" & dane$wageCat == 4 &
                            dane$female == 0) /
  sum(dane$wageCat == 4 & dane$female == 0)
P.A.pod.war.B_.C_.2 <- sum(dane$experCat == "High" & dane$wageCat != 4 &
                             dane$female == 0) /
  sum(dane$wageCat != 4 & dane$female == 0)

 (P.A.pod.war.B.2 < P.A.pod.war.B_.2) ==
  (P.A.pod.war.B.C.2 > P.A.pod.war.B_.C.2 &

     P.A.pod.war.B.C_.2 > P.A.pod.war.B_.C_.2)
```

[1] FALSE

Test 3: A = region, B = nonwhite, C = married

Podział zmiennych:

- **A:** region (region zamieszkania: West, North Central, South, Other),
- **B:** nonwhite (osoba niebiałoskóra, zmienna binarna: 1 - tak, 0 - nie),
- **C:** married (osoba zamężna, zmienna binarna: 1 - tak, 0 - nie).

Wynik: Paradoks Simpsona nie wystąpił.

```
# Test 3: A = region, B = nonwhite, C = married
P.A.pod.war.B.3 <- sum(dane$region == "West" & dane$nonwhite == 1) /
  sum(dane$nonwhite == 1)
P.A.pod.war.B_.3 <- sum(dane$region == "West" & dane$nonwhite == 0) /
  sum(dane$nonwhite == 0)
P.A.pod.war.B.C.3 <- sum(dane$region == "West" & dane$nonwhite == 1 &
                           dane$married == 1) /
  sum(dane$nonwhite == 1 & dane$married == 1)
P.A.pod.war.B_.C.3 <- sum(dane$region == "West" & dane$nonwhite == 0 &
                            dane$married == 1) /
  sum(dane$nonwhite == 0 & dane$married == 1)
P.A.pod.war.B.C_.3 <- sum(dane$region == "West" & dane$nonwhite == 1 &
```

```
                                 dane$married == 0) /
  sum(dane$nonwhite == 1 & dane$married == 0)
P.A.pod.war.B_.C_.3 <- sum(dane$region == "West" & dane$nonwhite == 0 &
                                  dane$married == 0) /
  sum(dane$nonwhite == 0 & dane$married == 0)

(P.A.pod.war.B.3 < P.A.pod.war.B_.3) ==
  (P.A.pod.war.B.C.3 > P.A.pod.war.B_.C.3 &

     P.A.pod.war.B.C_.3 > P.A.pod.war.B_.C_.3)
```

[1] FALSE

Przeprowadzone testy wykazały, że w żadnym z badanych przypadków paradoks Simpsona nie wystąpił. Należy jednak zauważyć, że brak wystąpienia paradoksu w tych konkretnych podziałach zmiennych nie gwarantuje, że paradoks nie wystąpi w przypadku innych podziałów. Dlatego analiza powinna być przeprowadzana z uwzględnieniem różnych kombinacji zmiennych.

2 Zadanie 2

Na potrzeby analizy przyjęto następującą interpretację cyfr odpowiadających zmiennym:

- 1 - `wageCat`
- 2 - `educCat`
- 3 - `female`

Poniżej znajdują się opisy analizowanych relacji między zmiennymi:

(a) **[1 3]**

- Zmienna 2 ma rozkład równomierny i jest niezależna od pozostałych zmiennych.
- Zmienne 1 oraz 3 mają dowolny rozkład i są niezależne od siebie nawzajem oraz od zmiennej 2.

(b) **[13]**

- Zmienna 2 ma rozkład równomierny i jest niezależna od pozostałych zmiennych.
- Zmienne 1 oraz 3 są od siebie zależne i mają dowolny rozkład.

(c) **[1 2 3]**

- Każda ze zmiennych ma dowolny rozkład i są od siebie nawzajem niezależne.

(d) **[12 3]**

- Zmienna 3 ma dowolny rozkład i jest niezależna od pozostałych.
- Zmienne 1 i 2 są od siebie zależne i mają dowolny rozkład.

(e) **[12 13]**

- Przy ustalonej wartości zmiennej 1, zmienne 2 i 3 są niezależne, czyli są warunkowo niezależne.

(f) [**1 23**]

- Zmienna 1 ma dowolny rozkład i jest niezależna od zmiennych 2 i 3.
- Zmienne 2 i 3 są od siebie zależne.

3 Zadanie 3

Funkcja `glm` (Generalized Linear Model) pozwala na dopasowanie uogólnionych modeli liniowych, rozszerzających klasyczne modele liniowe o możliwość modelowania zmiennych zależnych z różnych rodzin rozkładów. Jest szeroko stosowana w analizach statystycznych, takich jak regresja logistyczna czy regresja Poissona. Model w funkcji `glm` definiujemy za pomocą formuły w postaci: `y ~ x1 + x2 + x3`. Formuła ta wskazuje, że zmienna zależna `y` jest modelowana jako funkcja zmiennych niezależnych `x1`, `x2` i `x3`. Parametr `family` określa rodzinę rozkładów, które najlepiej odpowiadają charakterowi zmiennej zależnej. Dostępne opcje to: `gaussian` (rozkład normalny, domyślny, stosowany w klasycznej regresji liniowej), `binomial` (rozkład dwumianowy, stosowany w regresji logistycznej dla zmiennych binarnych), `poisson` (rozkład Poissona, używany w przypadku danych licznikowych) oraz inne, takie jak `Gamma` czy `inverse.gaussian`. Na przykład, dopasowanie regresji logistycznej można przeprowadzić za pomocą następującego kodu:

```
model <- glm(y ~ x1 + x2, family = binomial, data = my_data)
```

Funkcja `loglin` służy do dopasowywania modeli logarytmiczno-liniowych (log-linear models), stosowanych głównie w analizach danych tabelarycznych, takich jak tablice kontyngencji. Modele te pozwalają na analizę zależności między kategorycznymi zmiennymi w tabelach wielowymiarowych. Model w funkcji `loglin` definiujemy za pomocą listy marginesów, które mają być uwzględnione w modelu. Każda zmienna musi być zidentyfikowana przez jej pozycję w tablicy kontyngencji. Modele log-liniowe w `loglin` zakładają rozkład Poissona dla danych tabelarycznych. Na przykład, analiza modelu logarytmiczno-liniowego dla tablicy kontyngencji może wyglądać następująco:

```
data(Titanic)
loglin(Titanic, margin = list(1, 2, c(1, 2)))
```

Funkcja `loglm` jest bardziej elastycznym odpowiednikiem `loglin`, pozwalającym na deklarację modeli logarytmiczno-liniowych za pomocą formuły, podobnie jak w `glm`. Jest często wykorzystywana do analiz danych tabelarycznych z kategorycznymi zmiennymi. Model definiujemy za pomocą formuły w postaci: `Freq ~ A + B + A:B`. Oznacza to, że zmienna zależna `Freq` (liczba obserwacji) jest modelowana jako funkcja zmiennych kategorycznych `A` i `B` oraz ich interakcji `A:B`. Podobnie jak w przypadku `loglin`, rodzina rozkładów zakłada rozkład Poissona dla danych tabelarycznych. Na przykład, dopasowanie modelu logarytmiczno-liniowego dla tablicy kontyngencji można przeprowadzić w następujący sposób:

```
library(MASS)
data(Titanic)
loglm(Freq ~ Class + Sex + Class:Sex, data = Titanic)
```

Podsumowując, funkcje `glm`, `loglin` i `loglm` mają swoje specyficzne zastosowania: `glm` pozwala na szerokie modelowanie danych w oparciu o różne rodziny rozkładów, podczas gdy `loglin` i `loglm` są dedykowane analizom danych tabelarycznych, takich jak tablice kontyngencji.

4 Zadanie 4

Model [12 3]

- **Podpunkt (a):** Prawdopodobieństwo, że zarobki kobiety o najwyższym poziomie wykształcenia należą do najwyższej kategorii:

$$P(\text{wageCat} = 4 \mid \text{female} = 1, \text{ educCat} = \text{"High"}) = 0.1111$$

- **Podpunkt (b):** Prawdopodobieństwo, że zarobki mężczyzny o najwyższym poziomie wykształcenia należą do najwyższej kategorii:

$$P(\text{wageCat} = 4 \mid \text{female} = 0, \text{ educCat} = \text{"High"}) = 0.3759$$

- **Podpunkt (c):** Prawdopodobieństwo, że kobieta o najwyższej kategorii zarobków ma najwyższy poziom wykształcenia:

$$P(\text{educCat} = \text{"High"} \mid \text{wageCat} = 4, \text{female} = 1) = 0.3289$$

- **Podpunkt (d):** Prawdopodobieństwo, że mężczyzna o najwyższej kategorii zarobków ma najniższą kategorię wykształcenia:

$$P(\text{educCat} = \text{"Low"} \mid \text{wageCat} = 4, \text{female} = 0) = 0.5970$$

- **Podpunkt (e):** Prawdopodobieństwo, że osoba o najwyższym poziomie wykształcenia ma zarobki na najwyższym poziomie:

$$P(\text{wageCat} = 4 \mid \text{educCat} = \text{"High"}) = 0.1347$$

- **Podpunkt (f):** Prawdopodobieństwo, że osoba o najwyższym poziomie wykształcenia ma zarobki na najniższym poziomie:

$$P(\text{wageCat} = 1 \mid \text{educCat} = \text{"High"}) = 0.1567$$

Model [12 13]

- **Podpunkt (a):** Prawdopodobieństwo, że zarobki kobiety o najwyższym poziomie wykształcenia należą do najwyższej kategorii:

$$P(\text{wageCat} = 4 \mid \text{female} = 1, \text{ educCat} = \text{"High"}) = 0.2298$$

- **Podpunkt (b):** Prawdopodobieństwo, że zarobki mężczyzny o najwyższym poziomie wykształcenia należą do najwyższej kategorii:

$$P(\text{wageCat} = 4 \mid \text{female} = 0, \text{ educCat} = \text{"High"}) = 0.5339$$

- **Podpunkt (c):** Prawdopodobieństwo, że kobieta o najwyższej kategorii zarobków ma najwyższy poziom wykształcenia:

$$P(\text{educCat} = \text{"High"} \mid \text{wageCat} = 4, \text{female} = 1) = 0.5496$$

- **Podpunkt (d):** Prawdopodobieństwo, że mężczyzna o najwyższej kategorii zarobków ma najniższą kategorię wykształcenia:

$$P(\text{educCat} = \text{"Low"} \mid \text{wageCat} = 4, \text{female} = 0) = 0.3817$$

- **Podpunkt (e):** Prawdopodobieństwo, że osoba o najwyższym poziomie wykształcenia ma zarobki na najwyższym poziomie:

$$P(\text{wageCat} = 4 \mid \text{educCat} = \text{"High"}) = 0.2251$$

- **Podpunkt (f):** Prawdopodobieństwo, że osoba o najwyższym poziomie wykształcenia ma zarobki na najniższym poziomie:

$$P(\text{wageCat} = 1 \mid \text{educCat} = \text{"High"}) = 0.0501$$

Podsumowanie

Wyniki modelu [12 13] różnią się od wyników modelu [12 3] w zakresie wartości prawdopodobieństw. Model [12 13], dzięki uwzględnieniu interakcji między `wageCat` a `educCat`, lepiej odzwierciedla zależności między zarobkami a poziomem wykształcenia. Wartości prawdopodobieństw są wyższe dla osób z najwyższym wykształceniem i zarobkami w najwyższej kategorii, co sugeruje, że model [12 13] może być bardziej odpowiedni dla analizy tych danych.

5 Zadanie 5

Zmienne w analizie

- 1: wageCat - 2: educCat - 3: female - 5: married - 6: region - 8: smsa

Wyniki testów hipotez

(a) Zmienne losowe `wageCat`, `female` i `educCat` są wzajemnie niezależne.

```
testuj.model(tabela_1, list(c(1), c(2), c(3)), list(c(1, 2, 3)))
# Wynik: p = 4.85841e-25

testuj.model(tabela_1, list(c(1), c(2), c(3)), list(c(1, 2), c(1, 3)))
# Wynik: p = 9.388752e-28
```

Odrzucamy H_0 w obu przypadkach. Lepsze są modele alternatywne H_1, które uwzględniają zależności między zmiennymi `wageCat`, `female` i `educCat`.

Wniosek: Zależności między tymi zmiennymi są istotne.

(b) Zmienna losowa `wageCat` jest niezależna od pary zmiennych `female` i `educCat`.

```
testuj.model(tabela_1, list(c(1), c(2, 3)), list(c(1, 2, 3)))
# Wynik: p = 1.132846e-22

testuj.model(tabela_1, list(c(1), c(2, 3)), list(c(1, 2), c(1, 3)))
# Wynik: p = 1.666627e-25
```

Odrzucamy H_0 w obu przypadkach. Lepsze są modele alternatywne H_1, co wskazuje na istotną zależność `wageCat` od `female` i `educCat`.

Wniosek: Zmienna `wageCat` jest istotnie zależna od zmiennych `female` i `educCat`.

(c) Zmienna losowa `wageCat` jest niezależna od zmiennej `educCat`, przy ustalonej zmiennej `female`.

```
testuj.model(tabela_1, list(c(1, 2), c(2, 3)), list(c(1, 2, 3)))
# Wynik: p = 6.446744e-09

testuj.model(tabela_1, list(c(1, 2), c(2, 3)), list(c(1, 2), c(1, 3)))
# Wynik: p = 4.845518e-11
```

Odrzucamy H_0 w obu przypadkach. Zmienna `wageCat` jest istotnie zależna od `educCat`, nawet przy ustalonej wartości zmiennej `female`.

Wniosek: Zależność między zmiennymi `wageCat` i `educCat` jest istotna.

(d) Zmienna losowa `wageCat` jest niezależna od zmiennej `female`, przy ustalonej zmiennej `educCat`.

```
testuj.model(tabela_1, list(c(1, 3), c(2, 3)), list(c(1, 2, 3)))
# Wynik: p = 1.343257e-11

testuj.model(tabela_1, list(c(1, 3), c(2, 3)), list(c(1, 2), c(1, 3)))
# Wynik: p = 4.481184e-15
```

Odrzucamy H_0 w obu przypadkach. Zmienna `wageCat` jest istotnie zależna od `female`, nawet przy ustalonej wartości zmiennej `educCat`.

Wniosek: Zależność między zmiennymi `wageCat` i `female` jest istotna.

(e) Zmienne losowe `wageCat`, `married` i `region` są wzajemnie niezależne.

```
testuj.model(tabela_2, list(c(1), c(2), c(3)), list(c(1, 2, 3)))
# Wynik: p = 7.921775e-06
```

Odrzucamy H_0. Lepszym modelem jest H_1, który uwzględnia zależności między `wageCat`, `married` i `region`.

Wniosek: Zależności między tymi zmiennymi są istotne.

(f) Zmienna losowa `wageCat` jest niezależna od pary zmiennych `married` i `region`.

```
testuj.model(tabela_2, list(c(1), c(2, 3)), list(c(1, 2, 3)))
# Wynik: p = 4.646635e-06
```

Odrzucamy H_0. Lepszym modelem jest H_1, który uwzględnia zależność `wageCat` od `married` i `region`.

Wniosek: Zmienna `wageCat` jest istotnie zależna od zmiennych `married` i `region`.

(g) Zmienna losowa `wageCat` jest niezależna od zmiennej `region`, przy ustalonej zmiennej `married`.

```
testuj.model(tabela_2, list(c(1, 2), c(2, 3)), list(c(1, 2, 3)))
# Wynik: p = 0.0857768
```

Nie mamy podstaw do odrzucenia H_0. Możemy przyjąć, że `wageCat` jest niezależna od `region`, przy ustalonej wartości zmiennej `married`.

Wniosek: Nie wykazano istotnej zależności między zmiennymi `wageCat` i `region` przy uwzględnieniu zmiennej `married`.

(h) Zmienna losowa `wageCat` jest niezależna od zmiennej `female`, przy ustalonej zmiennej `married`.

```
testuj.model(tabela_3, list(c(1, 3), c(2, 3)), list(c(1, 2, 3)))
# Wynik: p = 1.731536e-16
```

Odrzucamy H_0. Lepszym modelem jest H_1, który uwzględnia zależność `wageCat` od `female` przy ustalonej wartości zmiennej married

6 Zadanie 6

```
#A
tabela_6_1 <-  table(dane[,c(4,5,6)])
testuj.model(tabela_6_1,list(c(1,2),c(2,3)),list(c(1,2),c(1,3),c(2,3)))
```

[1] 0.1394362

```
testuj.model(tabela_6_1,list(c(1,3),c(2,3)),list(c(1,2),c(1,3),c(2,3)))
```

[1] 0.6200855

```
#B
tabela_6_2 <-  table(dane[,c(1,3,5)])
testuj.model(tabela_6_2,list(c(1,2),c(2,3)),list(c(1,2),c(1,3),c(2,3)))
```

[1] 1.692414e-16

```
testuj.model(tabela_6_2,list(c(1,3),c(2,3)),list(c(1,2),c(1,3),c(2,3)))
```

[1] 6.702515e-06

```
#C
tabela_6_3 <-  table(dane[,c(4,6,8)])
testuj.model(tabela_6_3,list(c(1,2),c(2,3)),list(c(1,2),c(1,3),c(2,3)))
```

[1] 0.06926906

```
testuj.model(tabela_6_3,list(c(1,3),c(2,3)),list(c(1,2),c(1,3),c(2,3)))
```

[1] 0.1108664

Założenia

Paradoks Simpsona zachodzi, jeśli w przypadku przynajmniej jednej hipotezy nie odrzucimy H_0. W tym zadaniu analizujemy trzy różne trójki zmiennych:

- **Tabela A:** Zmienna 4 (smsa), Zmienna 5 (married), Zmienna 6 (region).
- **Tabela B:** Zmienna 1 (wageCat), Zmienna 3 (female), Zmienna 5 (married).
- **Tabela C:** Zmienna 4 (smsa), Zmienna 6 (region), Zmienna 8 (educCat).

Wyniki analizy

Tabela A:

```
Test 1: p = 0.1394
Test 2: p = 0.6201
```

W obu testach $p > 0.05$, co oznacza, że nie mamy podstaw do odrzucenia H_0. Paradoks Simpsona **nie zachodzi**.

Tabela B:

```
Test 1: p = 1.6924e-16
Test 2: p = 6.7025e-06
```

W obu przypadkach $p < 0.05$, co oznacza, że odrzucamy H_0. Paradoks Simpsona **może zajść**.

Tabela C:

```
Test 1: p = 0.0693
Test 2: p = 0.1109
```

W obu testach $p > 0.05$, co oznacza, że nie mamy podstaw do odrzucenia H_0. Paradoks Simpsona **nie zachodzi**.

7 Zadanie 7

7.1 a)

Zadanie 7 polega na wybraniu jak najlepszego modelu log-liniowgo do zmiennych wageCat, educCat, female i region. W podpunkcie a) musieliśmy dokonać wyboru w oparciu o testy, co za tym idzie, dokonywać bardzo dużo decyzji o odrzucaniu pojedynczych interakcji lub decydowaniu o ich istotności. Kroki były następujące: Na samym początku sprawdziliśmy czy model pełny, czyli zawierający wszystkie interakcje (włącznie z tą 3-ciego rzędu) będzie najlepszy. Otrzymaliśmy p-value około 0.07, zatem w zależności od przyjętego poziomu ufności mogliśmy przyjąć, że model maksymalny jest najlepszy i zakończyć pracę, albo uznać, że interakcja czterech zmiennych nie jest istotna i szukać lepszego modelu log-liniowego. Przyjęliśmy drugą opcję. Dalej widać 4 hipotezy w której odrzucaliśmy pojedynczo wszystkie interakcje 2-giego rzędu, dowiadując się, że żadna nie jest istotna. Następnie testowaliśmy, czy model zawierający wyłącznie interakcje 1-szego rzędu będzie lepszy i był. Potem zrobiliśmy 6 testów, w każdym wyrzucając jedną interakcję. wniosek był taki, iż istotne są tylko interakcje (1,2) oraz (1,3), zatem następną hipotezą (która się potwierdziła) było to , że model [12 13 4] jest dobry. Na końcu próbowaliśmy go jeszcze zmniejszać, jednak żaden mniejszy model nie był lepszy. Ostatecznie kryterium testów uzyskaliśmy odpowiedź, że szukany model log-liniowy to [12 13 4].

```r
testuj.model(tabela_7,
             list(c(1, 2, 3), c(1, 2, 4), c(1, 3, 4), c(2, 3, 4)),
             list(c(1, 2, 3, 4),
                  c(1, 2, 3), c(1, 2, 4), c(1, 3, 4), c(2, 3, 4)))
```

[1] 0.07509711

```r
#maksymalny model jest gorszy


testuj.model(tabela_7, list(c(1, 2, 3), c(1, 2, 4), c(1, 3, 4)),
             list(c(1, 2, 3), c(1, 2, 4), c(1, 3, 4), c(2, 3, 4)))
```

[1] 0.39128

```r
testuj.model(tabela_7, list(c(1, 2, 3), c(1, 2, 4), c(2, 3, 4)),
             list(c(1, 2, 3), c(1, 2, 4), c(1, 3, 4), c(2, 3, 4)))
```

[1] 0.5315836

```r
testuj.model(tabela_7, list(c(1, 2, 3), c(1, 3, 4), c(2, 3, 4)),
             list(c(1, 2, 3), c(1, 2, 4), c(1, 3, 4), c(2, 3, 4)))
```

[1] 0.3281343

```r
testuj.model(tabela_7, list(c(2, 3, 4), c(1, 3, 4), c(1, 2, 4)),
             list(c(1, 2, 3), c(1, 2, 4), c(1, 3, 4), c(2, 3, 4)))
```

[1] 0.7255517

```r
#żadna potrójna interakcja nie jest ważna


testuj.model(tabela_7, list(c(1, 2),c(1, 3), c(1, 4),c(2, 3),c(2, 4), c(3, 4)),
             list(c(1, 2, 3), c(1, 2, 4), c(1, 3, 4), c(2, 3, 4)))
```

[1] 0.6458193

```r
testuj.model(tabela_7, list(c(1, 3), c(1, 4),c(2, 3),c(2, 4), c(3, 4)),
             list(c(1, 2),c(1, 3), c(1, 4),c(2, 3),c(2, 4), c(3, 4)))
```

[1] 1.218595e-10

```r
testuj.model(tabela_7, list(c(1, 2), c(1, 4),c(2, 3),c(2, 4), c(3, 4)),
             list(c(1, 2),c(1, 3), c(1, 4),c(2, 3),c(2, 4), c(3, 4)))
```

[1] 2.373274e-15

```
testuj.model(tabela_7, list(c(1, 2),c(1, 3),c(2, 3),c(2, 4), c(3, 4)),
             list(c(1, 2),c(1, 3), c(1, 4),c(2, 3),c(2, 4), c(3, 4)))
```

[1] 0.2138326

```
testuj.model(tabela_7, list(c(1, 2),c(1, 3),c(1,4),c(2, 4), c(3, 4)),
             list(c(1, 2),c(1, 3), c(1, 4),c(2, 3),c(2, 4), c(3, 4)))
```

[1] 0.07552046

```
testuj.model(tabela_7, list(c(1, 2),c(1, 3),c(1,4),c(2, 3), c(3, 4)),
             list(c(1, 2),c(1, 3), c(1, 4),c(2, 3),c(2, 4), c(3, 4)))
```

[1] 0.7969971

```
testuj.model(tabela_7, list(c(1, 2),c(1, 3),c(1,4),c(2, 3),c(2, 4)),
             list(c(1, 2),c(1, 3), c(1, 4),c(2, 3),c(2, 4), c(3, 4)))
```

[1] 0.08603074

```
#więc ważne są 1,2 i 1,3 i 4

testuj.model(tabela_7, list(c(1, 2),c(1,3), 4),
             list(c(1, 2),c(1, 3), c(1, 4),c(2, 3),c(2, 4), c(3, 4)))
```

[1] 0.220728

```
testuj.model(tabela_7, list(2,c(1,3), 4),
             list(c(1, 2),c(1,3), 4))
```

[1] 2.231232e-13

```
testuj.model(tabela_7, list(c(1, 2),3, 4),
             list(c(1, 2),c(1,3), 4))
```

[1] 6.167442e-17

```
testuj.model(tabela_7, list(c(1, 2),c(1,3)),
             list(c(1, 2),c(1,3), 4))
```

[1] 3.371963e-08

7.2 b)

Formula of backward model: Freq educCat + female + region + wageCat + educCat:wageCat +
educCat:female + female:wageCat **Formula of forward model:** Freq educCat + region **Formula of both model:** Freq educCat + region

AIC (Akaike Information Criterion) to jedno z najczęściej stosowanych kryteriów oceny jakości modelu statystycznego. AIC penalizuje modele za ich złożoność, ale jednocześnie nagradza za dopasowanie do danych. Wzór na AIC to:

$$AIC = 2k - 2\ln(L)$$

gdzie k to liczba parametrów w modelu, a L to funkcja wiarygodności modelu. Niższa wartość AIC wskazuje na lepszy model.

W ramach wyboru modelu log-liniowego do zmiennych ‘wageCat‘, ‘educCat‘, ‘female‘, ‘region‘ zastosowano trzy metody selekcji zmiennych:

- **Backward selection**: Metoda ta polega na rozpoczęciu od pełnego modelu i stopniowym usuwaniu zmiennych, które najmniej przyczyniają się do dopasowania modelu, aż do momentu, gdy dalsze usuwanie zmiennych pogarsza jakość modelu. Otrzymany model to:

  educCat+female+region+wageCat+educCat:wageCat+educCat:female+female:wageCat

- **Forward selection**: Metoda polega na rozpoczęciu od modelu zawierającego tylko stałą, a następnie dodawaniu zmiennych, które poprawiają dopasowanie modelu. Ostateczny model to:

  educCat + region

- **Both (stepwise) selection**: Metoda łączy cechy obu poprzednich: zmienne są dodawane do modelu, ale również usuwane, jeśli nie poprawiają istotnie dopasowania. Model uzyskany tą metodą to:

  educCat + region

Spośród tych modeli, wybieramy model o najniższym AIC, który będzie najlepszym kompromisem pomiędzy jakością dopasowania a prostotą modelu.

7.3 c)

Formula of backward model: Freq educCat + female + region + wageCat + educCat:wageCat +
female:wageCat **Formula of forward model:** Freq educCat + region **Formula of both model:** Freq educCat + region

BIC (Bayesian Information Criterion) jest kolejnym kryterium oceny jakości modelu, które różni się od AIC poprzez silniejszą penalizację za liczbę parametrów. BIC jest szczególnie użyteczne przy selekcji modeli w dużych próbach, ponieważ kara za złożoność modelu jest większa niż w przypadku AIC. Wzór na BIC to:

$$BIC = \ln(n)k - 2\ln(L)$$

gdzie n to liczba obserwacji, k to liczba parametrów, a L to funkcja wiarygodności modelu. Niższa wartość BIC wskazuje na lepszy model.

Podobnie jak w przypadku AIC, zastosowano trzy metody selekcji zmiennych:

- **Backward selection**: Ostateczny model uzyskany tą metodą to:

  educCat + female + region + wageCat + educCat:wageCat + female:wageCat

- **Forward selection**: Model wybrany tą metodą to:

$$\text{educCat} + \text{region}$$

- **Both (stepwise) selection**: Model uzyskany tą metodą to:

$$\text{educCat} + \text{region}$$

Tak jak w przypadku AIC, wybór najlepszego modelu następuje na podstawie najmniejszej wartości BIC. Dla tego zadania modele uzyskane metodą forward i both były identyczne i zawierały tylko zmienne 'Freq', 'educCat' i 'region'.