Miary związku i analiza korespodencji

Aleksandra Palka

10 grudnia 2023

1 Wstęp

W projekcie zostały przeanalizowane dane pod względem występowania zależności między nimi i sile tych związków. Wykorzystane zostały testy niezależności, miary współzmienności i analiza korespondecji.

2 Zadanie 1- testowanie niezależności

Korzystamy z funkcji *chisq.test* w pakiecie R, na poziomie istotności 0.05 by zweryfikować hipotezę o niezależności stopnia zadowolenia z pracy i wynagrodzenia dla danych przedstawionych w Tabeli 1.

Wynagrodzenie	**b. niezadow.**	**niezadow.**	**zadow.**	**b. zadow.**	**Suma**
poniżej 6000	32	44	60	70	206
6000-15000	22	38	104	125	289
15000-25000	13	48	61	113	235
powyżej 25000	3	18	54	96	171
Suma	**62**	**108**	**319**	**412**	**901**

Tabela 1: Stopień zadowolenia z pracy w zależności od wynagrodzenia

```
tabela_1<-matrix(c(32,22,13,3,44,38,48,18,60,104,61,54,70,125,113,96),ncol=4)
rownames(tabela_1)<-c("<6000","6000-15000","15000-25000","25000<")
colnames(tabela_1)<-c("b.niez","niezaw","zaw","b.zaw")

test<-chisq.test(tabela_1)
test_z_poprawka<-chisq.test(tabela_1,correct=TRUE)
test_simulate<-chisq.test(tabela_1,simulate.p.value = TRUE)
```

Po przeprowadzeniu testu Chi-squared w wersji z poprawką Yatesa i bez niej w obu przypadkach dostajemy taką samą p-wartość, mniejszą niż poziom istotności 0.05 dlatego odrzucamy hipotezę o niezależności i wnioskujemy, że stopień zadowolenia z pracy jest zależny od wynagrodzenia. Przeprowadzamy też test Chi-squared korzystając z parametru *simulate.p.value=TRU*E, p-wartość na podstawei 2000 symulacji Monte Carlo jest równa 0.0004998, co jest większe niż ta policzona dokładnie. Nadal jest mniejsza niż poziom istotności i oznacza odrzucenie hipotezy.
Odrzucenie hipotezy o niezależności jest równoznaczne z odrzuceniem hipotezy o jednorodnośći rozkładów warunkowych, zatem nie są one takie same w poszczególnych grupach.

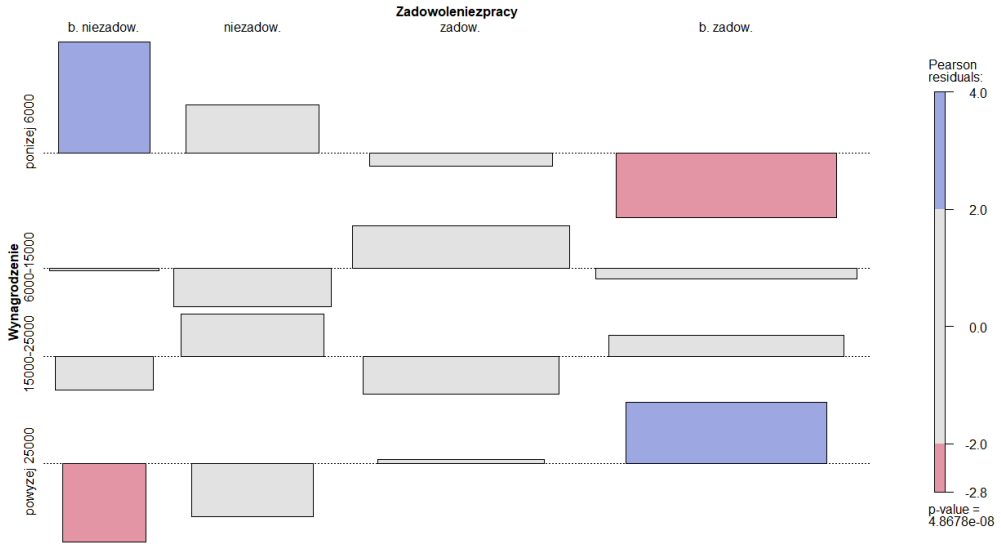
Dla danych z tej tabeli narysujemy również *association plot* korzystając z funkcji *assoc* z pakiety *vcd*. Pozwala on zobaczyć odchylenia, czyli które komórki tabeli mają największy wpływ na odrzucenie hipotezy zerowej.

```
library(vcd)
assoc(tabela_1,shade=TRUE)
```

```
> test

        Pearson's Chi-squared test

data:  tabela_1
X-squared = 51.83, df = 9, p-value = 4.868e-08

> test_z_poprawka

        Pearson's Chi-squared test

data:  tabela_1
X-squared = 51.83, df = 9, p-value = 4.868e-08
```

```
> test_simulate

        Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data:  tabela_1
X-squared = 51.83, df = NA, p-value = 0.0004998
```

Jak widać na wykresie (rys.1.) największą różnicę powodują odpowiedzi osób o skrajnych zarobkach. Wśród osób zarabiających najmniej jest jest znacznie więcej bardzo niezadowolonych niż średnio dla wszystkich badanych i mniej bardzo zadowolonych niż średnio. Dla grupy zarabiającej najlepiej obserwujemy dokładnie odwrotną sytuację.



Rysunek 1: Wykres z funkcji *assoc* dla danych z zadania 1.

3 Zadanie 2 - miary związku

W oparciu o dane w pliku dotyczące reakcji pacjentów na różne dawki dwóch rodzajów leku obliczamy wartości miar współzmienności. Dane składają się ze zmiennych:

- *Dawka* opisującej ilośc podanego leku w skali logarytmicznej,
- *Rodzaj* - 0 lub 1 w zależności od producenta leku,
- *Reakcja* - 1 gdy nastąpiła poprawa po podaniu leku i 0 gdy nie nastąpiła,
- *Miejsce* to miejsce podanie leku - szpital (1) lub dom (0).

3.1 Miara współzmienności dla zmiennych *Reakcja* i *Dawka*.

Zmienna *Dawka* określa ilość podanej dawki leku (w skali logarytmicznej), jest zmienną porządkową. Jako miarę zmienności wykorzystamy współczynnik γ i obliczymy go za pomocą funkcji *GoodmanKruskalGamma* z pakietu *DescTools* na poziomie ufności 0.95.

```
library(DescTools)
reakcja <- read.csv2(file="Reakcja.csv",header=TRUE)
```

```
> GoodmanKruskalGamma(ftable(reakcja,col.vars="Reakcja",row.vars = "Dawka"),direction = "column")
[1] 0.4879575
```

Obliczona wartość wynosi ok. 0.5 i jest dodatnia, co sugeruje zauważalną dodatnią zależność pomiędzy badanymi zmiennymi. Wraz ze zwiększaniem się dawki można było zaobserwować wzrost ilości pozytywnych reakcji na lek.

3.2 Miara współzmienności dla zmiennych *Reakcja* i *Miejsce*.

Zmienna *Miejsce* przyjmuje wartości 0-pacjent leczony w domu i 1-w szpitalu. Jest zatem zmienną nominalną i użyjemy współczynnika τ Goodmana-Kruskala. Do wyliczenia użyjemy funkcji *GoodmanKruskalTau* z pakietu *DescTools* na poziomie ufności 0.95.

```
> GoodmanKruskalTau(ftable(reakcja,col.vars="Reakcja",row.vars = "Miejsce"),direction = "column")
[1] 0.08022077
```

Obliczona wartość wynosi ok. 0.08. Zależność zmiennych między sobą jest niewielka, jednak sugeruje, że trochę częściej pozytywne reakcje na lek można było zauważyć u osób leczonych w szpitalu.

4 Zadanie 3 - analiza danych o wynagrodzeniu

Korzystamy z danych zawartych w Tabela 1 dotyczących zadowolenia z pracy i zarobków badanych osób.

```
tabela1 = matrix(c(32, 44, 60, 70, 22, 38 ,104 ,125, 13,48 ,61 ,113 , 3 ,18 ,54 ,96),nrow=4,byrow=TRUE)
dimnames(tabela1) <- list(Wynagrodzenie=c("ponizej 6000 ", "6000-15000 ", "15000-25000" , "powyzej 25000"),
                Zadowoleniezpracy=c("b. niezadow.", "niezadow.", "zadow." ,"b. zadow."))
```

4.1 Miara współzmienności

Chcemy policzyć miarę współzmienności zmiennych *Wynagrodzenie* i *Stopień zadowolenia z pracy*, czyli sprawdzić czy występuje między nimi jakaś zależność. Skorzystamy ponownie ze współczynnika γ na poziomie ufności 0.95.

```
> GoodmanKruskalGamma(tabela1)
[1] 0.218102
```

Dodatni wynik w tych granicach oznacza, żę dane są od siebie dodatnio zależne. Osoby o wyższych zarobkach są średnio bardziej zadowolone z pracy w porównaniu do osób z mniej płatną pracą.

4.2 Analiza korespondencji

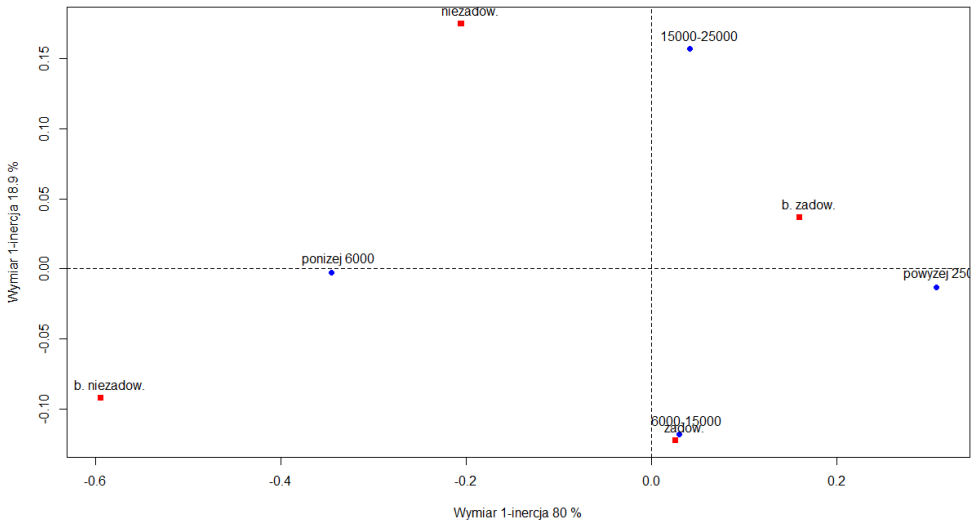
Miara współzmienności pozwala określić siłę zależności między zmiennymi, ale nie dostarcza informacji o jej charakterze. W celu dokładniejszego określenia powiązań między zmiennymi wykorzstamy analizę korespodencji. Do obliczneia współrzednych punktów skorzystamy z funkcji zaimplementowanej na podstawie [1].

```
#funkcja znajdujaca macierze F i G
FG = function(dane){
  c = t(((addmargins(dane)[nrow(dane)+1,]/sum(dane))[-ncol(dane)-1]))
  r = t(((addmargins(dane)[,ncol(dane)+1]/sum(dane))[-nrow(dane)-1]))
  Dr = diag(c(r))
  Dc = diag(c(c))
  P = dane/sum(dane)
  A = solve(Dr^(1/2)) %*% (P-t(r) %*% c) %*% solve(Dc^(1/2))
  A_rozklad = svd(A)
  U = A_rozklad$u
  V = A_rozklad$v
  D = diag(A_rozklad$d)
  F = solve(Dr^(1/2)) %*% U %*% D
  G = solve(Dc^(1/2)) %*% V %*% D
  inercja = sum(D^2)
  inercje_procent = A_rozklad$d^2/inercja
  return (list("F"=F,"G"=G,"inercja"=inercja,"procent"=inercje_procent))
}
```

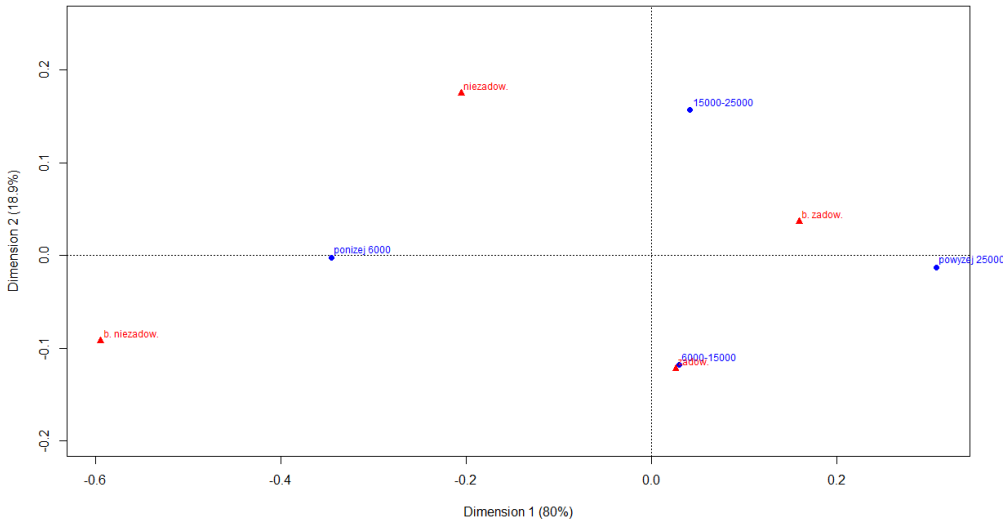
```
# tworzenie wykresu
wsF = FG(tabela1)$F
wsG = FG(tabela1)$G
```

```
inercja = FG(tabela1)$inercja
inercja_wymiarow= FG(tabela1)$procent
xlimit = c(min(min(wsF[,1]),min(wsG[,1])), max(max(wsF[,1]),max(wsG[,1])))
ylimit = xlim = c(min(min(wsF[,2]),min(wsG[,2])), max(max(wsF[,2]),max(wsG[,2])))
xlabel = paste("Wymiar 1-inercja",toString(round(inercja_wymiarow[1]*100,1)),"%")
ylabel = paste("Wymiar 1-inercja",toString(round(inercja_wymiarow[2]*100,1)),"%")
plot(wsF, col=c("blue"),pch=16,xlim=xlimit,ylim=ylimit,xlab=xlabel,ylab=ylabel)
abline(h=0,v=0,lty=2, lwd=1)
points(wsG, col=c("red"),pch=15)
zarobki_nazwy = c("ponizej 6000 ", "6000-15000 ", "15000-25000" , "powyzej 25000")
zadow_nazwy = c("b. niezadow.", "niezadow.", "zadow." ,"b. zadow.")
for (i in 1:nrow(wsF)){
  text(x=wsF[i,1]+0.01,y=wsF[i,2]+0.01,zarobki_nazwy[i])
}
for (i in 1:nrow(wsG)){
  text(x=wsG[i,1]+0.01,y=wsG[i,2]+0.01,zadow_nazwy[i])
}

inercja
```

```
# funkcja z pakietu ca
library(ca)
plot(ca(tabela1))
```



Rysunek 2: Wykres pokazujący zależność zmiennych dla zadania 6 korzystając z własnej funkcji.



Rysunek 3: Wykres pokazujący zależność zmiennych dla zadania 6korzystając z pakiety *ca*.

Oba wykresy (rys.2 i 3) są podobne,współrzędne punktów są zgodne, funckja została zaiplementowana poprawnie.

Inercja całkowita, inaczej bezwładność dla tych danych wynosi ~ 0.057, przy czym maksymalna możliwa wartość to $min\{R-1, C-1\} = 3$ (gdzie C-liczba kolumn, a R-wierszy). Inercje możemy interpretować jako miarę rozproszenia układu, więc to rozproszenie jest bardzo małe, punkty znajdują się w dość niewielkich odległościach od siebie.

Na podstawie wykresu możemy uzyskać informacje o charakterystyce zależności badanych zmiennych. Ogólnie im bliżej siebie położone są punkty, tym bardziej są podobne.
Wymiar 1 odpowiada za 80% wariancji danych, a wymiar 2-gi za 18.9%. Oznacza to, że przedstawiając dane na wykresie straciliśmy bardzo niewiele zmienności danych.
Najbardziej odstającą od środka odpowiedzią było duże niezadowolenie z pracy, to właśnie ta cecha najbardziej różnicuje dane. Również osoby o skrajnych zarobkach mocno odstawały od profilu przeciętnego - te punkty znajdują się po przeciwnych stronach osi OX. Zatem odpowiedzi na pytanie znacząco różniły się pomiędzy grupami. Najniższe zarobki wiązały się z odpowiedziami najbardziej różnymi od przeciętnych. Poza tym można zauważyć podobieństwo wśród odpowiedzi osób o zarobkach w dwóch środkowych przedziałach- są one równocześnie najbardziej zbliżone do profilu średniego.
Analizując rozmieszczenie punktów możemy zauważyć ogólną zależność - osoby o niższych zarobkach częściej odpowiadały, że są niezadowolone lub nawet bardzo niezadowolone ze swojej pracy. Nie oznacza to, że większość z tych osób tak odpowiedziała, lecz, że procentowy stosunek tych odpowiedzi był w tej grupie wyższy niz w innych. Zarobki od 6 do 15 tyś często wiązały się z odpowiedzią o zadowoleniu z pracy. Wśród tej grupy i osób zarabiających od 15 do 25 tyś. odpowiedzi były dość podoobne do średnich dla wszystkich badanyhc(mała odległość od środka), są wśród nich osoby zadowolone jak i niezadowolone. Dla najlepiej zarabiających widać, że przedstawiciele tej grupy przeważnie nie byli niezadowoleni.

5 Zadanie 4 - analiza przykładowych danych o lekach

Korzystamy z danych zawartych w Tabela 2, wyników ankiety dotyczącej najczęściej stosowanego leku przeciwbólowego. Odpowiedzi zostały dodatkowo podzielone ze względu na wiek ankietowanych.

Wiek	**do lat 35**	**od 36 do 55**	**powyżej 55**	**Suma**
Ibuprom	35	0	0	35
Apap	22	22	0	44
Paracetamol	15	15	15	45
Ibuprofen	0	40	10	50
Panadol	18	3	5	26
Suma	**90**	**80**	**30**	**200**

Tabela 2: Wiek ankietowanych i preferencje dotyczące leków

```
leki = matrix(c( 35, 0, 0, 22, 22, 0, 15, 15, 15, 0, 40, 10, 18, 3, 5 ),nrow=5,byrow=TRUE)
dimnames(leki) <- list(Lek=c("Ibuprom","Apap","Paracetamol","Ibuprofen","Panadol"),
                Wiek=c("do lat 35" ,"od 36 do 55" ,"powyzej 55"))
```

5.1 Miary współzmienności

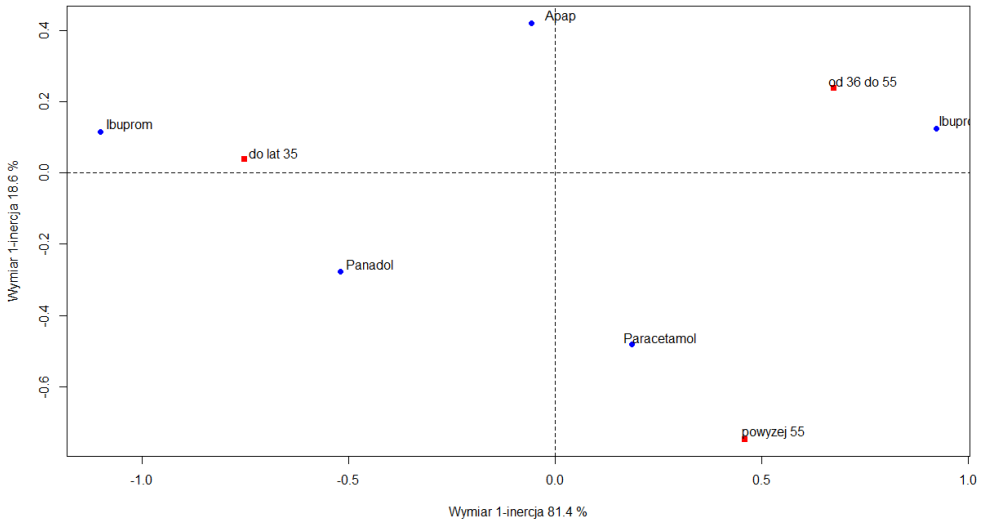
Do sprawdzenia na poziomie ufności 0.95 zależności wybieranego leku od wieku wykorzystamy współczynnik τ .

```
> GoodmanKruskalTau(ftable(leki,col.vars = "Lek",row.vars = "Wiek"),direction="column")
[1] 0.1512141
```

Obliczona wartość wskazuje na zależność, aczkolwiek że nie jest ona bardzo silna.

5.2 Analiza korespondencji

```
#tworzenie wykresu
wsF = FG(leki)$F
wsG = FG(leki)$G
inercja = FG(leki)$inercja
inercja_wymiarow= FG(leki)$procent
xlimit = c(min(min(wsF[,1]),min(wsG[,1])), max(max(wsF[,1]),max(wsG[,1])))
ylimit = xlim = c(min(min(wsF[,2]),min(wsG[,2])), max(max(wsF[,2]),max(wsG[,2])))
xlabel = paste("Wymiar 1-inercja",toString(round(inercja_wymiarow[1]*100,1)),"%")
ylabel = paste("Wymiar 1-inercja",toString(round(inercja_wymiarow[2]*100,1)),"%")
plot(wsF, col=c("blue"),pch=16,xlim=xlimit,ylim=ylimit,xlab=xlabel,ylab=ylabel)
abline(h=0,v=0,lty=2, lwd=1)
points(wsG, col=c("red"),pch=15)
leki_nazwy = c("Ibuprom","Apap","Paracetamol","Ibuprofen","Panadol")
wiek_nazwy = c("do lat 35" ,"od 36 do 55" ,"powyzej 55")
for (i in 1:nrow(wsF)){
  text(x=wsF[i,1]+0.07,y=wsF[i,2]+0.02,leki_nazwy[i])
}
for (i in 1:nrow(wsG)){
  text(x=wsG[i,1]+0.07,y=wsG[i,2]+0.02,wiek_nazwy[i])
}

inercja
```



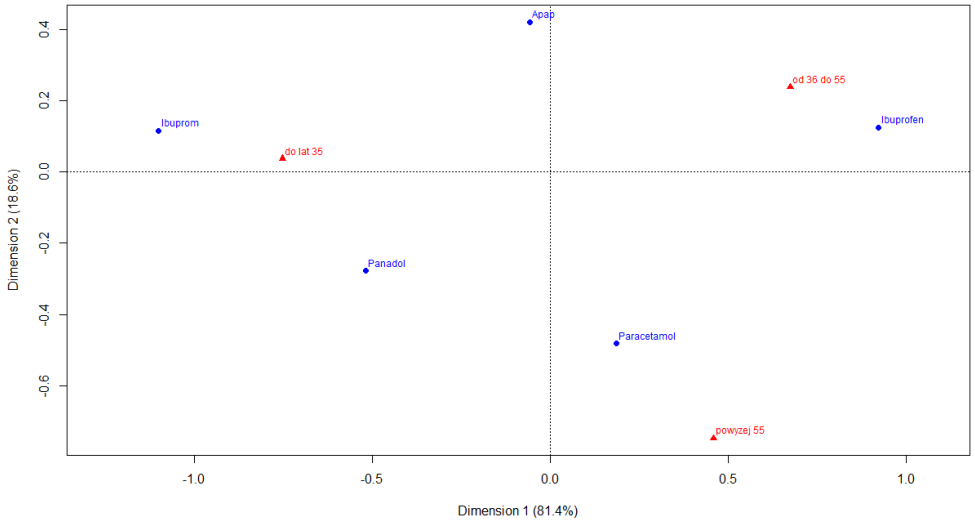
Rysunek 4: Wykres pokazujący zależność zmiennych dla zadania 7 korzystając z własnej funkcji.

Inercja całkowita, inaczej bezwładność dla tych danych wynosi ~ 0.57, przy czym maksymalna możliwa wartość to $min\{R-1, C-1\} = 2$. Rozproszenie układu nie jest bardzo duże.
Pierwszy wymiar odpowiada za 81.4% bezwładności, a drugi-18.6%, co oznacza, że na wykresie przedstawiona jest znaczna większość zmienności danych.

Patrząc na punkty odpowiadające wiekowi respondentów widzimy, że odpowiedzi na pytanie rożniły się w grupach wiekowych. Osoby młodsze poniżej 35 lat wybierały najczęściej inny lek niż starsze. Osoby w wieku średnim i starsze również odpowiadały inaczej na pytanie, lecz te róznice nie są aż tak duże jak w przypadku najmłodszych badanych.
Wśród leków najbardziej skrajne na mapie są punkty odpowiadające wyborowi Ibupromu i Ibuprofenu. Procent wyboru tych leków wśród różnych grup znacząco się różnił, czyli wśród grupy, która wybierała Ibuprom niewiele wybierało Ibuprofen i na odwrót. Duży odsetek wyboru Ibupromu wiązał się też z częstym wyborem Pandolu przez innych przedstawicieli grupy.
Ibuprom oraz Panadol wybierały głównie osoby młode, a Ibuprofen był często wybieranym lekiem przez osoby w średnim wieku. Starsi decydowali częściej od innych na Paracetamol.
Najbliżej profilu środkowego położone są Apap, i Paracetamol - to te leki najmnniej rożnicowały grupy, tj. choć były preferowane wbardziej śród starszych, młodsi również po nie sięgali.

Bibliografia

[1] M.Greenacre O. Nenadic. “Correspondance analysis in R, with two- and three-dimensional graphics: the ca package”. W: *Journal of Statistica Software* (2007).



Rysunek 5: Wykres pokazujący zależność zmiennych dla zadania 7 korzystając z pakiety *ca*.