Zaawansowane modele liniowe raport nr 3

Dominik Mika

13 października 2021

Symulacje

Zadanie 1

W pierwszym zadaniu będziemy badać rozkłady statystyki T oraz χ^2. Służą one do testowania hipotez:

$$H_0 : \text{dane pochodzą z rozkładu Poissona}$$

vs

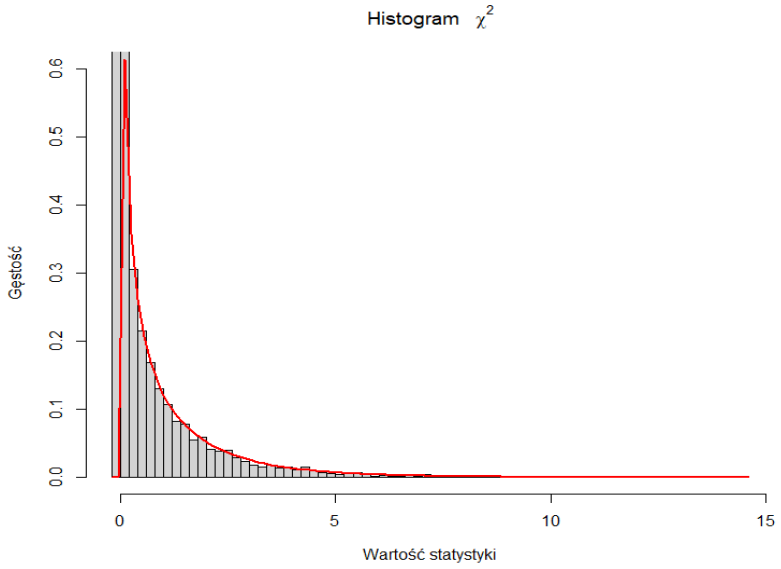
$$H_1 : \text{dane pochodzą z rozkładu ujemnego dwumianowego.}$$

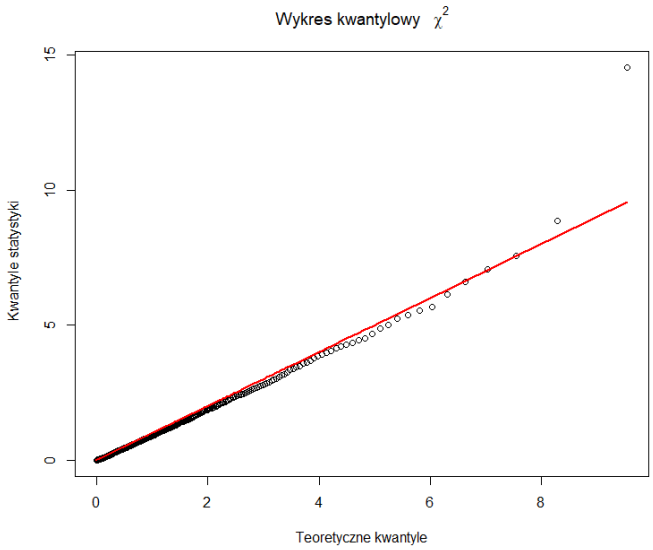
Wyglądają one następująco:

1. $\chi^2 = D(M_1) - D(M_2)$
2. $T = \frac{\hat{\alpha}}{\text{Var}(\hat{\alpha})}$

W tym celu wygenerujemy macierz $X \in \mathbb{M}_{1000\times 2}$, t.że jej elementami są i.i.d zmienne z rozkładu $N(0, \sigma = 1/\sqrt{1000})$. Następnie wyznaczymy ciąg predyktorów liniowych $\eta = X\beta$, gdzie $\beta = (3, 3)$ i na ich podstawie wygenerujemy 10000 niezależnych replikacji wektora odpowiedzi y przy założeniu hipotezy zerowej. Na koniec przedstawimy histogramy oraz wykresy kwantylowe tych statystyk.

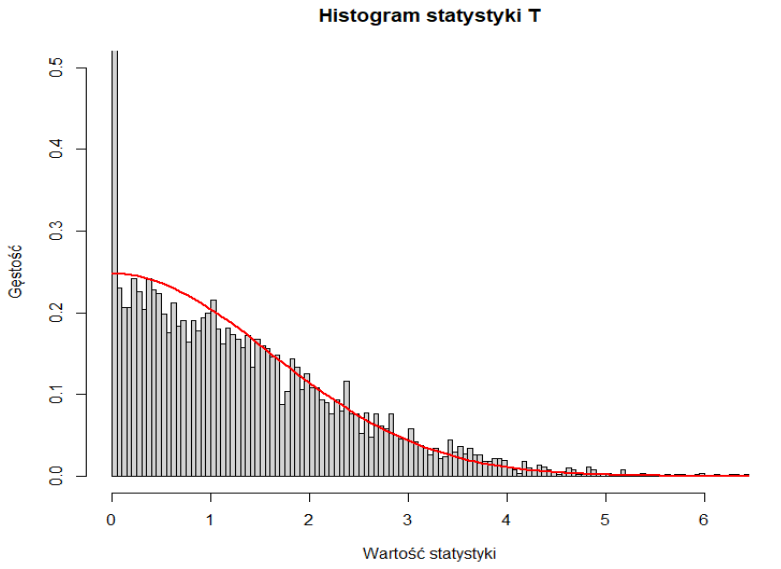
Poniżej znajdują się wykresy dla statystyki chi-kwadrat:

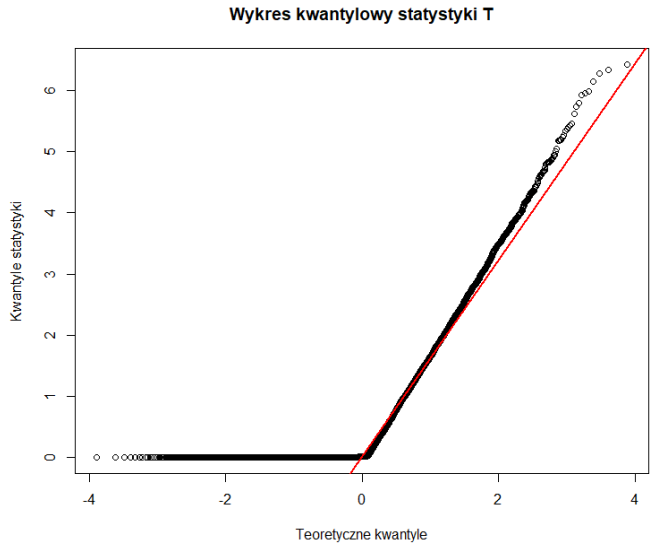




Na histogramie widzimy, że krzywa gęstości dość dobrze dopasowuje się do histogramu. Jedynie na początku widzimy, że histogram mocno wychodzi ponad krzywą. Wynika to z tego, że dopasowana gęstość to gęstość rozkładu chi-kwadrat, a nasza statystyka pochodzi z mieszanki tego rozkładu i rozkładu skoncentrowanego w zerze. Natomiast na wykresie kwantylowym widzimy, że punkty w miarę układają się na prostej. Prosta ma nieco zbyt duże nachylenie. Powodem jest obserwacja odstająca, znajdująca się w prawym górnym rogu.

Natomiast poniżej znajdują się wykresy dla statystyki T:





W tym przypadku również widzimy, że krzywa dobrze dopasowuje się do histogramu oprócz początku, gdzie wartość 0 występuje zdecydowanie zbyt często w porównaniu do rozkładu normalnego $N(0,1)$. Ponownie wynika to z tego, że statystyka T pochodzi z mieszanki rozkładu normalnego i skoncentrowanego w zerze. Również na wykresie kwantylowym widzimy, że punkty w miarę do-

brze dopasowują się do krzywej. W celu otrzymania bardziej dopasowanej moglibyśmy manipulować wartością $\hat{\sigma}$.

Analiza danych

Zadanie 2

W tej części dokonamy analizy zbioru danych *Deb and Trivedi*. Będziemy chcieli zbadać związek pomiedzy liczbą wizyt w gabinecie lekarskim (zmienna zależną, kolumna "ofp") i zmiennymi niezależnymi opisujacymi pacjenta:

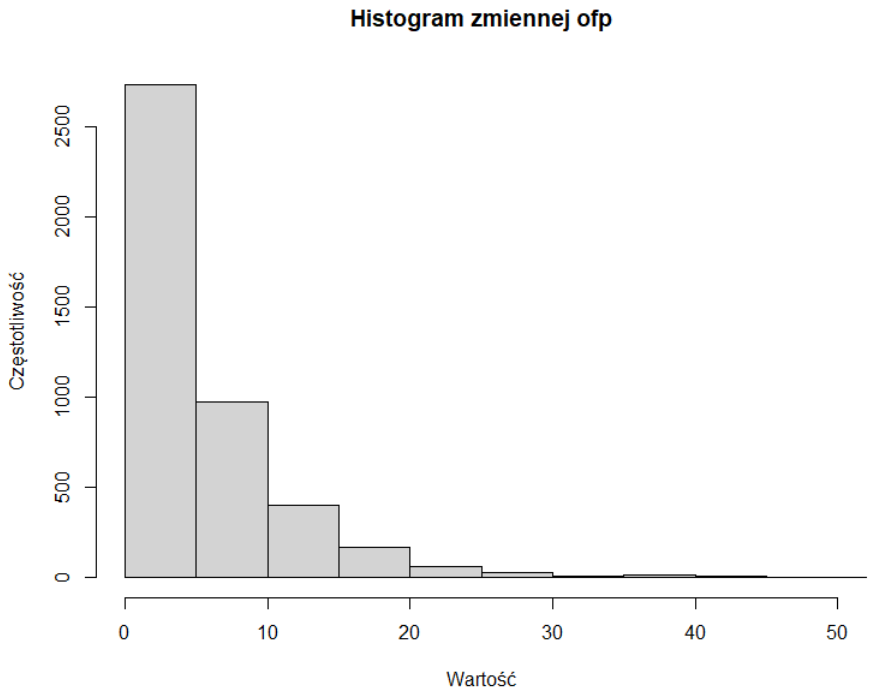
- "hosp"– liczba pobytów w szpitalu,
- "health"– zmienna opisująca subiektywny odczucie pacjenta o jego zdrowiu,
- "numchron"– liczba przewlekłych stanów chorobowych,
- "gender"– płeć,
- "school"– liczba lat edukacji,
- "privins"– indykator opisujący to czy pacjent ma dodatkowe prywatne ubezpieczenie zdrowotne.

Zadanie 3

W tym zadaniu dokonamy wstępnej analizy oraz wizualizacji nasze zbioru danych.

- Najpierw narysujemy histogram zmiennej "ofp", aby zbadać czy występuje zjawisko nadmiernej dyspersji lub inflacji w zerze.

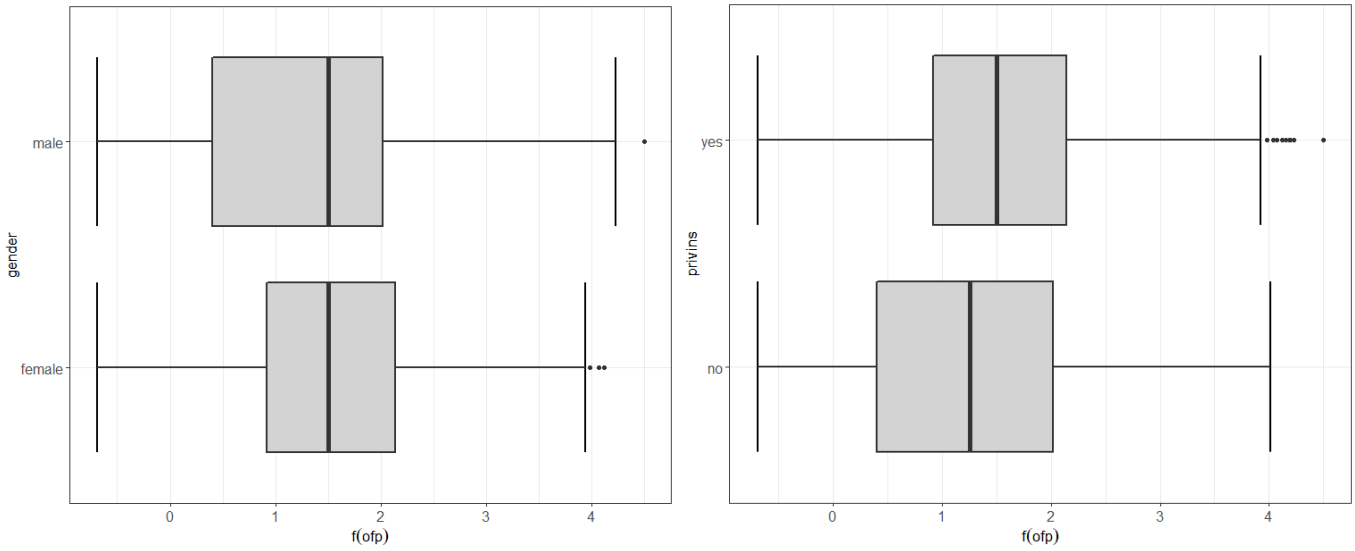
  Histogram znajduje się poniżej:

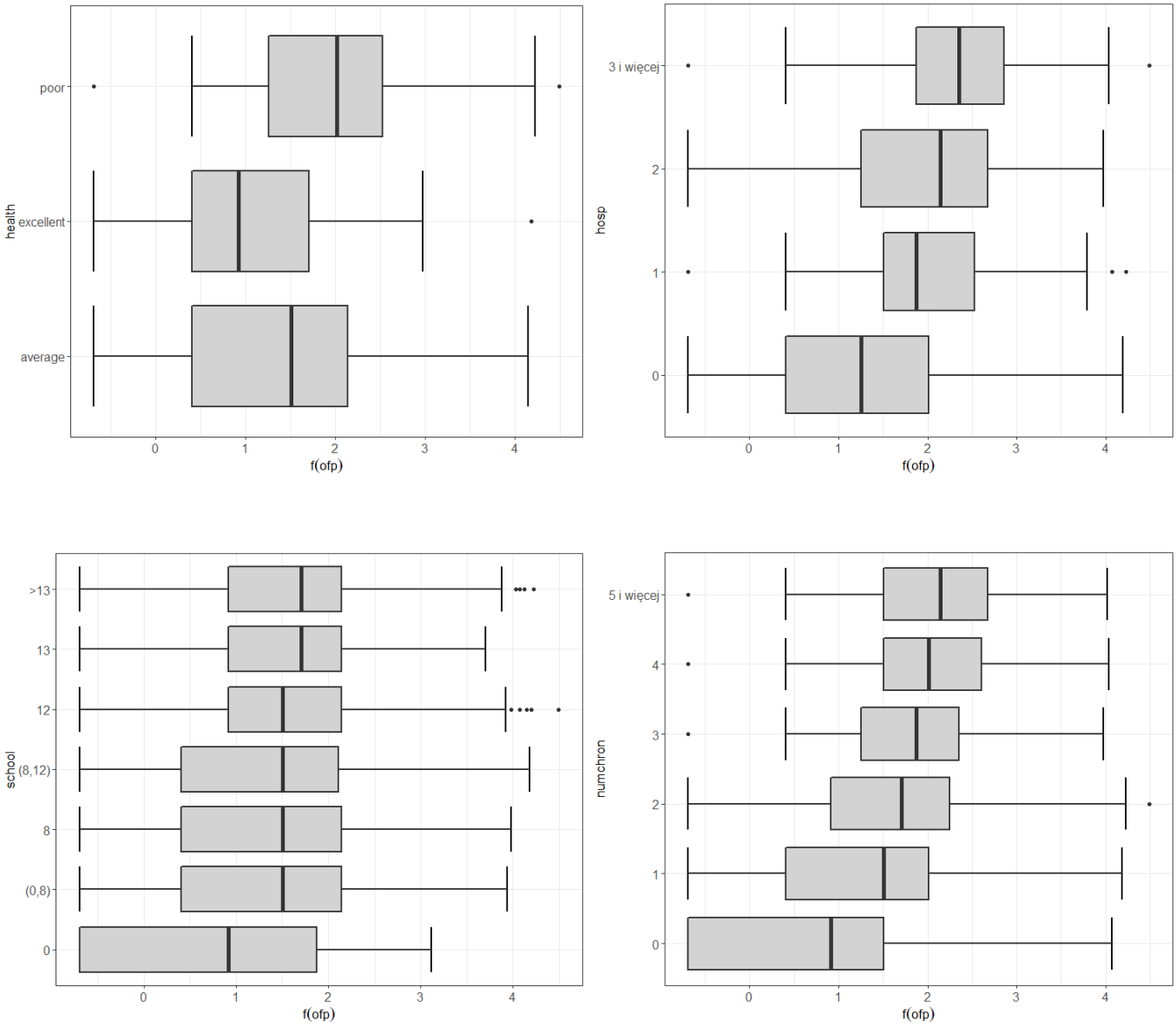


Pierwsze co możemy zauważyć, to zdecydowanie większa wysokość słupka odpowiadającemu wartości 0. Możemy, więc przypuszczać, że występuje tutaj inflacja w zerze. Natomiast bezpośrednio z tego wykresu nie potrafimy wywnioskować, czy zachodzi nadmierna dyspersja.

- Następnie ze względu na znaczącą liczbę zer wprowadzimy zmienną pomocniczą $f(ofp) = \log(ofp + 0.5)$ dzięki której łatwiej będzie zbadać wzajemne zależności pomiędzy "ofp"i regresorami.
- Na koniec przedstawię wykresy pudełkowe wszystkich zmiennych objaśniających. Ze względu na to, że niektóre z tych zmiennych przyjmują dużo wartości, które rzadko występują, dokonam odpowiednie grupowania w celu uproszczenia analizy.

Poniżej znajdują się wykresy dla wszystkich zmiennych:





Na wszystkich wykresach widzimy, że wszystkie z tych zmiennych prawdopodobnie są istotne. Niektóre „pudełka" są do siebie bardzo podobne, ale najczęściej widzimy, że rozstępy kwartylowe nieco się różnią. W przypadku zmiennej *school* widzimy, że wykresy praktycznie się nie różnią dla liczby lat edukacji należącej do przedziału $(0, 12)$. Wynika to prawdopodobnie z tego, że w tym przedziale znajdują się osoby z wykształceniem podstawowym. Dla osób z liczbą lat edukacji ≥ 12 pudełka są również podobne ale nieco się różnią.

Zadanie 4

Na koniec spróbujemy wybrać najlepszy model dopasowujący się do naszych danych. Skorzystamy z kilku różnych modeli, a następnie w każdym z nich jeżeli to konieczne dokonamy redukcji nieistotnych zmiennych co potwierdzimy odpowiednimi testami.

Następnie dla wszystkich tych modeli dokonamy tabelaryzacji wyników. Podamy parametry regresji, jeżeli takie istnieją to podamy parametry związane z nadmierną dyspersją lub inflacją w zerze,

logarytm funkcji wiarogodności, AIC, BIC oraz oczekiwaną liczbę zer generowanych przez model.

Model Poissona

W modelu Poissona okazuje się, że wszystkie zmienne są istotne. Dla wszystkich zmiennych p-wartość jest mniejsza niż 10^{-16}. Stąd w tym przypadku zostawiamy pełny model.

Model ujemny dwumianowy

W przypadku tego modelu mamy podobną sytuację do poprzedniej. Niektóre p-wartości są nieco większe, ale wszystkie nadal są mniejsze niż 0.05. Stąd ponownie model pełny jest najbardziej adekwatny.

Model ZIPR

W przypadku tego modelu wszystkie parametry związane ze zliczaniem zdarzenia również są istotne. Natomiast w przypadku modelowania inflacji w zerze okazuje się, że zmienna *health* jest nieistotna. W celu sprawdzenia naszych przypuszczeń wykonamy test oparty na statystyce chi-kwadrat. Wynosi ona 2.431, a jej p-wartość 0.2965. Stąd na poziomie istotności $\alpha = 0.05$ nie odrzucilibyśmy hipotezy zerowej, czyli model zredukowany jest wystarczający.

Model ZINBR

W tym modelu również wszystkie zmienne związane ze zliczaniem zdarzenia są istotne, ale okazuje się, że zmienne *health* i *hosp* są nieistotne dla modelowania inflacji w zerze. By potwierdzić to znów wykonamy test chi-kwadrat. Statystyka wynosi 7.217, a p-wartość 0.0653, stąd na poziomie istotności $\alpha = 0.05$ znów nie odrzucamy hipotezy zerowej, czyli model zredukowany jest wystarczający.

Model Poissona z barierą

W tym modelu mamy bardzo podobną sytuację do modelu ZIPR, czyli chcemy odrzucić zmienną *health* z inflacji w zerze. Statystyka wynosi 3.987, a p-wartość 0.136, czyli model zredukowany jest wystarczający.

Model ujemny dwumianowy z barierą

Tutaj również jak w poprzednim zadaniu odrzucamy zmienną chcemy *health* w infalcji w zerze. Satystyka testowa wynosi 3.987, a p-wartość 0.136, czyli znów model zredukowany jest wystarczający

Podsumowanie wyników

Dla wszystkich tych modeli stabelaryzujemy nasze wyniki. Oznaczymy parametry β_i jako te związane ze zliczaniem zdarzenia, parametry γ_i jako te związane z modelowanie inflacji w zerze, a α to parametr związany z nadmierną dyspersją. Natomiast indeksy parametrów regresji odnoszą się do odpowiednich zmiennych lub stanów zmiennych. Opis przedstawiam poniżej:

1. $\hat{\beta}_0, \hat{\gamma}_0$- intercept
2. $\hat{\beta}_1, \hat{\gamma}_1$- hosp
3. $\hat{\beta}_2, \hat{\gamma}_2$- healthexcellent
4. $\hat{\beta}_3, \hat{\gamma}_3$- healthpoor
5. $\hat{\beta}_4, \hat{\gamma}_4$- numchron
6. $\hat{\beta}_5, \hat{\gamma}_5$- gendermale
7. $\hat{\beta}_6, \hat{\gamma}_6$- school
8. $\hat{\beta}_7, \hat{\gamma}_7$- privinsyes

Poniżej znajduje się tabelka dla modeli niemodelujących inflacji w zerze:

	Model Poissona	Model ujemny dwumianowy
$\hat{\beta}_0$	1.029	0.929
$\hat{\beta}_1$	0.165	0.218
$\hat{\beta}_2$	-0.362	-0.342
$\hat{\beta}_3$	0.248	0.305
$\hat{\beta}_4$	0.147	0.175
$\hat{\beta}_5$	-0.112	-0.126
$\hat{\beta}_6$	0.026	0.027
$\hat{\beta}_7$	0.202	0.224
Liczba parametrów	8	9
$\hat{\alpha}$	-	0.829
AIC	35959.23	24359.11
BIC	36010.35	24416.62
Logarytm wiarogodności	-17971.61	-12170.55
Oczekiwana liczba zer	46.71	608.01

Porównując kryteria informacyjne dla tych dwóch modeli widzimy, że różnica jest dość znacząca dla dwóch modeli. Wiemy, że model Poissona nie modeluje nadmiernej dyspersji, a model ujemny dwumianowy już tak. Możemy, więc podejrzewać, że takie zjawisko zachodzi.

Natomiast poniższa tabelka zawiera podsumowanie modeli modelujących inflację w zerze:

	Model ZIPR	Model ZINBR	Model Poissona z barierą	Model ujemny dwumianowy z barierą
$\hat{\beta}_0$	1.406	1.198	1.406	1.198
$\hat{\beta}_1$	0.159	0.211	0.159	0.212
$\hat{\beta}_2$	-0.307	-0.322	-0.304	-0.332
$\hat{\beta}_3$	0.253	0.285	0.254	0.316
$\hat{\beta}_4$	0.102	0.129	0.102	0.126
$\hat{\beta}_5$	-0.062	-0.084	-0.062	-0.068
$\hat{\beta}_6$	0.019	0.021	0.019	0.021
$\hat{\beta}_7$	0.081	0.118	0.081	0.100
$\hat{\gamma}_0$	-0.059	-0.070	0.016	0.016
$\hat{\gamma}_1$	-0.307	-	0.318	0.318
$\hat{\gamma}_2$	-	-	-	-
$\hat{\gamma}_3$	-	-	-	-
$\hat{\gamma}_4$	-0.540	1.247	0.548	0.548
$\hat{\gamma}_5$	0.418	0.556	-0.419	-0.419
$\hat{\gamma}_6$	-0.056	-0.084	0.057	0.057
$\hat{\gamma}_7$	-0.754	-1.243	0.746	0.746
Liczba parametrów	14	14	14	15
$\hat{\alpha}$	-	0.673	-	0.717
AIC	32298.49	24216.51	32300.88	24210.14
BIC	32387.96	24305.98	32390.35	24306.00
Logarytm wiarogodności	-16135.24	-12094.25	-16136.44	-12090.07
Oczekiwana liczba zer	682.3	707.22	683	683

Tutaj też możemy zauważyć, że modele z rozkładem ujemny dwumianowy zmiennej odpowiedzi, która odnosi się do zliczania wizyt w gabinecie pod względem kryteriów informacyjnych wypadają dużo lepiej. Możemy, więc powiedzieć, że prawdopodobnie w naszych danych występuje zjawisko nadmiernej dysperji. Lepiej również wypadły modele modelujące inflację w zerze, czyli ponownie możemy założyć, że takie zjawisko występuje.

Podsumowując widzimy, że pod względem AIC najlepszy model to ten ujemny dwumianowy z barierą. Natomiast pod względem BIC najlepszy jest model ZINBR. Wynika to z tego, że model z barierą ma najwięcej parametrów za co w przypadku BIC płacimy karę. Do wybrania lepszego modelu możemy posłużyć się również oczekiwaną liczbą zer, a tutaj widzimy, że lepiej radzi sobie model z barierą, ponieważ liczba zer w naszym zbiorze to dokładnie 683.