Lista 2

Aneta Przydróżna

Praca nad zbiorem danych, opisującym relacje między prawdopodobieństwem przyjęcia na studia a wynikami z testów rachunkowych i poziomem niepewności.

Do estymacji macierzy kowariancji wektora estymatorów w modelu regresji logistycznej posłuży macierz informacji Fishera, która po odwróceniu daje szukaną macierz. Pierwiastek z elementów na przekątnej daje bardzo bliskie wartości estymatorom odchyleń standardowych, zwracanych przez R. Różnice zaczynają się na poziomie 10^{-2}. Wszystki wyestymowane ręcznie wartości są nieznacznie większe od tych wyprodukowanych przez R.

$I^{-1} = (X^T S X)^{-1}$, gdzie S jest macierzą diagonalną z wyrazami $s_{ii} = \hat{p}_i(1 - \hat{p}_i)$ na przekątnej.

```
##               ręcznie   przez.R
## (Intercept) 6.7992310 6.7985192
## numeracy    0.2480977 0.2480840
## anxiety     0.4804650 0.4804027
```

Testowanie hipotezy zerowej mówiącej, że obie zmienne nie mają wpływu na zmienną odpowiedzi.

$H_0 : \beta_1 = \beta_2 = 0$

Do testowania posłużą wartości null deviance oraz residual deviance, nazwane odpowiednio ND i RD. Przy założeniu braku istotności zmiennych dla dużej liczby obserwacji, różnica tych dwóch dewiancji ma rozkład $\mathcal{X}^2$ o liczbie stopni swobody równej liczbie zmiennych niezależnych w modelu, czyli w tym przypadku 2.Gdy wartość statystyki ND-RD przekroczy wartość testu Chi kwadrat, wtedy odrzucona zostaje hipoteza zerowa.W tym przypadku ND-RD=39.7, a $\mathcal{X}^2{}_2(0.95)$=3.8. Korzystając z wartości wyrażenia '1- dystrybuanta' otrzymujemy liczbę bardzo bliską 0, co również sugeruje, że zmienne objaśniające X_1 i X_2 są istotne.

Testowanie hipotezy zerowej mówiącej, że rozkład danych jest zgodny z założonym modelem.

Tym razem statystyką jest wartość RD=28.3, która jest znacznie mniejsza od kwantyla, dla poziomu istotności 0.05 i 47 stopni swobody. Wyrażenie '1- dystrybuanta' wynosi 0.99, co jest znacznie większe od 0.05, więc tym razem przyjęta zostaje hipoteza H_0.Można stwierdzić, że model jest dobrze dopasowany.

ϵ w metodzie Newtona

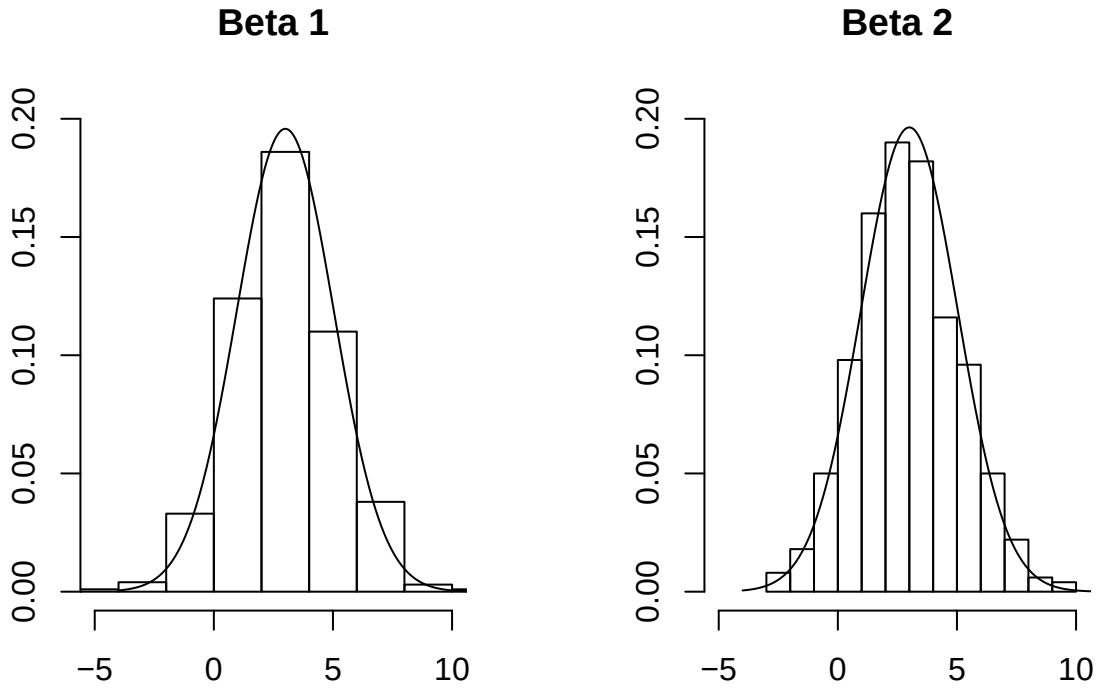
Po wykonaniu funkcji summary od modelu w R, otrzymujemy liczbę iteracji Fishera, aczkolwiek w regresji logistycznej stosowany jest algorytm Newtona-Raphsona. Dana liczba to k-ty krok zbliżania się do poszukiwanego, minimalnego β, do momentu, aż różnica parametru w k+1 i k-tym kroku będzie mniejsza niż ϵ. Poniżej znajduje się porównanie modeli stworzonych dla następujących ϵ-ów: 10^{-1}, 10^{-2}, 10^{-3}i 10^{-6}.

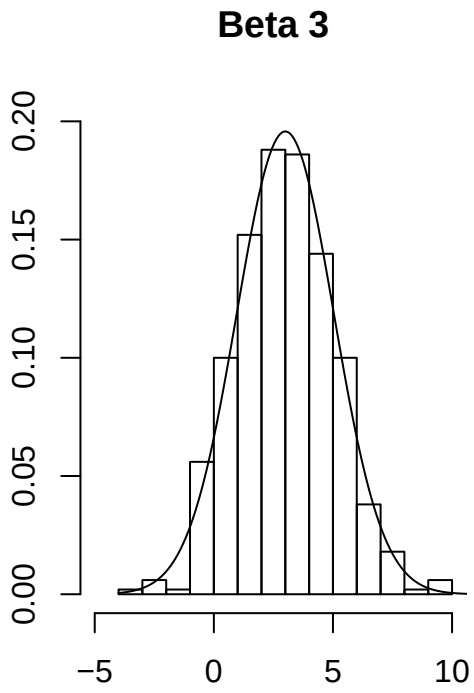
```
##          0.1       0.01      0.001  0.000001
## 1 12.8900764 14.0924744 14.2368342 14.238581
## 2  0.5375846  0.5735451  0.5773136  0.577352
## 3 -1.2639531 -1.3713061 -1.3839203 -1.384069
## 4  3.0000000  4.0000000  5.0000000  6.000000
```

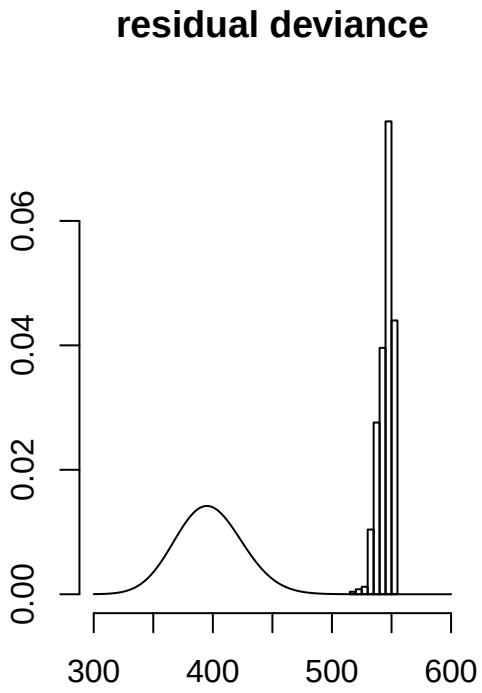
Estymowane parametry zwiększają swoją wartość bezwzględną przy zmniejszającej się wartości ϵ. Można też zauważyć, że zmiany są coraz subtelniejsze wraz z zawężającym się progiem wielkości różnic kolejnych kroków. Czwarty wiersz pokazuje ilość kroków algorytmu, dla odpowiednich ϵ-óW.

Regresja logistyczna z macierzą X_{400x3}, której elementy są zmiennymi losowymi z rozkładu $N(0, \sigma^2 = \frac{1}{n})$ oraz wektorem $\beta = (3, 3, 3)$.

Macierz informacji Fishera: $I = (X^T SX)$, gdzie S jest macierzą diagonalną z wyrazami $s_{ii} = \hat{p}_i(1 - \hat{p}_i)$ na przekątnej. Wyestymowany wektor $\hat{\beta} = (X^T X)^{-1} X^T Y$ dzięki metodzie największej wiarygodności posłuży do utworzenia asymptotycznej macierzy kowariancji.







Histogramy β_1 , β_2 ,β_3 powinny przyjmować kształt rozkładu normalnego z wartością średnią 3 i wariancją równą odpowiedniej wartości z przekątnej odwróconej macierzy Fishera.Ten rozkład został zaznaczony ciągłą linią, natomiast słupki wyznaczają częstotliwość przyjmowania wartości bliskich 3, które były wyestymowane 500 razy. Histogram residual deviance przyjmuje wartości bliskie 550, rozkład chi kwadrat z n-p=497 stopniami swobody, do którego histogram powinien się dopasować, gdyby rozkład danych był zgodny z założonym modelem, ma wartość średnią bliską 400 i jest bardziej rozłożysty.

Poniżej znajdują się obciążenia β

```
##        beta_1     beta_2     beta_3
## 1 -0.03679753 -0.0375894 0.01559782
```

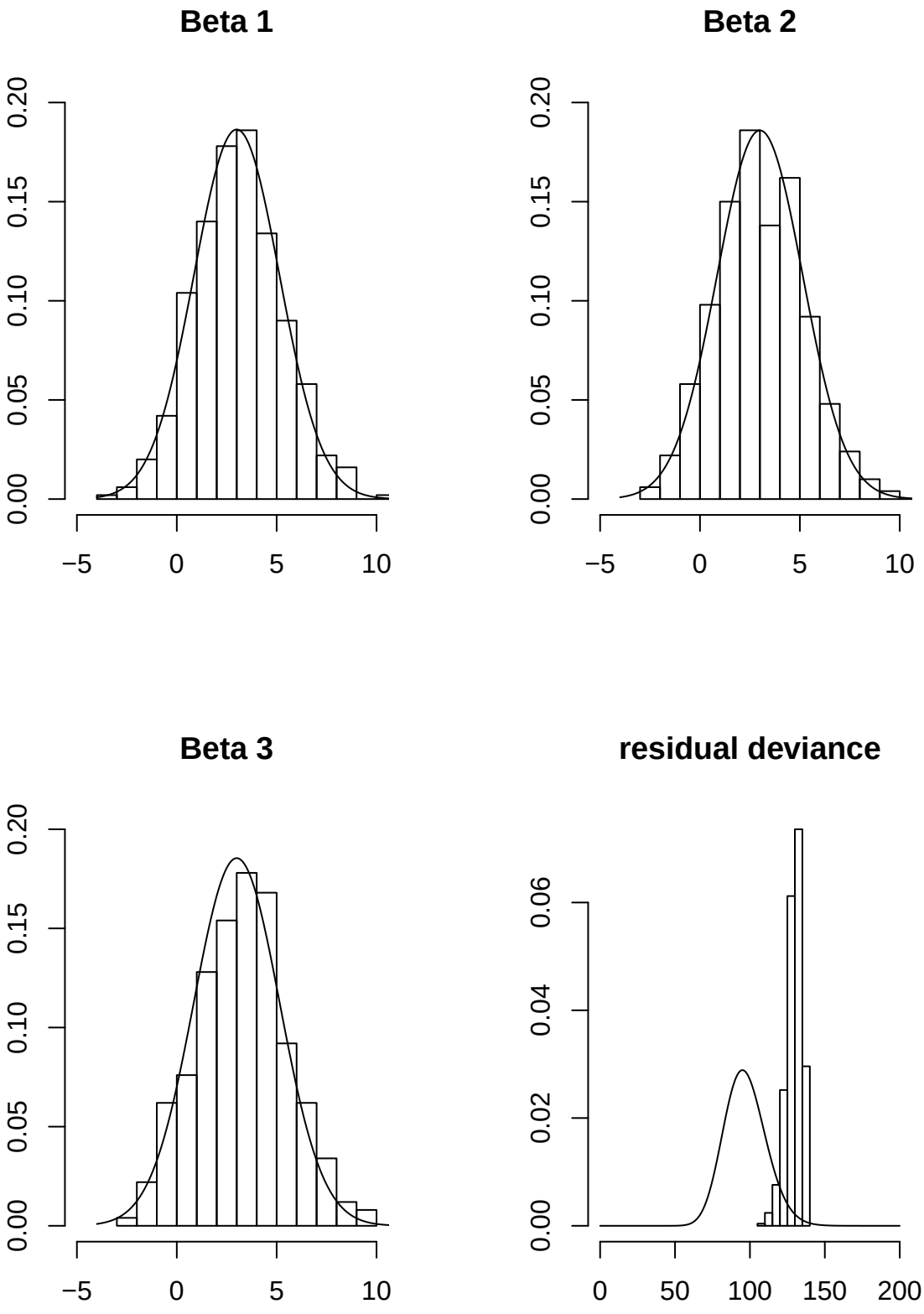
```
##              [,1]      [,2]         [,3]
## [1,] 4.6111940193 0.1356447 0.0005935101
## [2,] 0.1356447402 4.4295715 0.4058149599
## [3,] 0.0005935101 0.4058150 4.0450600965
```

```
##            [,1]       [,2]       [,3]
## [1,] 4.15340691 0.04085715 0.05121462
## [2,] 0.04085715 4.12783741 0.03088491
## [3,] 0.05121462 0.03088491 4.15489275
```

```
##             [,1]        [,2]        [,3]
## [1,] -0.45778711 -0.09478759  0.05062111
## [2,] -0.09478759 -0.30173408 -0.37493005
## [3,]  0.05062111 -0.37493005  0.10983266
```

Pierwsza macierz została stworzona z wyestymowanego 500 razy wektora β, natomiast druga jest asymptotyczną macierzą kowariancji. Trzecia to macierz różnic.

Regresja logistyczna z macierzą X_{100x3}.



Dla n=100 histogramy β_1 , β_2 ,β_3 nie zmieniły się za bardzo, natomiast dla residual deviance gęstość chi kwadrat i rozkład estymowanych danych są znacznie bliżej siebie.

Poniżej znajdują się obciążenia β, które są większe niż powyżej,przez własność dążenia estymatorów do wartości parametru równego 3, wraz ze wzrostem n. Dlatego, gdy n spadło z 400 na 100, obciążenie wzrosło.

```
##       beta_1      beta_2    beta_3
## 1 0.09569586 0.004927774 0.2731382
```

```
##           [,1]       [,2]       [,3]
## [1,] 4.7115508  0.1931715  0.2772884
## [2,] 0.1931715  4.9377209 -0.2272179
## [3,] 0.2772884 -0.2272179  5.1132595
```

```
##           [,1]      [,2]      [,3]
## [1,] 4.5806720 0.1645425 0.2254804
## [2,] 0.1645425 4.6030143 0.1986827
## [3,] 0.2254804 0.1986827 4.6290986
```

```
##             [,1]        [,2]        [,3]
## [1,] -0.13087879 -0.02862899 -0.05180798
## [2,] -0.02862899 -0.33470659  0.42590064
## [3,] -0.05180798  0.42590064 -0.48416091
```

Trzy tabelki, znowu przedstawiają kolejno: macież kowariancji estymatorów, asymptotycznąmacierz oraz macierz różnic.

Tym razem zrobię regresję logistyczną z macierzą X_{400x3}, której elementy są zmiennymi losowymi z rozkładu $N(0, \sigma^2 = \frac{1}{n}S)$, gdzie $S_{ii} = 1$, a $S_{ij} = 0.3$, dla i różnego od j.

Wykresy estymatorów β się nie zmieniły, podobnie jak histogram residual deviance dla n=400.

```
##       beta_1       beta_2     beta_3
## 1 0.08125032 -0.006194345 0.04975074
```

```
##             [,1]       [,2]       [,3]
## [1,] -0.07121663 -0.1045220 -0.1471461
## [2,] -0.10452197 -0.3263343  0.4006627
## [3,] -0.14714607  0.4006627 -0.5215594
```

Obciążenia są najmniejsze, jak do tej pory. Różnice macierzy wahają się od ok. 0 do 0.6.

Powyższy schemat dla n=100.

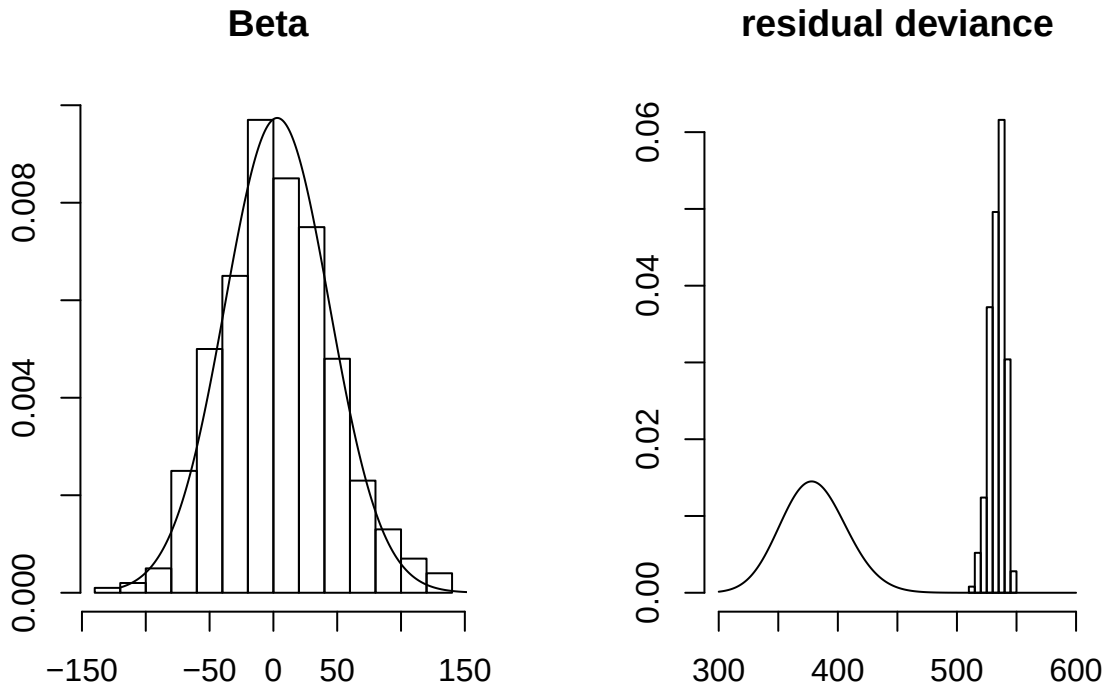
Tutaj histogramy wyglądają podobnie jak wcześniej, gdy n również było równe 100.

Macierz różnic:

```
##      beta_1      beta_2    beta_3
## 1 0.1121226 -0.04003224 0.1226744
```

```
##             [,1]        [,2]        [,3]
## [1,] -0.06825864 -0.21115159 -0.04636236
## [2,] -0.21115159  0.20801124  0.07714837
## [3,] -0.04636236  0.07714837 -0.26351928
```

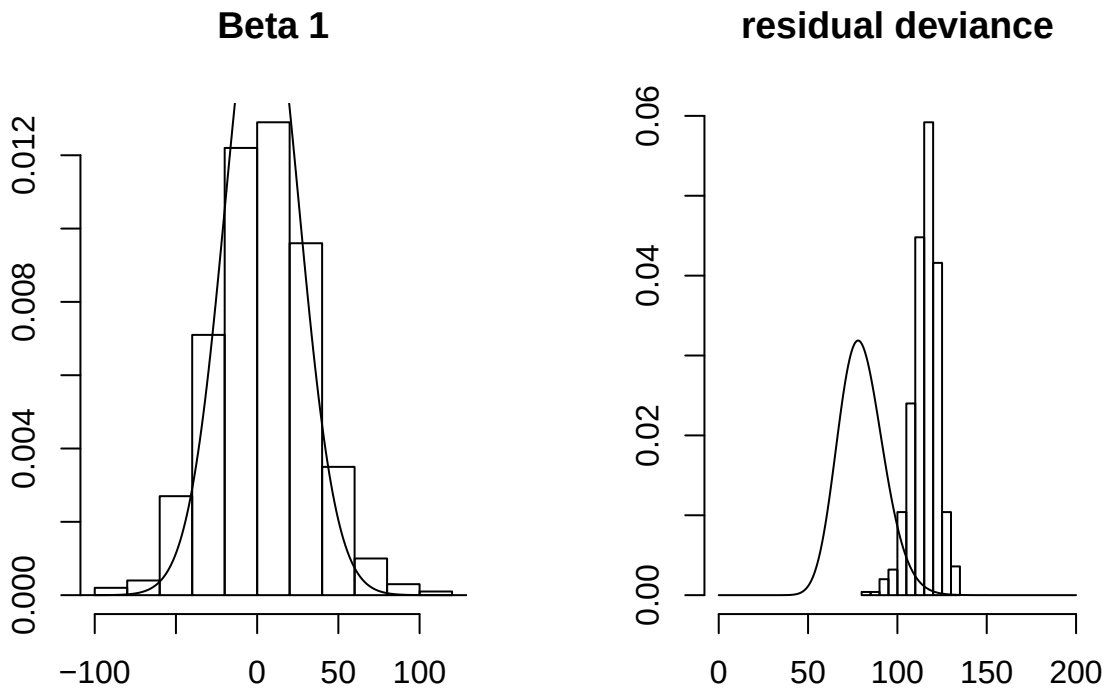
Powyższe doświadczenie dla macierzy X_{400x20}, której elementy są zmiennymi losowymi z rozkładu $N(0, \sigma^2 = \frac{1}{n})$,




```
##  [1] -0.1581696  2.0707055 -1.0810797 -0.2411390 -1.3234152  2.9798691
##  [7]  0.8844539  1.8509061 -1.4253419  0.7636388 -1.6597056 -1.1876808
## [13] -0.1434046 -0.9955314 -1.9149560  0.6967827  0.2381866  0.2938478
## [19] -0.5143020 -3.1524983
```

Histogram rozkładu estymatorów wygląda tak samo jak dla p=3, ale oś x-ów zwiększyła bardzo swój zakres, a oś y zawężyła. Rozkład residual deviance bez zmian. Obciążenia estymatorów parametrów są znacznie większe, przekraczające nawet 3.

Powyższy schemat dla n=100.




```
##  [1]  0.4443214  1.1154478  0.5249883 -0.9139326  0.9628674  2.7412044
##  [7] -0.6767133  2.3953572 -0.2256922  3.4540203  1.8329609 -0.1612431
## [13]  0.4485260  1.6033302  2.2804521  0.9902342  1.6562522  0.3237976
## [19] -0.9460048 -0.1640130
```

Ostatni wykres residual deviance wskazuje, że model jest najlepiej dobrany, ponieważ rozkład słupków i linia gęstości są sobie najbliższe.