Zaawansowane Modele Liniowe - Lista 3

Uogólnienia regresji Poissona

Wstęp

Regresja ujemna dwumianowa

W modelu regresji Poissona zakładaliśmy, że średnie i wariancje poszczególnych obserwacji są sobie równe. W przypadku, gdy założenie to jest złamane i wariancje są większe niż średnie, mówimy o zjawisku **nadmiernej dyspersji**. Ma wtedy sens założenie, że dane pochodzą z rozkładu **ujemnego dwumianowego** z parametrami $\alpha \geq 0, \mu > 0$ i funkcją masy prawdopodobieństwa

$$P(Y_i = y) = \frac{\Gamma(y + \alpha_i^{-1})}{\Gamma(y+1)\Gamma(\alpha_i^{-1})} \left(\frac{\alpha_i^{-1}}{\alpha_i^{-1} + \mu_i}\right)^{\alpha_i^{-1}} \left(\frac{\mu_i}{\alpha_i^{-1} + \mu_i}\right)^{y},$$

który dla małych wartości α_i przybliża rozkład Poissona z parametrem μ_i (zbiega do niego, gdy $\alpha_i \to 0$), równocześnie zachowując **tę samą wartość oczekiwaną** $E[Y_i] = \mu_i$, ale **większą wariancję**

$$Var[Y_i] = \mu_i + \alpha_i \mu_i^2 > \mu_i.$$

W modelu regresji ujemnej dwumianowej zakładamy (tak samo jak w modelu regresji Poissona), że dla każdej z n niezależnych obserwacji

$$log(\mu_i) = X\beta,$$

gdzie X jest $n \times p$ macierzą planu rozszerzoną o wektor jedynek i $\beta \in R^p$. Regresja ujemna dwumianowa z ustalonym α należy do rodziny wykładniczej, więc stosują się do niej wszystkie twierdzenia obowiązujące dla uogólnionych modeli liniowych. Ponadto większość z nich zachodzi również, gdy estymujemy α.

Modele z inflacją

Występowanie nadmiernej względem modelu Poissona liczby zer w zmiennej objaśnianej nazywamy **inflacją w zerze**. Taka sytuacja ma miejsce, gdy w pewnym podzbiorze populacji badane zjawisko po prostu nie występuje (przykładem mogą być osoby niepalące). Do modelowania takich zjawisk użyjemy modelu **ZIPR** (Zero Inflated Poisson Regression). Analogicznie możemy rozważyć użycie modelu **ZINB**, gdy chcemy modelować z użyciem rozkładu ujemnego dwumianowego (gdy występuje też nadmierna dyspersja) ze zjawiskiem inflacji w zerze. W modelach z inflacją zakładamy, że obserwacje są niezależnymi realizacjami zmiennych losowych pochodzących z mieszanki odpowiedniego rozkładu i rozkladu dwupunktowego (zwiększa się liczba szacowanych parametrów, a szacowane są one metodą największej wiarygodności; model logistyczny na początku decyduje do której podpopulacji nalezy obserwacja). Testowanie globalnej hipotezy o tym czy występuje zjawisko inflacji w zerze przebiega z użyciem statystyki Deviance w sposób przypominający ten opisany w zadaniu 1.

Przy inflacji w zerze oraz nadmiernej dyspersji możemy użyć również tzw. **modelu z barierą** (znowu zakładamy istnienie dwóch podpopulacji gdzie w jednej nie występuje badana cecha, ale mogą one mieć rozkład inny niż dwupunktowy).

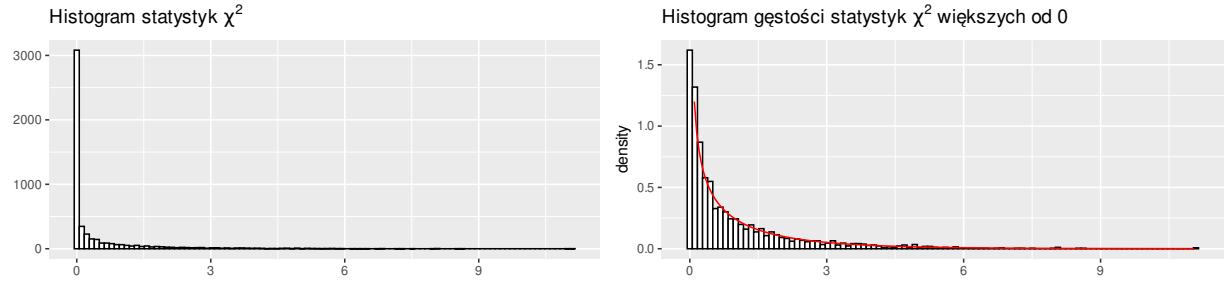
Zadanie 1

W tym zadaniu generujemy 10000-krotnie dane z modelu regresji Poissona i dopasowujemy do nich modele Poissona oraz ujemny-dwumianowy w celu weryfikacji hipotezy

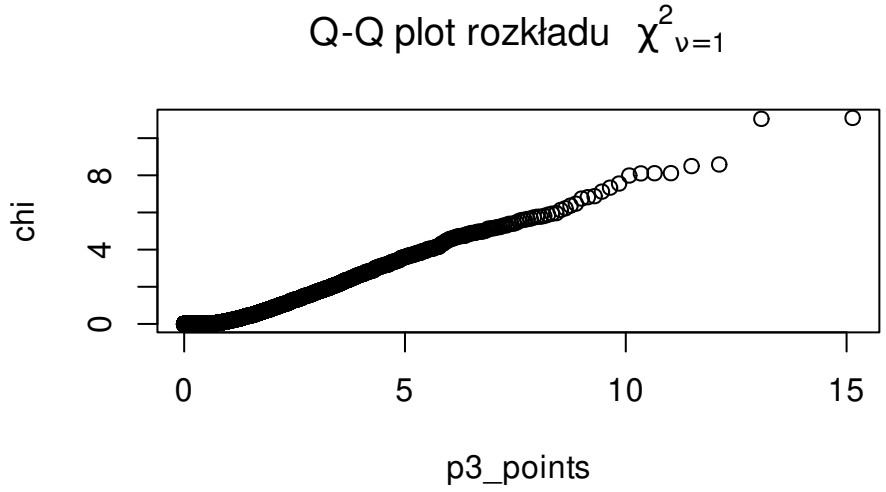
H_0 : Dane pochodzą z modelu Poissona ($\alpha = 0$) vs H_1 : Dane pochodzą z modelu regresji ujemnej dwumianowej ($\alpha > 0$).

Wiadomo, że przy hipotezie zerowej statystyka $\chi^2 = -2(l(M_0) - l(M_1))$ ma asymptotyczny rozkład będący mieszanką rozkładu skoncentrowanego w zerze (50%) oraz χ^2 z 1 stopniem swobody (50%), zatem odrzucimy H_0 na poziomie istotności $q$dla wartosci statystyki χ^2 większych od kwantyla rzędu $1 - 2q$ z rozkładu χ^2_1. Oznacza to również, że mniej więcej w połowie przypadków $\alpha = 0$.

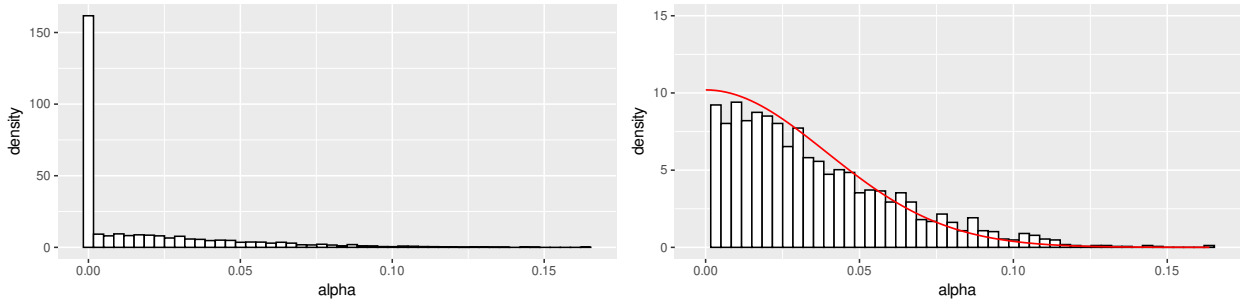
Wykresy



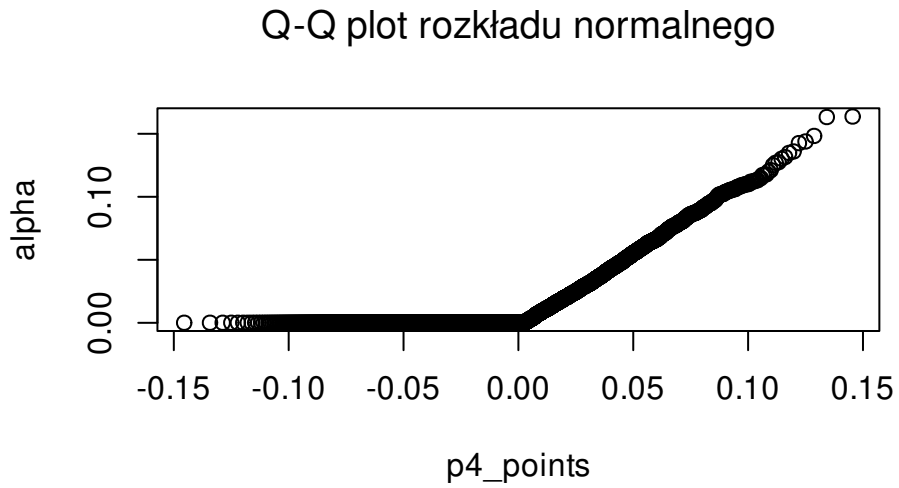
Komentarz: Wykres po lewej przedstawia rozkład wszystkich punktów. Około połowa z nich jest równa 0. Wykres po prawej przedstawia wszystkie niezerowe punkty - zgodnie z teorią rozkład ich przypomina rozkład χ^2_1 oznaczony czerwoną linią. Analiza przedstawionego poniżej wykresu kwantylowo-kwantylowego potwierdza zgodność wyników z teorią.



Wykresy: estymator $\hat{\alpha}$.



Komentarz: W tym przypadku również uzyskujemy wykresy podobne do tych z wykładu - mniej więcej połowa uzyskanych estymatorów α jest równa zero (na tyle blisko 0, że tak je traktujemy), a pozostałe mają rozkład normalny. Ponizej wykres kwantylowo-kwantylowy dla wszystkich wartości $\hat{\alpha}$, włącznie z zerowymi (widać je jako prostą linię na poziomie 0, a pozostałe obserwacje układają się zgodnie z przewidywaniami na prostej.)



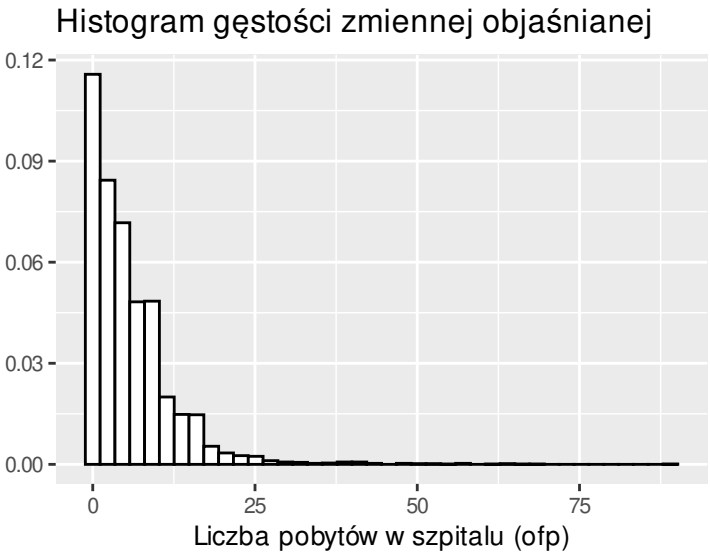
Zadania 2-3

Zajmiemy się analizą danych medycznych ze zbioru Deb i Trivedi (1997), gdzie zmienną objaśnianą będzie liczba pobytów w szpitalu (`ofp`).

Table 1: Pierwsze 3 wiersze danych

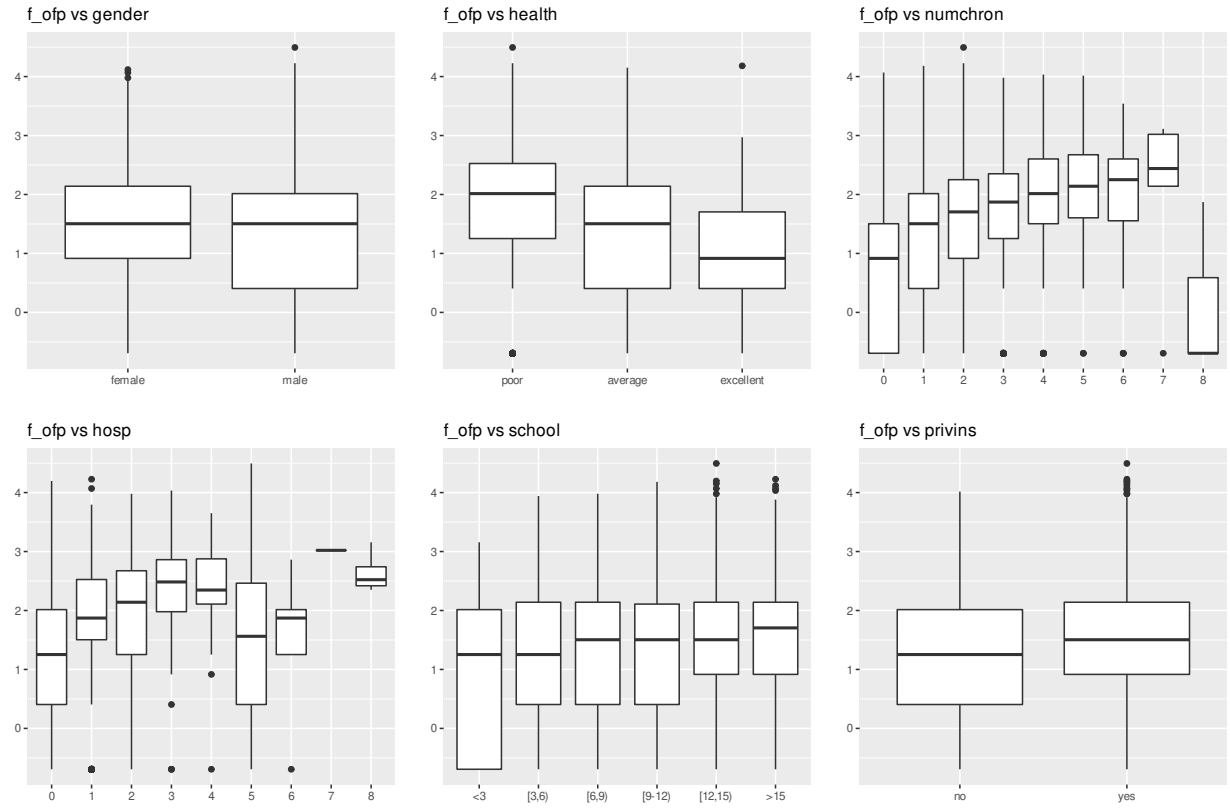
hosp	health	numchron	gender	school	privins	ofp
1	average	2	male	6	yes	5
0	average	2	female	10	yes	1
3	poor	4	female	10	no	13

Wstępna analiza



Komentarz: Wystąpień wartości 0 jest prawie 1200 z 4406 wszystkich obserwacji. Może to wskazywać na zjawisko inflacji w zerze. Ze względu na duża liczbę 0 wprowadzimy pomocniczą zmienną $f(ofp) = log(ofp + 0.5)$ przez ciągłe przekształcenie `ofp`. Następnie porządkujemy zmienne kategoryczne i przygotowujemy wykresy pudełkowe.

Boxploty



	Poiss	nb	ZIPR	ZINBR	Poiss z bar	nb z bar
Beta						
intercept	1.268293e+00	1.171420e+00	1.675342e+00	1.451990e+00	1.675496e+00	1.500942e+00
hosp	1.635207e-01	2.174552e-01	1.585420e-01	2.012500e-01	1.585207e-01	2.117786e-01
health	-2.791316e-01	-3.192941e-01	-2.666990e-01	-2.958589e-01	-2.663483e-01	-3.215577e-01
numchron	1.455485e-01	1.746608e-01	1.013657e-01	1.288651e-01	1.012657e-01	1.263264e-01
gendermale	-1.125886e-01	-1.267754e-01	-6.244150e-02	-8.066570e-02	-6.235060e-02	-6.846430e-02
school	2.621980e-02	2.686900e-02	1.913950e-02	2.138990e-02	1.907410e-02	2.071210e-02
privinsyes	2.039601e-01	2.255360e-01	8.166300e-02	1.271852e-01	8.195950e-02	1.006386e-01
Gamma						
intercept	NA	NA	5.256851e-01	1.908350e+00	-5.095388e-01	-5.095388e-01
hosp	NA	NA	-2.989955e-01	-8.037221e-01	3.074322e-01	3.074322e-01
health	NA	NA	1.236902e-01	5.145070e-02	-1.583978e-01	-1.583978e-01
numchron	NA	NA	-5.264021e-01	-1.246568e+00	5.298137e-01	5.298137e-01
gendermale	NA	NA	4.153458e-01	6.449119e-01	-4.157623e-01	-4.157623e-01
school	NA	NA	-5.717100e-02	-8.505250e-02	5.896270e-02	5.896270e-02
privinsyes	NA	NA	-7.605529e-01	-1.172275e+00	7.558355e-01	7.558355e-01
thety	1.206476e+00	1.482476e+00	1.395428e+00	1.206476e+00	1.482476e+00	1.395428e+00
liczba_param	7.000000e+00	7.000000e+00	1.400000e+01	1.400000e+01	1.400000e+01	1.400000e+01
aic	3.596751e+04	2.435732e+04	3.596751e+04	2.435732e+04	3.596751e+04	2.435732e+04
bic	3.601224e+04	2.440845e+04	3.238889e+04	2.430746e+04	3.239007e+04	2.430385e+04
f_wiaro	-1.797675e+04	-1.217066e+04	-1.613571e+04	-1.209080e+04	-1.613630e+04	-1.208900e+04

Komentarz:

- regresory różnią się wpływem jaki wywierają na zmienną `f_ofp`.
- zmienne (pogrupowana) `school`, `gender` i `privins` wydają się nie mieć większego wpływu na odpowiedź, choć widać pewne różnice w rozrzutach.
- pozostałe zmienne sprawiają wrażenie istotnych, w przypadku niektórych związek jest podobny do liniowego.

Zadanie 4

W tym zadaniu zbudujemy różne modele opisane w raporcie i porównamy ich dopasowanie do danych. Przewidujemy zmienną `ofp`. Dopasujemy podstawowe wersje wszystkich 6 modeli z listy, a następnie porównamy ich wersje ze wszystkimi zmiennymi i bez zmiennych potencjalnie nieistotnych na podstawie wykresów, osobno każdej z: `gender`, `privins` i `school`. Testami opartymi o statystykę Deviance sprawdzimy, czy redukcja była słuszna.

Przeprowadzenie kilku podstawowych testów nie wykazało nieistotnych zmiennych, jedynie wskazało różnice pomiędzy np. modelem Poissona a ZINBR. Wyniki poniżej dotyczą modeli pełnych.