Raport z listy 3 ZML

Erwin Jasic

10 maja 2021

Cel raportu

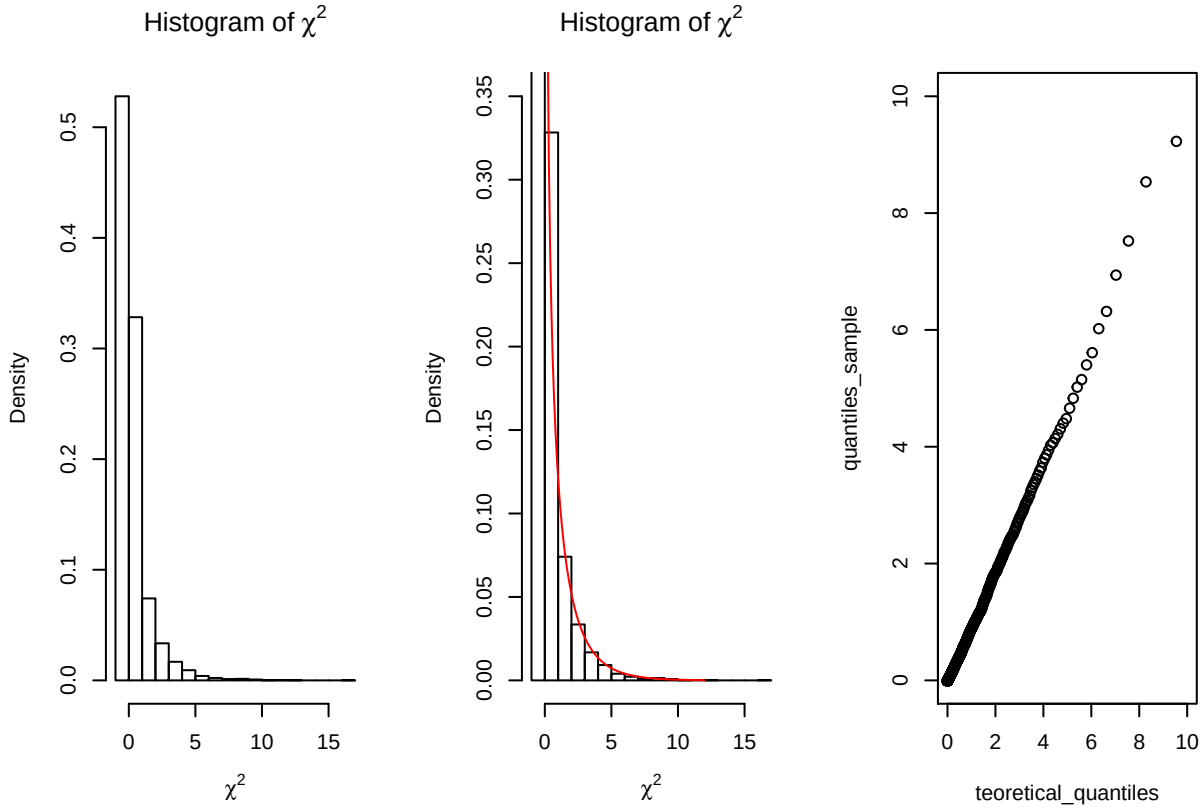
W poniższym raporcie zajmiemy się praktycznym zastosowaniem teorii z wykładu dotyczącego uogólnionej regresji poissona. W pierwszej części raportu zajmiemy się pewną symulacją, w której zobaczymy jak można ze sobą porównywać pewne zjawiska na losowych danych, a w drugiej części zajmiemy się analizą rzeczywistych danych.

Symulacje:

Zadanie 1

W tym zadaniu skupimy się na zachowaniu statystyki χ^2 oraz $\hat{\alpha}$ w przypadku testowania czy dane pochodzą z rozkładu Poissona, przeciwko temu, że dane pochodzą z rozkładu ujemnego dwumianowego. Taki test można przeprowadzić za pomocą statystyki $\chi^2 = D(M_1) - D(M_2)$ lub statystyki $T = \frac{\hat{\alpha}}{\sqrt{Var(\hat{\alpha})}}$.

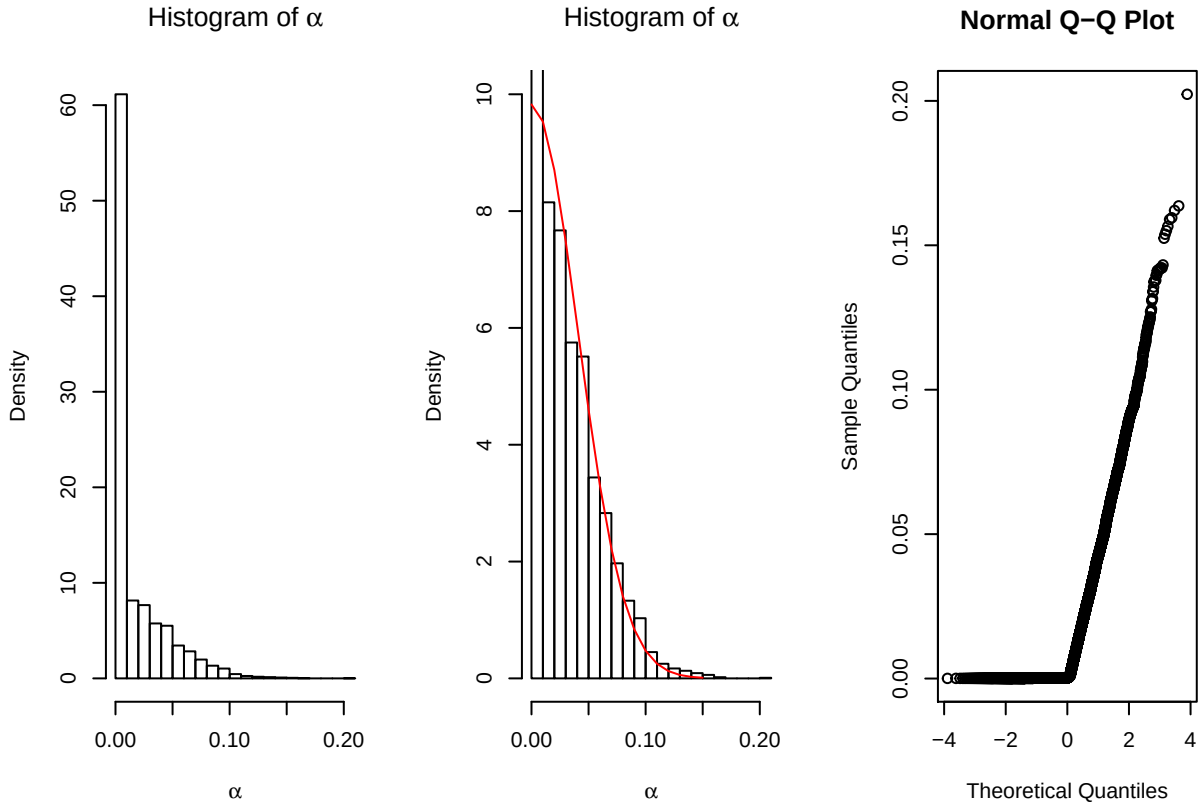
Najpierw wygenerujemy dane zgodnie z treścią zadania, a następnie wykonamy doświadczenie 10000 razy przy założeniu hipotezy zerowej, by otrzymać wiarygodne wyniki. Dla każdej z 10000 replikacji dopasujemy jeden z dwóch modeli i obliczymy ciąg statystyk χ^2 zgodnie ze wzorem podanym w treści zadania.



Na powyższych histogramach widzimy, że około połowy masy wykresu jest skoncentrowana w zerze co jest zgodne z teorią przedstawioną na wykładzie. Tę relacje dokładnie opisuje następujący wzór

$$\chi^2 = 0.5F_0 + 0.5F_1,$$

gdzie F_0 jest dystrybuantą zmiennej losowej skoncentrowanej w punkcie 0 ($P(X = 0) = 1$), a F_1 jest dystrybuantą zmiennej losowej z rozkładu χ^2 z 1 stopniem swobody. Na drugim histogramie widzimy dobre dopasowanie przeskalowanej gęstości χ^2 dla danych większych od zera. Na ostatnim wykresie zauważamy dobre dopasowanie kwantyli z próby do kwantyli z asymptotycznego rozkładu.



Podobnie jak dla χ^2 widzimy dominujący słupek w wartości róWnej 0 na osi x. Możemy zauważyć, że dane na drugim wykresie opadają zgodnie z gęstością rozkładu normalnego, zatem tutaj również zgadza się to z teorią z wykładu. Ostatni wykres przedstawia wykres kwantylowo-kwantylowy, gdzie około połowa obserwacji ma wartości bliskie 0, a druga połowa ma rozkład zgodny z rozkładem normalnym.

Analiza danych

Zadanie 2

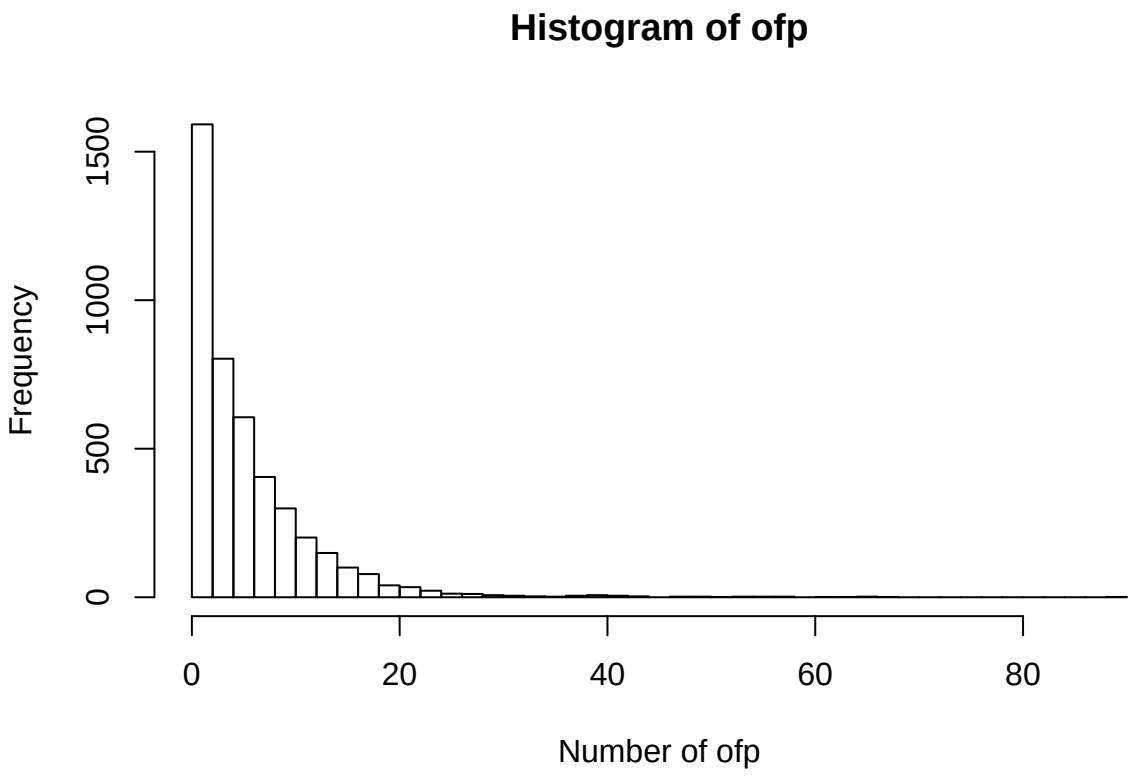
To zadanie polega tylko na wczytaniu danych 'DebTrivedi.csv', które będziemy analizować powyżej. Za zmienną zależną przyjmujemy kolumnę "ofp", która oznacza liczbę wizyt w gabinecie lekarskim pacjenta.

Głównie będziemy się interesować jak poniższe zmienne niezależne wpływają na zmienną zależną:

- "hosp" – liczba pobytów w szpitalu,
- "health" – zmienna opisująca subiektywne 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

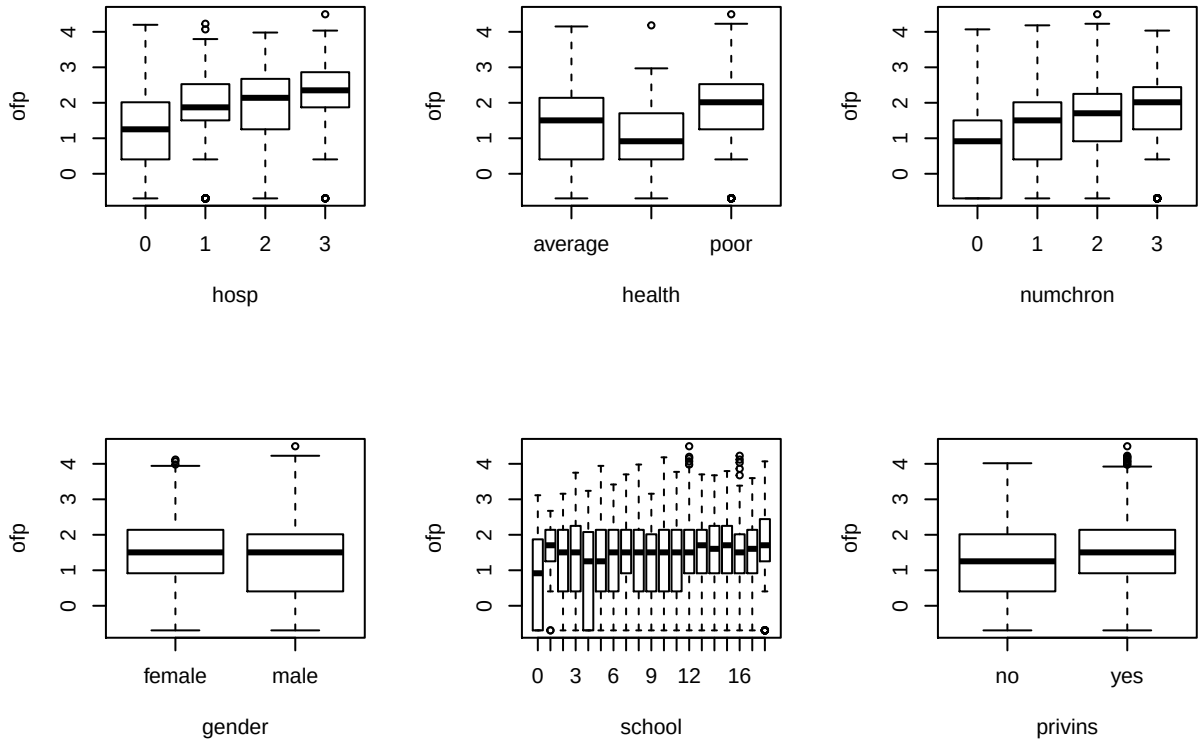
W poniższym zadaniu wykonamy wstępną analizę naszych danych. Najpierw popatrzmy na rozkład zmiennej zależnej.



Histogram wskazuje na to, że występuje zjawisko inflacji w zerze. Z histogramu nie potrfimy odczytać czy występuje zjawisko nadmiernej dyspersji.

Aby poradzić sobie z tym zjawiskiem wprowadzimy zmienną pomocniczą zgodnie ze wskazówką z treści drugiej kropki w zadaniu 3.

Teraz narysujmy boxploty między zmienną zależną, a każdą ze zmiennej niezależnej wymienionej w zadaniu 2. (Uwaga: Na pierwszym i trzecim wykresie w pierwszym wierszu liczba 3 na osi x oznacza, że zjawisko wystąpiło 3 lub więcej razy.)



Na podstawie powyższych wykresów możemy stwierdzić, że im większa jest liczba pobytów w szpitalu (“hosp”), tym więcej razy taki pacjent odwiedza gabinet lekarski (“ofp”). Sytuacja analogiczna ma miejsce, gdy badamy związek między liczbą przewlekłych stanów chorobowych (“numchron”), a zmienną zależną (“ofp”). Z drugiej strony wydaje się, że żaden z boxplotów z drugiego wierszu nie wykazuje wyraźnej zależności.

Zadanie 4

W tym zadaniu dopasujemy do danych sześć różnych modeli:

- Model Poissona – glm(),
- Model ujemny dwumianowy – glm.nb(),
- Model ZIPR – zeroinfl(),
- Model ZINBR – zeroinfl(),
- Model Poissona z barierą – hurdle(),
- Model ujemny dwumianowy z barierą – hurdle()

oraz porównamy je ze sobą pod różnymi względami. Zapiszemy wyniki dla poszczególnych parametrów w tabeli dla większej przejrzystości. Uzupełnimy tabelę zgodnie z treścią zadania. Przydatne funkcje z których będziemy korzystać w celu otrzymania wyników to np. *summary*() z której odczytamy, które zmienne niezależne są istotne w modelu. Przydatnymi funkcjami będą również $AIC()$, $BIC()$ i $logLike()$.

	Poisson	NB.Poisson	ZIPR	ZINBR	Hurdle	Hurdle Binomial
liczba param. w modelu	8.00000	8.0000	6.000	5.0000	6.00	6.00
logLike	-17971.61281	-12170.5536	-16135.244	-12094.2541	-16136.44	-12090.07
AIC	35959.22562	24359.1072	32298.487	24216.5083	32300.88	24210.14
BIC	36010.35140	24416.6237	32387.957	24305.9784	32390.35	24306.00
oczekiwana liczba zer	46.71402	608.0085	682.298	707.2184	683.00	683.00

Na podstawie powyższej tabeli możemy stwierdzić na podstawie statystyk AIC oraz BIC, że modele: ujemny dwumianowy, ZINBR oraz ujemny dwumianowy z barierą najlepiej dopasowują się do danych (mają najmniejsze wartości dla statystyk AIC i BIC). Wszystkie modele poza zwykłą regresją Poissona oczekują dużej ilości zer, co może wskazywać, że regresja Poissona jest słabą regresją dla tych danych (wstępna analiza danych (histogram) z zadania 3 pokazuje, że oczekujemy raczej dużej liczby zer).

Podsumowanie

Dzięki wykonaniu poleceń dotyczących uogólnionej regresji Poissona dowiedzieliśmy się jak wykorzystywać uzyskaną teoretyczną wiedzę z wykładu w praktycznych zadaniach takich jak symulacje czy rzeczywista analiza danych. Zobaczyliśmy, że mamy do dyspozycji wiele narzędzi do analizy i wyciągania odpowiednich wniosków na podstawie wykresów, czy tabeli. Przetestowaliśmy, że wzory z wykładu działają w praktyce (w szczególności zadanie 1 i wzór na χ^2).