Zaawansowane modele liniowe - raport 3

Łukasz Rębisz

2023-06-12

Cel raportu

Celem raportu jest analiza **uogólnionych modeli regresji Poissona**. Za pomocą symulacji zbadamy model regresji ujemnej dwumianowej (model obejmujący klasyczny model Poissona jak również model ze zjawiskiem nadmiernej dyspersji) pod względem dopasowania rozkładów statystyk testowych do wartości teoretycznych.

Następnie, analizując dane dotyczące liczby wizyt w gabinecie lekarskim w zależności od kilku czynników, porównamy ze sobą następujące modele:

- model Poissona,
- model ujemny dwumianowy,
- model ZIPR - model regresji Poissona z inflacją w zerze,
- model ZINBR - model ZIPR oparty o rozkład ujemny dwumianowy,
- model Poissona z barierą,
- model ujemny dwumianowy z barierą.

Symulacje

Rozważmy problem testowania:

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

przeciwko

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

Zbadanie powyższej hipotezy można przeprowadzić za pomocą statystyk:

i)
$$\chi^2 = D(M_1) - D(M_2) = -2\left(l_1(\hat{\beta}^{(1)}) - l_2(\hat{\beta}^{(2)})\right),$$
gdzie l_i jest logarytmem funkcji wiarogodności dla modelu $i = 1, 2$ obliczonym w punkcie odpowiadającym estymatorowi $\hat{\beta}^{(i)}$;

ii)
$$T = \frac{\hat{\alpha}}{Var(\hat{\alpha})}.$$

Wygenerujmy losową macierz $X \in \mathbb{M}_{1000\times 2}$, taką że $X_{ij} \sim^{i.i.d.} N(0, \sigma = 1/\sqrt{1000})$. Następnie wyznaczmy ciąg predyktorów liniowych $\eta = X\beta$ dla wektora $\beta = (3, 3)$ i na ich podstawie wygenerujmy 10000 niezależenych replikacji wektora odpowiedzi Y przy założeniu hipotezy zerowej H_0.

Dla każdej replikacji wektora odpowiedzi Y dopasujmy do otrzymanych danych **model regresji ujemnej dwumianowej** oraz **model regresji Poissona**. Na podstawie otrzymanych modeli obliczmy wartości statystyk χ^2 oraz T zgodnie z powyższymi wzorami.

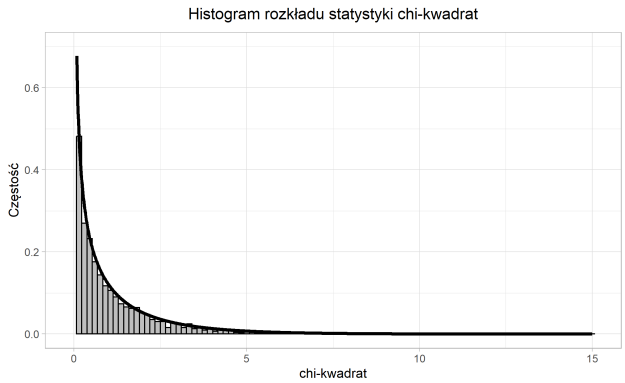
Rozkład statystyki *chi-kwadrat*

Teoretyczny rozkład statystyki χ^2:

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

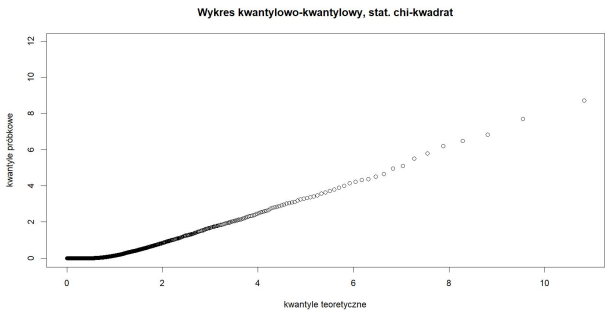
gdzie F_0 jest dystrybuantą zmiennej losowej skoncentrowanej w punkcie 0, a F_1 jest dystrybuantą zmiennej losowej z rozkładu chi-kwadrat z 1 stopniem swobody.

Porównajmy empiryczny rozkład statystyki χ^2 z powyższym rozkładem teoretycznym.



Rysunek 1: histogram rozkładu statystyki *chi-kwadrat* z zaznaczoną krzywą rozkładu teoretycznego

Analiza powyższego wykresu wskazuje na dobre dopasowanie rozkładu empirycznego statystyki χ^2 do krzywej rozkładu teoretycznego dla danych większych od zera oraz dominację piku w zerze. Wynik jest zgodny z oczekiwaniami, ponieważ w przypadku prawdziwości hipotezy zerowej dane pochodzą z rozkładu Poissona. Wówczas statystyka χ^2 ma połowę swej masy skoncetrowaną w zerze, druga połowa podlega rozkładowi χ^2 z 1 stopniem swobody.



Rysunek 2: wykres kwantylowo-kwantylowy dla statystyki *chi-kwadrat*

Analiza powyższego wykresu kwantylowo-kwantylowego wskazuje na dość dobre (zwłaszcza dla wartości większych od zera) dopasowanie kwantyli empirycznych do kwantyli z asymptotycznego rozkładu.

Rozkład statystki *alfa*

Wiemy, że dla $\alpha > 0$ statystyka α ma asymptotycznie rozkład normalny:

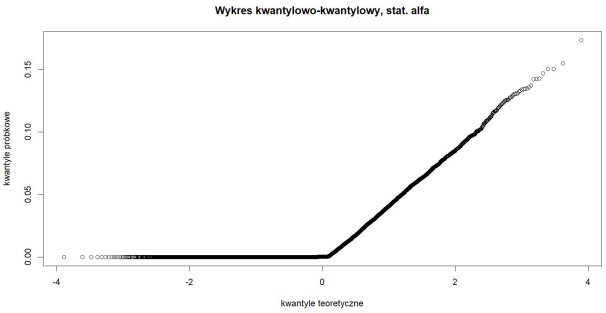
$$\alpha \sim N(0, Var(\alpha)).$$

Dla $\alpha = 0$ (brzeg dziedziny parametru; rozkład Poissona) rozkład statystyki α dla prostej dodatniej pozostaje bez zmian, ale masa probabilistyczna rozrzucona po ujemnej półprostej zostaje skoncentrowana w zerze.

Porównajmy empiryczny rozkład statystyki α z powyższym rozkładem teoretycznym.

Rysunek 3: histogram rozkładu statystyki *alfa* z zaznaczoną krzywą rozkładu teoretycznego

Powyższy histogram rozkładu statystyki α wskazuje na dominację wartości bliskich zeru oraz opadanie danych wraz ze wzrostem wartości α odpowiadające zachowaniu się gęstości rozkładu normalnego.

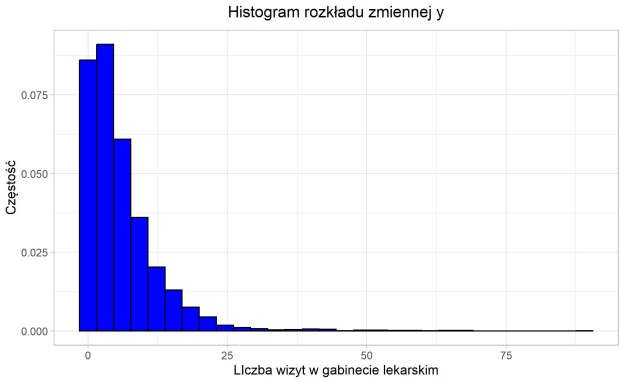


Rysunek 4: wykres kwantylowo-kwantylowy dla statystyki *alfa*

Analiza powyższego wykresu kwantylowo-kwantylowego pokazuje, że około połowa obserwacji (prawa część wykresu) układa się wzdłuż prostej. Wskazuje to na bliskość tej części danych (dla $\alpha > 0$) do rozkładu normalnego. Natomiast druga połowa oberwacji (lewa, ujemna część wykresu) koncetruje się wokół wartości bliskich zeru.

Analiza danych

W poniższej analizie zbadamy związek pomiędzy zmienną $y :=$ *liczbie wizyt w gabinecie lekarskim* a zmiennymi niezależnymi opisującymi pacjenta: *liczbą hospitalizacji*, *subiektywnym stanem zdrowia*, *liczbą przewlekłych stanów chorobowych*, *płcią*, *liczbą lat edukacji*, oraz posiadaniem *prywatnego ubezpieczenia zdrowotnego*.



Rysunek 5: Histogram zmiennej y

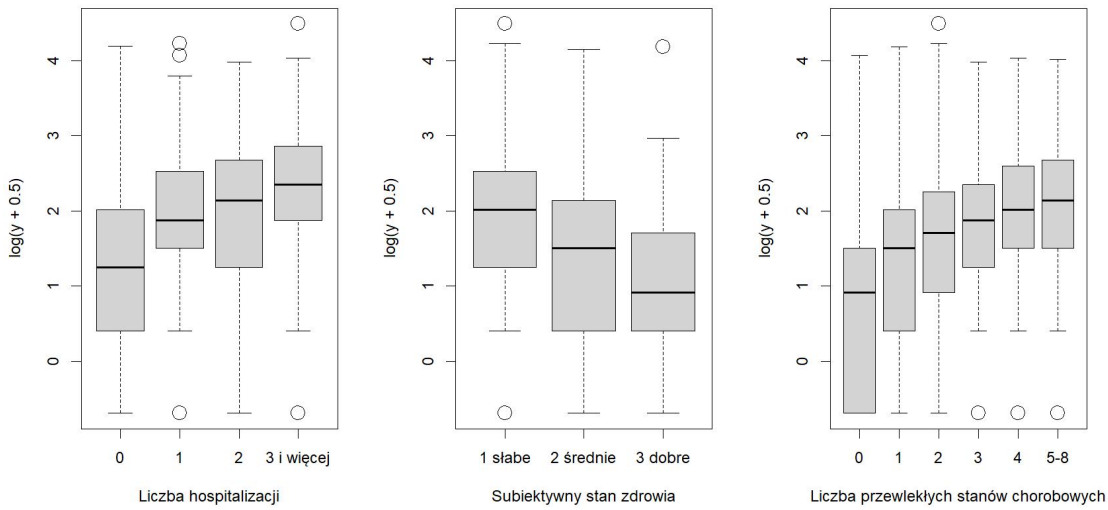
Histogram dla zmiennej objaśnianej y wskazuje na obecność zjawiska nadmiernej **inflacji** w zerze (duża liczba wartości bliskich 0) oraz nadmiernej **dyspersji** (ciężki ogon rozkładu).

Ze względu na znaczącą liczbę zer wprowadźmy zmienną pomocniczą

$$f(y) = \log(y + 0.5),$$

dzięki której łatwiej będzie zbadać zależność pomiędzy zmienną y i regresorami (funkcja f jest monotoniczna - zachowuje uporządkowanie pomiędzy punktami).

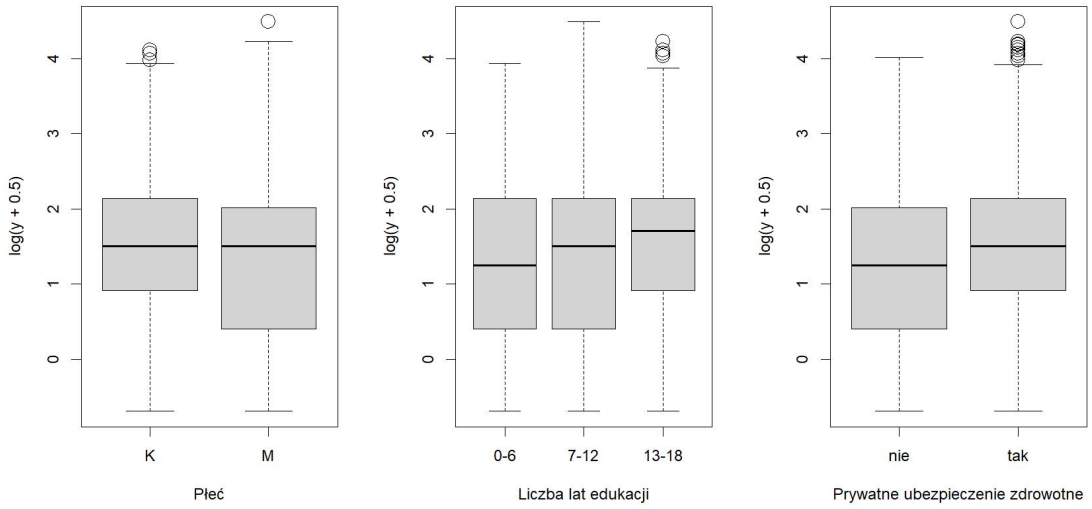
Narysujmy dla każdego regresora osobno wykresy pudełkowe dla zmiennej $f(y)$ w rozbiciu ze względu na przyjmowane wartości przez dany regresor.



Rysunek 6: wykresy pudełkowe zmiennej $f(y) = \log(y + 0.5)$ w zależności od zmiennych: *liczba hospitalizacji*, *subiektywny stan zdrowia* oraz *liczba przewlekłych stanów chorobowych*

Analiza powyższych wykresów pozwala wysunąć następujące wnioski:

- Zmienna objaśniana zależy od *liczby hospitalizacji*. Obserwujemy wzrost wartości zmiennej $\log(y + 0.5)$ wraz ze wzrostem *liczby hospitalizacji*. Szczególnie zjawisko to jest wyraźne, gdy porównujemy mediany dla danych grup.
- Zmienna *subiektywny stan zdrowia* również wpływa na wartość zmiennej wynikowej: im lepsza ocena *stanu zdrowia* pacjenta (w opinii samego pacjenta), tym mniejsza jest wartość zmiennej $\log(y + 0.5)$.
- W przypadku *liczby przewlekłych stanów chorobowych* obserwujemy (zwłaszcza porównując ze sobą mediany dla danych grup) średni wzrost wartości zmiennej wynikowej wraz ze wzrostem badanego regresora. W szczególności pacjenci, którzy nigdy nie doświadczyli *przewlekłego stanu chorobowego* wykazują się bardzo małą wartością zmiennej wynikowej (mediana bliska 1, duża część obserwacji bliska zera).



Rysunek 7: wykresy pudełkowe zmiennej $f(y) = \log(y + 0.5)$ w zależności od zmiennych: *płeć*, *liczba lat edukacji* oraz posiadanie *prywatnego ubezpieczenia zdrowotnego*

Powyższe wykresy pokazują, że:

- *Płeć* nie ma wyraźnego wpływu na zmienną wynikową. Zarówno grupa kobiet jak i mężczyzn charakteryzuje się medianą bliską 1.5. Wyniki dla grupy kobiet są jednakże bardziej zbliżone do siebie: dominująca część obserwacji koncentruje się wokół wartości mediany.
- Zmienna *liczba lat edukacji* również nie wpływa znacząco na wartość zmiennej wynikowej: wszystkie grupy charakteryzują się podobnymi wartościami mediany.
- Posiadanie przez pacjentów *prywatnego ubezpieczenia zdrowotnego* ma wpływ na zmienną wynikową. Dla grupy pacjentów z *prywatym ubezpieczeniem zdrowotnym* mediana wartości zmiennej wynikowej jest większa.

Modele liniowe

Dopasujmy do badanych danych następujące modele: Poissona, ujemny dwumianowy, ZIPR, ZINBR, modele Poissona oraz ujemny dwumianowy z barierami bez wyrazu wolnego (*interceptu*).

Spośród badanych modeli wszystkie zmienne są istotne (T testy) dla modelu Poissona, ujemnego dwumianowego, modelu ZIPR oraz modeli z barierą (Poissona oraz ujemnego dwumianowego).

Jedynie w przypadku modelu ZINBR obserwujemy nieistotność zmiennych *liczba hospitalizacji* oraz *płeć* w części modelu odpowiadającej zjawisku inflacji w zerze. Porównajmy model bez tych zmiennych w tej części z pełnym modelem ZINBR, wykonując test Deviance.

Otrzymana statystyka $\chi^2 \approx 7.26 > 6 = (\chi^2)_2^{-1}(1 - 0.05)$, a zatem odrzucamy hipotezę zerową mówiącą o tym, że model bez powyższych zmiennych jest lepszy od pełnego modelu.

Wniosek: dalszej analizie poddamy pełne modele.

Tabela 1: wyniki dla poszczególnych modeli

	m. Poissona	m. NB	m. ZIPR	m. ZINBR	m. Poissona z barierą	m. NB z barierą
Estymator β_1	0.20	0.27	0.20	0.26	0.20	0.27
Estymator β_2	0.15	0.14	0.26	0.21	0.26	0.21
Estymator β_3	0.23	0.27	0.21	0.25	0.21	0.24
Estymator β_4	0.23	0.27	0.23	0.26	0.23	0.25
Estymator β_5	0.05	0.05	0.05	0.04	0.05	0.05
Estymator β_6	0.34	0.30	0.25	0.27	0.26	0.21
Estymator γ_1	–	–	−0.29	−10.55	0.30	0.30
Estymator γ_2	–	–	0.13	0.33	−0.19	−0.19
Estymator γ_3	–	–	−0.52	−2.39	0.52	0.52
Estymator γ_4	–	–	−0.39	−0.56	0.41	0.41
Estymator γ_5	–	–	−0.05	−0.14	0.06	0.06
Estymator γ_6	–	–	−0.75	−1.06	0.75	0.75
Estymator α	–	0.9	–	0.83	–	0.97
Liczba parametrów	6	7	12	13	12	13
log f. wiarogodności	−18710.56	−12301.66	−17290.25	−12268.48	−17289.45	−12256.26
AIC	37433.11	24617.31	34604.49	24562.95	34602.89	24538.52
BIC	37471.46	24662.05	34681.18	24646.03	34679.58	24621.6
$E[f(0)]$	88.37	688.57	688.44	744.92	681.66	681.66

Analiza powyższych wyników pokazuje, że:

- Wartości estymatorów β_i są dla danego $i = 1, ..., 6$ bliskie sobie bez względu na zastosowany model.
- Wartości estymatorów γ_i zasadniczo różnią się pomiędzy danymi modelami. Wartości te pochodzą bowiem z różnych konstrukcji spowodowanych występowaniem bariery w zerze (wartości praktycznie równe) lub też zjawiskiem inflacji w zerze (model ZIPR i ZINBR): wartości różne.
- Parametr α odpowiadający za modelowanie zjawiska nadmiernej dyspersji (poprzez rozkład ujemny dwumianowy) przyjmuje podobne wartości dla modeli, w którym został zastosowany.
- Najmniejsze wartości AIC oraz BIC otrzymaliśmy dla **modelu ujemnego dwumianowego z barierią**. Z tego powodu uznajemy ten model za **optymalny** dla badanych danych. Również modele **NB** i **ZINBR** charakteryzują się małymi wartościami AIC i BIC. Wszystkie trzy wymienione modele obejmują zjawisko nadmiernej dyspersji (modelowanej przez rozkład ujemny dwumianowy). Wskazuje to na istotność tego zjawiska w analizowanych danych.
- Oczekiwana liczba zer generowanych przez model (suma funkcji rozkładu prawdopodobieństwa obliczona w zerze dla wszystkich obserwacji) poza małą wartością dla rozkładu Poissona oraz największą wartością dla modelu ZINBR charakteryzuje się podobnym wynikiem równym ok. 680-690 zer dla pozostałych modeli. Prawdziwa liczba zer dla danych wynosi 683.