Raport 3 - Zaawansowane modele liniowe

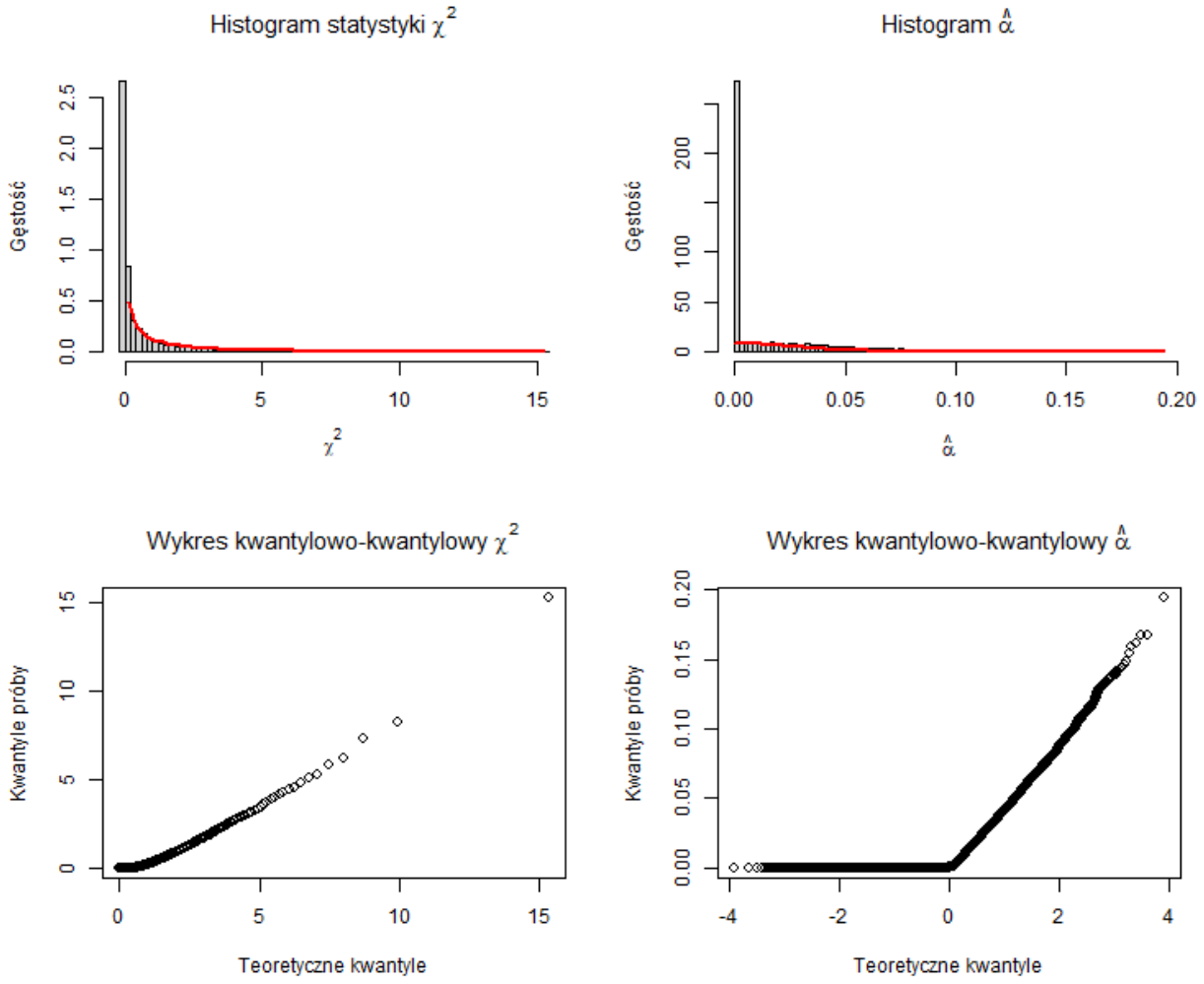
Magdalena Potok

2024-06-04

Celem niniejszego raportu jest zbadanie uogólnionych regresji Poissona, czyli modeli, które radzą sobie ze zjawiskami nadmiernej dyspersji oraz inflacji w zerze. Pierwsza część raportu to symulacje, w której przyjrzymy się zachowaniom poszczególnych parametrów i ich estymatorów w sytuacji, gdy dobrane modele nie przewidują tych zjawisk. W części drugiej raportu najpierw została przeprowdzona analiza danych, a następnie, na podstawie analizy statystycznej, został wybrany model najlepiej opisujący dane.

Symulacje

Wygenerowana została macierz $\mathbb{X} \in \mathbb{M}_{1000x2}$, taka, że $X_{ij} \sim N(0, \sigma = \frac{1}{\sqrt{1000}})$ i.i.d. Następnie wygenerowany został ciąg predyktorów liniowych $\eta = X\beta$, dla wektora $\beta = (3, 3)$ i 10000 niezależnych replikacji wektora odpowiedzi y. Na tej podstawie wyznaczyliśmy statystykę χ^2 oraz $\hat{\alpha}$. Na poniższych wykresach widzimy rozkłady statytstyki χ^2 oraz $\hat{\alpha}$ z dopasowaną gęstością teoretyczną (czerwona krzywa). A pod nimi znajdują się ich wykresy kwantylo-kwantylowe.



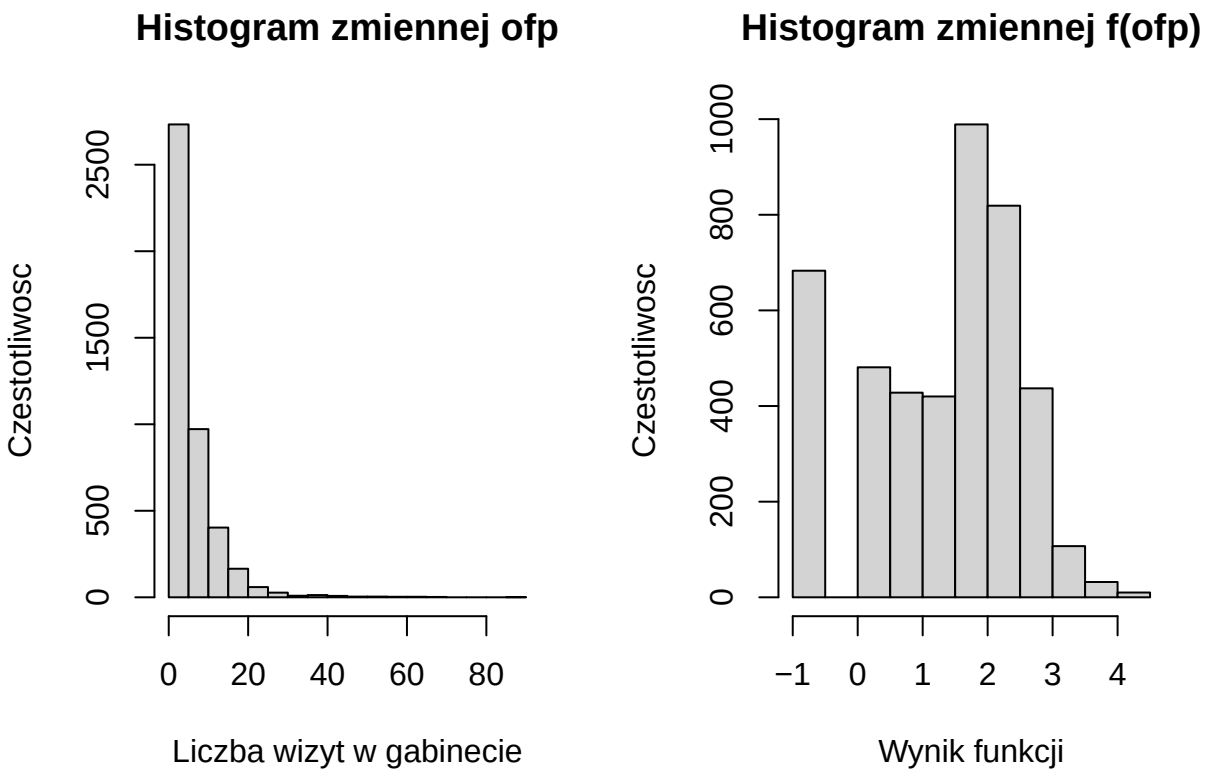
Rysunek 1. Histogramy oraz wykresy kwantylo-kwantylowe dla $\hat{\alpha}$ i χ^2.

Z histogramów możemy odczytać, że empiryczny rozkład obu statystyk jest bardzo zbliżony do teoretycznego - wyjątek stanowi słupek dla wartości bliskich 0. Koncentracja masy rozkładu w tych okolicach sugeruje występowanie zjawiska nadmiernej dyspersji.
Z wykresów kwantylo-kwantylowych możemy zawyżyć, że $\hat{\alpha}$ po prawej części wykresu układa się wzdłuż prostej, co oznacza, że dane są biskie rozkładowi normalnemu, jednak lewa strona koncentruje się wokół 0. Wykres χ^2 sugeruje dobre dopasowanie kwantyli generowanej próby do asymptotycznego rozkładu.

Analiza danych

W drugiej części raportu będziemy zajmować się danymi z pliku "DebTrivedi", które zawierają informacje dotyczące 4406 osób w wieku wyższym lub równym 66, które są objęte publicznym pogramem ubezpieczeniowym. Celem jest zmodelowanie zapotrzebowania na opiekę mdeyczną.
Będziemy badać związek pomiędzy liczbą wizyt w gabinecie lekarskim (zmienna zalezna "ofp") a zmiennymi niezależnym opisującymi pacjenta:
- "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.

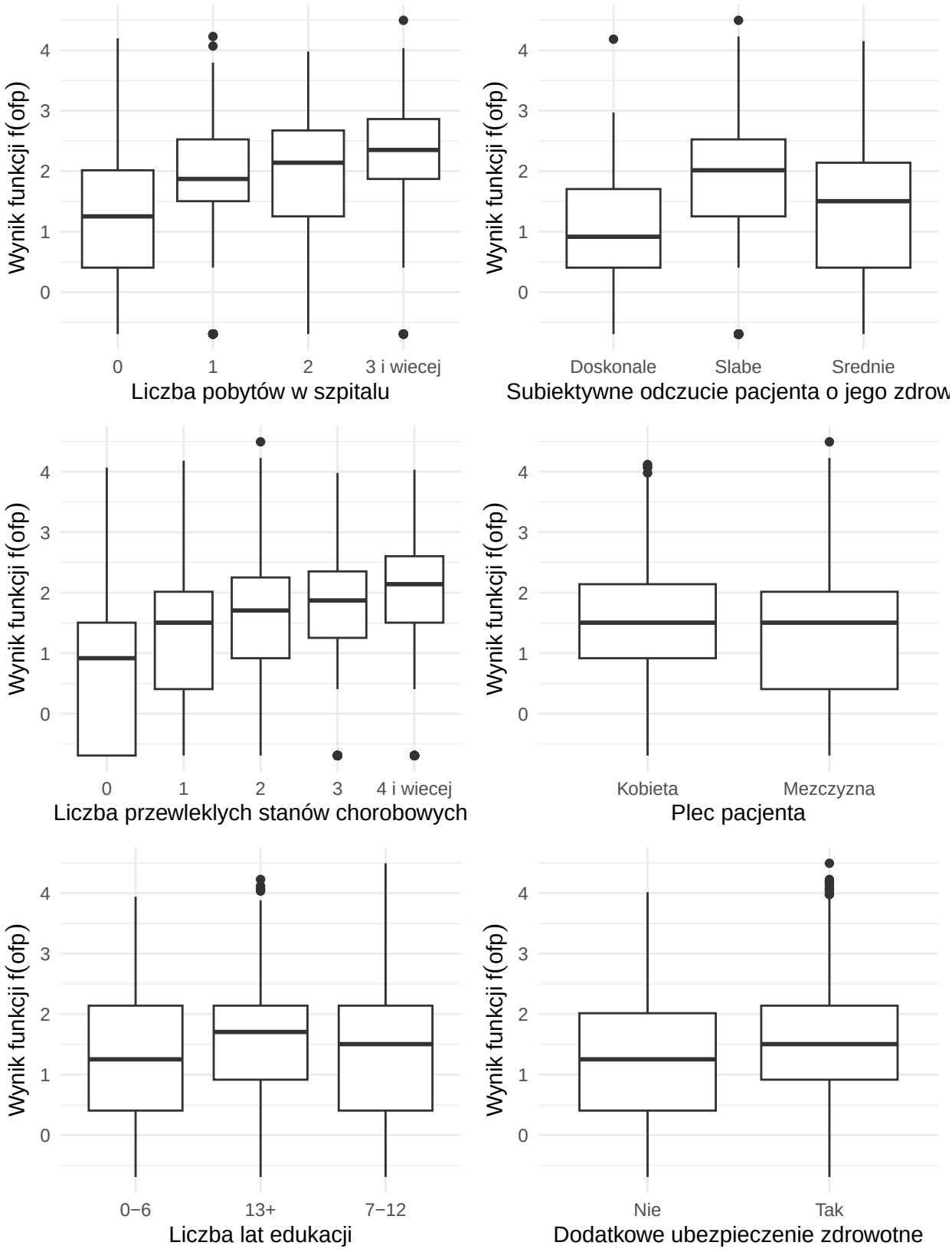
Przejdziemy teraz do wstępnej analizy danych.



Rysunek 2. Histogram liczby wizyt w gabinecie oraz histogram przekształcenia liczby wizyt przez funckję f.

Na histograme po lewej stronie można dostrzec, że dominującym słupkiem jest ten, dla wartości bliskich zeru − świadczy to o obecności zdarzenia nadmiernej reprezentacji wartości 0, czyli tkzw. inflacji w zerze. Wartość wariancji badanej zmiennej objaśniającej wynosi 45.687, natomiast średnia wynosi 5.774. Te wartości są znacząco różne, co oznacza, że zachodzi również zjawisko nadmiernej dyspersji. Aby poradzić sobie z tym problemem wprowadzimy zmienną pomocniczą $f(ofp) = log(ofp + 0.5)$. Histogram nowej zmiennej przedstawiony jest po prawej stronie. Tym razem wyliczone wartości wynoszą: wariancja = 1.257 oraz średnia =1.313 − widać, że te wartości są do siebie zbliżone, a histrogram nie jest skupiony w 0.

Dla każdego regresora osobno zostały narysowane boxploty dla zmiennej objaśniającej $f(ofp)$.



Rysunek 3. Wykresy pudełkowe dla każdego regresora osobno.

Z wykresów pudełkowych możemy wyczytać następujące wnioski:

- Osoby, które nie były nigdy w szpitalu mają najniższą medianę liczby wizyt w gabinecie, jest to grupa osób, która najrzadziej odwiedza gabinet. Możemy zauważyć, że im więcej wizyt w szpitalu, tym mediana jest wyższa.
- Mediana liczby wizyt w gabinecie lekarskim jest najniższa dla grupy osób, które uznały swój stan zdrowia jako doskonały. Największy rozstrzał odpowiedzi jest dla grupy osób, które uznały swój stan zdrowia jako średni. Grupa, która uważa, że słabo się czuje ma najwyższą medianę wizyt w gabinecie.
- Im więcej przewlekłych chorób u pacjenta, tym częstsze wizyty w gabinecie lekarskim. Największy rozstrzał odpowiedzi jest wśród osób, które nie mają żadnej choroby.
- Mediana liczby wizyt lekarskich jest niezmienna ze względu na płeć, jednak u mężczyzn jest większy rozstęp międzykwartylowy.
- Liczba lat edukacji nie ma wpływu na liczbę wizyt, wykresy wyglądają bardzo podobnie.
- Możemy zauważyć, że dodatkowe ubezpieczenie też nie ma znaczącego wpływu, minimalnie wyższa mediana liczby wizyt dla osób, które takowe posiadają.

Do danych dopasowane zostało 6 modeli, następnie zostały zredukowane o niepotrzebne zmienne. W modelach Poissona oraz w ujemnym dwumianowym nie została usunięta żadna zmienna. Modele ZIPR, ZINBR oraz model Possiona z barierą został zredukowane w parametrach γ o intercept oraz zmienną "health". Z kolei model ZINBR w parametrach β został zredukowany o zmienne "gender" oraz "privins", a w parametrach γ o intercept oraz "health". Powyższe zmienne zostały usunięte na podstawie p-wartości w podsumowaniach modeli. Aby sprawdzić poprawność usuniętych zmiennych porównamy wartości kryteriów informacyjnych zredukowanych modeli oraz pełnych.

Tabela 1. Wartości kryteriów informacyjnych dla modeli pełnych i zredukowanych.

	ZIPR	ZINBR	Poisson z barierą	ujemny dwumianowy z barierą
AIC pełny	32300.06	24215.29	32300.90	24210.16
AIC zred.	32298.49	24211.44	32300.88	24215.57
BIC pełny	32402.31	24323.93	32403.15	24318.80
BIC zred.	32387.96	24307.30	32390.35	24298.65

Z tabeli możemy odczytać, że wartości kryterium BIC są niższe zawsze dla zredukowanego modelu. W przypadku kryterium AIC wszystkie modele zredukowane oprócz modelu ujemnego dwumianowego z barierą mają niższe wartości. Jednak w tym przypadku ta różnica jest niewielka, więc do dalszej analizy wybierzemy model zredukowany.

Analiza tabeli 2. (następna strona) wskazuje na optymalne dopasowanie do danych przez zredukowany model ujemny dwumianowy z barierą. Kryteria AIC i BIC przyjmują najniższe wartości dla tego modelu, co potwierdza jego adekwatność – model ZINBR przyjmuje bardzo zbliżone wartości, co wskazuje na również dobre dopasowanie do danych. Te wnioski zgadzają się ze wstępnymi obserwacjami, które sugerowały obecność nadmiernej dyspersji oraz inflacji w zerze na podstawie histogramu liczby wizyt w gabinecie lekarskim na rysunku 2. Również dla modelu ZINBR oraz modelu ujemniego dwumianowego z barierą logarytm funkcji wiarogodności ma najniższą wartość, co wskazuje na lepsze dopasowanie do danych. Jeżeli chodzi o oczekiwaną liczbę zer, to w danych zmienna ofp przyjmuje 683 zera, dokładnie taka wartość wyszła dla modelu ujemnego dwumianowego z barierą oraz modelu Poissona z barierą.
Z drugiej strony, model Poissona wykazuje najgorsze dopasowanie według analizy tabeli. Wartości kryteriów AIC i BIC dla tego modelu są najwyższe i niekorzystne. Ten wynik nie zaskakuje, biorąc pod uwagę obserwacje dotyczące nadmiernej dyspersji i inflacji w zerze. Stąd wniosek, że model Poissona prawdopodobnie nie jest wystarczający do opisania tych danych.

Tabela 2. Wyliczone wartości dla każdego wyżej wybranego modelu.

	Poisson	Ujemny dwumianowy	ZIPR	ZINBR	Poisson z barierą	Ujemny dwumianowy z barierą
Parametry beta						
intercept	1.03	0.93	1.41	1.2	1.41	1.22
hosp	0.16	0.22	0.16	0.2	0.16	0.21
healthexcellent	-0.36	-0.34	-0.31	-0.32	-0.3	-0.34
healthpoor	0.25	0.31	0.25	0.29	0.25	0.31
numchron	0.15	0.17	0.1	0.13	0.1	0.13
gendermale	-0.11	-0.13	-0.06	-0.08	-0.06	-
school	0.03	0.03	0.02	0.02	0.02	0.02
privinsyes	0.2	0.22	0.08	0.12	0.08	-
Parametry gamma						
intercept'	-	-	-	-	-	-
hosp'	-	-	-0.31	-	0.32	0.32
healthexcellent'	-	-	-	-	-	-
healthpoor'	-	-	-	-	-	-
numchron'	-	-	-0.54	-1.25	0.55	0.55
gendermale'	-	-	0.42	0.56	-0.42	-0.42
school'	-	-	-0.06	-0.08	0.06	0.06
privinsyes'	-	-	-0.75	-1.24	0.75	0.75
theta	-	1.21	-	1.49	-	1.39
liczba parametrów	8	8	13	12	13	11
logarytm funkcji wiarogodności	-17971.61	-12170.55	-16140	-12090	-16140	-12090
AIC	35959	24359	32298.49	24216.51	32300.88	24215.57
BIC	36010.35	24416.62	32387.96	24305.98	32390.35	24298.65
oczekiwana liczba zer	46.71	608.01	710.2	749.14	683	683

Podsumowanie

- Z histogramów parametrów z części symulacyjnej można było odczytać zjawisko inflacji w zerze - największy słupek był dla wartości bliskich 0.
- Wykres kwantylo-kwantylowy $\hat{\alpha}$ sugerował, że istnieje wiele replikacji, w których $\hat{\alpha}$ ma wartości bliskie 0, a dla pozostałych danych empiryczne kwantyle dobrze pasują do teoretycznych.
- W przypadku wykresy kwantylo-kwantylowego χ^2 punkty lezą blisko lini prostej, co oznacza, że empiryczny rozkład dobrze pasuje do teoretycznego.
- Analiza wstępna pobranych danych wykazała występowanie zjawisk nadmiernej dyspersji oraz inflacji w zerze.
- Najlepszym dopasowaniem do danych okazał się zredukowany o niepotrzebne zmiene model ujemny dwumianowy z barierą, a niewiele gorszy od niego okazał się model ZINBR. Najgorzej dopasowanym do danych okazał się model Poissona.