Raport 2 — Zaawansowane modele liniowe

Magdalena Potok

2024-05-07

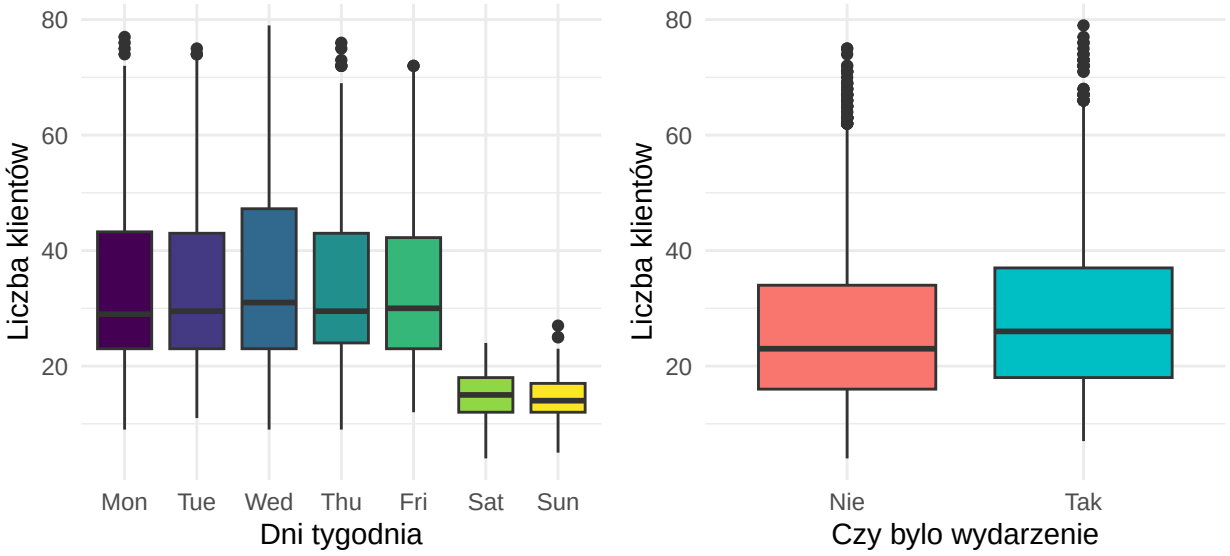
Celem raportu jest przeanalizowanie danych na podstawie regresji Poissona.

Zadanie 1

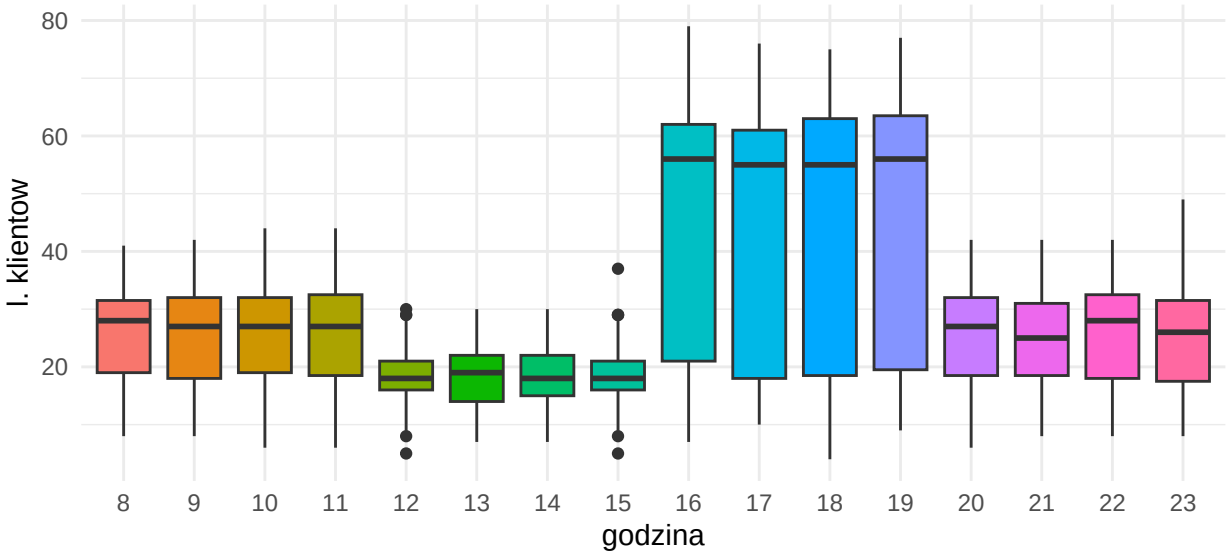
Plik **sklep** zawiera dane o liczbie klientów przychodzących do pewnego sklepu w okresie około 3-ech miesięcy, w zbiorze znajdują się kolumny: "*no.klients*" – liczba klientów obsłużonych w danej godzinie, "*day*" – dzień tygodnia, "*hour*" – godzina oraz "*events*" – informacja, czy w danym dniu miało miejsce jakieś wydarzenie sportowe, gdzie 0 oznacza, że nie, a 1, że tak.
W tym raporcie zostaną przeanalizowane te dane za pomocą regresji Poisssona traktując liczbę obsłużonych klientów jako zmienną objaśniającą, a pozostałe zmienne jako potencjalne predyktory.

Zadanie 2

Sporządzimy wykresy boxplot dla zmiennej y w zależności od każdego predyktora osobno.



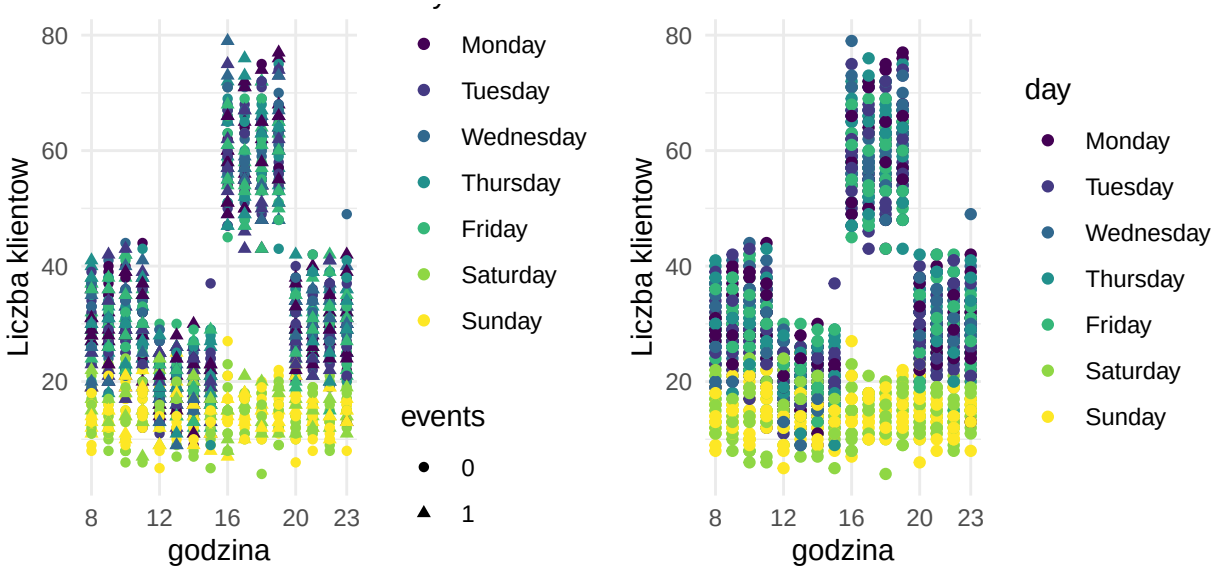
Rysunek 1.: Boxploty zmiennej y w zależności od dni tygodnia oraz "events".



Rysunek 2: Boxplot zmiennej y w zależności od godzin.

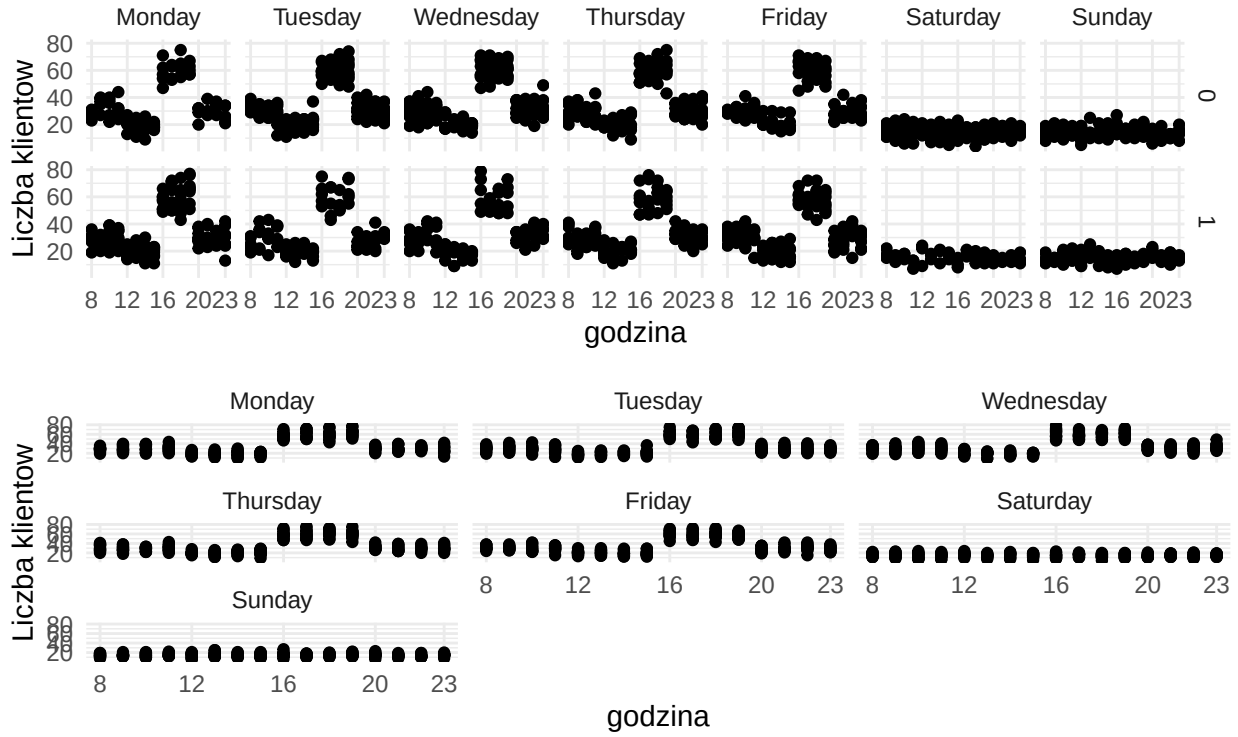
Z pierwszego wykresu łatwo odczytać, że w tygodniu panuje przeważnie większy ruch, niż w weekendy, świadczy o tym rozstawienie pudełek na wykresie typu boxplot. Z wykresu, gdzie predyktorem jest wydarzenie, możemy odczytac, że ten parametr nie ma istotnego wpływu na liczbę klientów w sklepie. Ostatni wykres, czyli ilość klientów w zależności od godziny, informuje nas, że największy ruch jest w godzinach 16-19, a najmniejszy w godzinach 12-15.

Sporządzimy teraz wykresy qqplot.



Rysunek 3: Wykresy qplot zależności liczby klientów od godziny i dnia tygodnia z rozbiciem na zmienną "events" oraz bez.

Z wykresu po lewej stronie możemy zauważyć, że zmienna “events” nie wpływa na liczbę klientów, ilość punktów zaznaczonych trójkątami nie układa się w żaden specjalny, różniący się od okrągłego znaku, sposób. Wykres po prawej stronie, który nie uwzględnia rozróżnienia zmiennej events, jest dużo bardziej przejrzysty i z niego możemy odczytać liczbę klientów w danych dniach i o ustalonych godzinach.



Rysunek 4: Wykresy qplot zależności liczby klientów od godziny i dnia tygodnia z rozbiciem na podgrupy ze względu na zmienne objaśniające.

Tutaj ponownie z pierwszego wykresu dostrzegamy brak wpływu zmiennej "events". Z obu wykresów możemy dostrzec, że liczba klientów w ciągu tygodnia ma 4 charakterystyczne grupy godzin. W weekend natomiast kształtuje się rozkład jednostajny.

Zadanie 3

Model Poissona z interakcją pomiędzy wszystykimi regresorami posiada 224 zmiennych. Model z interakcją bez regresora wydarzenie sportowe posiada 112 zmiennych, co oznacza, że $224 - 112 = 112$ zmiennych w modelu z interakcją zależy od regresora wydarzenie sportowe. Aby zbadać, czy zmienna wydarzenie sportowe jest istotna posłużyliśmy się komendą anova i przeprowadzony został test statystyczny *chi*-kwadrat porównując model zawierający wszystkie interakcje z modelem bez regresora wydarzenie sportowe. P-wartość tego testu wyszła $0.3755 > 0.05$, oznacza to, że nie możemy odrzucić hipotezy, że ten regresor jest nieistotny. Do zbadania, czy interakcje są istotne zbudowany został model bez interakcji i również przeprowadzony został test anova. Wynik tym razem wyszedł bardzo mały, mniejszy niz 0.05, oznacza to, że interakcje są istotne.

Zadanie 4

Nowy model ma 8 zmiennych. Najbogatszym modelem z poprzedniego zadania jest model, który posiada interakcję między wszystkimi regresorami. Ponownie do przeprowadzenia tego testu posłużyliśmy się komendą anova, p-wartość tego testu wynosi w przybliżeniu 0.869, co oznacza, że nie możemy odrzucić hipotezy zerowej, takiej, że nowy model nie różni się statystycznie od poprzedniego. Wychodzi na to, że model, który ma 8 zmiennych jest tak samo wydajny, jak model, który ma 224 zmienne.

Zadanie 5

Tabela 1: Podgrupy modelu grupującego na godziny i typ dnia.

Grupy	Dzień roboczy 8:00-11:59	Dzień roboczy 12:00-15:59	Dzień roboczy 16:00-19:59	Dzień roboczy 20:00-23:59	Dzień weekendowy 8:00-11:59	Dzień weekendowy 12:00-15:59	Dzień weekendowy 16:00-19:59	Dzień weekendowy 20:00-23:59
Średnie	30.01	19.71	59.64	29.98	14.79	14.95	14.87	14.38
Postać predyktora	$\beta_0 + \beta_4$	β_0	$\beta_0 + \beta_2$	$\beta_0 + \beta_3$	$\beta_0 + \beta_1 + \beta_4 + \beta_7$	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2 + \beta_5$	$\beta_0 + \beta_1 + \beta_3 + \beta_6$
Wartość predyktora	3.401	2.981	4.088	3.401	2.694	2.705	2.699	2.665

Z tabeli możemy odczytać, że w weekend klienci przychodzą z tą samą częstotliwością niezależnie od godziny. W następnym zadaniu sprawdzimy, czy predyktory liniowe odpowiadające podgrupom godzin weekendowych są takie same − do tego posłużymy się testem Walda.

Zadanie 6

Korzystając z postaci prekdytora dla dni weekendowych możemy wyznaczyć $\eta_5 = \beta_0 + \beta_1 + \beta_4 + \beta_7$, $\eta_6 = \beta_0 + \beta_1$, $\eta_7 = \beta_0 + \beta_1 + \beta_2 + \beta_5$ oraz $\eta_8 = \beta_0 + \beta_1 + \beta_3 + \beta_6$. Będziemy badać hipotezę następującej postaci:

$$H_0 : \eta_5 = \eta_6 = \eta_7 = \eta_8 \quad vs \quad H_1 : \exists i \neq j : \eta_i \neq \eta_j.$$

Do przetestowania tych hipotez posłużymy się statystyką Walda. Statystyka testowa wynosi 1.38, przy prawdziwości H_0 zbiega ona do rozkładu χ^2 z 3 stopniami swobody. Odrzucamy hipotezę zerową, jeśli dla $\alpha = 0.05$ $W > F^{-1}(1 - \alpha, 3) = 7.81$, a więc nie ma podstaw do odrzucenia hipotezy zerowej.

Zadanie 7

Tabela 2: Grafik pracy sklepu.

	Poniedziałek	Wtorek	Środa	Czwartek	Piątek	Sobota	Niedziela
Pracownik 1	8-12	8-12	8-12	8-12	8-12	8-16	8-16
Pracownik 2	8-16	8-16	8-16	8-16	8-16	-	-
Pracownik 3	16-24	16-24	16-24	16-24	16-24	-	-
Pracownik 4	16-24	16-24	16-24	16-24	16-24	-	-
Pracownik 5	16-20	16-20	16-20	16-20	16-20	16-24	16-24

Pracownicy $2, 3$ i 4 pracują po 40 godzin tygodniowo, natomiast pracownicy 1 i 5 po 36 godzin.

Wnioski

- Predyktor “events” nie ma istotnego wpływu na ilość klientów.
- Typ dnia oraz godziny mają wpływ na zmienną objaśniającą, czyli ilość klientów.
- W weekend klienci przychodzą z tą samą częstotliwością o różnych porach.
- Uproszczony model tak samo dobrze przewiduje dane, a posiada znacznie mniej zmiennych