Zaawansowane Modele Liniowe - raport 2

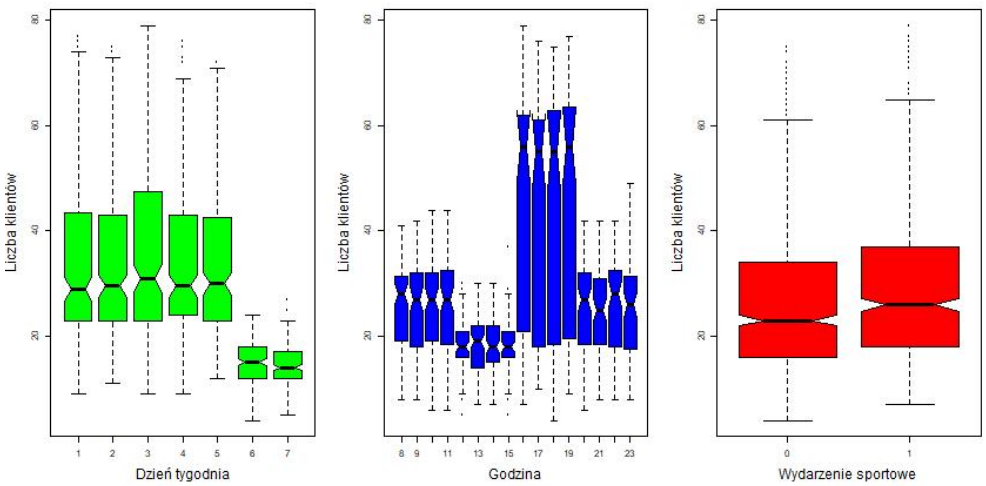
Łukasz Rębisz

2023-05-25

Cel raportu

Celem raportu jest przeanalizowanie za pomocą **regresji Poissona** danych opisujących $y = \textit{liczbę klientów}$ pewnego sklepu w okresie około trzech miesięcy w zależności od *dnia tygodnia*, *godziny* oraz występowania w danym czasie *wydarzenia sportowego*.

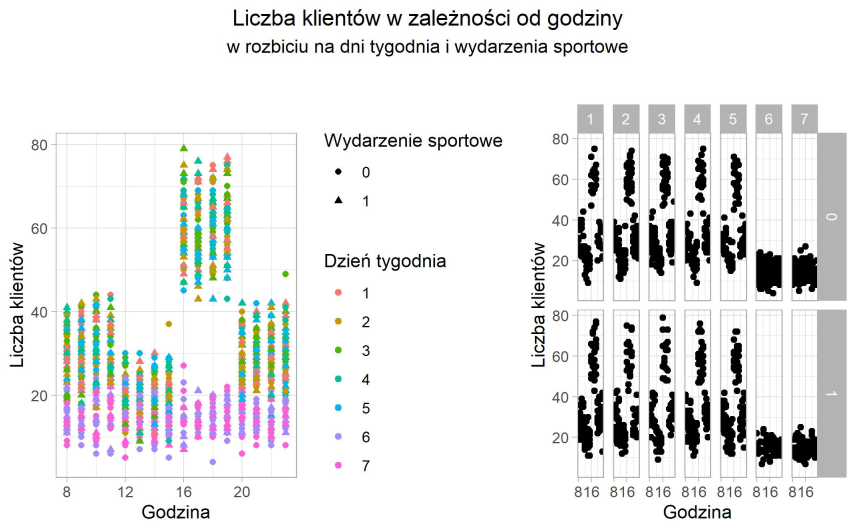
Analiza danych



Rysunek 1: *Boxploty* zmiennej y w zależności od każdego z predyktorów

Analiza powyższych wykresów pokazuje, że:

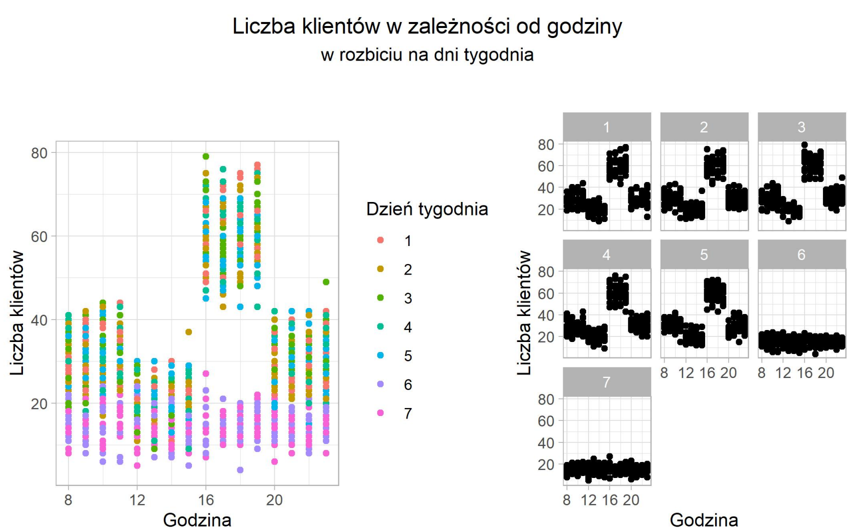
- W ciągu tygodnia występuje wyraźny podział na dwa okresy: dni robocze oraz weekendy. W czasie weekendów średnia liczba klientów jest wyraźnie mniejsza niż w ciągu tygodnia roboczego. Wewnątrz obu grup rozkłady średniej liczby klientów są praktycznie identyczne.
- *Liczba klientów* w zależności od *godziny* przedstawia następujący schemat: szczyt występuje w godzinach popołudniowych 16-19, najmniej klientów przychodziło do sklepu w godzinach 12-15, natomiat poranki (8-11) i wieczory (20-23) charakteryzowały się średnią *liczbą klientów*.
- *Liczba klientów* nie zależy w wyraźny sposób od występowania *wydzarzeń sportowych*. Mediany dla obu grup mają podobną wartość, również kształt wykresów jet analogiczny z nieznacznym wskazaniem na wydarzenia sportowe (wówczas średnia liczba klientów jest nieznacznie większa).



Rysunek 2: Wykresy *y* od *godziny* w rozbiciu na *dzień* i *wydarzenie sportowe*

Powyższe wykresy przedstawiają zależność *liczby klientów* od *godziny* w rozbiciu na *dni tygodnia* oraz *wydarzenia sportowe*. Analiza wykresów pokazuje, że:

- *Wydarzenia* sportowe NIE mają istotnego wpływu na *liczbę klientów*: na lewym wykresie punkty oznaczające występowanie wydarzeń (koła lub trójkąty) występują na przemian - brak wyraźnego wzoru w ich rozkładzie. Natomiast prawy wykres pokazuje, że bez względu na występowanie *wydarzeń sporotwych* rozkład *liczby klientów* danego *dnia tygodnia* w zależności od *godziny* jest analogiczny.



Rysunek 3: Wykresy *y* od *godziny* w rozbiciu na *dzień*

Pominięcie rozróżnienia na występowanie *wydarzeń sportowych* pozwala wysunąć wniosek:

- Rozkład *liczby klientów* w zależności od *godziny* w weekendy jest jednostajny (pojawienie się klientów jest tak samo prawdopodobne w każdej godzinie) w przeciwieństwie do dni roboczych, gdzie wyraźnie zauważamy każdego dnia opisane powyżej godziny szczytu oraz najmniejszego zainteresowania. Ponadto, średnia *liczba klientów* w soboty i niedziele jest wyraźnie mniejsza niż w tygodniu roboczym.

Wnioski:

- Zmienna *wydarzenia sportowe* NIE ma istotnego wpływu na *liczbę klientów*.
- Istnieje możliwość pogrupowania *dni* (tydzień roboczy - weekend) oraz *godzin* tak, aby zredukować liczbę zmiennych.
- Rozkład *liczby klientów* w weekendy w zależności od *godziny* jest jednostajny, w ciągu tygodnia roboczego występują natomiast godziny szczytu oraz najmniejszego zainteresowania.

Model Poissona

Skonstruujmy model Poissona z interakcją pomiędzy wszystkimi regresorami, traktując je jako faktory.

Liczba zmiennych modelu z interakcją wynosi 224.

Od regresora *wydarzenia sportowe* zależy 112 zmiennych.

Istotność zmiennej *wydarzenia sportowe* - test Deviance:

$\chi^2 = D(\text{Model bez wydarzeń sportowych}) - D(\text{Model z wydarzeniami}) =$

$= 116.13 < 137.7 = F^{-1}_{\chi^2_{112}}(1 - 0.05).$

Zatem nie mamy podstaw do odrzucenia hipotezy zerowej mówiącej, że model bez *wydarzeń sportowych* jest lepszy od modelu z *wydarzeniami*. Zatem zmienna *wydarzenia sportowe* NIE jest istotna.

Istotność interakcji - test Deviance:

$\chi^2 = D(\text{Model bez interakcji}) - D(\text{Model z interakcjami}) =$

$= 1051.57 > 235.08 = F^{-1}_{\chi^2_{201}}(1 - 0.05).$

Zatem odrzucamy hipotezę zerową mówiącą, że model bez interakcji jest lepszy od modelu z interakcją. Zatem interakcja pomiędzy zmiennymi jest istotna.

Podział zmiennych *dzień tygodnia* oraz *godzina*

Stwórzmy dwie nowe zmienne:

- pierwszą opisującą czy dzień jest weekendowy,
- drugą grupującą godziny każdego dnia w bloki 4-godzinne.

Model regresji Poissona z interakcją pomiędzy powyższymi zmiennymi ma 8 zmiennych.

Porównanie z modelem z interakcją skonstruowanym w poprzednim podpunkcie:

$\chi^2 = D(\text{Model z interakcją oparty na nowych zmiennych}) - D(\text{Model z interakcją, poprzednie zmienne}) =$

$= 192.85 < 251.29 = F^{-1}_{\chi^2}(1 - 0.05, 216).$

Zatem nie mamy podstaw do odrzucenia hipotezy zerowej mówiącej, że model z interakcajmi oparty na nowych zmiennych (*weekend, pora dnia*) jest lepszy od modelu z interakcjami opartego na wyjściowych zmiennych. Zatem nowy model jest niegorszy od „najbogatszego” modelu z poprzedniego zadania. Pomiędzy badanymi modelami NIE ma więc istotnej statystycznej różnicy. Możemy zatem stosować model z interakcjami oparty na nowych zmiennych.

Tabela 1: tabela predykcji dla nowego modelu

1.	rob. 8-11	rob. 12-15	rob. 16-19	rob. 20-23	w. 8-11	w. 12-15	w. 16-19	w. 20-23
2.	30.0	19.7	59.6	30.0	14.8	15.0	14.9	14.4
3.	$\hat{\beta}_0$	$\hat{\beta}_0 + \hat{\beta}_2$	$\hat{\beta}_0 + \hat{\beta}_3$	$\hat{\beta}_0 + \hat{\beta}_4$	$\hat{\beta}_0 + \hat{\beta}_1$	$\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_5$	$\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_3 + \hat{\beta}_6$	$\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_4 + \hat{\beta}_7$
4.	3.401	2.981	4.088	3.401	2.694	2.705	2.699	2.665
5.	30.0	19.7	59.6	30.0	14.8	15.0	14.9	14.4

Wiersze powyższej tabeli przedstawiają następujące dane:

1) Informacja o wszystkich grupach.
2) Średnia liczba klientów w podgrupie na godzinę.
3) Predyktor liniowy oparty na zmiennych *weekend, pora dnia* postaci:

$$\eta_i = \hat{\beta}_0 + x_{i,1}\hat{\beta}_1 + ... + x_{i,7}\hat{\beta}_7,$$

   gdzie $x_{i,1}, ..., x_{i,7} \in \{0, 1\}$ oraz $x_{i,1}$ - zmienna *weekend*, $x_{i,2}$ - *druga pora dnia*, $x_{i,3}$ - *trzecia pora dnia*, $x_{i,4}$ - *czwarta pora dnia*, $x_{i,5}$ - *druga pora dnia, dzień weekendowy*, $x_{i,6}$ - *trzecia pora dnia, dzień weekendowy*, $x_{i,7}$ - *czwarta pora dnia, dzień weekendowy.*

4) Wartość predyktora liniowego $\eta_i = \hat{\beta}_0 + x_{i,1}\hat{\beta}_1 + ... + x_{i,7}\hat{\beta}_7$.
5) Wartość $\lambda_i = \exp(\eta_i) = \exp(\hat{\beta}_0 + x_{i,1}\hat{\beta}_1 + ... + x_{i,7}\hat{\beta}_7)$.

Powyższe wyniki pokazują, że:

- Otrzymane na podstawie modelu średnie wartości *liczby klientów* (5. wiersz) pokrywają się z wartościami średnimi dla badanych danych (2. wiersz).
- W rozkładzie liczby klientów w dniach roboczych wyraźnie widać godziny szczytu (16-19) oraz najmniejszego zainteresowania (12-15).
- W weekendy klienci przychodzą z tą samą częstotliwością (średnio 14-15 osób na godzinę) bez względu na porę dnia.

Test Walda

Analiza wstępna i wyniki powyższej tabeli sugerują, że w weekend klienci przychodzą z tą samą częstotliwością o różnych godzinach. Przetestujmy zatem hipotezę, czy predyktory liniowe odpowiadające podgrupom godzin weekendowych są takie same.

Mamy 4 takie podgrupy i testujemy równość każdego predyktora z każdym, korzystając z **testu Walda**:

$$\eta_1 = \beta_0 + \beta_1$$
$$\eta_2 = \beta_0 + \beta_1 + \beta_2 + \beta_5$$
$$\eta_3 = \beta_0 + \beta_1 + \beta_3 + \beta_6$$
$$\eta_4 = \beta_0 + \beta_1 + \beta_4 + \beta_7$$

Równość $\eta_1 = \eta_2 = \eta_3 = \eta_4$ prowadzi do układu równań:

$$\beta_2 + \beta_5 = 0, \quad \beta_3 + \beta_6 = 0, \quad \beta_4 + \beta_7 = 0.$$

Równoważnie poważszy układ równań można zapisać w postaci macierzowej:

$$A\beta = 0,$$

gdzie $A \in M_{3\times 8}$ jest postaci:

$$A = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}.$$

Załóżmy, że $\hat{\beta} \xrightarrow{d} N(\beta, \Sigma)$ oraz że $det(\Sigma) \neq 0$. Wówczas na mocy twierdzenia Walda otrzymujemy, że przy H_0 statystyka Walda

$$W = (A\hat{\beta})'(A\Sigma A')^{-1}(A\hat{\beta})$$

zbiega wg rozkładu do rozkładu chi-kwadrat z 3 stopniami swobody.

Statystka Walda $W = 0.013 < 7.815 = F^{-1}_{\chi^2_3}(0.95)$, zatem nie odrzucamy na poziomie istotności $\alpha = 0.05$ hipotezy zerowej mówiącej o równości parametrów η_i odpowiadającym podgrupom godzin weekendowych.

Grafik pracy

Załóżmy, że jeden pracownik jest w stanie obsłużyć do 20 klientów w ciągu godziny. Na podstawie wyników tabeli zaplanujmy optymalną liczbę pracowników oraz grafik pracy. Ponadto każdy z pracowników może pracować maksymalnie 8h w ciągu dnia.

Tabela 2: tabela przedstawiająca optymalny grafik pracy

$$\begin{array}{r} 1. \\ 2. \\ 3. \\ 4. \end{array} \begin{pmatrix} \text{rob. 8-11} & \text{rob. 12-15} & \text{rob. 16-19} & \text{rob. 20-23} & \text{w. 8-11} & \text{w. 12-15} & \text{w. 16-19} & \text{w. 20-23} \\ 30.0 & 19.7 & 59.6 & 30.0 & 14.8 & 15.0 & 14.9 & 14.4 \\ 2 & 1 & 3 & 2 & 1 & 1 & 1 & 1 \\ P1+P2 & P5 & P3+P4+P5 & P3+P4 & P1 & P1 & P2 & P2 \end{pmatrix}$$

Wiersze powyższej tabeli przedstawiają następujące dane:

1) Informacja o wszystkich grupach.
2) Średnia liczba klientów w podgrupie na godzinę.
3) Potrzebna liczba pracowników przy założeniu, że w ciągu godziny 1 pracownik może obsłużyć max 20 klientów.
4) Grafik pracy rozpisany na poszczególnych pracowników.

Przy tak rozpisanym grafiku pracy żaden pracownik nie ma danego dnia „okienek". Ponadto łączna liczba godzin pracy tygodniowo dla poszególnych pracowników wynosi:

$$P1 = P2 = 36\text{h}, \quad P3 = P4 = P5 = 40\text{h}.$$

Podsumowanie

Wstępna analiza otrzymanych danych pozwoliła na wyeliminowanie z modelu nieistotnej zmiennej *wydarzenia sportowe*. Porównanie różnych modeli **regresji Poissona** poprzez test Deviance pozwoliło wybrać optymalny model oparty o nowe zmienne: *weekend* oraz *pora dnia*. Ostatnim etapem raportu była predykcja na podstawie optymalnego modelu. Dzięki niej stwierdziliśmy, że rozkład predyktorów liniowych dla dni weekendowych jest jednostajny oraz opisaliśmy rozkład w dni robocze (godziny szczytu, najmniejszego zainteresowania). Powyższe analizy pozwoliły stworzyć optymalny grafik pracy.