Zaawansowane Modele Liniowe

Lista 3
Regresja Poissona

1. Pobierz plik "sklep" z danymi i wczytaj do R. W tym zbiorze mamy informację o liczbie klientów przychodzących do pewnego sklepu w okresie około 3–ech miesięcy. W zbiorze znajdują się cztery kolumny:

- no.klients – liczba klientów obsłużonych w danej godzinie (wart.: 0,1,2,...),
- day – dzień tygodnia (wart.: poniedziałek, wtorek,..., niedziela),
- hour – godzina (wart.: 8,9,...,23),
- events – informacja o tym czy w danym dniu miało miejsce jakieś wydarzenie sportowe (wart.: 0 - nie, 1 - tak),

Twoim zadaniem będzie przeanalizowanie danych za pomocą regresji Poissona traktując liczbę obsłużonych klientów jako zmienną objaśnianą (y), a pozostałe zmienne jako potencjalne predyktory.

2. Dokonaj wstępnej analizy danych:

- Sporządź boxplot zmiennej y w zależności od każdego predyktora osobno. (funkcja *boxplot()*). Opisz występujące prawidłowości.
- Zainstaluj pakiet *ggplot2* (jest to świetny pakiet do ilustracji danych) i uzyskaj za pomocą funkcji qqplot następujące wykresy:

  qplot(hour,no.klients, shape = as.factor(events),col = day, data = sklep)

  Powyższy wykres przedstawia zależność y od *godziny* w rozbiciu na *dzień* (różne kolory; parametr "*col*") i *wydarzenie sportowe* (kształt znacznika; parametr "*shape*").

  qplot(hour,no.klients, facets = events $\sim$ day, data = sklep)

  Powyższy wykres korzysta z parametru *facets*, który umożliwił rozbicie poprzedniego wykresu na podgrupy ze względu na zmienne *wydarzenie sportowe* (oś pionowa – znaczniki z prawej strony) oraz dzień tygodnia (oś pozioma). Porównując oba wykresy jakie są twoje obserwacje odnośnie wpływu poszczególnych zmiennych? Czy zmienna *wydarzenie sportowe* cokolwiek różnicuje?

qplot(hour,no.klients, color = day, data = sklep)
qplot(hour,no.klients, facets = ~day, data = sklep)

Powyższe wykresy przedstawia zależność y od *godziny* w rozbiciu na dzień tygodnia. Czy na podstawie tych wykresów potrafisz powiedzieć coś więcej o rozrzucie danych w boxplocie y vs. *godzina* np. dla godzin od 16 do 19? Czy widzisz jakieś prawidłowości? Czy dostrzegasz możliwość pogrupowania dni tygodni i/lub godzin w ciągu dnia ze względu na podobne zachowanie klientów?

3. Analiza wstępna sugeruje przynajmniej trzy rzeczy:

- zmienna *wydarzenie sportowe* nie ma wpływu na y.
- istnieje możliwość pogrupowania *dni* oraz *godzin* tak by zredukować liczbę zmiennych.
- jednostajny rozkład klientów w weekend.

Skonstruuj model Poissona z interakcją pomiędzy wszystkimi regresorami traktując je jako faktory. Ile zmiennych ma taki model? Ile zmiennych w modelu zależy od regresora *wydarzenie sportowe*?
Zbadaj czy zmienna *wydarzenie sportowe* jest istotna.
Zbadaj czy interakcje są istotne?

4. Stwórz dwie nowe zmienne. Pierwszą opisującą to czy dzień jest dniem roboczym czy weekendowym. Druga grupująca godziny każdego dnia w bloki 4–o godzinne. Skonstruuj model Poissona z interakcją pomiędzy nowymi zmiennymi traktując je jako faktory.
Ile zmiennych ma nowy model?
Zbadaj czy nowy model różni się statystycznie od "najbogatszego" z poprzedniego zadania?

5. Stwórz tabelę otrzymaną w oparciu o model z zadania 4 składającą się z czterech wierszy.
W pierwszym wierszu zamieść informację o wszystkich podgrupach, do których trafiają poszczególne godziny w różnych dniach tygodnia, np. dzień roboczy 8:00-11:59; dzień weekendowy 16:00-19:59 itd.
W drugim wierszu zamieść średnią liczbę obsłużonych klientów (na godzinę) odpowiadającą podgrupie z pierwszego wiersza.

W trzecim i czwartym wierszu wpisz postać predyktora liniowego $\eta_i = \hat{\beta}_0 + x_{i1}\hat{\beta}_1 + ... + x_{ip-1}\hat{\beta}_{p-1}$ oraz jego wartość, odpowiadającego podgrupie z pierwszego wiersza.

Uwaga 1. Przy wypełnianiu 3 i 4 wiersza zwróć uwagę na dwie rzeczy. Po pierwsze elementy macierzy planu przyjmują tylko dwie wartości $x_{i1} \in \{0,1\}$, dlatego predyktor liniowy dla każdej podgrupy jest sumą pewnych elemetów wektora $(\hat{\beta}_0, ..., \hat{\beta}_{p-1})'$, $\eta_i = \hat{\beta}_0 + \sum_{j \in A} \hat{\beta}_j$ gdzie $A = \{j : x_{ij} = 1\}$. Po drugie jeżeli dwie obserwacje j-ta i k-ta trafiają do tej samej podgrupy to odpowiadające im wiersze w macierzy planu są takie same i wówczas $\eta_j = \eta_k$. Wynika z tego że η_i dla każdej podgrupy jest wyznaczona jednoznacznie.

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

Uwaga 1. Mamy 4 takie podgrupy i testujemy równość każdego predyktora z każdym. Skorzystaj z testu Walda.

7. Właściciel sklepu poprosił, abyś na podtawie wyników tabeli zaplanował optymalną liczbę pracowników oraz grafik pracy. W tym celu załóż, że w ciągu godziny jeden pracownik jest w stanie obsłużyć do 20 klientów.