ADZD - Lista 2

Regresja Poissona, Test Walda

Wstęp

Poniższy raport skupia się na zastosowaniu modelu *regresji Poissona* do analizy pewnych danych. W takim modelu zmienna objaśniana odpowiada liczbie zdarzeń i przyjmuje wartości ze zbioru liczb naturalnych $Y_i \in \{1, 2, 3, ...\}$ dla $i = 1, ...n$ i zakładamy, że pochodzi ona z rozkładu Poissona z parametrem $\lambda_i > 0$:

$$P(Y_i = k) = e^{-\lambda_i} \frac{\lambda_i^k}{k!}.$$

Zachodzi $E[Y_i] = Var(Y_i) = \lambda_i$. Zmienne Y_i są niezależne, a związek między wartością oczekiwaną Y_i a predyktorami opisuje równanie

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i,1} + ...\beta_{p-1} x_{i,p-1} = \beta_0 + \beta \cdot X_i.$$

Podobnie jak w przypadku regresji logistycznej, estymator wektora współczynników $\hat{\beta} \in R^p$ wyznaczany jest za pomocą algorytów optymalizacyjnych. Testowanie istotności współczynników i dopasowania modelu do danych przebiega analogicznie do tego w regresji logistycznej, tzn. w oparciu o asymptotyczny rozkład wektora parametrów $\hat{\beta} \to_d N(\beta, J^{-1})$. Macierz $S(\beta)$ występująca w faktoryzacji macierzy J w przypadku regresji Poissona jest macierzą diagonalną taką, że

$$S(\beta)_{i,i} = \lambda_i.$$

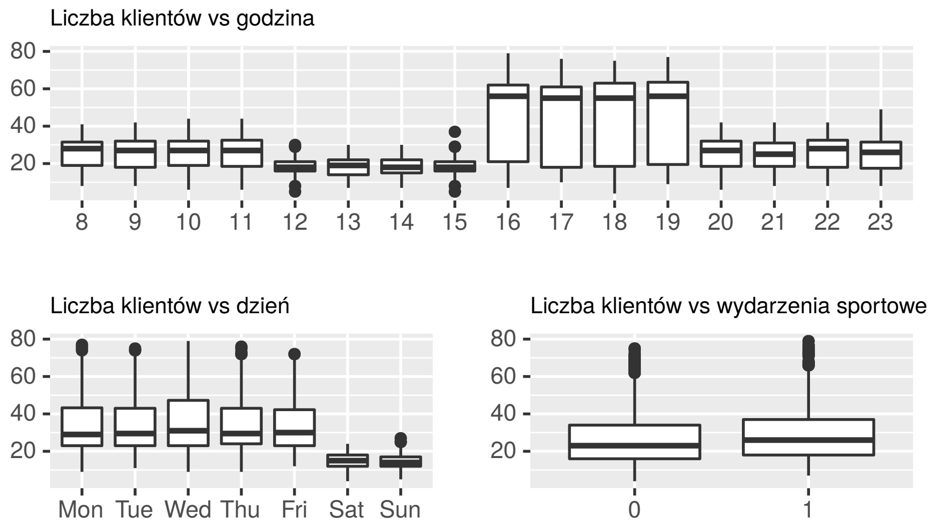
Nieznane λ_i zastępujemy ich estymatorami, czyli przewidzianymi przez model $E[Y_i]$.

Uwaga o kateogrycznych regresorach: W przypadku, gdy któreś z regresorów $X_1, ...X_{p-1}$ to zmienne kategoryczne, dopasowanie modelu w oparciu o powyższe równanie nie będzie miało sensu (np. nie chcemy żeby wtorek był traktowany jako średnia z poniedziałku i środy). W takiej sytuacji standardowym rozwiązaniem jest zakodowanie każdej z takich zmiennych w formie *one-hot encoding*, tzn. zmienną X_i przyjmującą k możliwych wartości zamienić na k wektorów binarnych odpowiadających występowaniu kolejnych wartości danej cechy (lub tzw. *dummy encoding* - bardzo podobne ale koduje zmienną z k poziomami jako $k - 1$ wektorów (ostatni to same 0)). Funkcja `glm` w R robi to automatycznie, gdy rozpozna kategoryczne zmienne. Zmienne objaśniające w naszym zbiorze danych (`hour`, `events`, `day`) są zmiennymi kategorycznymi i po wczytaniu powinny zostać przekonwertowane na typ `factor`.

```
## 'data.frame':    1456 obs. of  4 variables:
##  $ hour      : Factor w/ 16 levels "8","9","10","11",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ day       : Ord.factor w/ 7 levels "Mon"<"Tue"<"Wed"<..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ events    : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ no.klients: int  26 37 36 32 13 22 23 22 57 53 ...
```

Zadania 1,2 (Wstępna analiza danych)

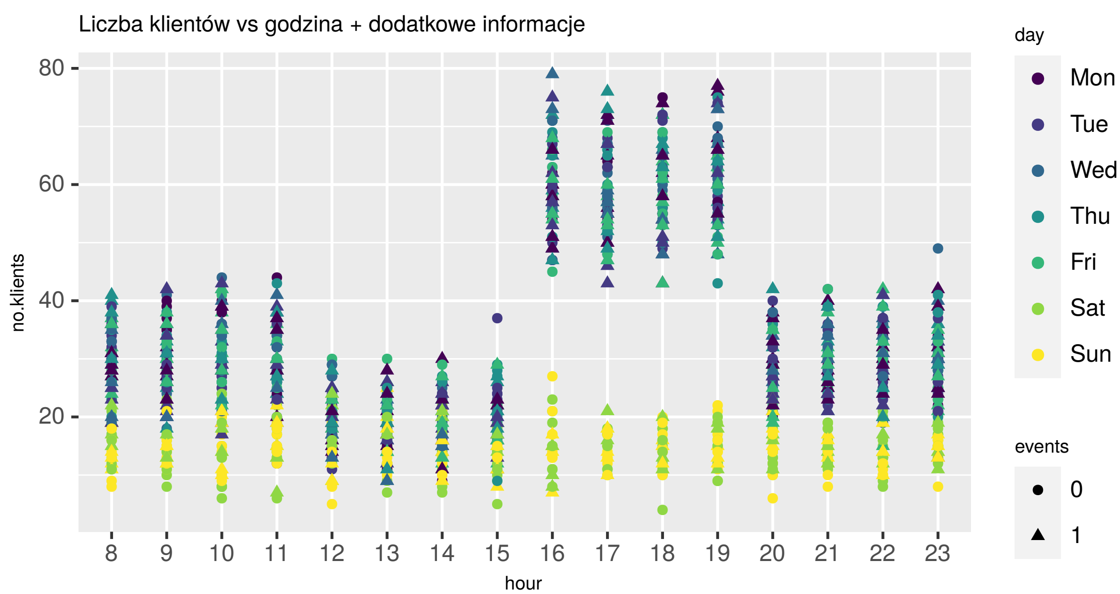
Za pomocą wykresów pudełkowych analizujemy zależność zmiennej zależnej od każdego z 3 regresorów.

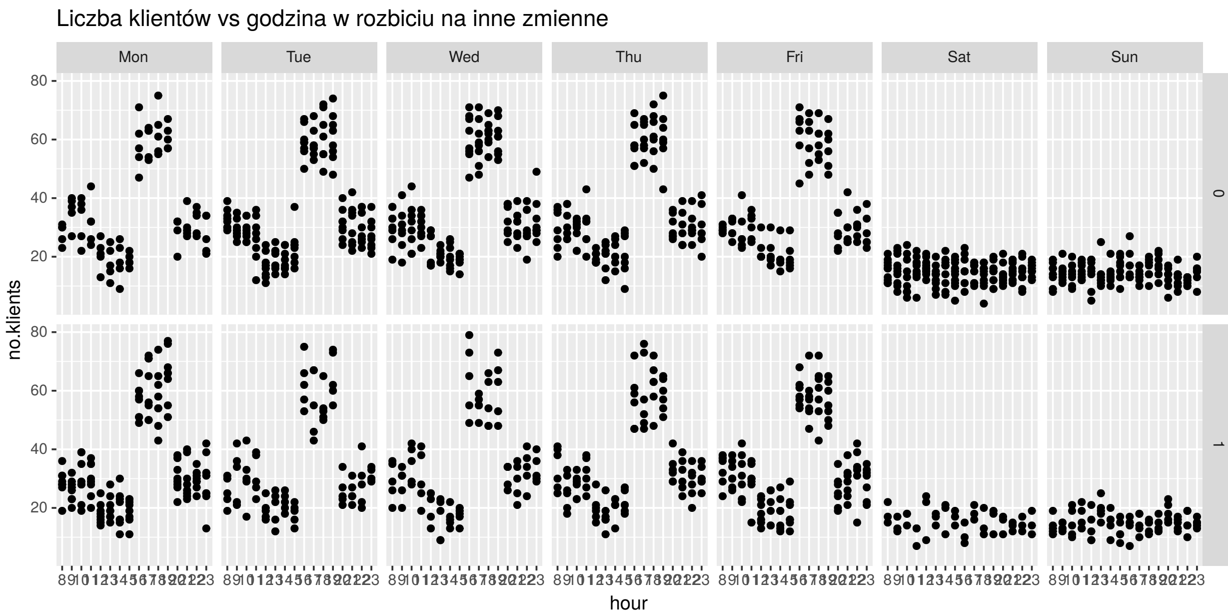


Komentarz:

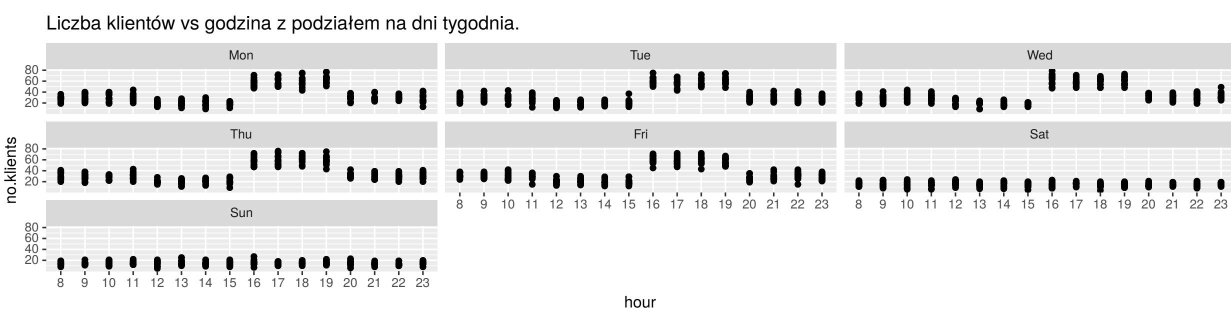
- W podziale ze względu na dni tygodnia można wyróżnić 2 grupy: dni robocze i weekend. Pierwsza grupa charakteryzuje się wyższą średnią, ale również rozrzutem.
- W podziale ze względu na godziny można wyróżnić 3 grupy o podobnych średnich i rozrzutach.
- Nie widać znaczących różnic ze względu na zmienną `events` (wydarzenie sportowe).
- Zależność rozrzutu i średniej opisana w pierwszym podpunkcie jest w przybliżeniu zgodna z założeniami modelu Poissona, w którym wariancja powinna równać się średniej.

W dalszej części analizujemy zależność zmiennej objaśnianej od więcej niż 1 regresora na raz.





Komentarz: Można zauważyć, że wykresy nie różnią się ze względu na wartość zmiennej `events` - potencjalnie więc można będzie usunąć ją ze zbioru regresorów. Przeanalizujemy ponownie wykres pokolorowany ze względu na dzień tygodnia ignorując zmienną `events` oraz dodatkowy wykres typu `facets` wykonany bez podziału na `events`.



Komentarz:

- W godzinach od 16 do 19 włącznie widać znaczącą różnicę między dniami roboczymi a weekendem.
- Rozkład liczby klientów w tygodniu zależy od godziny, podczas gdy w weekendy utrzymuje się na mniej więcej stałym poziomie niezależnie od pory dnia.
- Na wykresach dla dni roboczych można wyróżnic 2 grupy godzin: od 16 do 19 (o znacząco wyższej niż pozostałe liczbie klientów) oraz pozostałe.

Zadanie 3

Wykresy dodatkowe zarówno potwierdziły obserwacje z wykresów pudełkowych, ale również dostarczyły dodatkowych informacji. W oparciu o te analizy można spodziewać się braku istotności zmiennej `events` oraz zakodować dni i godziny w postaci pogrupowanej (np. dni - weekend / robocze) redukując w ten sposób ilość zmiennych w formie one-hot. Dodatkowo, ponieważ zauważamy pewną interakcję pomiędzy dniem tygodnia i godziną, użyjemy modelu z interakcją. Tak utworzony model (używamy dummy encoding) ma aż $(2-1)+(7-1)+(16-1)+(2-1)(7-1)+(2-1)(16-1)+(7-1)(16-1)+(2-1)(7-1)(16-1) = 223$ parametry nie licząc interceptu.

```r
model_zad3 = glm(no.klients ~ day*events*hour, data = sklep, family = poisson())
model_zad3$coefficients %>% length()
```

```
## [1] 224
```

$(2-1)(1+6+15+7\cdot 16) = 112$ z nich zniknęłyby, gdyby pozbyć się zmiennej `events` ze zbioru.

```r
grepl("events", names(model_zad3$coefficients)) %>% sum()
```

```
## [1] 112
```

Istotność zmiennej `events` testujemy z użyciem statystki *Deviance* i testu χ^2. Porównamy modele z interakcją M_0 skonstruowany bez zmiennej `events` i M_1 ze zmienną (czyli model skonstruowany już wyżej).

```
## Analysis of Deviance Table
##
## Model 1: no.klients ~ day * hour
## Model 2: no.klients ~ day * events * hour
##   Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
## 1      1344     1475.7
## 2      1232     1359.6 112   116.13   0.3755
```

Test zwrócił p-wartość ok. 0.38 - znacznie większą niż zadany pozom istotności $\alpha = 0.05$, stąd nie odrzucamy hipotezy zerowej co znaczy że zmienna nie jest istotna. W ramach ostatniego podpunktu w analogiczny sposób porównamy model ze wszystkimi zmiennymi bez i z interakcją.

```
## Analysis of Deviance Table
##
## Model 1: no.klients ~ day + events + hour
## Model 2: no.klients ~ day * events * hour
##   Resid. Df Resid. Dev  Df Deviance  Pr(>Chi)
## 1      1433     2411.2
## 2      1232     1359.6 201   1051.6 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tym razem test wykazał istotność zmiennych, w dodatku nawet na poziomie jeszcze mniejszym niż 0.05 - wiemy, że interakcje w istotny sposób wpływają na model.

Zadanie 4

W tym zadaniu konstruujemy nowe zmienne: `day_weekend` dzielącą dni na weekendowe i inne oraz `hour_block` dzielącą dzień na 4-godzinne bloki. Nowy model uwzględnia tylko intercept, te 2 zmienne i interakcje między nimi; ma tylko $(2-1)+(4-1)+(2-1)(4-1) = 7$ zmiennych. P-wartości wyznaczone przez `summary` wskazują istotność 6 z nich (nieistotna zmienne to `hour_block4` oraz jej interakcja z `day_weekend1`). Wykorzystamy statystykę *Deviance* do zbadania, czy modele różnią się statystyczne. Test taki pozwalał na testowanie hipotezy

$$H_0 : \forall(i \in A)\beta_i = 0 \quad \text{vs} \quad H_1 : \exists(i \in A)\beta_i \neq 0.$$

Statystką testową jest $\chi^2_{|A|} = D(M_0) - D(M_1)$, gdzie M_0 to model z hipotezy zerowej (tutaj model z zadania 4), M_1 to model z hipotezy alternatywnej (tutaj model z zadania 3), a $D(M)$ to *Deviance* dla danego modelu. Liczba stopni swobody jest równa różnicy ilości zmiennych w modelach $|A| = 223 - 7 = 216$.

Funkcja `anova` zwróciła wartość *Deviance* i odpowiadającą jej p-wartość odpowiednio 192.85 i 0.87. Ponieważ p-wartość jest bardzo duża, nie mamy podstaw do odrzucenia hipotezy zerowej, tzn. zakładamy że modele nie różnią się.

Zadanie 5

Zmienna `day_weekend` przyjmuje wartość 1 dla soboty i niedzieli, 0 w przeciwnym wypadku. Zmienna `hour_block` przyjmuje wartości $1, 2, 3, 4$ które odpowiadają odpowiednio blokom godzinowym:

- od 8:00 do 12:00
- od 12:00 do 16:00
- od 16:00 do 20:00
- od 20:00 do 24:00

Grupujemy ziór danych równocześnie ze względu na zmienne `day_weekend` i `hour_block` i obliczamy średnie liczby klientów na godzinę w każdej z nich. Model jest postaci $Y = \beta_0 + \beta_1 X_1 + ... + \beta_7 X_7$, gdzie $X_1, ...X_7$ odpowiadają odpowiednio zmiennym: `hour_block2`, `hour_block3`, `hour_block4`, `day_weekend1`, `hour_block2:day_weekend1`, `hour_block3:day_weekend1`, `hour_block4:day_weekend1`.

Table 1: Średnie liczby klientów w grupach w zadaniu 5

day_weekend	0	0	0	0	1	1	1	1
hour_block	1	2	3	4	1	2	3	4
mean_klients	30.00769	19.71154	59.64231	29.98077	14.78846	14.95192	14.86538	14.37500

Widzimy, że np.:

- pierwsza kolumna w tak utworzonej tabeli odpowiada wektorowi samych 0 dla wszystkich wymienionych zmiennych,
- druga kolumna odpowiada `hour_block2 = 1` i 0 wszędzie indziej
- ostatnia kolumna to `hour_block4:day_weekend1 = 1` i 0 wszędzie indziej.

Na podstawie takiej analizy możemy ustalić jakie kombinacje liniowe współczynników odpowiadają każdej z kolumn. Wiersz `betas` zawiera symboliczny zapis kombinacji liniowej, `predictor` jej wartość obliczoną przez podstawienie wyznaczonych przez `glm` wartości $\hat{\beta}$. W modelu regresji Poissona wartość kombinacji liniowej jest równa logarytmowi wartości oczekiwanej Y, a więc przewidywane średnie możemy uzyskać nakładając na wartości w ostatnim wierszu funkcję `exp` - wynik dołączony został jako dodatkowy wiersz tabeli (`pred_mean_klients`).

Table 2: Zadanie 5 - wyniki (zaokrąglone do 3 miejsca po przecinku)

day_weekend	0	0	0	0	1	1	1	1
hour_block	1	2	3	4	1	2	3	4
mean_klients	30.008	19.712	59.642	29.981	14.788	14.952	14.865	14.375
betas	b0	b0 + b1	b0 + b2	b0 + b3	b0 + b4	b0 + b1 + b4 + b5	b0 + b2 + b4 + b6	b0 + b3 + b4 + b7
predictor	3.401	2.981	4.088	3.401	2.694	2.705	2.699	2.665
pred_mean_klients	30.008	19.712	59.642	29.981	14.788	14.952	14.865	14.375

Jak widać, przewidziane przez model średnie są z dokładnością do 3 miejsca po przecinku takie same jak prawdziwe. Wektor różnic pomiędzy nimi to:

```
## [1] -2.307692e-06  1.538462e-06  2.307692e-06  7.692307e-07 -1.538462e-06
## [6] -3.076924e-06 -4.615385e-06  0.000000e+00
```

Zadanie 6

W tym zadaniu skorzystamy z **testu Walda** do przetestowania czy predyktory dla poszczególnych dni weekendowych rzeczywiście są takie same, gdzie

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

tzn. testujemy hipotezę postaci

$$H_0 : \eta_1 = \eta_2 = \eta_3 = \eta_4 \quad \text{vs} \quad H_1 :\sim H_0.$$

Hipotezę zerową można zapisać jako koniunkcję $\binom{4}{2} = 6$ warunków postaci $\eta_i = \eta_j$. Upraszczamy warunki:

- $\eta_1 = \eta_2 \iff \beta_1 = -\beta_5$
- $\eta_1 = \eta_3 \iff \beta_2 = -\beta_6$
- $\eta_1 = \eta_4 \iff \beta_3 = -\beta_7$
- $\eta_2 = \eta_3 \iff \beta_1 + \beta_5 = \beta_2 + \beta_6$
- $\eta_2 = \eta_4 \iff \beta_1 + \beta_5 = \beta_3 + \beta_7$
- $\eta_3 = \eta_4 \iff \beta_2 + \beta_6 = \beta_3 + \beta_7,$

co sprowadza się do trzech równań: $\beta_1 + \beta_5 = 0, \beta_2 + \beta_6 = 0, \beta_3 + \beta_7 = 0$, a w formie macierzowej $A\beta = 0$ dla

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Przy założeniach modelu regresji Poissona i prawdziwej hipotezie zerowej statystyka

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

zbiega wg. rozkładu do statystyki χ^2_3. Test Walda odrzuci H_0 dla wartości W większych od kwantyla rzędu $1-\alpha$ rozkładu chi-kwadrat z 3 (liczba wierszy A) stopniami swobody. (Macierz Σ jest macierzą asymptotycznej kowariancji $\hat{\beta}$, czyli odwróconą macierzą informacji Fishera.)

```
W
```

```
##              [,1]
## [1,] 5.162398e-07
```

```
W > chisq_kwantyl # jak tak to odrzucamy H_0!
```

```
##       [,1]
## [1,] FALSE
```

Wartość statystyki testowej nie jest większa niż odpowiedni kwantyl, więc nie odrzucimy hipotezy zerowej. Możemy zakładać, że wszystkie średnie dla soboty i niedzieli rzeczywiście są takie same.

Zadanie 7

W oparciu o tabelę z zadania 5 ustalimy optymalną liczbę pracowników z podziałem na poszczególne dni i pory dnia. Zakładamy, że każdy pracownik może obsłużyć do 20 klientów w ciągu godziny i na tej podstawie wyznaczamy najpierw minimalną liczbę pracowników potrzebną do obsłużenia wszystkich klientów. Założymy też, że priorytetem dla sklepu jest obsłużenie maksymalnej liczby klientów (patrząc na średnie).

Table 3: Zadanie 7 - ilu pracowników potrzeba do obsłużenia wszystkich klientów?

day_weekend	0	0	0	0	1	1	1	1
hour_block	1	2	3	4	1	2	3	4
mean_klients	30.00769	19.71154	59.64231	29.98077	14.78846	14.95192	14.86538	14.37500
min_pracownicy	2	1	3	2	1	1	1	1

Największa liczba pracowników potrzebna w tym samym momencie to 3, więc w sklepie będzie 3 pracowników - każdy z nich pojawi się na zmianie od poniedziałku do piątku w czasie odpowiadającym trzeciemu blokowi, tzn. od 16:00 do 20:00. Sklep jest czynny od 8:00 do 24:00 przez 7 dni w tygodniu, co łącznie daje 112 godzin roboczych do rozdysponowania. W przypadku 3 pracowników przy równym podziale pracy każdy z nich przepracuje niecałe 38h tygodniowo, co w przybliżeniu odpowiada pełnemu etatowi. Można jednak zauważyć, że przy takiej liczbie pracowników któryś z nich byłby skazany na 4-godzinne okno pomiędzy 12 a 16 w dni robocze, co zazwyczaj nikomu nie odpowiada. Nie chcemy również, żeby ktokolwiek pracował więcej niż 8h dziennie. Przy 4 pracownikach niemożliwe jest ułożenie grafiku tak, żeby równocześnie nikt nie miał wspomnianego wyżej okna i równocześnie nikt nie pracował przez 3 bloki, czyli 12 godzin. Rozsądną liczbą pracowników przy takich założeniach będzie 5.

Dni robocze planujemy tak, że w kolejnych blokach godzinowych pracują odpowiednio

- 8:00-12:00: A,B
- 12:00-16:00: B
- 16:00-20:00: C,D,E
- 20:00-24:00: D,E

W weekendy potrzebny jest tylko 1 pracownik. Przykładowo w oba dni od 8:00 do 16:00 może być to pracownik A, a od 16:00 do 24:00 pracownik C.