Uniwersytet Ekonomiczny we Wrocławiu  
ul. Komandorska 118/120  
53-345 Wrocław

Michał Trznadel (189405)  
Adam Wasilewski (189409)

**Projekt zaliczeniowy**

Przedmiot: Statystyczna analiza danych  
Prowadzący:  
dr hab. Marcin Pełka  
dr Anna Sulima  
Kierunek: Informatyka w biznesie

Wrocław 19.05.2023r.

# Regresja liniowa

Dla regresji liniowej dane zostały dobrane ze strony <https://stat.gov.pl/>. Do wykonania regresji liniowej wykorzystane zostały dane:

* Liczba ludności w wieku 20-24 lata w roku 2021 z podziałem na 16 województw.
* Przeciętne miesięczne wynagrodzenie brutto w roku 2021 z podziałem na 16 województw.
* Liczba nowoutworzonych miejsc pracy w roku 2021 z podziałem na 16 województw.

Celem stworzenia tego modelu było poszukiwanie zależności między objaśnianą (ilością ludzi w wieku 20-24 lata), a objaśniającymi (przeciętne wynagrodzenie miesięczne brutto, liczba nowoutworzonych miejsc pracy). Po wykonaniu budowy modelu liniowego w języku R funkcją lm() otrzymaliśmy wyniki:

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

W wyników przedstawionych powyżej widać, że wynagrodzenia mają ujemny wpływ na ludność, zaś liczba miejsc pracy ma silny wpływ dodatni. Wartość t-value dla wszystkich zmiennych jest stosunkowo daleko od zera co może wskazywać na istnienie związku między nimi. Wartość p-value jest mniejsza od 0,05 co upewnia nas, że istnieje związek między zmiennymi. R-kwadrat przedstawia miarę dopasowania modelu do rzeczywistych danych. W naszym przykładzie R2=0,9477 przez co możemy przyjąć że wyjaśnia obserwowaną wariancję zmiennej odpowiedzi.

# Regresja logistyczna

Dla regresji logistycznej dane zostały dobrane ze strony <https://stat.gov.pl/>. Do wykonania regresji logistycznej wykorzystane zostały dane:

* Liczba zgonów wg płci kobiet ogółem w 2021 roku z podziałem na województwa.
* Liczba zgonów wg płci mężczyzn ogółem w 2021 roku z podziałem na województwa.
* Liczba kobiet ogółem w 2021 roku z podziałem na województwa.
* Liczba mężczyzn ogółem w 2021 roku z podziałem na województwa.

Następnie została dołączona do wyników kolumna z wartościami 0 – kobieta, 1 – mężczyzna dla danych pobranych z GUS-u. Kolejnym krokiem było połączenie danych liczby zgonów i liczby ludzi w jedną tabelę, a następnie wykonanie regresji logistycznej funkcją glm()

Obraz zawierający tekst, zrzut ekranu, Czcionka, dokument

Opis wygenerowany automatycznie

Wzrost predykcyjnej liczba zgonów o jedną jednostkę wiąże się ze średnią zmianą o   
9,010\*10-4 w dzienniku szans zmiennej odpowiedzi płeć na wartość 1, w przypadku wzrost ludności o jedną jednostkę wiąże się ze średnią zmianą o -1,201\*10-5 w dzienniku szans zmiennej odpowiedzi płeć na wartość 1. Zmienne liczba zgonów i ludność są statystycznie istotne, ponieważ wyliczone dla nich wartości p są mniejsze od 0,05.

# Analiza głównych składowych

Do analizy głównych składowych dane zostały pobrane ze strony <https://stat.gov.pl/>. Do wykonania analizy głównych składowych wykorzystane zostały dane:

* Wysokość stopy bezrobocia w Polsce w latach 2010 – 2015
* Liczba nowoutworzonych miejsc pracy w Polsce w latach 2010 – 2015
* Liczba przedszkoli w Polsce w latach 2010 – 2015
* Liczba absolwentów szkół wyższych w latach 2010 – 2015

Obraz zawierający tekst, Czcionka, zrzut ekranu

Opis wygenerowany automatycznie

|  |  |
| --- | --- |
|  |  |

Pierwsza składowa wyjaśnia 79,44% wariacji wspólnej wszystkich zmiennych, druga składowa wyjaśnia 99,99%, a trzecia i czwarta składowa wyjaśnia 100% wariacji, dlatego należy przyjąć dwie główne składowe. Z wykresu biplot() widać wysoką korelację dodatnią między stopą bezrobocia, a przedszkolami, można zauważyć, że nowoutworzone miejsca pracy i absolwenci tworzą spory kąt rozwarty co wskazuje na ujemną korelację między nimi, dodatkowo na wykresie widać, że nowe miejsca pracy mają największy wpływ na stopę bezrobocia (comp.1), na nowe miejsca pracy (comp.2) największy wpływ ma liczba absolwentów.

# Analiza skupień

Do analizy skupień dane zostały pobrane ze strony <https://stat.gov.pl/>. Do analizy wykorzystano metodę spektralną. Wykorzystane zostały dane:

* Liczba ludności w poszczególnych województwach w roku 2020,
* Zaległości w opłatach za czynsz i inne opłaty mieszkaniowe w poszczególnych województwach w roku 2020,
* Koszty eksploatacji mieszkań w poszczególnych województwach w roku 2020,
* Powierzchnia mieszkań w poszczególnych województwach w roku 2020

Obraz zawierający tekst, wyświetlacz, zrzut ekranu, oprogramowanie

Opis wygenerowany automatycznieObraz zawierający tekst, zrzut ekranu, wyświetlacz, oprogramowanie

Opis wygenerowany automatycznie

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

Metoda spektralna podzieliła województwa na 2 klasy. Do pierwszej klasy należą województwa Łódzkie, Mazowieckie i Śląskie. Klasa pierwsza charakteryzuje się wysokimi zaległościami w płatnościach do wspólnot mieszkaniowych, wysoką powierzchnią mieszkań i wysokimi kosztami eksploatacji. Do drugiej kategorii należą mieszkania o nieco niższych wartościach tych trzech zmiennych. Na wykresie widoczne jest oddalenie wartości dla klasy pierwszej.