

## ĆWICZENIE 6

### REGRESJA DLA NIECIĄGŁEJ ZMIENNEJ OBJAŚNIANEJ

#### Cel

Zapoznanie z modelem regresji nieliniowej. Wykonanie przykładowych obliczeń dotyczących modelu regresji nieliniowej dla danych dyskretnych ze szczególnym uwzględnieniem poprawnej interpretacji wyników.

#### Wprowadzenie teoretyczne

Regresja dla dyskretnej zmiennej objaśnianej jest w pewnym sensie podobna do regresji liniowej. Podstawową różnicę stanowi tutaj właśnie fakt, że zmienna objaśniana nie jest zmienną ciągłą. Regresja ta może być wykorzystywana do prognozowania zmiennej wynikowej, która przyjmuje dwie lub więcej wartości wynikowych. W celu wykorzystania tej regresji, dyskretna zmienna objaśniana jest zamieniana na zmienną ciągłą poprzez funkcję prawdopodobieństwa.

Najczęściej wykorzystywanymi w praktyce modelami z dyskretną zmienną objaśnianą są:

- model logistyczny („logit”),
- model probitowy („probit”),
- model „log-log”.

Pozwalają one opisać zależności między częstościami występowania poszczególnych wariantów zmiennej objaśnianej a wybranymi zmiennymi objaśniającymi.

W praktyce często mamy do czynienia z obserwacjami zmiennej objaśnianej, które można podzielić tylko na dwie kategorie. Wynik każdej obserwacji  $y_1, y_2, \dots, y_n$  może być interpretowany jako sukces lub porażka. Wówczas zmienne  $y_1, y_2, \dots, y_n$  nazywane są **obserwacjami binarnymi**, **danymi binarnymi** lub **reakcjami binarnymi**. Przyjmuje się, że  $y_1, y_2, \dots, y_n$  mają rozkład Bernoulliego  $B(1, p_i)$ , gdzie  $p_i$  można interpretować jako prawdopodobieństwo sukcesu. Rozkład obserwacji  $y_1, y_2, \dots, y_n$  jest określony za pomocą funkcji prawdopodobieństwa

$$f(y_i, p_i) = p_i^{y_i} (1 - p_i)^{1 - y_i}.$$

**Model logistyczny** jest postaci

$$\log \frac{p}{1 - p} = \alpha_0 + \alpha_1 x_1 + \dots + \alpha_k x_k,$$

gdzie  $x_1, x_2, \dots, x_k$  są zmiennymi objaśniającymi,  $\alpha_1, \alpha_2, \dots, \alpha_k$  są parametrami modelu. Ponadto zachodzi związek

$$p = \frac{\exp(\alpha_0 + \alpha_1 x_1 + \dots + \alpha_k x_k)}{1 + \exp(\alpha_0 + \alpha_1 x_1 + \dots + \alpha_k x_k)}.$$

**Model probitowy** jest postaci

$$\Phi^{-1}(p) = \alpha_0 + \alpha_1 x_1 + \dots + \alpha_k x_k,$$

gdzie  $\Phi$  oznacza dystrybuantę rozkładu normalnego  $N(0,1)$ .

**Model „log-log”** jest postaci

$$\log(-\log(1 - p)) = \alpha_0 + \alpha_1 x_1 + \dots + \alpha_k x_k.$$

### Zadanie do wykonania

1. Plik *dane.txt* zawiera informacje o stanie cywilnym, wieku, płci i statusie zawodowym dwustu czterech osób. Stosując model regresji logistycznej:
  - sprawdzić zależność statusu zawodowego od pozostałych zmiennych,
  - sprawdzić zależność stanu cywilnego od pozostałych zmiennych,
  - obliczyć prawdopodobieństwo, że osoba będąca tzw. singlem jest osobą niepracującą,
  - obliczyć prawdopodobieństwo, że badana osoba jest kobietą i nie pracuje,
  - obliczyć prawdopodobieństwo, że osoba w wieku 43 lat jest stanu wolnego.

Sposób kodowania danych w pliku *dane.txt* jest następujący:

- st.cywilny: 1 – w związku, 0 – singiel,
- plec: 1 – kobieta, 0 – mężczyzna,
- praca: 1 – osoba ucząca się lub pracująca, 0 – osoba nie pracuje.

Źródła:

Magiera R. „Modele i metody statystyki matematycznej” GiS, Wrocław 2002