10.1 Interpretacja modelu, testowanie, selekcja zmiennych.

Zbiór danych *SAheart.data* (South African Heart Disease) zawiera dane dotyczące zapadalności na zawał serca wśród mężczyzn pomiędzy 15 a 64 rokiem życia. Zmienna **chd** oznacza że wystąpił (watość 1) lub nie wystąpił (wartość 0) zawał serca.

- Dopasuj model regresji logistycznej.
- Które zmienne są istotne statystycznie w modelu pełnym?
- Oblicz iloraz szans (ang. odds ratio) w modelu logistycznym w przypadku kiedy wartości wszystkich zmiennych są ustalone, natomiast zwiększamy wiek pacjenta o jeden rok.
- Używając metody eliminacji wstecznej z kryterium AIC oraz BIC dokonać selekcji zmiennych (funkcja `step`).
- Używając testu dewiancyjnego, określ, czy w modelu istnieją zmienne istotne.
- Dopasuj model z dodatkową zmienną age^2. Używając testu dewiancyjnego określ, czy zmienna age^2 jest istotna.

10.2 Problem liniowej separowalności klas. Dane *earthquake.txt.* dotyczą klasyfikacji wstrząsów (zmienna **popn**) na podstawie danych sejsmologicznych (zmienne **body** i **surface**).

- Wykonaj wykres rozproszenia dla zmiennych **body** i **surface** z zaznaczeniem przynależności do klas.
- Dopasuj model regresji logistycznej opisujący zależność zmiennej **popn** od zmiennych **body** i **surface**. Jak wyjaśnić fakt że p-wartości statystyk Walda wskazują na nieistotność zmiennych objaśniających?

10.3 Przykład symulacyjny.

- Wygeneruj dane z modelu logistycznego

$$y_i \sim Bern(p_i),$$

gdzie

$$p_i = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2})]},$$

dla $i = 1, \ldots, n$, $x_{1,i}$, $x_{2,i} \sim N(0,1)$, $n = 50$. Parametry $\beta_0 = 0.5$, $\beta_1 = \beta_2 = 1$. Dopasuj model logistyczny dla wygenerowanych danych i oblicz estymatory współczynników. Powtórz eksperyment $L = 50$ razy i na tej podstawie oszacuj błąd średniokwadratowy

$$MSE = E(||\hat{\beta} - \beta||^2),$$

gdzie $||\cdot||$ jest normą euklidesową.

- Powtórz eksperyment dla $n = 50, 60, 70, \ldots, 300$ i narysuj wykres pokazujący zależność MSE od n.