Regresja logistyczna

Elżbieta Kukla

20 lutego 2011

Katatstrofa Challenger'a

Katastrofa Challenger'a

Co nas interesuje?

- Interesuje nas prawdopodobieństwo uszkodzenia pierścieni w zależności od temperatury powietrza w chwili startu wahadłowca i obliczenie tego prawdopodobieństwa, gdy temperatura wynosi $31^o F$.
- Najprostsze podejście oparte na modelu liniowym generuje wiele kłopotów.

Inne podejście

- Niech $Y_i \sim B(n_i, p_i)$, $i = 1, \dots, n$, tzn.

$$P(Y_i = y_i) = \binom{n_i}{y_i} p_i^{y_i}(1 - p_i)^{n_i - y_i}.$$

- Zakładamy, że Y_i są niezależne.
- Y_i zależy od q predyktorów $(x_{i1}, \dots, x_{iq})$.
- Poszukujemy modelu, który opisuje związek $x_1, \dots, x_q$ z p.

Inne podejście

- Postępując tak jak w modelu liniowym, konstruujemy liniowy predyktor:
$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_q x_{iq}.$$
- Ustalenie $\eta_i = p_i$ nie jest odpowiednie – chcemy, by $0 \leqslant p_i \leqslant 1$.
- Używamy funkcji wiążącej g takiej że $\eta_i = g(p_i)$.
- Poszukujemy funkcji g – monotonicznej i takiej że $0 \leqslant g^{-1}(\eta) \leqslant 1$ dla dowolnego η.

Trzy popularne fukncje

Mamy trzy popularne funkcje:

1. Logit: $\eta = \log(\frac{p}{1-p})$ (regresja logistyczna).
2. Probit: $\eta = \phi^{-1}(p)$, gdzie ϕ jest dystrybuantą rozkładu normalnego (regresja probitowa).
3. *complementary log-log*: $\eta = \log(-log(1-p))$.

Idea użycia funkcji wiążącej jest jedną z głównych idei uogólnionych modeli liniowych (GLM).

Regresja logistyczna

Rozkład Bernoulliego:

$$P(Y_i = y_i) = f_i(y_i) = \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i}$$

$$\log f_i(y_i) = y_i \log(p_i) + (n_i - y_i) \log(1 - p_i) + \log \binom{n_i}{y_i}$$

$$\log f_i(y_i) = y_i \log(\frac{p_i}{1 - p_i}) + n_i \log(1 - p_i) + \log \binom{n_i}{y_i}$$

Mamy rodzinę wykładniczą, bo powyższe wyrażenie ma postać:

$$\log f_i(y_i) = \frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi).$$

Regresja logistyczna

$$\log f_i(y_i) = \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)$$

$$\log f_i(y_i) = y_i \log(\frac{p_i}{1-p_i}) + n_i \log(1-p_i) + \log \binom{n_i}{y_i}$$

Zauważamy, że:

- $\theta_i = \log(\frac{p_i}{1-p_i}) = \eta_i \qquad \text{(logit)}$
- $p_i = \frac{e^{\theta_i}}{1+e^{\theta_i}}$ oraz $1 - p_i = \frac{1}{1+e^{\theta_i}}$
- $b(\theta_i) = n_i \log(1 + e^{\theta_i})$
- $c(y_i, \phi) = \log \binom{n_i}{y_i}$
- $a_i(\phi) = \phi$ i $\phi = 1$
- $E(Y_i) = \mu_i = b'(\theta_i) = n_i \frac{e^{\theta_i}}{1+e^{\theta_i}} = n_i p_i$
- $Var(Y_i) = v_i = a_i(\phi) b''(\theta_i) = n_i \frac{e^{\theta_i}}{(1+e^{\theta_i})^2} = n_i p_i (1-p_i)$

Estymacja metodą największej wiarogodności

Parametry modelu szacujemy używając metody największej wiarogodności.
Logarytm funkcji wiarogodności jest dany przez:

$$l(\beta) = \sum_{i=1}^{n} [y_i \eta_i - n_i \log(1 + e^{\eta_i}) + \log \binom{n_i}{y_i}].$$

Maksymalizujemy to wyrażenie w celu otrzymania estymatorów $\hat{\beta}$ korzystając z metody Fisher scoring.

orings {faraway}

Space Shuttle Challenger O-rings

Description

The 1986 crash of the space shuttle Challenger was linked to failure of O-ring seals in the rocket engines. Data was collected on the 23 previous shuttle missions. The launch temperature on the day of the crash was 31F.

Usage

data(orings)

Format

A data frame with 23 observations on the following 2 variables.

temp – temperature at launch in degrees F

damage – number of damage incidents out of 6 possible

Nasz model

Nasz model:

$$\eta_i = \log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_i.$$

gdzie
x_i – temperatura powietrza (zmienna objaśniająca).