

Regresja logistyczna

Regresja logistyczna, inaczej model logitowy jest to dopasowanie krzywej logistycznej $1 / (1 + \exp(-a - b \cdot x))$ do danych postaci – np. $(s, x) = (\text{ślepy/zdrowy}, \text{wiek})$. Parametry a i b wyznaczone są z zasady największej wiarygodności. Postępowanie takie, oznaczmy je (1), może być rozumiane jako poprawa prostych metod przybliżania p-stwa ślepoty jako funkcji wieku. Najprostsza jest chyba następująca metoda (2):

- podziel dane na grupy wiekowe,
- oszacuj p-stwa ślepoty w grupach,
- wykreśl łamaną na płaszczyźnie (wiek, p-stwo ślepoty)

Metoda (2) nie wykorzystuje pełnej informacji, bo kwantuje wiek - arbitralnie dzieli dane na klasy i interpoluje krzywą między klasami. Zamiast prymitywnej interpolacji możemy dopasowywać do skwantowanych danych krzywą logistyczną za pomocą metody najmniejszych kwadratów. Na tym polega następująca metoda (3). Niech x_i oznacza wiek, n_i – liczbę obserwacji oraz p_i - częstość występowania ślepoty w klasie $i=1, \dots, 5$. Przyjmijmy ponadto $y_i = \log(p_i / (1 - p_i))$. Wykorzystując linearyzację tak jak w metodzie delta otrzymujemy $\text{var}(y_i) \sim 1 / (n_i \cdot p_i \cdot (1 - p_i))$. Mając niezależne obserwacje y_i o znanych wariancjach, możemy teraz wyznaczyć estymatory najmniejszych kwadratów a i b minimalizując sumę ważonych kwadratów (napisze wektorowo jak w R): $\text{sum}(n_i \cdot p_i \cdot (1 - p_i) \cdot (y_i - a - b \cdot x_i)^2)$.

Rozwiązanie (3) jest analityczne i bardzo bliskie rozwiązaniu wyznaczonemu z zasady największej wiarygodności dla skwantowanych danych - tak były rozwiązywane modele logistyczne zanim nie pojawiło się ogólne sformułowanie estymacji największej wiarygodności dla uogólnionych modeli liniowych, w których maksymalizuje się łączną wiarygodność dla surowych 0/1 danych. Logistyczna funkcja odpowiedzi jest w tym przypadku otrzymana z uniwersalnej zasady przy wykorzystaniu pełnej informacji.