Metody nieparametryczne w statystyce

Tomasz Wójtowicz

Wydział Zarządzania
AGH Akademia Górniczo-Hutnicza im. S. Staszica w Krakowie

Regresja logistyczna

Regresja logistyczna

Rozważmy model

$$y = f(x_1, x_2, \dots, x_k)$$

gdzie y jest zmienną jakościową.

Jeżeli y przyjmuje dwie wartości, np. y_1 i y_2, to możemy zapisać:

$$y = \begin{cases} 0 & \text{gdy } y_1 \\ 1 & \text{gdy } y_2 \end{cases}$$

Czy można wtedy rozważać model linowy:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k?$$

Regresja logistyczna

Model linowy:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k?$$

Wady:

- interpretacja wartości jako prawdopodobieństw,
- duża wrażliwość modelu na zmiany danych,
- problemy, gdy y przyjmuje więcej niż dwie wartości.

Lepiej rozważyć model, w którym $f(x_1, x_2, \dots, x_k)$ przyjmuje tylko wartości z przedziału $[0,1]$.

Regresja logistyczna

Rozważmy najprostszą sytuację z jedną zmienną objaśniającą:

$$y = f(x)$$

Interesuje nas model opisujący następujące prawdopodobieństwo:

$$P(y = 1|x)$$

Oznaczmy je przez

$$p(x)$$

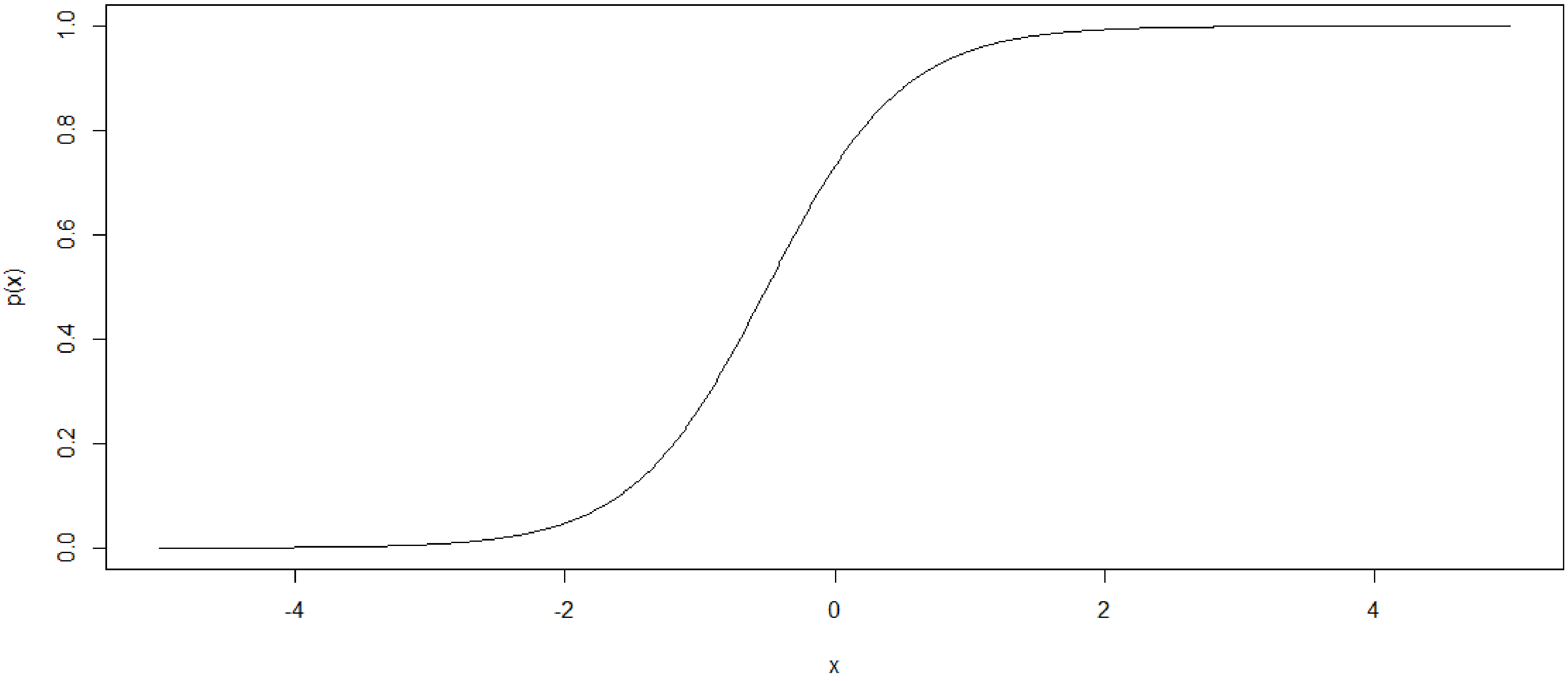
Przyjmijmy, że ma prawdopodobieństwo $p(x)$ ma postać:

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

czyli jest opisane za pomocą funkcji logistycznej.

Regresja logistyczna

Przykład:



Regresja logistyczna

Po przekształceniach otrzymujemy:

$$\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x}$$

Wielkość

$$\frac{p(x)}{1-p(x)}$$

nazywamy **szansą** (odd), że y przyjmie wartość 1.

Przykład:

Jeżeli $p(x) = 0{,}2$ to szansa jest równa $1:4$.

Jeżeli $p(x) = 0{,}9$ to szansa jest równa $9:1$

Regresja logistyczna

Równoważnie możemy zapisać:

$$\ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x$$

Wartość po lewej stronie: *logit*.

Zależności:

$$p(x) \in (0,1)$$

$$\frac{p(x)}{1-p(x)} \in (0,+\infty)$$

$$\ln\left(\frac{p(x)}{1-p(x)}\right) \in (-\infty,+\infty)$$

Regresja logistyczna

Interpretacja parametrów modelu:

- funkcję logistyczną można zapisać w postaci:
$$\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} = \frac{1}{1+e^{-\beta_0}+e^{-\beta_1 x}}$$
- interpretuje się głównie β_1,
- jeżeli $\beta_1 > 0$, to x ma stymulujący wpływ na y (zwiększa prawdopodobieństwo przyjęcia wartości 1),
- jeżeli $\beta_1 < 0$, to x ma ograniczający wpływ na y (zwiększa prawd. przyjęcia wartości 1),
- jeżeli $\beta_1 = 0$, to x nie wpływa na y.

Regresja logistyczna

Estymacja parametrów: metoda największej wiarygodności.

Wiemy, że dla danego x:

$y = 1$ z prawdopodobieństwem $p(x)$

$y = 0$ z prawdopodobieństwem $1 - p(x)$

Jeżeli mamy dane: $(x_1, y_1), \dots, (x_n, y_n)$ to funkcja wiarygodności ma postać:

$$l(\beta_0, \beta_1) = l(x_1, \dots, x_n, y_1, .., y_n; \beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \cdot \prod_{i:y_i=0} \big(1 - p(x_i)\big)$$

$$l(\beta_0, \beta_1) = \prod_{i=1}^{n} \big(p(x_i)\big)^{y_i} \big(1 - p(x_i)\big)^{1-y_i}$$

Regresja logistyczna

Niech

$$L(\beta_0, \beta_1) = \ln\big(l(\beta_0, \beta_1)\big)$$

wtedy:

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n} y_i \cdot \ln\big(p(x_i)\big) + (1 - y_i) \cdot \ln\big(1 - p(x_i)\big)$$

Po wyestymowaniu parametrów modelu możemy m.in. testować ich istotność.

W szczególności:

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

Regresja logistyczna

Regresję logistyczną możemy zastosować również w przypadku większej liczby zmiennych objaśniających. Wtedy:

$$\ln\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k = X\beta$$

czyli

$$p(X) = \frac{e^{X\beta}}{1+e^{X\beta}}$$

Regresja logistyczna

Przykład – karty kredytowe

dane: Default z pakietu ISLR