

ロジスティック回帰分析

ある現象が起こる確率 P を

$$P = \frac{1}{1 + \exp(- (a_0 + a_1x_1 + a_2x_2 + \cdots + a_mx_m))} \quad \dots \textcircled{1}$$

とする。

$$\begin{aligned} 1 - P &= 1 - \frac{1}{1 + \exp(- (a_0 + a_1x_1 + a_2x_2 + \cdots + a_mx_m))} \\ &= \frac{\exp(- (a_0 + a_1x_1 + a_2x_2 + \cdots + a_mx_m))}{1 + \exp(- (a_0 + a_1x_1 + a_2x_2 + \cdots + a_mx_m))} \quad \dots \textcircled{2} \end{aligned}$$

となるので、

$$\begin{aligned} \frac{P}{1-P} &= \frac{1}{\exp(- (a_0 + a_1x_1 + a_2x_2 + \cdots + a_mx_m))} \\ &= \exp(a_0 + a_1x_1 + a_2x_2 + \cdots + a_mx_m) \quad \dots \textcircled{3} \end{aligned}$$

従って、

$$P = (1 - P)\exp(a_0 + a_1x_1 + a_2x_2 + \cdots + a_mx_m)$$

となり、

$$\begin{aligned} P(1 + \exp(a_0 + a_1x_1 + a_2x_2 + \cdots + a_mx_m)) \\ = \exp(a_0 + a_1x_1 + a_2x_2 + \cdots + a_mx_m) \end{aligned}$$

となるので、

$$p = \frac{\exp(a_0 + a_1x_1 + a_2x_2 + \cdots + a_mx_m)}{1 + \exp(a_0 + a_1x_1 + a_2x_2 + \cdots + a_mx_m)} \quad \dots \textcircled{4}$$

となる。

③式より $0 < P < 1$ であれば

$$\ln\left(\frac{p}{1-p}\right) = (a_0 + a_1x_1 + a_2x_2 + \cdots + a_mx_m)$$

($\ln(x)$ は x の自然対数である)

でありそのまま最小二乗法を適用して計算できる。

しかし、 $P = 0$ 、 $P = 1$ が起こりうる場合は、そのまま適用できない。

n 個の確率値を P_i ($i = 1 \dots n$) とし、

$$[X] = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,m} \end{bmatrix} \quad \{a\} = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{pmatrix}$$

$$[X] \text{ の第 } i \text{ 行を } \{x_i\}^t = \{1 \quad x_{i,1} \quad x_{i,2} \quad \cdots \quad x_{i,m}\}$$

とすると、④より

$$P_i = \frac{\exp(\{x_i\}^t \{a\})}{1 + \exp(\{x_i\}^t \{a\})} \quad \dots \textcircled{5}$$

と表現できる。

$$\begin{aligned} y_i = 1 & \text{ の時の 確率を } P_i \\ y_i = 0 & \text{ の時の 確率を } (1 - P_i) \end{aligned}$$

とする。 合わせて表現すると、

$$Q_i = P_i^{y_i} (1 - P_i)^{(1-y_i)}$$

$$L(a_0, a_1, \dots, a_m) = \prod_{i=1}^n Q_i = \prod_{i=1}^n P_i^{y_i} (1 - P_i)^{(1-y_i)}$$

となり、

$$\begin{aligned} \ln(L(a_0, a_1, \dots, a_m)) &= \ln\left(\prod_{i=1}^n P_i^{y_i} (1 - P_i)^{(1-y_i)}\right) \\ &= \sum_{i=1}^n y_i \ln(P_i) + \sum_{i=1}^n (1 - y_i) \ln(1 - P_i) \end{aligned} \quad \dots \textcircled{6}$$

この $\ln\{L(a_0, a_1, \dots, a_m)\}$ を最大にする a_0, a_1, \dots, a_m を探す。

$$\textcircled{5} \text{ より } P_i = \frac{\exp(\{x_i\}^t \{a\})}{1 + \exp(\{x_i\}^t \{a\})} \quad \text{なので、}$$

$$\ln(P_i) = \ln\left(\frac{\exp(\{x_i\}^t \{a\})}{1 + \exp(\{x_i\}^t \{a\})}\right) = \{x_i\}^t \{a\} - \ln(1 + \exp(\{x_i\}^t \{a\}))$$

$$\begin{aligned}\ln(1 - P_i) &= \ln\left(1 - \frac{\exp(\{\boldsymbol{x}_i\}^t \{a\})}{1 + \exp(\{\boldsymbol{x}_i\}^t \{a\})}\right) = \ln\left(\frac{1}{1 + \exp(\{\boldsymbol{x}_i\}^t \{a\})}\right) \\ &= \ln(1) - \ln(1 + \exp(\{\boldsymbol{x}_i\}^t \{a\})) = -\ln(1 + \exp(\{\boldsymbol{x}_i\}^t \{a\}))\end{aligned}$$

従って

$$\begin{aligned}A_i &\equiv \boldsymbol{y}_i \ln(P_i) = \boldsymbol{y}_i \left(\{\boldsymbol{x}_i\}^t \{a\} - \ln(1 + \exp(\{\boldsymbol{x}_i\}^t \{a\})) \right) \\ &= \boldsymbol{y}_i \{\boldsymbol{x}_i\}^t \{a\} - \boldsymbol{y}_i \ln(1 + \exp(\{\boldsymbol{x}_i\}^t \{a\}))\end{aligned}$$

$$\begin{aligned}B_i &\equiv (1 - \boldsymbol{y}_i) \ln(1 - P_i) = (\boldsymbol{y}_i - 1) \left\{ \ln(1 + \exp(\{\boldsymbol{x}_i\}^t \{a\})) \right\} \\ &= \boldsymbol{y}_i \ln(1 + \exp(\{\boldsymbol{x}_i\}^t \{a\})) - \ln(1 + \exp(\{\boldsymbol{x}_i\}^t \{a\}))\end{aligned}$$

とすると、

$$A_i + B_i = \boldsymbol{y}_i \{\boldsymbol{x}_i\}^t \{a\} - \ln(1 + \exp(\{\boldsymbol{x}_i\}^t \{a\}))$$

となり 目的関数の対数尤度は

$$\begin{aligned}\ln\{L(\boldsymbol{a}_0, \boldsymbol{a}_1, \dots, \boldsymbol{a}_m)\} &= \sum_{i=1}^n (A_i + B_i) \\ &= \sum_{i=1}^n \left(\boldsymbol{y}_i \{\boldsymbol{x}_i\}^t \{a\} - \ln(1 + \exp(\{\boldsymbol{x}_i\}^t \{a\})) \right)\end{aligned}$$

となる。

対数尤度を用いて、最適解 $\boldsymbol{a}_0 \sim \boldsymbol{a}_m$ を求める解法は、ここでは

- ・ ニュートン法
- ・ *BFGS* 法による準ニュートン法

の2法を用いている。

オッズ、 オッズ比 について

説明変数が $x_1, x_2 \cdots x_m$ とし、0 または 1 の2値データを取る目的変数 Y について、 $Y=1$ である確率を P_0 としたとき、③により以下が成り立つ。

$$\frac{P_0}{1-P_0} = \exp(a_0 + a_1x_1 + a_2x_2 + \cdots + a_mx_m)$$

ここで、上記の説明変数のうちの1つの x_1 が1増加し、 $x_2 \cdots x_m$ については変わらないという場合で、 $Y=1$ である確率を P_1 とすると、やはり ③により以下が成り立つ。

$$\frac{P_1}{1-P_1} = \exp(a_0 + a_1(x_1 + 1) + a_2x_2 + \cdots + a_mx_m)$$

ともに対数を取ると、

$$\ln \frac{P_0}{1-P_0} = a_0 + a_1x_1 + a_2x_2 + \cdots + a_mx_m$$

$$\ln \frac{P_1}{1-P_1} = a_0 + a_1x_1 + a_1 + a_2x_2 + \cdots + a_mx_m$$

であるから

$$\ln \frac{P_1}{1-P_1} - \ln \frac{P_0}{1-P_0} = a_1 \quad \text{つまり} \quad \ln \frac{P_1/(1-P_1)}{P_0/(1-P_0)} = a_1 \quad \text{より} \quad \frac{P_1/(1-P_1)}{P_0/(1-P_0)} = \exp(a_1)$$

となる。

$\frac{P_0}{1-P_0}$ を $x_1, x_2 \cdots x_m$ の時のオッズ、 $\frac{P_1}{1-P_1}$ を $x_1 + 1, x_2 \cdots x_m$ の時のオッズという。

そして、2つのオッズの比 $\frac{P_1/(1-P_1)}{P_0/(1-P_0)}$ を オッズ比といい、この値が1より大きいほど、

説明変数の目的変数に対する影響が大きいことを 意味する。

上記の例では、説明変数 x_1 のカテゴリ値が 0 から 1 への変化に伴って、オッズが

オッズ比 ($\frac{P_1/1-P_1}{P_0/1-P_0}$) 倍となる。

オッズ比が1より大きければ、目的変数が1になる確率が高くなることを示し、逆にオッズ比が1より小さければ、目的変数が1になる確率が低くなることを示している。