2項ロジスティック回帰分析 1

渡邊直樹

2021 年 9 月 29 日

2 項ロジスティック回帰モデル

2 項ロジスティク回帰（binomial logistic regression）とは，商品を購入するか否か，離職するか否かといった 2 つのカテゴリーに分けられる被説明変数を自然対数を使って変換し，それを他のいくつかの説明変数の加重和によって表すことで，被説明変数と説明変数の間に想定される仮説を検証するための統計手法である．説明変数に関する新たなデータが与えられたとき，既存データを用いて推定されたモデルに基づいて，どちらのカテゴリーが選択されるかに関する確率を求めることができる．本稿では，フォーマルな定式化ではなく，Excel で 2 項ロジスティック回帰を行う際に必要な操作を説明する．

m 個の説明変数について，t 人分データがあるとする．このとき，個人 i の意思決定 Y_i に関するロジスティック回帰モデルは

$$\ln \frac{Y_i}{1-Y_i} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_m X_{mi} + u_i \tag{1}$$

と書ける．$\ln(x)$ は x のネイピア数を底とする対数関数である．ここで，u_i は個人 i の特異性に起因する誤差項であり，説明変数の加重和では説明できない Y_i の変動はこの誤差項の確率的な変動によるものと考えよう．たとえば，個人 i がある商品を購入しないことを $Y_i = 0$，購入することを $Y_i = 1$ で表すとしよう．この $Y_i/(1-Y_i)$ をオッズ比（odds ratio）という[1]．実際には，$Y_i = 1$ のときにはオッズ比の分母が 0 になってしまうので，推計式では $Y_i = 1$ のときには 1 に非常に近い数値（たとえば，0.999）で代用する．

[1] $Y_i/(1-Y_i)$ をオッズと呼んでもよいが，実際の分析では（後述のように）0 と 1 の値をそのまま使うことはできず，近似確率として 0.0001 や 0.999 を使用しているので，ここではそれらを 2 群における生起確率とみなしてオッズ比と呼んでいる．

オッズ比の自然対数をとった値を使って，(1) 式を線形回帰することでパラメータの推定値 $\hat{\beta}_k$（$k = 1, \ldots, m$）を求める．Excel における実数 x の自然対数を求める関数は LN(x) である．各推計式における個人 i に関する残差（residual）は

$$e_i = \ln \frac{Y_i}{1 - Y_i} - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_m X_{mi}) \tag{2}$$

である．

Excel のアドインに組み込まれている分析ツール「回帰分析」には各係数の推定値の p 値は表示されるが，AIC の値は計算されない．ここで，n はデータを取得した個人の数，k_j はモデル j において推定される係数の数であるとしよう．モデル j の AIC は次の式で計算される．

$$\mathrm{AIC}_j = n \ln(\frac{\sum e_i^2}{n}) + 2k_j. \tag{3}$$

モデル j における個人 i の残差 e_i は，Excel の分析ツール「回帰分析」の残差を計算するオプションに該当するチェックボックスにチェックを入れることで，計算結果が出力される．それを使って残差自乗和を計算し，自然対数を取るとよい．セルに $=16*\mathrm{LN}(\text{モデル } j \text{ の残差自乗和}/16)+2*k_j$ と入力すると，AIC_j が計算される．正確な選択確率の導出には適切なモデルの選択が欠かせない．AIC の値がより低いモデルを選択することは適切なモデルを選択するための基準の一つである．

モデルが確定すると，個人 i' が $Y_{i'} = 1$ を選択する確率は

$$y_{i'} = \hat{\beta}_0 + \hat{\beta}_1 X_{1i'} + \hat{\beta}_2 X_{2i'} + \cdots + \hat{\beta}_m X_{mi'} \tag{4}$$

から，自然対数の定義より，次のようにして求められる．

$$\mathrm{Prob}(Y_{i'} = 1) = \frac{\mathrm{EXP}(y_{i'})}{1 + \mathrm{EXP}(y_{i'})}. \tag{5}$$

ここで，EXP(x) はネイピア数 e を底とする指数関数を表す Excel の記法である[2]．以上の数値を Excel で計算するのは，確かに面倒．ほとんどの統計パッケージではロジスティック回帰を行う際に LN と EXP の計算を自動実行してくれる．(Easy R では，マウスでのクリックだけで R を操作できる．)

[2] $y_{i'} = \ln \frac{Y_{i'}}{1-Y_{i'}}$ より，$e^{y_{i'}} = \frac{Y_{i'}}{1-Y_{i'}}$ なので，これを変形して，$(1 + e^{y_{i'}})Y_{i'} = e^{y_{i'}}$ となることを確認してほしい．