

研究レポート No.3 ～ロジスティック回帰分析～

2011年10月4日 株式会社アイズファクトリー <http://www.isfactory.co.jp/>

概要

ロジスティック回帰分析は、結果が2値の場合に、その結果の起きる確率を予測できる統計的回帰モデルである。例えばダイレクトメールの反応では、結果が反応の有無という2値で測られる。サンプル（DM配送先）には、反応の有無の他に、例えば性別、年齢、収入等の属性情報があるとする。過去実績データにて反応の有無を属性情報で回帰することにより、各サンプルの反応確率を予測するモデルを作成する。このモデルによって、反応の有無が未知である新規データの各サンプルに対して反応確率を予測することができる。本レポートでは、ロジスティック回帰分析についての基本的な解説と、実際の使用上での留意点について紹介する。

1. はじめに

商品案内やカタログ等のDMを送付して販促に役立てようと考へた場合、送付に対する反応（以降、成功と称する）数を最大にするためには全員に送付すれば良い。しかしながら費用対効果を考へた場合は、成功率順に一定の人数に絞り込んでから送付するのが効率的である。では、成功率は計算可能だろうか。このような分析を可能にするのがロジスティック回帰分析である。

ロジスティック回帰では、送付リスト全員に送付するのに対して、成功数をできるだけ減らさずに送付数を絞り込むことができるのでコストダウンに効果である。ロジスティック回帰以外の手法で送付リストを絞り込んで送付していた場合と比べると、成功率順に送付することで成功率が増えるので、売上向上に効果的である。

DM送付の場合、ロジスティック回帰でのモデル作成および予測の手順は次のようになる。まず過去の成功/失敗の結果が必要である。これに個人の属性、例えば性別、年齢、収入等を紐付けることで分析用データを作成する。作成したデータから各サンプルの成功率を予測するロジスティック回帰のモデルを作成する。成功/失敗の付いていない新規の送付リストに対しても、モデルと個人の属性から成功率を求めることができる。モデル作成時に使用した個人属性は、予測を行う送付リスト側でも全て必要となる。

DM送付以外にも、金融機関での個人属性による与信調査や、医療機関での血液データによる疾病の有無等、結果が2値である様々な分野に対して、各サンプルの成功率を求めることができる。

以下では2値変数に対するロジスティック回帰の基本部分についてレビューを行う。3値以上の変数に対してもロジスティック回帰を当てはめることができるが、ここでは扱わない。第2章では線形モデルを拡張した一般化線形モデルでの回帰について説明する。第3章では回帰係数の推定手法として最尤法を説明する。第4章でまとめを行う。

2. 一般化線形モデル

まず線形モデルを考へる。線形モデルとは、

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + e \quad \dots(2.1)$$

の形で書き表されるモデルである。ここで e は平均0、分散 σ^2 の正規分布に従う誤差である。この意味で、目的変数

Y は確率変数であり、平均 μ 分散 σ^2 の正規分布に従い

$$E(Y) = \mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n \quad \dots(2.2)$$

と記述できる。ここで E は期待値を取る記号である。成功/失敗を予測するモデルでは、 Y の観測値 y は成功/失敗に応じてそれぞれ1か0の値を取る。 β は回帰係数であり、最尤法などで計算される。 x は説明変数であり、連続変数もしくは2値化した質的変数である。連続変数とは月収の様に数値の差が意味を持つ変数である。質的変数とは男女の様に大小の比較ができない変数である。質的変数の2値化の例として赤、青、黄の3値を取る変数を考へる。この変数は青か否かで1変数、黄か否かで1変数、赤か否かは他の2色の選択で決まるので、2つの2値変数で表現することができる。

次に線形回帰でサンプルの成功率を計算することを考へる。 μ は Y の期待値であることから、 μ の値は0以上1以下となるべきである。一方、(2.2)式の右辺では x の値次第で0未満や1超となるため、確率を表すモデルとしては不適当である。そのため線形モデルを拡張した一般化線形モデルが必要となる。

一般化線形モデルへの拡張[1,2]では、 Y の従う確率分布を一般化する。線形モデルでは、目的変数 Y は μ を平均とする正規分布である必要があった。これを後述する指数型分布族で表現することができる。また μ を単調かつ微分可能な関数 g を用いて説明変数の線形関数で表すことができ、
$$g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n \equiv \langle \beta x \rangle \quad \dots(2.3)$$
というモデルを構築することができる。これを一般化線形モデルと呼ぶ。関数 g はサンプルごとに期待値 μ と線形部分を結び付けることからリンク関数と呼ばれる。

一般化線形モデルでは目的変数 Y の従う確率分布を指数型分布族に拡張する。指数型分布族とは、パラメータ θ を用いて

$$f(y; \theta) = \exp(a(y)b(\theta) + c(\theta) + d(y)) \quad \dots(2.4)$$

と記述できる分布である。 $a(y) = y$ の時、分布は正準形と呼ばれ、この時 $b(\theta)$ は分布の自然パラメータと呼ばれる。また、成功/失敗という2値を取るモデルは、二項分布の確率分布に従う。二項分布の確率関数は

$$f(y; \theta) = \exp\left(y \ln(\theta/(1-\theta)) + n \ln(1-\theta) + \ln\binom{n}{y}\right) \quad \dots(2.5)$$

の様に指数型分布族として記述できることから、一般化線形モデルに拡張可能である。 $\binom{n}{y}$ は、 n 回の試行から y 回の成功を選択する組合せの数である。

(2.3)式のリンク関数として、二項分布の自然パラメータ $g(\mu) = \ln(\mu/(1-\mu)) \dots (2.6)$

を用いるモデルをロジスティック回帰と呼ぶ。自然パラメータをリンク関数とすることで、簡単な関数によりパラメータ μ の分布の制約条件を満たすことができる。二項分布では、許される μ の範囲は $0 \sim 1$ に限定されるが、この条件は、(2.3)式と(2.6)式とから $\mu = 1/(1 + \exp(-\langle \beta x \rangle))$ となるので、自然に満たされる。ここで $\langle \beta x \rangle$ は(2.3)式の線形部分である。また(2.6)式は対数オッズという自然な解釈を持っている。

3. 最尤法

回帰係数 β の推定は、最尤推定法を用いることが多い。この手法は、母集団の分布が分かっているがその母数が未知である場合に、尤度を最大化することで母数を推定する手法である。ロジスティック回帰では母集団の分布が二項分布である。尤度とは母集団分布の基で観測データが得られる確率や確率密度のことである。尤度を最大化することで回帰係数 β の値を決めることができる。これにより、二項分布での母数である成功確率を各サンプルで算出できる。計算の簡単化のため、一般に尤度関数は対数化を行う。解析的に解くのは一般に不可能であり、反復的[3]にNewton-Raphson法を用いて数値的に解くことになる。

以下では、尤度関数の具体的な計算方法を示す。(2.3)式と(2.6)式に基づいて回帰を行う。

$$\ln(\mu/(1-\mu)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n \dots (4.1)$$

この式を用い、回帰係数 β の値として初期値に適当な値を入れ、全体の尤度が最大となるように数値的に解くことで回帰係数 β を求める。まず各サンプルの尤度は、確率分布として二項分布を仮定しているため以下の様に記述できる。

$$\binom{1}{y} \mu^y (1-\mu)^{1-y} \dots (4.2)$$

具体的には、あるサンプルが成功例であれば $y=1$ なので、その尤度は μ となる。一方あるサンプルが失敗例であれば $y=0$ なので、その尤度は $1-\mu$ となる。成功例の場合、予測正解率 μ が高く1に近い値を取り、失敗例であれば予測正解率 μ が低く、 $1-\mu$ が1に近い値となる。いずれの場合でも、尤度が1に近い値となるように回帰係数 β を決定すればよい。全体の尤度は、各サンプルの尤度の積となる。尤度は対数を取ることで、次式のように記述できる。

$$\sum \ln \left(\binom{1}{y} \mu^y (1-\mu)^{1-y} \right) \dots (4.3)$$

ここで、和は全てのサンプルについて取る。(4.3)式の値が最大となる条件により、回帰係数 β が決まる。

ここまでで、ロジスティック回帰のモデル作成が出来た。実際に予測する段では、成功/失敗のデータの無い任意の個人に対して、(4.3)式を最大化する条件で決まった β と個人の属性値 x に(4.1)

式を適用し、予測成功率 μ を算出することになる。

4. まとめ

ロジスティック回帰は、DM送付の様に結果が成功/失敗の様な2値で測られるデータについて、個人の性別、年齢、収入等の属性を用いて回帰し、各サンプルに成功率を付与する手法である。この回帰結果を用いると、成功/失敗の付与されていない新規データについて、モデルの作成で使用した個人の属性を基に成功率を算出することができる。理論的には、ロジスティック回帰は線形モデルを拡張した一般化線形モデルの枠組の中で記述される。回帰係数は、最尤法等により数値的に求めることができる。

ロジスティック回帰はいくつかの点で非常にパワフルなモデルであると言える。例えば次の様な点が挙げられる。

- サンプルについて、単純な成功/失敗の予測ではなく、成功率を予測することができる。そのため、各サンプルに序列を付けることができる。
- 量的変数だけではなく質的変数も分析できる。また多数の変数を分析に使用することができる。
- 効果の小さい変数がモデルに入っている場合、当該変数の回帰係数が小さくなるため、モデルとして頑強である。

このようにロジスティック回帰は有用なモデルであるが、実際にデータにあてはめる際には精度の向上のために気を付けなくてはいけない点がいくつか存在する。例えば以下の様なことが起きる。

- 似た様な振舞いの変数を複数入れると、精度が極めて悪化することがある。
- 量的変数に対数化等の変数変換を行うことで、分析精度が急激に上がることがある。
- 質的変数のカテゴリ化の方法により、分析精度が良くも悪くもなる。
- 全サンプル数に対して変数もしくはカテゴリ総数が多すぎると、オーバーフィッティングとなり分析精度が悪化する。

このような事情があるため、分析精度を向上させるためには、分析の前段階であるデータクレンジングの処理により、適切な変換やカテゴリのチューニング、使用変数の選択を適切に行うことが重要となる。データクレンジングの他、適切なセグメント分け等の手法によっても、分析精度を大幅に向上させることができる。

5. 参考文献

- [1] Annette J. Dobson, 一般化線形モデル入門, 共立出版
- [2] J. A. Nelder, R. W. M. Wedderburn, Generalized linear models, Journal of Royal. Statistical Society A135, 370-383, 1972.
- [3] A. Charnes, E. L. Frome, P. L. Yu, The equivalence of generalized least squares and maximum likelihood estimates in the exponential family, Journal of the American Statistical Association, 71, 169-171