研究員の眼

統計分析を理解しよう
-ロジスティック回帰分析の概要-

生活研究部 准主任研究員　金　明中
(03)3512-1825　kim@nli-research.co.jp

ロジスティック回帰分析とは

最近、回帰分析の中でよく使われているのがロジスティック回帰分析（Logistic Regression Analysis）（以下、ロジスティック分析）である[1]。被説明変数が量的データである一般的な回帰分析は、説明変数と被説明変数の間の線形関係を仮定しており、一般線形モデル（Ordinary Linear Model）と呼ばれている。しかしながら社会のすべての現象が線形的な関係ではないので、非線形的な関係に対する分析も必要である。また、現実的には被説明変数が量的（Quantitative）データではなく質的（Qualitative）データであるケースも多い。例えば、所得がいくらぐらいである時、家を所有するか、給料がどのぐらいある時、車を買うか、年収がどのぐらいである時、結婚するかなど説明変数は量的データあるものの、被説明変数は「家を所有している、家を所有していない」のような質的データになっている場合がある[2]。

このように被説明変数が質的データであっても分析ができるよう一般線形モデルを拡張したのが一般化線形モデル（GLM：Generalized Linear Model）である。一般線形モデルが、被説明変数が正規分布をしている時のみを扱っていることに比べて、一般化線形モデルは、正規分布以外の分布（二項分布、ポアソン分布等）に従う被説明変数を予測する時にも使われる。また、一般線形モデルでは被説明変数と説明変数の線形的な関係を推計することに対して、一般化線形モデルは2値変数を扱えるようにするために被説明変数を適切な関数に変えた $f(x)$ と説明変数の関係を推計する。このような一般化線形モデルで最も使われている分析方法がロジスティック分析である。

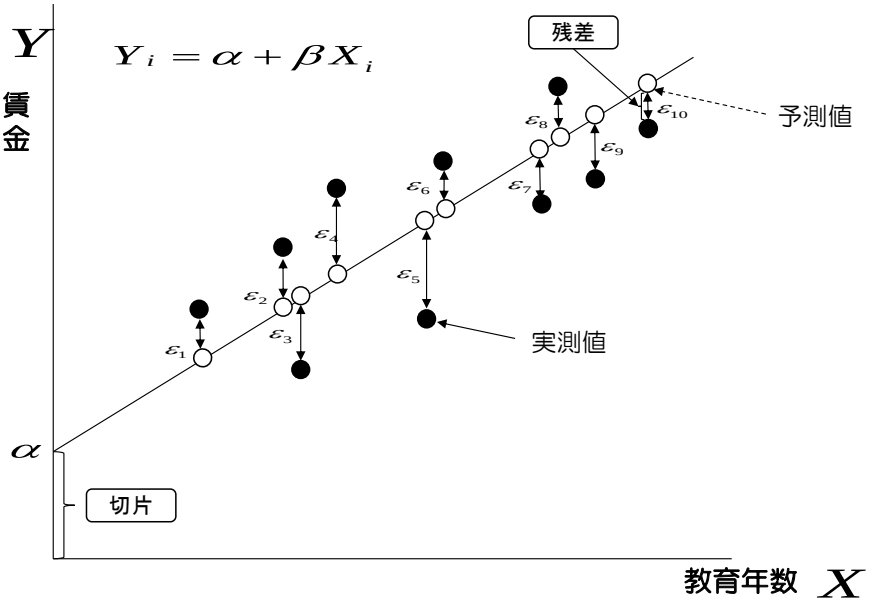
被説明変数が2値変数である場合には二項ロジスティック分析を、3項（カテゴリー）以上の場合は多項ロジスティック分析を、そして、順序変数である場合には順序ロジスティック分析を行う。

[1] 回帰分析の概要については、金 明中（2018）「回帰分析を理解しよう！－回帰分析の由来と概念、そして分析結果の評価について－」研究員の眼 、2018年5月16日を参照すること。
[2] 量的データとは、データの連続性があり、足したり引いたり演算ができ、演算しても数値として意味のあるデータである。一方、質的データは、分類や種類を区別するためのデータ（性別、学歴カテゴリ、地域カテゴリ等）であり、そのまま足したり引いたり演算ができず演算をしても意味のないデータである。

質的データを一般線形モデルで推計する誤り

一般線形モデルでは、説明変数が「1 単位」変化した際に被説明変数がどのぐらい変化したのかを把握しており、実測値と予測値の差である「残差（residuals）[3]」の二乗和が最小になるように最小二乗法（OLS：Ordinary Least Squares）を用いて分析を行う（図表 1）。

図表 1 回帰直線の概念



資料）金 明中（2018）「回帰分析を理解しよう！－回帰分析の由来と概念、そして分析結果の評価について－」研究員の眼 2018 年 5 月 16 日

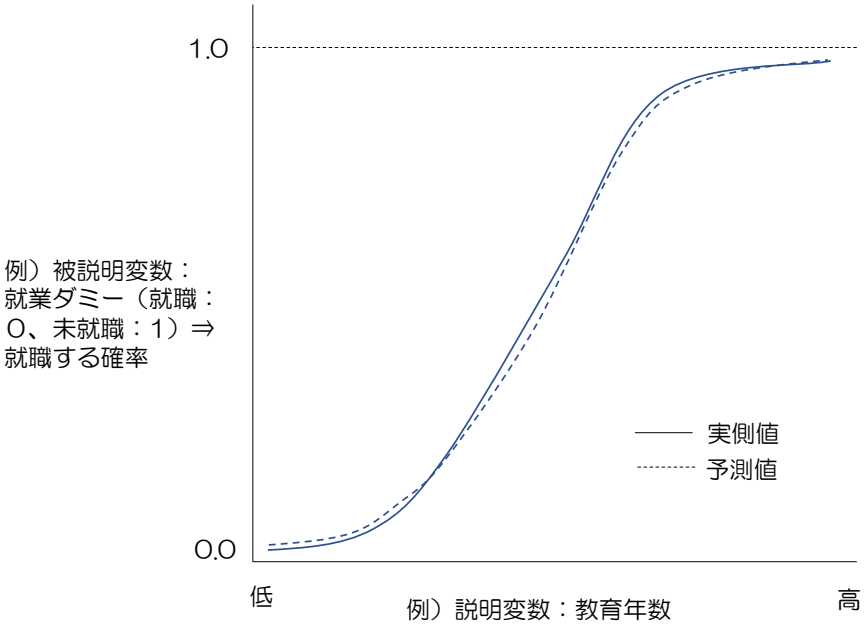
一方、被説明変数が質的データであるロジスティック分析は、図表 1 のように説明変数と被説明変数の間に線形関係が存在せず、説明変数が「1 単位」変化した際に被説明変数がどのぐらい変化したのかを把握することが難しい。従って、ロジスティック分析では最小二乗法ではない最尤推定法（Maximum Likelihood Estimation）いう手段で係数の値を求める。最尤推定法は、実測値を最も説明できる尤度（likelihood）を推定する（尤度を最大化する）方法である。蓑谷（1997）は、「尤度関数の値を最大にする推定係数のもとで観測結果が得られると考える」と説明している。尤度とは統計学では「もっともらしさ」の意味であり、得られた推定係数（パラメーター）で実測値が得られる最大の確率のことだと説明できる。

山澤（2004）は「尤度とは、同時確率密度関数の解釈を変えたものである。確率密度関数は、確率変数がどの程度の確率で表れるかを関数にしたもので、同時確率密度関数は、複数の確率変数が同時に起こる確率である」と説明している。

[3] 「誤差」は、母集団の真の回帰式から算出される値（真値）と実際に測定された値（実測値）との差を表す。一方、「残差」は標本集団のデータを用いて推計された回帰式から得られた値（予測値）と実際に測定された値（実測値）との差を表す。従って、誤差は計算で求められないが、残差は計算で求められる。誤差＝実測値－真値、残差＝実測値－予測値。

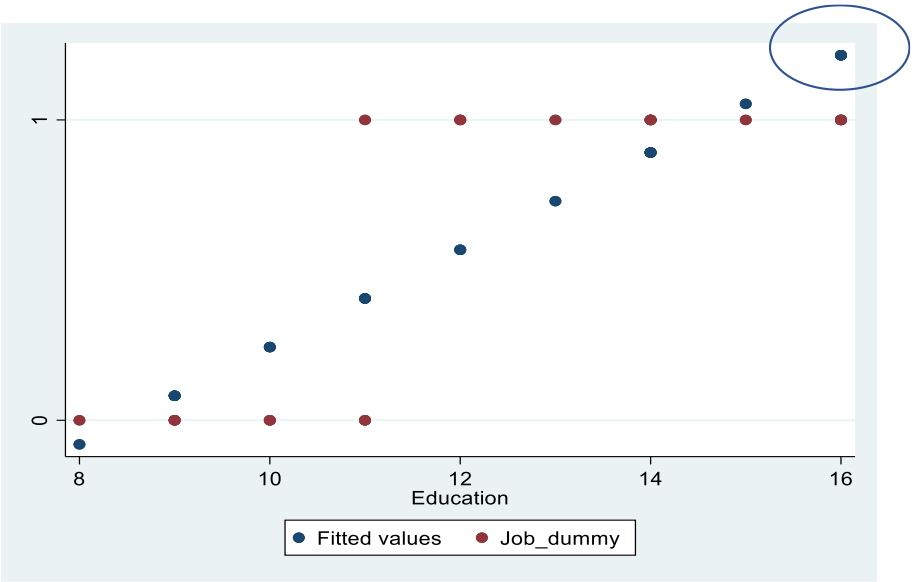
ロジスティック分析は、ある事件（event）が発生するかしないかを直接予測することではなく、その事件が発生する確率を予測する。従って、被説明変数の値は 0 と 1 の間の数値になる。分析結果、被説明変数の値、すなわち確率が 0.5 より大きいとその事件が発生すると予測し、0.5 を下回るとその事件が発生しないと予測する。ロジスティック分析における説明変数と非説明変数との関係を示すと図表 2 のような S 字形の曲線になる。

図表２ロジスティック回帰分析の曲線



回帰分析の初心者が最も陥りやすい間違いは、質的データである被説明変数を量的データとして扱い、一般線形モデルによる回帰分析を行うことである。使用者がコマンドを入力し、Stata 等の統計分析ソフトを回すと、パソコンはカテゴリのような質的データを量的データとして認識し、推計を行ってしまう。その結果、被説明変数が質的変数である場合には、被説明変数は確率概念として 0 から 1 の間の数値を推計すべきなのに、図表 3 のように被説明変数の値が 1 を超えるという誤りが発生してしまう。

図表３ロジスティック回帰分析を一般線形モデで推計した際の結果の例



オッズとオッズ比を理解しよう

ロジスティック分析をする際によく登場するのがオッズやオッズ比である。まず、オッズとは、ある事象が起こる可能性で、発生しない確率（1−p）に対する発生する確率（p）の比率である（(式 1)）。

$$(式\ 1)\quad オッズ = \frac{p(Y=1)}{1-p(Y=1)}$$

オッズは 0 から∞の値が得られ、オッズが 1 より大きいと発生する確率が発生しない確率より大きいことを、逆に 1 より小さいと発生しない確率が発生する確率より大きいことを意味する。また、オッズが 1 になると、事象の発生する確率と発生しない確率が同じくなる。

一方、オッズ比とは二つのオッズの比率であり、例えばオッズ比を利用すると、自動車を所有している世帯主が自動車を所有していない世帯主に比べて何倍住宅を所有しているかが計算できる。図表 4 は自動車の所有有無と住宅の所有有無を示しているクロス表であり、①自動車を所有している世帯主が住宅を所有しているオッズ A と、②自動車を所有していない世帯主が住宅を所有しているオッズ B を求めることが可能である。

図４ 自動車の所有有無と住宅の所有有無に関するクロス表

住宅＼自動車	所有	未所有	合計
所有	60（a）	40（b）	100（a+b）
未所有	30（c）	70（d）	100（c+d）
合計	90（a+c）	110（b+d）	200（n=a+b+c+d）

①自動車を所有している世帯主が住宅を所有しているオッズ A

(式 2) $$\frac{\text{自動車を所有している世帯主が住宅を所有している確率}}{\text{自動車を所有している世帯主が住宅を所有していない確率}} = \frac{\frac{a}{(a+c)}}{\frac{c}{(a+c)}} = \frac{a}{c} = \frac{60}{30} = 2.0$$

②自動車を所有していない世帯主が住宅を所有しているオッズ B

(式 3) $$\frac{\text{自動車を所有していない世帯主が住宅を所有している確率}}{\text{自動車を所有していない世帯主が住宅を所有していない確率}} = \frac{\frac{b}{(b+d)}}{\frac{d}{(b+d)}} = \frac{b}{d} = \frac{40}{70} = 0.571$$

ここから、オッズ比は（式 4）のように計算することができる。もちろん（式 4）は（式 2）と（式 3）を用いて（式 5）のように計算することも可能である。この結果から自動車を所有している世帯主が自動車を所有していない世帯主に比べて 3.5 倍住宅を所有していると推計することができる。

(式 4) $$\text{オッズ比} = \frac{\text{オッズA}}{\text{オッズB}} = \frac{2.0}{0.571} = 3.5$$

(式 5) $$\frac{\text{自動車を所有している世帯主が住宅を所有しているオッズ}}{\text{自動車を所有していない世帯主が住宅を所有しているオッズ}} = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{ad}{bc} = \frac{60\times70}{40\times30} = 3.5$$

確率の値は 0 から 1 の間の数値であるが、この数値に基づいて計算されたオッズは 0 から∞の値を持つ。従って確率が 0 である場合、オッズは 0 であり、確率が 1 に近くなるとオッズは無限大（∞）になる。一方、発生する確率と発生しない確率が 0.5 で同じである場合にはオッズは 1 になる。

但し、オッズ比が 1 より小さい（回帰係数が「－」）結果が出た場合は、求めた可能性が減少したことを意味するので解釈に注意が必要である。例えば、被説明変数として就業ダミー（就業を 1、未就業を 0）を用いて説明変数が「子供の数」が就業に与える影響を分析した結果、回帰係数が「－1.0416」が出て、オッズ比は「0.35289」が得られたと仮定しよう。この結果は子供の数が一人増えると、就

業する可能性が 0．35289 倍増加すると読み取ることができるものの、実際は子供の数が増えると就業する可能性が低くなることを意味する。しかしながら、初心者の場合は「0.35289」という正の数値を誤って解釈することも多いだろう。そこで、このような誤りを最大限防止するためにエクセルの数式（(式 6)）を利用して値を変換することも一つの方法である。例えば、回帰係数「－1.0416」を（式 6）に入れて計算すると「－64．7」という負の数値が得られる。つまり、この結果は子供の数が一人増えると、就業する可能性が 64．7%減少することを意味するのであるが、負の数値であるため解釈による誤りを防ぐことができる。

（式 6）$=\left(\exp(\text{回帰係数})-1\right)\times 100$　　⇒　－64．7

ロジット変換

次はロジットについて簡単に説明したい。ロジットは上記で説明したオッズ比に対数を取ったものである。ロジット変換をすると、0 と 1 という質的データを持つ被説明変数の値は「－∞」から「＋∞」に代わることになる。そこで、まるで連続性のある量的データのように扱うことができる（(式 7)）。

（式 7）$-\infty < \text{ロジット}=\ln(\text{オッズ}) < +\infty$

（式 8）$\ln\left(\dfrac{p(Y=1)}{1-p(Y=1)}\right)=a+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n$

但し、ロジットの値は解釈が難しいので、（式 9）のように確率の値に変換する。

（式 9）$p(Y=1)=\dfrac{e^L}{1+e^L}: L=\log it,\quad e\approx 2.718$

（式 9）は次のような式の展開で導出された。

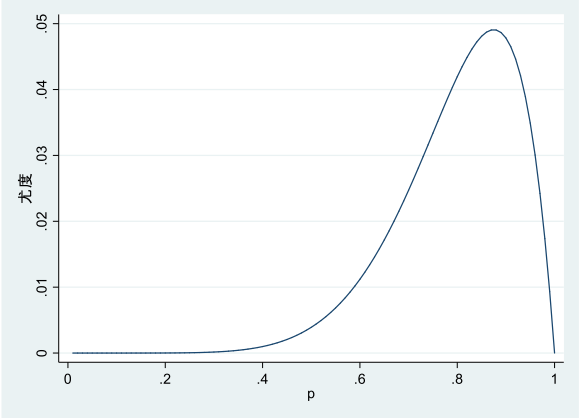
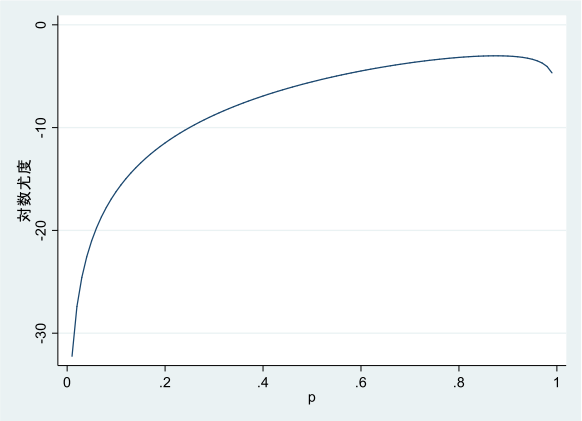
$$\ln\left(\frac{p(Y=1)}{1-p(Y=1)}\right) \quad \Rightarrow \quad e^L=\frac{p(Y=1)}{1-p(Y=1)} \quad \Rightarrow \quad p(Y=1)(1+e^L)=e^L \quad \Rightarrow \quad p(Y=1)=\frac{e^L}{1+e^L}$$

このように変換されたロジットは、線形モデルとして推計することができる。但し、回帰係数を推定する際には最小二乗法ではなく最尤推定法を使う。尤度関数は（式 10）の通りである。

（式 10）$L(\pi\,|\,h,n)={}_nC_h\pi^h(1-\pi)^{n-h}$

ここでnはサンプル・サイズ、hは成功する回数、πは成功する確率を意味する。例えば、合格率が 80%で 10 人が応募して、7 人が合格する確率πを求めると、約 20．1%になる。例えば、サンプル・サイズ（n）と成功する回数（h）が不変であれば、尤度（$L(\pi\,|\,h,n)$）を最大にするπを求めるこ

とが大事である。そこで、πの値を 0.01 から 0.99 まで入力した後に、その値を$L(\pi|h,n)$に代入し、尤度を最大にする値を求めてみた。すると、図表 5 のようにπ=0.87 の際に尤度が最大になる。従って回帰係数は尤度を最大化する値で推定され、（式 10）にπの値を入れると求められる。但し、計算が複雑であるので一般的には対数を取った対数尤度（log likelihood）がよく使われる（図表 6）。対数尤度は反復作業をして最大値を求める。

図表５二項変数の尤度関数の例：$L(\pi|7,10)$	図表６二項変数の対数尤度関数の例：$L(\pi|7,10)$
	

結びに代えて

一般的にロジット分析は回帰係数を求める分析であり、ロジスティック分析はオッズ比を求める分析として知られている。ロジット分析やロジスティック分析をする際に最も注意すべきことは、①質的データである被説明変数を量的データとして扱い、一般線形モデルによる回帰分析を行うことと、②分析から得られた値（例えば回帰係数やオッズ比）を間違って解釈しないことである[4]。本文で説明した基本概念を理解し、ロジスティック分析等を有効に活用して頂くことを願うところである。

参考文献

- 金 明中（2018）「回帰分析を理解しよう！－回帰分析の由来と概念、そして分析結果の評価について－」研究員の眼 2018 年 5 月 16 日
- 金 明中（2019）「統計分析を理解しよう-よく使われている統計分析方法の概要-」研究員の眼、2019 年 6 月 28 日
- 蓑谷 千凰彦（1997）『計量経済学』多賀出版
- 山澤 成康（2004）『実践計量経済学入門』日本評論社
- キム ギョンソク・キム ギョンヒ（2004）『Stata を用いた統計実務』統計庁 STATA 研究会

[4] 統計分析の概要については、金 明中（2019）「統計分析を理解しよう-よく使われている統計分析方法の概要-」研究員の眼、2019 年 6 月 28 日を参照すること。

