基礎研レター

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　－

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

1――回帰分析の由来

統計的分析方法の中で最も使われている一つが回帰分析である。回帰分析とは、説明変数が被説明変数に与える影響の大きさを把握し、説明変数の特定値に対応する被説明変数の値を予測するモデルを算出する方法である。より簡単に言うと、ある変数の値で、他の変数の値を予測し、両者の関連性を確認する分析方法だと言える。

回帰分析が世の中に登場するまでには少なくとも 4 人の学者の貢献があった。まず、フランスの数学者ルジャンドル（Adrien-Marie Legendre 、1752～1833）は、回帰分析の代表的な手法である「最小二乗法」のアイディアを最初（1805 年）に発表した[1]。最小二乗法（OLS, Ordinary Least Squares）とは、残差（観測値と予測値の差）の二乗和を最小にする推計方法である（詳細は次の節で説明）。その後、最小二乗法を発展させたのがドイツの数学者（天文学者で物理学者でもある）であるガウス（Carl Friedrich Gauss、1777-1855）である。ガウスは、惑星の軌道を予測する計算方法として最小二乗法を用いた。それから、イギリスの遺伝学者であるゴルトン（Francis Galton、1822-1911）は、親と子どもの身長を分析し、非正常的に身長が大きい子どもと小さい子どもの身長は全人口の平均身長に回帰する傾向があることを見つけた。ゴルトンはこの現象を平均からの回帰（regression to the mean）と呼び、回帰という言葉が始めて分析の中で使われるようになった。また、ゴルトンの友人であるピアソン（Karl Pearson, 1857-1936）は 1,000 人以上のデータを集めて、身長が高いお父さんグループの子どもの平均身長はお父さんより小さく、身長が低いお父さんグループの子どもの平均身長はお父さんより大きいという「普遍的回帰の法則（law of universal regression）」を証明した。この結果は親の身長がいくら高くても子どもの身長は子どもの世帯の平均の近接する傾向があることを意味している。

[1] ガウスは、最小二乗法の原理を 1794 年に初めて発見したと主張しており、ルジャンドルとの間に最小二乗法の発見をめぐって論争が続いたそうだ。

2――回帰分析の概念

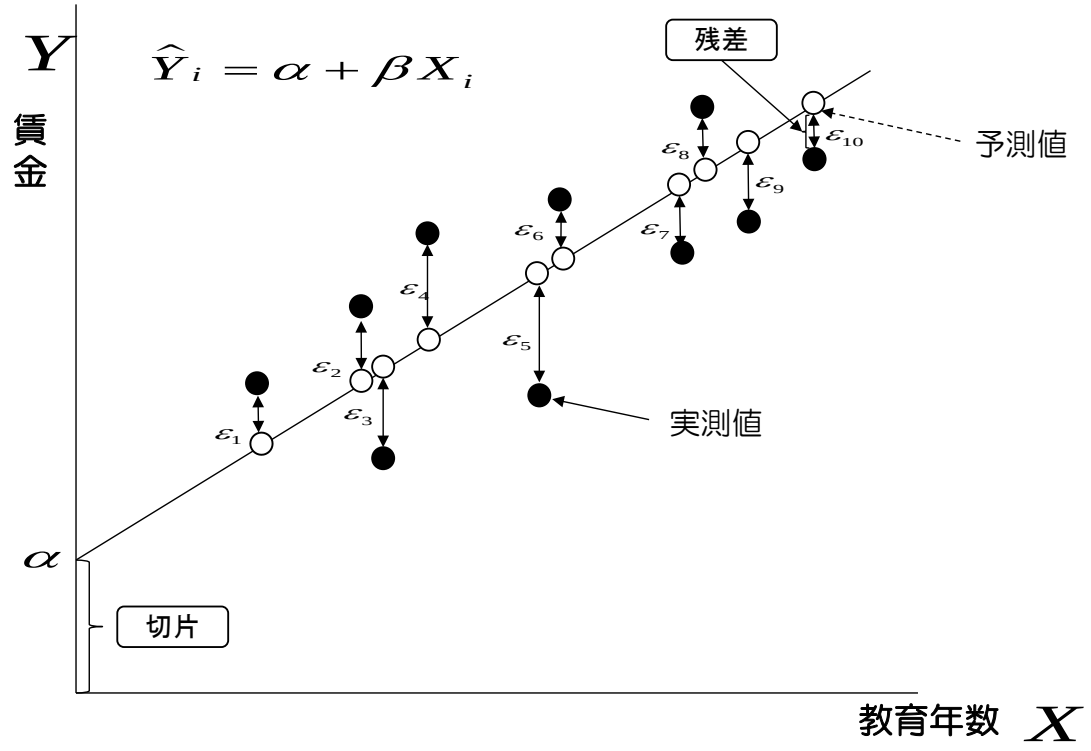
一般的には予測される変数を被説明変数（従属変数、目的変数とも呼ぶ）と呼び、予測のために使われる変数を説明変数（独立変数）と呼ぶ。また、被説明変数を予測する際に使われる説明変数が一つであると単回帰分析（simple regression model）であり、説明変数が二つ以上であると重回帰分析（multiple regression model）である。

式 1)は、教育年数と賃金の多寡の関係をみた単回帰分析の方程式である。教育年数は影響を与える説明変数（X）、賃金は影響を受ける被説明変数（Y）と置くこととする。そして、μ_iは誤差項で、賃金の変化のうち教育年数で説明できない部分、つまり教育年数以外に賃金に影響を与える要因の合計である。

式 1） 賃金＝$a+b\times$教育年数$+\mu$ → $Y_i=\alpha+\beta X_i+\mu_i$

被説明変数 → Y＝賃金、説明変数 → X＝教育年数、誤差項 → μ

図表 1 回帰直線の概念



αは方程式の切片とも呼び、βは回帰係数である。そこで、式 1)の方程式によると、教育年数以外に賃金に影響を与える要因ε_iが固定されている場合、教育年数 1 年の変化は賃金をβだけ変化させることになる。

式 1)の方程式の目標は見えないパラメータ（parameter）であるαとβを推計することであり、より正確なパラメータを推計する方法として使われているのが最小二乗法である。最小二乗法は、観測

値と予測値の差である「残差（residuals）[2]」の二乗和が最小になるようにαとβを推計する方法である。例えば、図表1の黒丸（●）は実際に測定された実測値であり、その中に描かれている近似線は実測値との残差を最も小さくするために推計された予測値である白丸（○）を繋げた近似線、つまり回帰直線である。一方、被説明変数の予測値は、$\hat{Y}$のように書けるので、式 1)の予測式は式 2)のようになる。

式2） $\hat{Y}_i = \alpha + \beta X_i$

従って、式1）と式2）を用いて、残差ε_iを求めると式3)のように表すことができ、その残差の合計は式4)のようになる。

式3）残差： $\varepsilon_i = Y_i - \hat{Y}_i$

式4）残差の合計： $\sum_{i=1}^{n} \varepsilon_i$

しかしながら、残差は正になるケースもあり、負になるケースもあるので、このまま残差を合計すると、残差は相殺され、残差の合計は小さくなってしまうという問題点がある。そこで、最小二乗法では、残差の二乗を求めてその合計が最小になるようにαとβを推計している（式5)）。

式5）残差の二乗の合計： $\sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i)^2$

また、βは、XとYの共分散をXの分散で割ったものであり[3]（式6)）、αはYの平均$\overline{Y}$からXの平均$\overline{X}$にβを乗じた値を差し引いて求める(式7))。

---

[2] 「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

[3] 共分散は、XとYという2つの変数の関連性を表す指標で、2つの変数の偏差の積の平均を計算して求める。

$$X\text{と}Y\text{の共分散} = S_{XY} = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{n}$$

式 6）回帰係数：$\beta = \dfrac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=!}^{n}(X_i - \overline{X})^2}$

式 7）切片：$\alpha = \overline{Y} - \beta\overline{X}$

3――分析結果の判断

パソコンが普及していなかった時代には上記のような式を用いて定数項や回帰係数を求めたものの、最近では Stata や SPSS、そして R などのような統計パッケージを使えば簡単に分析結果を出すことができるようになった。従って、我らがすべきことはその分析結果が統計的に有意であるかどうかを判断することである。

まず、分析の結果から確認できるのが R^2 で表示される決定係数（coefficient of determination）であり、これは説明変数が被説明変数をどれくらい説明できるかを表す。決定係数は 0 から 1 の範囲内の値を取り、決定係数が 1 に近いほど説明力が高いことを意味する。しかしながら、社会科学関連の分析では決定係数が低い場合が頻繁にある。その理由としては被説明変数に影響を与えると思われるすべての変数が利用できないことや、分析者が選択した一部の変数のみが説明変数として利用されている点などが挙げられる。そして、線形モデルの場合、決定係数は相関係数の二乗に等しいので、例えば、決定係数が 0.2 だとしても、これを相関係数に直すと 0.45 に当たるので決して低い数値だとは言えず、ある程度は説明力があると解釈できる。

次に分析結果が統計的に意味のある分析かどうかを確認するために使うのか有意確率と有意水準である。分析の前には、仮説が正しいかどうかの判断のために帰無仮説を立てるのが一般的である。帰無仮説は、例えば「A という説明変数は被説明変数に何の影響も与えていない」あるいは「A と B の平均には差がない」のように、たいていは否定されることを期待して立てられる。一方、検定しようとする帰無仮説に対立する仮説が対立仮説であり、これは帰無仮説を棄却するために使われる。帰無仮説と対立仮説を式として表すと式 8）と式 9)の通りである。

式 8）帰無仮説：$H_0 = H_1$

式 9）対立仮説：$H_0 \neq H_1$

回帰分析の目的は、説明変数が被説明変数にどのぐらいの影響を与えているかを調べることにある。つまり、分析者の主張が正しいことを証明するためには帰無仮説を棄却させ、対立仮説が有意であることを統計的に検定する必要があり、その統計的検定に使われるのが有意確率と有意水準である。まず、有意確率とは、 p値 と表記され、帰無仮説のもとで得られた検定統計量が実現する確率である。有意確率を分かりやすく言えば、分析の結果から帰無仮説が正しいと説明できる確率、つまり、帰無仮説のような結果が起きる確率のことである。一般的に p値 が小さければ小さいほど帰無仮説が棄却される確率が高くなる。では、どのぐらいの水準で帰無仮説が棄却できるだろうか。この際に使われ

るのが有意水準である。有意水準とは統計的仮説検定を行う場合に、帰無仮説を棄却するかどうかを判定する基準[4]であり、5% あるいは 1%という有意水準がよく使用される[5]。つまり、有意確率が事前に定めた有意水準より小さい場合には帰無仮説を棄却し、有意水準より大きい場合には帰無仮説を採択することになる。例えば、「男性と女性の賃金は差がない」という帰無仮説を有意水準 5%水準で棄却したということは、分析の結果から帰無仮説が正しいと説明できる確率（帰無仮説のような結果が起きる確率）は 5%未満であったので、対立仮説「男性と女性の賃金は差がある」を採択したことを意味する。つまり、帰無仮説は正しくない（回帰係数が 0 ではない）と判断するとともに、回帰係数は統計的に有意であると言えるのである。

そして、有意確率とともに確認すべきことが *t*値である。*t*値は、係数の値を係数の標準誤差[6]で除したものであり、説明変数が被説明変数に与える影響の大きさを表す。絶対値が大きければ大きいほど影響が強く、一般的には *t*値が 2 以上[7]なら、その説明変数が被説明変数を十分に説明していると判断される。面白いことは、*t*値は上記で説明した *p*値とは逆の関係にあることである。つまり、*t*値が大きいと *p*値は小さく、逆に *t*値が小さいと *p*値は大きいという結果になる。

式 10） $t = \dfrac{\text{係数}}{\text{係数の標準誤差}}$

4――結びに代えて

本稿では統計分析で最も使われている回帰分析の由来と概念、そして分析結果の判断について説明した。最近は、Stata や SPSS、そして R などのような統計パッケージを使い、半自動的に（？）実証分析を行うケースが多いので、実証分析の初心者の中には回帰分析の詳細を理解せず、*p*値と *t*値、そして回帰係数という分析結果だけを確認・利用するケースも少なくないだろう。もちろん、論文や報告書などでは分析結果やその解釈を最も大事にしているので、分析結果を重視することは当然のことかも知れない。しかしながら、その結果が出るまでの過程も大事である。パソコンの普及に伴って軽視されがちであったこのような過程を理解しようと努めることが分析結果をより明確に解釈できる近道であると信じている。本稿の内容がこれから回帰分析の過程を学ぶ人たちにとって、少しでも参考になることを願うところである。

---

[4] 松村 明（編集）『大辞林第三版』（2006）では、有意水準を「帰無仮説が真のときに統計量が棄却域に入る確率」と説明している。

[5] 状況によっては 10%が使われる場合もある。

[6] 標準誤差が小さいほど *t*値が大きくなり、説明力が高まる。

[7] 厳密には、*p*値が 0.1 の時に *t*値はおよそ 1.68 になるので、*t*値が 1.68 より大きければ、説明変数が被説明変数を説明していると判断できる（ぎりぎりセーフ）。

