

今回のテーマ

重回帰分析の応用として、既存の統計ソフト (IBM SPSS Statistics を想定) で分析可能な、幾つかの手法の特徴と使用方法について解説します。

1. 階層的重回帰分析

階層的重回帰分析とは、重回帰分析において変数の投入順序に「階層性」を持たせるテクニックである。これにより共変数の影響を統制したうえで関心のある変数の予測力を明示的に検証することができる。基本的に重回帰分析の拡張で、データの階層性を扱うことのできるモデルではない(莊島他,2018)。

通常重回帰分析では、 x_2 と x_3 を統制したうえでの x_1 の効果、 x_1 と x_2 を統制したうえでの x_2 の効果、 x_1 と x_2 を統制したうえでの x_3 の効果がそれぞれ検定され、想定される変数間の関係は図 1 のように表される。これに対し、階層的重回帰分析では図 2 のような変数間の関係を想定して分析が行われる。この例は、 y に対する x_2 の効果を x_3 が媒介することを意味している(三輪・林,2014)。

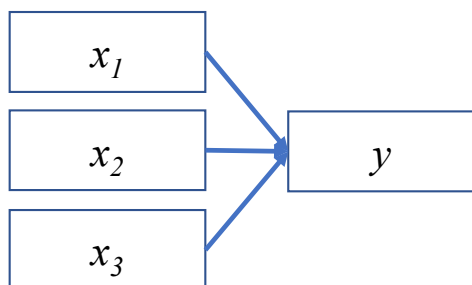


図 1 重回帰分析のモデル例

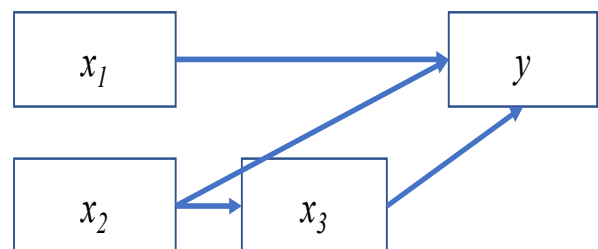


図 2 階層的重回帰分析のモデル

このような分析により、従属変数 y に対する独立変数 x の「総効果」を「直接効果」と第 3 変数 z を媒介とした「間接効果」に分解することが可能となる。図 2 のモデルでは、 $x_2 \rightarrow x_3$ のパスが直接効果、 $x_2 \rightarrow x_3 \rightarrow y$ のパスが間接効果を表す。

具体的な分析方法として、予測変数に対して新しい予測変数を加えたときにどの程度説明力が上昇したかにより検証し、説明力の増分が有意であれば新しく加えた変数は予測において重要であると結論づけることができる。説明力の増分の有意性の検定には、「決定係数の変化量は 0 である」ことを帰無仮説とした F 検定が用いられる(三輪・林,2014)。この場合加える変数は単一の変数に限定されず、複数の変数をまとめた変数セット全体でどの程度予測できるかが分

析可能である(荘島他,2018)。なお、注意点として、ステップワイズ法による変数選択の自動化を行う場合に、(1)標準誤差や有意確率が過小に推定されること、(2)偏回帰係数の推定値に無視しえない偏りが生じること、とあわせて、(3)サンプルサイズが小さい場合は安定した結果が得られないため注意が必要である。

また、各変数と交互作用項の相関による多重共線性の問題が考えられるので、変数間の相関が高い場合は、中心化処理を行う必要がある(荘島他,2018)。

実際の SPSS の操作方法、論文における活用例はサブゼミ時に紹介します。

2.一般化線形モデル

一般化線形モデルは、線形モデルを拡張したものである。通常の線形モデルは、一般に、独立変数の線形結合に平均 0、等分散の正規分布に従う誤差を加えたモデルとして定式される。

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

この場合、 y は平均 μ の等分散正規分布に従うことが仮定されているため、そうでない従属変数について分析するために、

(1) y は指数分布族(正規分布、Poisson 分布、二項分布¹など多くの分布が含まれる)に従う、(2) μ をリンク関数で変換するの 2 点を拡張し、以下の式で表す。

$$g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

この拡張により、データについて質的データや正規分布に従わないデータを対象の活用が可能になり、モデルについても線形関係以外でも逆数、対数、プロビット、ロジットなどに変換したモデルの使用が可能になる(寺島,2018)。

線形モデルでは最小二乗法により推定を行うが、一般化線形モデルの場合は、最尤法による推定を行う。最尤法は、得られたデータをもとに、それが得られる可能性(尤度 L)が最も高くなるパラメーター(確率分布の形を決める数値)を推定値とする方法である。モデル適合度の評価には、尤度を元にした適合度指標を用いる。代表的な適合度指標として、対数尤度、逸脱度、赤池情報量基準(AIC)、ベイズ情報量基準などがある。複数のモデルを比較し、モデル選択を行う場合によく利用されるのが AIC で、これは対数尤度を基づいた適合度で以下の式で表される。

$$AIC = -2LL + 2p$$

(但し、 LL は尤度 L の自然対数、 $-\infty < LL < 0$ 、 p はパラメーター)

この値が小さいほど良いモデルとされ、あてはまりに寄与しない独立変数を増やすと値が高くなる。また線形モデルにおける F 検定に対応するものに尤度

¹ 二項分布については、通常ロジスティック回帰分析、プロビット回帰分析を行うのが通例だが一般線形モデルは他にも負の二項分布、順序ロジスティック分析、Poisson 回帰分析等を包括したモデルである。

比検定があり、以下の式で表される。

$$LR\chi^2 = -2LL_0 - (-2LL_1)$$

一般化線形モデルではこの式を用いて、モデル全体の有効性の検定が行われる。概念としては、切片のみのモデルの尤度(L_0)と独立変数を追加したモデルの尤度(L_1)との比を取ったうえで、対数を取り、それに -2 をかけた形で表されている。切片のみのモデルと比較して独立変数を追加したモデルが、どの程度当てはまりの悪さを改善しているかに着目し、これに基づき「すべての母回帰係数が0:母集団において説明力がない」を帰無仮説とした検定を行う(寺島,2018)。

各回帰係数の検定には、ワルド検定が用いられる。この検定は、ワルド統計量

$$Wald\text{統計量} \left(\frac{b_j}{SE_j} \right)^2$$

が自由度1の χ^2 分布に従うことを利用し、帰無仮説「母回帰係数が0:母集団において説明力がない」について検定を行う。

SPSSによる分析方法と論文における活用例はサブゼミ時に紹介します。

3.傾向マッチング法

一般に、重回帰分析を行う場合、説明変数を選択できてもそれが従属変数にどのように影響するか、モデル化(特に因果関係の推定)が困難で、良い推定値を得るまで変数選択を変えながら何度も分析を行う場合が多々ある。傾向マッチング法は、(1)共変量のバランスの確認することでモデルの妥当性を評価が可能である点、(2)被説明数の出現率が少ない場合でも精度が担保できるといった点から近年注目されている手法である。

この手法は、主に医療統計分野で発展した手法である。例えば、新しい治療法の効果を分析したい場合に、治療法や効果の有無は、患者の特性や病歴など様々な共変量の影響を受けるため、治療を受けた人のみを検討しても、効果が検証できない。そこで治療を受けていない人から無作為にサンプル抽出して、交絡因子(調べようとする因子以外の因子で結果に影響するもの)の影響を取り除く手法が取られることがある。しかし、現実には無作為にサンプリングを行うのは困難で何らかのバイアスが生じる場合や、治療を受けた人と受けていない人で交絡因子の分布に大きな偏りが生じる場合がある。傾向マッチング法は、このような問題を解決するために考案された手法で、以下のプロセスで行う。

(1)ロジスティック回帰分析などを用いた傾向スコアの推定

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots$$

(2)治療を受けた人と治療を受けていない人のグループで、類似した傾向スコア

を持つサンプルをマッチング²³する。その後、マッチングした治療を受けた人の効果(Y_1)と受けていない人の効果(Y_0)の差を取り、平均を取る⁴。

$$ATE = E[Y_1 - Y_0]$$

マッチング後の治療を受けた人と受けていない人の交絡因子の相違を記載し、グループごとの統計量・標準化効果量などからバランスを確認する

傾向スコアの推定後、その大小によりいくつかの層を作成し、層ごとに解析した後、ひとつにまとめて解析する事もできる(大林,2016)

(3)マッチング後は、共変量の情報はすべて傾向スコアに反映されているため、線形混合モデル、一般化線形混合モデル、条件付きロジスティック回帰などを行って検証する。

4. マルチレベル分析

調査において、ある組織や地域などの中間グループを抽出し、さらにその中で個人を抽出する場合など階層化されたデータを分析したい場合がある。例えば複数の企業に所属する社員からなるデータを用いて、社員の給与水準に与える影響を分析したい場合には、社員レベルの要因に加え、所属している企業の要因が大きく影響する。しかし、通常の線形モデルで検証しようとした場合、(1)サンプルの独立性の仮定が満たされない、(2)従属変数への影響が上位レベル(所属企業への影響)によるものか、下位レベル(社員への影響)によるものかを区別できない、(3)グループ内のデータに相関があるせいで標準誤差が小さく推定されるといった問題がある(寺島,2018)。

線形モデルで最小二乗法によって切片と回帰係数を推定する場合、

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon$$

のモデルの係数 $\beta_0, \beta_1, \dots, \beta_k$ は、固定された値として推定されるため、これを固定効果と呼ぶ。グループ j に属するケース i の従属変数を y_{ij} 、独立変数を x_{ij} 、誤差を ε_{ij} とした時、以下の β_{0j}, β_{1j} は、グループ j に共通する切片と回帰係数である。

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \varepsilon_{ij}$$

ここで、切片 β_{0j} と係数 β_{1j} に着目し、切片の全体平均 γ_{00} と、グループごとの誤差 u_{0j} 、係数の全体平均 γ_{10} 、そのグループごとの誤差 u_{1j} で表した以下の式に、

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

グループレベルでの独立変数 ω_j を追加し、元の式に代入すると次の様に書ける。

² 類似した傾向サンプルが他方のグループに存在しない場合は、サンプルから削除する。

³ なおマッチングせずに治療を受けた人と受けない人の両方をサンプルに入れたうえで、傾向スコアとモデルに残っている(傾向スコアの推定に使用しなかった)共変量による多変量モデルを推定する方法もある。

⁴ はじめての傾向スコア分析(<https://www.lightstone.co.jp/stata/files/psa01.pdf>)

$$y_{ij} = (\gamma_{00} + \gamma_{01}\omega_j + u_{0j}) + (\gamma_{10} + \gamma_{11}\omega_j + u_{1j})x_{ij} + \varepsilon_{ij}$$

$$= \underbrace{\gamma_{00} + \gamma_{01}\omega_j + \gamma_{10}x_{ij} + \gamma_{11}\omega_j x_{ij}}_{\text{固定パート}} + \underbrace{u_{0j} + u_{1j}x_{ij} + \varepsilon_{ij}}_{\text{ランダムパート}}$$

このモデルの固定パートの切片と回帰係数の推定に加えて、ランダムパートの誤差の分散や共分散を推定するのがこの手法の目的で(寺島,2018)、切片のみにランダム効果を想定するか、回帰係数のみにランダム効果を測定するか、両方にランダム効果を想定するかによってランダム切片モデル、ランダムスロープモデル、ランダム切片・ランダムスロープモデルに分かれる。ただ、グループ間分散が大きくなければ、マルチレベル分析を適用することにあまり意味がないので級内相関係数(全体の分散に占めるグループの分散の割合、1に近いほどグループによる影響が大きいことを意味する)を用いてこの手法を適用するか検討する事が多い(寺島,2018)。

$$\text{級内相関係数} = \frac{\text{グループ間分散}}{\text{グループ間分散} + \text{グループ内分散}}$$

変数の解釈は通常重回帰分析と同様であるが、交互作用項を投入した際の多重共線性を避けるために中心化⁵された独立変数を投入するのが一般的である。マルチレベルの中心化では、全体平均からの偏差 $x_{ij} - \bar{x}$ に変換する全体平均中心化のほかにグループ別平均からの偏差 $x_{ij} - \bar{x}_j$ レベルの中心化がある。

マルチレベル分析は最尤法で推定するため、一般化線形モデルと同様、推定には対数尤度をもとにした適合度指標が用いられる。なおマルチレベル分析では制限つき最尤法が用いられる事が多く、SPSSの設定ではデフォルトになっている。実際のSPSSの使用例と関連論文はサブゼミ時に示す。

参考文献

- ・ 荘島宏二郎他(2017)、「研究委員会企画チュートリアルセミナー：統計モデルの違いを理解する」、『日本教育心理学会第59回総会発表論文集』pp.16-17
- ・ 寺島拓幸(2018)、「SPSSによる多変量データ分析」、東京出版(第5章)
- ・ 三輪哲・林雄亮(2014)、「SPSSによる応用多変量解析」、旺文社(第9章)
- ・ 久保拓哉(2012)、「データ解析のための統計モデリング入門」、岩波書店
- ・ 奥村康之(2021)、「傾向マッチングの書き方」、心理・医学系研究者のためのデータ解析環境Rによる統計学の研究会第8回研究会資料
- 石村貞夫・石村友二郎(2015)、「SPSSによる傾向スコアのマッチングの手順」、鶴見大学紀要、第52巻(4)
- 大林準(2016)、「ロジスティック回帰分析と傾向スコア」天理医学紀要、19(2)、pp.71-79

⁵ 独立変数の各値 x_i から平均値を引いて偏差にすること。