2021年6月18日　統計数理研究所　オープンハウス

統計的因果推論の視点による重回帰分析

岩崎　学　　統計思考院　特任教授

【はじめに】

重回帰分析 (multiple regression analysis) は，数ある統計的データ解析の諸手法の中で，様々な応用分野において最も広範に用いられている標準的な手法であるが，その半面，結果の解釈では最も誤りを犯しやすい解析法でもある．筆者の統計数理研究所でのミッションは，統計学・データサイエンスのすそ野を広げるための有為な人材を今以上に多く育成することであり，その種の人材には，重回帰分析というこの重要な統計手法に関する正しい知識を身に付け，今後の世界をデータで切り拓く人材を多く世に送り出したいと願っている．

ここでは，上記のような統計教育・人材育成に資すると思われる重回帰分析の理論と応用の勘所を，筆者のこれまでの経験に基づいて述べる．

本発表は，筆者の日本統計学会賞受賞記念論文である岩崎 (2021) に基づき，岩崎 (2019) でも扱っている．また，統計的因果推論に関しては，岩崎 (2015) あるいは近刊の訳書である阿部・岩崎 (2021) を見られたい．回帰分析に関する海外の文献としては，大判でありしかも534頁という大部の本（重くて場所をとる）である Gelman et al. (2020) が参考になろう．

【重回帰分析のモデルと諸種の仮定】

第 i 個体のアウトカムを表す確率変数 Y_i と p 個の説明変数 $x_{i1}, \ldots, x_{ip}$ に対し，重回帰モデルとその標準的な仮定は

$$Y_i = \beta_1 + \beta_2 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i \ (i = 1, \ldots, n), \quad E[\varepsilon_i] = 0,\ V[\varepsilon_i] = \sigma^2,\ Cov[\varepsilon_i, \varepsilon_j] = 0 \ (i \neq j),\ \varepsilon_i \sim N(0, \sigma^2)$$

である．ここで ε_i は，教科書的には誤差を表す確率変数であるが，「誤差」は天から降ってくるものではない．当然のことながら，分析対象となる現象は結構複雑なものであるはずで，Y_i に影響を与える変数は数多く，しかもその影響を表す関数形は複雑なものであろう．しかし，複雑であるとか難しいと言っていたのではそれ以上分析はできないため，分析者（データサイエンティスト）は，重要な説明変数を選択し，しかもそれらとアウトカムとの関係を表す関数形も比較的簡単なものとすべきであろう．そしてそこに含まれないもろもろのものを偶然変動と見なすのである．この「見なす」という観点が重要であり，何をそう見なすかが分析者の腕の見せ所であると言える．その意味で，筆者は誤差ではなく偶然変動項と呼ぶことにしている

もう1点，ε_i は説明変数 $x_{i1}, \ldots, x_{ip}$ と独立である必要がある．通常の回帰モデルでは説明変数は定数（確率変数ではない）であることから，記法的に常に $Cov[x_{ij}, \varepsilon_i] = 0$ であるが，現実問題では，説明変数と誤差項との独立性を担保しておく必要がある．これは計量経済分野では omitted variable として常に気にすべきとされる仮定であるが，他の分野ではあまり論点にならないので注意しなくてはならない．

【重回帰分析の目的】

重回帰分析の目的には，大きく分けて (i) 現象の記述，(ii) アウトカムの予測，(iii) 因果関係の確立，の3つがある．これらのうち (i) は，分析の初期段階では取り敢えず重要であり，推定したモデルを媒介にした当該現象の分かりやすい記述により，この段階で分析の目的が達せられることもあり得る．しかしその反面，やっただけになってしまう恐れもある．やはり (ii) の予測，あるいは (iii) の因果関係の確立が要求されるであろう．

(ii) の予測は，重回帰分析の大きな目的の一つである．筆者は，2変量間の関係を (a) 相関，(b) 回帰，(c) 因果，の3種類に分類して説明することが多く，岩崎 (2015, 2019) などの著書にもそう書いている．相関は双方的な概念であるが，予測と因果は一方向的な概念であり，重回帰分析の場合は一方向的であるので，(b) 回帰もしくは (c) 因果が関係してくる．そして，回帰と因果の違いとして，「因果」はその名の通り因果関係であるが，「回帰」は，必ずしも因果関係ではないかもしれないが予測には有用な関係，としている．

「予測」は回帰分析の重要な役割であるので論点は数多い．近年，ディープラーニングに代表される機械学習の方法論の進展は目覚ましく，使いやすいソフトウェアの提供もあって，その応用は爆発的に広がっている．いわゆる AI (Artificial Intelligence) の手法も予測の目的で用いられることが多い（将棋の AI は次の一手の予測に用いられるのがその例である）．したがって，重回帰分析の予測も，その種の機械学習の予測との比較によって論じるがよいと思われる．機械学習における予測の特徴は，予測法の中身がブラックボックスになっていることであろう（もちろん近年の研究では，ブラックボックスのホワイト化がそのテーマの一つであるが）．予測手法がブラックボックスであるということは，重回帰モデルで言うと，偏回帰係数の解釈をしないことにつながる．「予測」ではうまく予測さえできればいいのであって，どの説明変数を選択するのか，それらの変数が予測値にどのような影響を与えているのかも，機械学習ではブラックボックスなので，回帰分析でもある意味で関係がないと言える．したがってその意味で，経済学で問題となる多重共線性の問題も関係がない．しかし，偏回帰係数の推定値も分析結果として得られる訳であるから，予測に有用な変数はどれか，その変数の影響度はどの程度か，という知見を得るための偏回帰係数の解釈は，当然ながら多く見られる（そうすると多重共線性は問題となる）．

実は，分析の目的は (iii) 因果関係の確立であることが多いであろうし，それが現実問題では最も重要な知見であろう．ところが，データの取り方や分析の手順が「記述」や「回帰」であったにもかかわらず，その解釈に当たって「因果」としてしまう誤りが多いと思われる．例えば下の例では，散布図の各点は異なる商品であり，求められた単回帰式 $y = 34.77 + 3.38x_1$ の解釈として，Web更新回数が1回多い商品の売り上げは平均的に3.38だけ大きい，という解釈は妥当であるが，ある商品のWeb更新回数を「1 回増やせば売り上げは 3.38 だけ増える」という解釈は，厳密にいえば誤りである．記述もしくは予測の目的であったものを因果に解釈してしまっているのである．

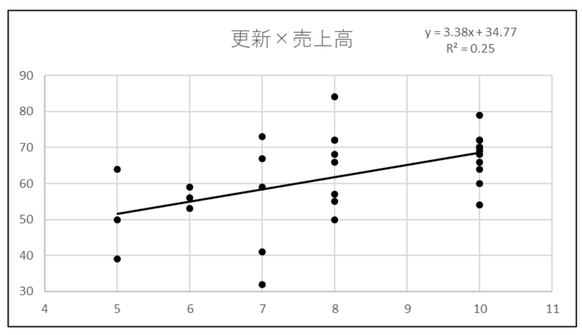
【説明変数の役割り】

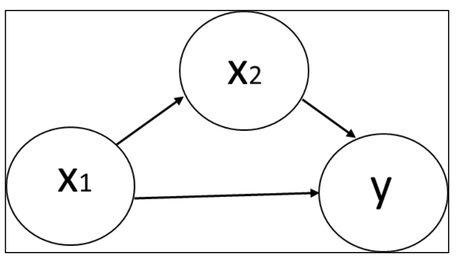
のである．重回帰モデルにおける各説明変数は $x_1, \ldots, x_p$ と単に記号でしか表現されていないが，実際の応用では，それぞれが固有の特徴と役割りを持つものであり，それらを的確に分析に反映させる必要がある．因果推論的には，アウトカム，処置変数，共変量，交絡変数，中間変数の区別をつけるのが重要とされる．下の例では，売り上げ (y) がアウトカム，Web更新回数 (x_1) が処置変数，Webアクセス (x_2) が中間変数となる．Web更新回数 (x_1) と売り上げ (y) の関係は $y = 34.77 + 3.38x_1$ であり，x_1 の係数は正であるので，Web更新回数が多ければ売り上げが多いという結果となる．ところがWebアクセス数 (x_2) を取り込んだモデルでは $y = -3.02 - 0.68x_1 + 0.89x_2$ と x_1 の係数が負になってしまう．これは，中間変数 (x_2) をモデルに取り込んでしまった誤りである．また，Webアクセス数 (x_2) と売り上げ (y) の関係を見たい場合には，Web更新数 (x_1) が交絡変数の役割となるので，重回帰モデルには x_1 を入れなくてはいけないことになる．データとして与えられている項目があるからといって，何でもかんでもモデルに取り込んでいいとは限らないし，逆にモデルに入れなければいけない変数もある．機械的にはいかない．

【例：Web 更新回数とWebアクセス回数が商品の売り上げに影響を与えるか】

簡単な例としてある商品の売り上げ (y) に対してWeb更新回数 (x_1) とWebアクセス数 (x_2) がどのように影響するかを取り上げる（岩崎 (2019)）．各変量の基本統計量およびWeb更新数 (x_1) と売り上げ (y) の散布図は以下のようである．そして重回帰式は $y = -3.02 - 0.68x_1 + 0.89x_2$ と求められる．また，各変量間の関係も以下に示した．

統計量	更新 (X1)	アクセス (X2)	売上高 (Y)
平均	8.07	79.37	62.03
標準偏差	1.74	12.05	11.81
相関	更新 (X1)	アクセス (X2)	売上高 (Y)
更新 (X1)	1	0.660	0.498
アクセス (X2)	0.660	1	0.841
売上高 Y)	0.498	0.841	1





【おわりに】

この発表は，1冊の書物を必要とする話題を限られたスペースでまとめたものであるが，本来はじっくり考えながら勉強すべき話題である．統計的データ解析で最重要な手法である重回帰分析は，誠に奥が深い．

【参考文献】

阿部貴行・岩崎　学（共訳）(2021) ローゼンバウム統計的因果推論入門．共立出版 (Rosenbaum, P. R. (2017) *Observation & Experiment*. Harvard University Press の日本語訳)．

岩崎　学 (2015) 統計的因果推論．「統計解析スタンダード」シリーズ．朝倉書店．

岩崎　学 (2019) 事例で学ぶ！あたらしいデータサイエンスの教科書．翔泳社．

岩崎　学 (2021) 統計的因果推論の視点による重回帰分析．日本統計学会誌シリーズJ，**50**，1-17．

Gelman, A., Hill, J. and Vertari, A. (2020) *Regression and Other Stories*. Advanced Methods for Social Research Series. Cambridge University Press.