【SLAA】

Field, A. (2009) Discovering Statistics Using SPSS(pp. 264-272)

8.1 What will this chapter tell me?

■本章は回帰分析の中でもカテゴリカルな結果を予測する (i.e., 従属変数が名義尺度である) ロジスティック回帰について概説をする

8.2 Background to logistic regression

■ロジスティック回帰は従属変数が名義尺度であれば、独立変数は連続変数でも名義尺度でもありうる

(例) laziness, pig-headedness, alcohol consumption, number of burps という独立変数から、あるデータの性別を予測する

(例) ある患者のデータ (独立変数) から、その患者にできた腫瘍が悪性か良性かを予測できる

■従属変数が 2 値データの場合を二項ロジスティック回帰分析、それ以上のカテゴリの場合は多項ロジスティック回帰分析と呼ぶことがある

8.3 What are the principles behind logistic regression

■単回帰分析の場合の回帰式は以下のようになる

$$Y_i = b_0 + b_1X_{1i} + \varepsilon_i$$

※Y_i：従属変数、b_0：切片、b_1：回帰係数、X_i：独立変数、ε_i：誤差

■(二項) ロジスティック回帰の場合回帰式は以下のようになる

$$P(Y) = \frac{1}{1+e^{-(b_0 + b_1 X_1)}}$$

※Y_i：従属変数、b_0：切片、b_1：回帰係数、X_i：独立変数、e：自然対数

■線形回帰とロジスティック回帰の違いは、前者は従属変数の値を予測するのに対して、後者は従属変数の確率を予測する (probability of Y)

■独立変数が複数になったとしても線形回帰式とロジスティック回帰式の関係性は同様である

■このような類似性があるのにも関わらず線形回帰をロジスティック回帰に用いることができない理由としては、線形回帰が独立変数と従属変数の関係が線形関係を前提としているためである

■線形回帰の章で線形関係の前提が満たされない場合の対処法として「対数変換」を提示したが、これは線形関係にない変数間の関係をロジットによって線形関係にするものである

■ロジスティック回帰はこの論理に従って、線形関係が満たされない場合の回帰を行う方法と言える

■ロジスティック回帰式には変形することが可能であるが、本書では Y が生じる可能性 (Y があるカテゴリに属しているかどうかの確率) を示す式を採用している

■線形回帰と同様、分析においては従属変数と独立変数の関係から各変数の (偏)回帰係数が算出されるが、この過程は最尤法を使って行われる

8.3.1 Assessing the model: the log-likelihood statistic

■ロジスティック回帰分析はある独立変数と従属変数の関係から、従属変数が生じる確率を予測する式を立てる。そしてあるケースにおいて従属変数が生じるかどうかを検討することができる

■線形回帰分析において回帰式のモデル説明率 (R^2) があるように、ロジスティック回帰においても回帰式のモデル説明率がある。ロジスティック回帰においては対数尤度によって導かれる

■対数尤度が示すモデル説明率 (式 8.5) は、予測された従属変数が生じる確率と実際の従属変数が生じる確率の関係性を示すものである

■対数尤度は線形回帰分析における残差の 2 乗値の概念と類似したものであり、対数尤度が大きいほど独立変数によって説明されない割合が高くなるため、その回帰式の適合度は高くないことを示す

■良い回帰モデルを選定するには、異なる回帰モデルに対する対数尤度の算出と各モデル間の対数尤度を比較する必要がある

■そのため、あるロジスティック回帰モデルをそのベースライン (偏回帰係数の値が 0 であるために切片のみからなるモデル) と比較する

■この場合のベースラインの切片は、従属変数のカテゴリの中でも最も生じる確率が高いものである

■そして式 (8.6) に示されているように、あるロジスティック回帰モデルとベースラインの差を 2 倍して得られる値はカイ二乗分布に従うとされている

■この時の自由度はパラメータの数であり、あるロジスティック回帰のパラメータの数 (独立変数＋1) とベースラインのパラメータの数 (常に 1) の差がカイ二乗分布の自由度となる

※カイ二乗値が有意であれば、立てたモデルによって説明率が向上している＝モデルの有効性がわかる

8.3.2 Assessing the model: R and R^2

■ロジスティック回帰分析においても、線形回帰分析のような R 値を算出することが可能である

■この場合の R 値は従属変数とそれぞれの独立変数との偏相関によって算出し、正の値であれば独立変数の値が大きくなるほど従属変数の確率が大きくなる、負の値であればその逆となることを示す

■そのため、R 値が大きいほどモデルの説明率が大きくなると解釈することが可能である

■R 値は (式 8.7) の形で算出することもできるが、この式に含まれる Wald 検定はある条件下においては不適切な値を取る場合があるため注意が必要である (この R 値を線形回帰で使用するのは妥当でない)

■線形回帰における R^2 に対応する値を算出する方法には議論があるが、SPSS にて呼ばれる Hosmer and Lemeshow's R^2_L を用いると比較的容易に算出することが可能である (式 8.8)

■この式においてはロジスティック回帰モデルの対数尤度をベースモデルの対数尤度でわることを意味しており、モデルを立てることによってベースラインからどの程度説明率が向上しているか (i.e., モデルの説明率) を示すことができる

■しかし、実際に SPSS で算出が行われる値は Cox and Snell's R^2_{cs} であり、(式 8.9) のようになる

■ただし、この式では理論的に最大値が 1 に達しないと考えられるため、それを修正した Nagelkerke's R^2_N の式も存在する

■いずれの式に関してもある程度似たような値が算出されるが、厳密には得られる値は異なっているため、解釈においてはモデルの実質的な有意性を示す基準として使用するとよい

8.3.3 Assessing the contribution of predictors: the Wald statistic

■線形回帰分析において、各独立変数の偏回帰係数の有意性は t 検定にて検定された

■ロジスティック回帰分析においてはその手法に対応するものとして Wald 検定があり (式 8.11)、この検定はカイ二乗分布に基づくものである

■線形回帰分析における t 検定と同様に、Wald 検定も各偏回帰係数が有意に 0 でないことを示す

■Wald 検定は各独立変数の偏回帰係数をその標準誤差でわった値であり、その点でも線形回帰分析の t 検定と類似している

■その一方で、偏回帰係数が極端に大きい値の場合には標準誤差が大きくなる傾向にあるため Wald 検定よりも尤度比の方が優れている場合もある

■この標準誤差の増大傾向は、本来は偏回帰係数が 0 でない (予測に貢献する) のにも関わらず、そうでないとしてしまうというタイプⅡエラーを誘発する

8.3.4 The odds ratio: Exp (*B*)

■ロジスティック回帰分析の解釈により重要な値として、オッズ比がある

■オッズ比は独立変数が増えることによって従属変数の確率がどのように変化するかを示す

※オッズ比とはある事象の起こり易さを 2 つの群で比較して示す統計学的な尺度であるとされ、各独立変数が従属変数に与える影響 (偏回帰係数) の大きさを示すと考えられる

■例として独立変数も二値データであるケースとして、避妊具の使用の有無 (独立変数) によって妊娠する (従属変数) 確率がどの程度予測されるかを考えてみる

■オッズとはある事象が起こりうる確率をその事象が起こらない確率でわった値であり、今回の例においては妊娠するオッズは妊娠する確率を妊娠しない確率でわった値となる

■避妊具を使用した場合と使用しなかった場合のそれぞれについてオッズを計算し、これらのオッズの割合、つまりオッズ比を算出する

■オッズ比の算出に際しては、(式 8.3) に基づく、今回の事例においては切片と回帰係数×独立変数 (0 or 1) の場合のロジスティック回帰式となる (式 8.12)

■そして、独立変数が 1 の場合のオッズ (回帰係数×独立変数＋切片) と独立変数の値が 0 の場合 (切片のみ＝ベースライン) のオッズでわることによって算出される

■このように、オッズ比は各独立変数含めた回帰式に占める元の回帰式とで従属変数の確率の割合を示すことになるため、(式 8.13) のオッズ比が 1 よりも大きい場合には各独立変数によって従属変数の確率が上昇することを示す

8.3.5 Methods of logistic regression

■線形回帰と同様にロジスティック回帰にも独立変数の投入の仕方によって複数の分析手法 (強制投入法、ステップワイズ法) が存在する

■強制投入法 (階層的投入法も含む？) は線形回帰の場合と同様に、(分析者が選択した) 独立変数を一度に回帰モデルに投入する方法である

■ステップワイズ法は大きく分けて変数増加法と変数減少法に大別される

■変数増加法においては、まず切片のみのベースラインの回帰式、次に統計的に従属変数に与える score (オッズ比？) が大きい独立変数の順に 1 つずつ投入された回帰式が作成され、その従属変数への影響が有意である独立変数がなくなるまでその作業を繰り返す

■ただし､その過程において回帰モデルに必要なくなった独立変数 (e.g., 相関の大きい変数など) については除去される場合もある

■変数が除去されるプロセスとしては､18.3.3 にあるような最尤比検定によって現在の回帰モデルとある独立変数を除去した場合の回帰モデルを比較した場合に、モデルの説明率が有意に減少するかどうかを調べる

■ある独立変数を除去することで現在の回帰モデルの説明率が有意に減少する場合、その変数はモデルに必要であるとして回帰式から除去されない

■それに対して、ある独立変数を除去しても現在の回帰モデルの説明率が有意に減少しない場合には、その変数を除去することに問題がないと判断されるため、回帰式から除去される

■最尤比検定の他にも、Forward: Conditional や Wald 検定が用いられる場合があるが、これらの手法の中でも最尤比検定が最も信頼性が高いと考えられる

■これに対し、変数減少法は変数増加法と同じ基準を用いて全く逆のプロセスを経ることで行われる。つまり、全てを投入したモデルから影響力が小さい独立変数を除去していくということである

■これらの中で分析においてどの手法を採用するかの選択については線形回帰の場合と類似しており、理論的根拠がある場合には強制投入法にて分析を行うのに対して、探索的な分析の場合にはステップワイズ法が推奨される (ただしステップワイズが必ずしも探索的とは限らない)

■また、ステップワイズ法にて分析を行う場合には、変数増加法よりは変数減少法が推奨される。これは、suppressor effects が生じることによってタイプⅡエラーを犯す可能性があるためである