判別分析

渡邊直樹

2021 年 10 月 6 日

はじめに

たとえば，医療において，ある患者に関するいくつかの検査指標から彼または彼女が特定の疾患に罹患しているか否かを判別したいとしよう．このとき，これまでに蓄積された各々の検査指標データに受検者が罹患者であるか非罹患者であるかというデータが紐づけられていれば，それらのデータと彼または彼女自身の検査指標を照合することで，その判別を行うことができるだろう．この分析目的を参照するならば，罹患者であるか否かを 0 か 1 で表し，それに対応するいくつかの検査指標の組み合わせをできるだけ「正しく 2 群に分類する」ことこそが本来の判別分析（discriminant analysis）である[1]．

しかし，正しく分類するのではなく，「データの当てはまりの良い式」を線形回帰や 2 項ロジスティック回帰を用いて求め，その理論値（fitted value）の大きさで「判定」することが，実務上の簡便法として，判別分析と呼ばれることがある．ここではこの簡便法に関してノートする．

線形判別分析では線形回帰が補助的に用いられる．しかし，被説明変数の予測値が 0 を下回ったり，1 を上回ったりすることがしばしば起こる．たとえば，特定の商品の購入確率を求めたい場合には，被説明変数の予測量は 0 と 1 の間に値を取らねばならない．この性質が保証されている推計方法の一つが 2 項ロジスティック回帰である．以下では，これらを用いた（簡便法としての）判別分析について，順を追ってその手続きと注意点を説明する．

[1] この例からも分かるように，顧客の購買行動に関する判別分析を行いたい場合，購入者の（性別，年齢，居住地域などに関する）属性データだけを入手しても，購買を控えた顧客の属性データもないと，顧客を 2 群に分ける要因を偏りなく探ることは極めて難しいことに注意してほしい．

線型判別分析

係数の推定

離職するか否かといった 2 つのカテゴリーに分けられ，それぞれ 0 と 1 という数値を与えられた被説明変数をいくつかの説明変数の加重和で表わそう．m 個の説明変数について，個人 1 から個人 n までの属性データが手許にあるとき，線形回帰モデルは

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_m X_{mi} + u_i \tag{1}$$

と書ける．ここで，u_i は誤差項であり，説明変数の加重和では説明できない個人 i の離職行動 Y_i を確率的な変動によるものとみなす．各期の u_i は正規分布に従うと仮定されることが多い．判別分析では (1) 式のような線形回帰モデルを線形判別関数という．説明変数の個数 $l(\leq m)$ や定数項 β_0 の有無により，いくつもの線形判別関数を設定できる．

(1) 式のパラメータの推定値を $\hat{\beta}_k$（$k = 1, \ldots, m$）とすると，

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_m X_{mi} \tag{2}$$

を判別スコア（discriminant score）という．また，個人 $n+1$ に関する属性データ $(X_{1n+1}, X_{2n+1}, \ldots, X_{mn+1})$ が与えられたとき，$\hat{Y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 X_{1n+1} + \hat{\beta}_2 X_{2n+1} + \cdots + \hat{\beta}_m X_{mn+1}$ を予測値（predicted value）という．

閾値

個人 i の判別スコアが 0.5 以上であれば $Y_i^* = 1$ に，それ以下であれば $Y_i^* = 0$ と判定するとしてみよう．個人 $n+1$ の場合，予測値を判別スコアとして用いる．実際，この基準はしばしば使われる．個人 1 から個人 n までの属性データによる判別の精度を次の式で確認しておこう．

$$\text{判別精度} = \frac{\text{正しく判定できた件数}}{\text{属性データを使用した個人の総数 } n}. \tag{3}$$

ここで，個人 i に関する正しい判定とは $Y^* = Y_i$ を意味する．

線形判別関数に組み込む説明変数を選択する際には，それらの係数の推定値に関する p 値や AIC だけではなく，判別精度がより高い説明変数の組み合わせという基準も参照することが重要である．

注意点

前述の基準で正しく判定できなかった個人の判別スコアは0.5からどれくらいまで離れたところまで観察されるだろうか？多くの場合，0.5周辺には正しく判定されなかった個人の判別スコアがいくつも見つかだろう．正しく判定できなかった個人の属性データの特徴を必ず確認しておこう．

人によっては，0と1の間に区間 $[a, b]$ を設定して，判別スコアが b 以上であれば $Y_i^* = 1$，a 以下であれば $Y_i^* = 0$ と判定し，判別スコアが区間 $[a, b]$ 内に値をとった場合には「判定不能」とすることもある．分析目的に合わせて，適宜，このような閾値の区間を設定してほしい．たとえば，$Y_i = 1$ となる個人の判別精度を重視するならば，閾値を1.0と設定し，判別スコアが1.0以上であれば $Y_i^* = 1$ と判定し，それ以下であれば判定不能とすることもある．判定不能な個人を含む場合，判別精度の分母は判別可能だった個人の件数 n' をとってもよいかもしれない．いずれにせよ，閾値と判別精度の基準は明示すべきである．

2項ロジスティック回帰分析を用いた判別分析

同じデータに対して，2項ロジスティック回帰モデルを（非線形）判別関数として適用することもできる．

$$\ln \frac{Y_i}{1 - Y_i} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_m X_{mi} + u_i \tag{4}$$

ここで，$\ln(x)$ は x のネイピア数を底とする対数関数である．$Y_i/(1 - Y_i)$ をオッズ比（odds ratio）という[2]．実際には，$Y_i = 1$ のときにはオッズ比の分母が0になってしまうので，推計式では $Y_i = 1$ のときには1に非常に近い数値（たとえば，0.999）で代用する．

各推計式における個人 i に関する残差（residual）は

$$e_i = \ln \frac{Y_i}{1 - Y_i} - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_m X_{mi})$$

である．モデルにおける説明変数が確定すると，個人 i' が $Y_{i'} = 1$ を選択する確率は

$$y_{i'} = \hat{\beta}_0 + \hat{\beta}_1 X_{1i'} + \hat{\beta}_2 X_{2i'} + \cdots + \hat{\beta}_m X_{mi'} \tag{5}$$

[2] $Y_i/(1 - Y_i)$ をオッズと呼んでもよいが，実際の分析では0と1の値をそのまま使うことはできず，近似確率として0.0001や0.999を使用しているので，ここではそれらを2群における生起確率とみなしてオッズ比と呼んでいる．

から，自然対数の定義より，次のようにして求められる．

$$\mathrm{Prob}(Y_{i'}=1)=\frac{\mathrm{EXP}(y_{i'})}{1+\mathrm{EXP}(y_{i'})}. \tag{6}$$

ここで，EXP(x) はネイピア数 e を底とする指数関数を表す Excel の記法である．確率 $\mathrm{Prob}(Y_i=1)$ を 2 項ロジスティック回帰を用いた判別分析における個人 i の判別スコアという．個人 $n+1$ に関する属性データが与えられたとき，確率 $\mathrm{Prob}(Y_{n+1}=1)$ を予測値という．

判定方法については，線形判別分析に関する説明に付した閾値の設定と注意点に関する説明が適用される．

判別精度の計算

ここでは Excel の使用を前提として，判別精度の計算方法の一例を記しておく．個人 i の判別スコアがシートにおけるセル（小声は C2 と付番しておく）に計算されているとする．閾値を 0.5 に設定すると右隣のセル（つまり，D2）に次のように打ち込み、Enter（return）キーを押す．

=IF(C2>0.5, 1, 0)

IF 関数は，IF(条件式, 条件に合致していれば（true であれば）1 を出力，条件に合致していなければ（false であれば）0 を出力）という形で新たな変数や文字列を生成する．条件式に等号を含めたいのであれば，$>$ の代わりに $>=$ とする．次に，このセルをコピーし，その下のセルにペーストして，すべての個人について閾値による判定を行なった結果を表示させる．このとき，IF 文の条件式においてC2 とは書かないこと．これは絶対参照といって，コピーする際に必ず C2 のセルを参照するので，D3 以下のセルにペーストしても，D2 の値がコピーされるだけになってしまう．

（注）元のデータにおいて，$Y_i=1$ と $Y_i=2$ といった具合に，0 と 1 ではなく，1 と 2 でカテゴリーに付番してあれば，閾値は 0.5 ではなく，1.5 に設定しなければならない．Y_i を IF 関数を使って 0 と 1 の表記に書き換えれば，閾値を 0.5 とする．

以上の作業で正しく判定できた件数をカウントできたので，(3) 式にその値を書き入れて，判別精度を計算する．さらに細かくみていくには $Y_i=1$ の時に正しく判別できたか，$Y_i=0$ のときはどうかにも気を遣いたい．