8. 判別分析

概要: 判別分析は、二つの母集団を設定し、あるサンプルがどちらの母集団に属するのかを推測する方法である。この解析においては、目的変数がカテゴリー型、説明変数が数値型のデータを扱う。まず、目的変数の結果がわかっている場合のデータを利用し、関連するパラメータを評価する。次に、新たな説明変数のセットが与えられる。各母集団との距離を定義し、説明変数のセットはその距離の短いほうの母集団に属すると推測する。さらに、用意した説明変数が適切かどうかの判断をする。

キーワード:判別関数;判別効率;変数選択;距離;マハラノビスの距離

8.1. 序

二つのグループがあるとする。あらかじめ種々のデータがあり、その場合にどちらのグループに属するかの結果がわかっているとする。そこにあるメンバーが来た場合、その関連するデータからそのメンバーはどちらのグループに属するのかを判別したい。それが、判別分析の主な主題である。

しかし、判別分析に取り組む前に考慮しなければならないことがある。新しいデータ(メンバー)が登場した場合、それがここで問題にするグループに属しているかどうかを判断しなくてはならない。つまり、判別するステップにすすんでいいかを判断する必要がある。ここで、グループに属しているならば、判別分析にすすんでいくことになる。しかし、そこでこのグループに属していないと判断されるならば、判別分析する必要がない。

ここでは、まず判別にすすむかどうかを議論する。

次に、判別分析にすすむ場合を議論する。判別分析では、新たなデータと各グループの距離を定義し、その距離が短いグループに属すると判断する。

以上の解析では説明変数は判別に全て有効である、という仮定を暗にしている。しかし、全ての説明変数が有効であるとは限らない。ここでは、その説明変数の有効性を判別効率を定義し、統計的に評価する方法を示す。

8.2. 判別するステップに進むかの判断

8.2.1. 1 変数の場合の距離

ここでは、まず 1 変数を考える。集合は平均 μ、分散 σ^2 で特徴づけられる。この場合、標準偏差 σ は $\sigma=\sqrt{\sigma^2}$ で与えられる。この集合に属するデータは正規分布 $N\left[\mu,\sigma^2\right]$ に従う

と仮定する。

ここに新たなデータ x が来たと考える。このデータはこの集団に属するのかどうかを判断したい。この場合、規格化数

$$u = \frac{x-\mu}{\sigma} \tag{1}$$

を考える。これは標準正規分布 $N\left[0,1^2\right]$ に従う。

したがって、このデータがこの集合に属するかどうかは確率と連動させることができる。推定確率 P を指定すれば、それに応じた P 値である z_P が以下のように評価される。

$$\begin{aligned} P &= \int_{-z_P}^{z_P} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz \\ &= Erf\left(\frac{z_P}{\sqrt{2}}\right) \end{aligned} \tag{2}$$

ここで Erf は誤差関数である。これから

$$z_P = \sqrt{2} Erf^{-1} P \tag{3}$$

となる。この z_p は P 値と呼ばれる。

$$\mu - z_P\sigma \le x \le \mu - z_P\sigma \tag{4}$$

であれば、データは集団に属し、その範囲外でればその集団に属さないと判断される。これは規格化変数でいうと

$$-z_P \le u \le z_P \tag{5}$$

と簡単に表記される。

この評価法は 1 変数の場合に有効であるが、2 変数以上に単純に拡張することができない。そこでも容易に拡張して評価できるようにしたい。そのためには、それを 2 乗した量で評価すると便利である。

この規格化変数の平方

$$u^2 = \left(\frac{x-\mu}{\sigma}\right)^2 \tag{6}$$

は自由度 1 の χ^2 分布に従う。したがって、

$$u^2 \le \chi^2(1,P) \tag{7}$$

として、この集団に属するかどうかを判断する。以後はこちらのほうの判断基準を採用する。

8.2.2. 2 変数の場合

次に二変数 X_1、X_2 の場合を扱う。この場合は、データはペアで (x_1,x_2) の形式で存在する。それぞれの平均を μ_1、μ_2、分散を σ_1^2、σ_2^2 と置く。この場合の確率密度分布は

$$
\begin{aligned}
&f\left(x_1,x_2\right)dx_1dx_2 \\
&=\frac{1}{2\pi\sqrt{\left(1-\rho^2\right)\sigma_{11}^{(2)}\sigma_{22}^{(2)}}}\exp\left[-\frac{1}{2\left(1-\rho^2\right)}\left(\frac{\left(x_1-\mu_1\right)^2}{\sigma_{11}^{(2)}}-2\rho\frac{\left(x_1-\mu_1\right)\left(x_2-\mu_2\right)}{\sqrt{\sigma_{11}^{(2)}\sigma_{22}^{(2)}}}+\frac{\left(x_2-\mu_2\right)^2}{\sigma_{22}^{(2)}}\right)\right]dx_1dx_2
\end{aligned}
\tag{8}
$$

となる。これは、正規分布 $N\left[\mu_1,\sigma_{11}^{(2)}\right]$ および $N\left[\mu_2,\sigma_{22}^{(2)}\right]$ に従い、お互いの相関係数が ρ である二変数 x_1 、 x_2 に対応する確率密度分布である。

ここで、さらに変数変換

$$
\begin{cases}
u_1=\dfrac{x_1-\mu_1}{\sqrt{\sigma_{11}^{(2)}}} \\
u_2=\dfrac{x_2-\mu_2}{\sqrt{\sigma_{22}^{(2)}}}
\end{cases}
\tag{9}
$$

と置くと

$$
\begin{aligned}
&f\left(x_1,x_2\right)dx_1dx_2 \\
&=\frac{1}{2\pi\sqrt{\left(1-\rho^2\right)\sigma_{11}^{(2)}\sigma_{22}^{(2)}}}\exp\left[-\frac{1}{2\left(1-\rho^2\right)}\left(\frac{\left(x_1-\mu_1\right)^2}{\sigma_{11}^{(2)}}-2\rho\frac{\left(x_1-\mu_1\right)\left(x_2-\mu_2\right)}{\sqrt{\sigma_{11}^{(2)}\sigma_{22}^{(2)}}}+\frac{\left(x_2-\mu_2\right)^2}{\sigma_{22}^{(2)}}\right)\right]dx_1dx_2 \\
&=\frac{1}{2\pi\sqrt{\left(1-\rho^2\right)\sigma_{11}^{(2)}\sigma_{22}^{(2)}}}\exp\left[-\frac{1}{2\left(1-\rho^2\right)}\left(u_1^2-2\rho u_1u_2+u_2^2\right)\right]\sigma_1\sigma_2du_1du_2 \\
&=\frac{1}{2\pi\sqrt{\left(1-\rho^2\right)}}\exp\left[-\frac{u_1^2-2\rho u_1u_2+u_2^2}{2\left(1-\rho^2\right)}\right]du_1du_2 \\
&=f\left(u_1,u_2\right)du_1du_2
\end{aligned}
\tag{10}
$$

となる。

これから

$$
f\left(u_1,u_2\right)=\frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left[-\frac{u_1^2-2\rho u_1u_2+u_2^2}{2\left(1-\rho^2\right)}\right]
\tag{11}
$$

を得る。

相関行列は

$$
R=\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}
\tag{12}
$$

と置けるから、

$$
f\left(u_1,u_2\right)=\frac{1}{2\pi\sqrt{|R|}}\exp\left(-\frac{D^2}{2}\right)
\tag{13}
$$

となる。ただし、

$$
\begin{aligned}
D^2 &= \left(u_1, u_2\right)\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}^{-1}\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \\
&= \left(u_1, u_2\right) R^{-1}\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}
\end{aligned}
\tag{14}
$$

である。

ここで、

$$
\begin{aligned}
D^2 &= \begin{pmatrix} u_1 & u_2 \end{pmatrix} R^{-1}\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \\
&= \frac{u_1^2 - 2\rho u_1 u_2 + u_2^2}{1-\rho^2} \\
&= \frac{1}{1-\rho^2}\left[\frac{\left(x_1-\mu_1\right)^2}{\sigma_{11}^{(2)}} + \frac{\left(x_2-\mu_2\right)^2}{\sigma_{22}^{(2)}} - 2\rho\frac{\left(x_1-\mu_1\right)}{\sqrt{\sigma_{11}^{(2)}}}\frac{\left(x_2-\mu_2\right)}{\sqrt{\sigma_{22}^{(2)}}}\right] \\
&= \frac{1}{\left(1-\rho^2\right)\sigma_{11}^{(2)}\sigma_{22}^{(2)}}\begin{pmatrix} x_1-\mu_1 & x_2-\mu_2 \end{pmatrix}\begin{pmatrix} \sigma_{22}^{(2)} & -\rho\sqrt{\sigma_{11}^{(2)}\sigma_{22}^{(2)}} \\ -\rho\sqrt{\sigma_{11}^{(2)}\sigma_{22}^{(2)}} & \sigma_{11}^{(2)} \end{pmatrix}\begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix} \\
&= \begin{pmatrix} x_1-\mu_1 & x_2-\mu_2 \end{pmatrix}\begin{pmatrix} \sigma_{11}^{(2)} & \sigma_{12}^{(2)} \\ \sigma_{12}^{(2)} & \sigma_{22}^{(2)} \end{pmatrix}^{-1}\begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix} \\
&= \left(\boldsymbol{x}-\boldsymbol{\mu}\right)^T \Sigma^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)
\end{aligned}
\tag{15}
$$

ただし

$$
\boldsymbol{x}-\boldsymbol{\mu} = \begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix} \tag{16}
$$

$$
\Sigma = \begin{pmatrix} \sigma_{11}^{(2)} & \sigma_{12}^{(2)} \\ \sigma_{12}^{(2)} & \sigma_{22}^{(2)} \end{pmatrix} \tag{17}
$$

である。

さらに

$$
\begin{aligned}
\left|\Sigma\right| &= \begin{vmatrix} \sigma_{11}^{(2)} & \sigma_{12}^{(2)} \\ \sigma_{12}^{(2)} & \sigma_{22}^{(2)} \end{vmatrix} \\
&= \sigma_{11}^{(2)}\sigma_{22}^{(2)} - \sigma_{12}^{(2)} \\
&= \sigma_{11}^{(2)}\sigma_{22}^{(2)}\left(1 - \frac{\sigma_{12}^{(2)}}{\sigma_{11}^{(2)}\sigma_{22}^{(2)}}\right) \\
&= \sigma_{11}^{(2)}\sigma_{22}^{(2)}\left(1-\rho^2\right) \\
&= \sigma_{11}^{(2)}\sigma_{22}^{(2)}\left|R\right|^2
\end{aligned}
\tag{18}
$$

以上より、

$$
\begin{aligned}
f(u_1,u_2)du_1du_2 &= \frac{1}{2\pi\sqrt{|R|}}\exp\left(-\frac{D^2}{2}\right)du_1du_2 \\
&= \frac{1}{2\pi\sqrt{|\Sigma|}}\exp\left(-\frac{D^2}{2}\right)dx_1dx_2 \\
&= f(x_1,x_2)dx_1dx_2
\end{aligned}
\tag{19}
$$

となる。ただし、

$$
\begin{aligned}
D^2 &= \begin{pmatrix} u_1 & u_2 \end{pmatrix} R^{-1} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \\
&= \begin{pmatrix} x_1-\mu_1 & x_2-\mu_2 \end{pmatrix} \Sigma^{-1} \begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}
\end{aligned}
\tag{20}
$$

である。つまり、D^2は変数に変換に依存せず、同じ表式になる。

この場合は、そのデータがこの集合に属するかは

$$
D^2 \le \chi^2(2,P) \tag{21}
$$

を満足しているかで判断できる。

8.2.3. 多変量の場合

次にm個の多変数$X_1,X_2,\cdots,X_m$の場合を扱う。この場合は、データはペアで$(x_1,x_2,\cdots,x_m)$で存在する。それぞれの平均を$\mu_1,\mu_2,\cdots,\mu_m$、分散を$\sigma_1^2,\sigma_2^2,\cdots,\sigma_m^2$と置く。

多変数の場合は、2変数の表式を拡張して以下となる。

$$
\begin{aligned}
& f(x_1,x_2,\cdots,x_m)dx_1dx_2\cdots dx_m \\
&= \frac{1}{\left(\sqrt{2\pi}\right)^m\sqrt{|\Sigma|}}\exp\left[-\frac{D^2}{2}\right]dx_1dx_2\cdots dx_m \\
&= f(u_1,u_2,\cdots,u_m)du_1du_2\cdots du_m \\
&= \frac{1}{\left(\sqrt{2\pi}\right)^m\sqrt{|R|}}\exp\left[-\frac{D^2}{2}\right]du_1du_2\cdots du_m
\end{aligned}
\tag{22}
$$

ただし

$$
\begin{aligned}
D^2 &= (\mathbf{x}-\boldsymbol{\mu})^T\,\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) \\
&= \mathbf{u}^T R^{-1}\mathbf{u}
\end{aligned}
\tag{23}
$$

$$
\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix} \tag{24}
$$

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{pmatrix} \tag{25}$$

$$\boldsymbol{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_m \end{pmatrix} \tag{26}$$

$$\Sigma = \begin{pmatrix} \sigma_{11}^{(2)} & \sigma_{12}^{(2)} & \cdots & \sigma_{1m}^{(2)} \\ \sigma_{21}^{(2)} & \sigma_{22}^{(2)} & \cdots & \sigma_{2m}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1}^{(2)} & \sigma_{m2}^{(2)} & \cdots & \sigma_{mm}^{(2)} \end{pmatrix} \tag{27}$$

$$R = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1m} \\ \rho_{21} & 1 & \cdots & \rho_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{m1} & \rho_{m2} & \cdots & 1 \end{pmatrix} \tag{28}$$

である。

この集合に属しているは

$$D^2 \le \chi^2[m, P] \tag{29}$$

を満足しているかで判断する。

8.3. 説明変数間の相互作用を考慮しない判別分析

前節の評価で、新たなデータがこの集合体に属していると判断されたとする。

具体的な例でいうと、ある m 個の検査によって二つの病気 G_1 と G_2 のいずれかであると診断を下さなければならないとする。この他の病気の可能性はないとする。

ここでは、あるある m 個の検査がそれぞれ独立であるとする。

あらかじめ病名の確定している患者群のデータがあるものとする。このデータをもとに病名を同定する。この際 G_1、G_2 の平均は異なるが、母分散、共分散行列 Σ は共通であるとする。

8.3.1. 一変数による判別

まず 1 変数による判別を扱う。

データ全体の平均を μ、分散を σ^2 とする。データの中にはグループ A とグループ B があり、それぞれの平均を μ_A、μ_B とする。

そこに新たなデータ x が来た際、これはグループ A なのか B なのかを判断したい。

この場合、二つのグループに連動した平方距離

$$D_A^2 = \frac{\left(x-\mu_A\right)^2}{\sigma^2} \tag{30}$$

$$D_B^2 = \frac{\left(x-\mu_B\right)^2}{\sigma^2} \tag{31}$$

を評価し、その平方距離の小さいほうのグループに属すると判断する。

8.3.2. 多変数による判別

ここでは、前節の結果を拡張し、多変数による判別を扱う。

データとしては $X_1, X_2, \cdots, X_m$ の m 個の種類を扱う。それぞれの、全体の平均、分散は $\mu_1, \mu_2, \cdots, \mu_m$ 、$\sigma_1^2, \sigma_2^2, \cdots, \sigma_m^2$ とする。各種類の変数においてグループ A、B の平均を μ_{iA}, μ_{iB} と置く。ただし、$i = 1,2,\cdots,m$ である。この場合、二つの平方距離を

$$D_A^2 = \sum_{i=1}^{m} \frac{\left(x_i-\mu_{iA}\right)^2}{\sigma_i^2} \tag{32}$$

$$D_B^2 = \sum_{i=1}^{m} \frac{\left(x_i-\mu_{iB}\right)^2}{\sigma_i^2} \tag{33}$$

とする。

8.4. 説明変数間の相互作用を考慮した判別分析

前節の評価で、新たなデータがこの集合体に属していると判断されたとする。

具体的な例でいうと、ある m 個の検査によって二つの病気 G_1 と G_2 のいずれかであると診断を下さなければならないとする。この他の病気の可能性はないとする。

あらかじめ病名の確定している患者群のデータがあるものとする。このデータをもとに病名を同定する。この際 G_1 、G_2 の平均は異なるが、母分散、共分散行列 Σ は共通であるとする。

8.4.1. 一変数による判別

まず 1 変数による判別を扱う。

データ全体の平均を μ 、分散を σ^2 とする。データの中にはグループ A とグループ B があり、それぞれの平均を μ_A 、μ_B とする。

そこに新たなデータ x が来た際、これはグループ A なのか B なのかを判断したい。

この場合、二つのグループに連動した平方距離

$$D_A^2 = \frac{\left(x-\mu_A\right)^2}{\sigma^2} \tag{34}$$

$$D_B^2 = \frac{\left(x-\mu_B\right)^2}{\sigma^2} \tag{35}$$

を評価し、その平方距離の小さいほうのグループに属すると判断する。

8.4.2. 多変数による判別

ここでは、前節の結果を拡張し、多変数による判別を扱う。

データとしては $X_1, X_2, \cdots, X_m$ の m 個の種類を扱う。それぞれの、全体の平均、分散は $\mu_1, \mu_2, \cdots, \mu_m$ 、$\sigma_1^2, \sigma_2^2, \cdots, \sigma_m^2$ とする。各種類の変数においてグループ A、B の平均を μ_{iA}, μ_{iB} と置く。ただし、$i = 1,2,\cdots,m$ である。この場合、二つの平方距離を

$$D_A^2 = \left(\boldsymbol{x}-\boldsymbol{\mu}_A\right)^T \Sigma^{-1} \left(\boldsymbol{x}-\boldsymbol{\mu}_A\right) \tag{36}$$

$$D_B^2 = \left(\boldsymbol{x}-\boldsymbol{\mu}_B\right)^T \Sigma^{-1} \left(\boldsymbol{x}-\boldsymbol{\mu}_B\right) \tag{37}$$

ただし、

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix} \tag{38}$$

$$\boldsymbol{\mu}_A = \begin{pmatrix} \mu_{1A} \\ \mu_{2A} \\ \vdots \\ \mu_{mA} \end{pmatrix} \tag{39}$$

$$\boldsymbol{\mu}_B = \begin{pmatrix} \mu_{1B} \\ \mu_{2B} \\ \vdots \\ \mu_{mB} \end{pmatrix} \tag{40}$$

である。

8.5. 重みを考慮した判別分析

これまでの解析では判別する項目の重みを考慮していない。判別する項目の中である特定のものがより重要である場合がある。この場合は、その重要度を考慮したい。その方法をここでは示す。

項目数は k 個あるとする。それぞれに項目においてそれらは規格化されているものとする。

するとグループ A との距離を以下のように判断する。

$$D_A^2 = \left(w_1\left(z_1 - \mu_{z_{A1}}\right) \quad w_2\left(z_2 - \mu_{z_{A2}}\right) \quad \cdots \quad w_k\left(z_k - \mu_{zAk}\right)\right) R^{-1} \begin{pmatrix} w_1\left(z_1 - \mu_{z_{A1}}\right) \\ w_2\left(z_2 - \mu_{z_{A2}}\right) \\ \vdots \\ w_k\left(z_k - \mu_{zAk}\right) \end{pmatrix} \tag{41}$$

ただし、

$$R = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1k} \\ r_{21} & 1 & \cdots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k1} & r_{k2} & \cdots & r_{kk} \end{pmatrix} \tag{42}$$

である。ここで、$\mu_{z_{Ai}}, r_{ij}$ は既存のデータから評価でいる。w_i は既存データによく一致するように定める。

8.6. 判別関数

ここでは、判別関数を導出する。

前節では二つのうちどちらのグループに属するのかを判別する平方距離を導入した。

ここではもっと直接的に評価できる判別関数を導出する。

まず、二変数から始め、複数変数に拡張する。

8.6.1. 二変数の判別関数

ここでは、まず簡単な場合、つまり変数の種類が 2 個の場合を扱う。

二変数の説明変数で二つのグループ A、B を判別することを考える。

あるデータ $\left(x_1, x_2\right)$ が来た際にそれがグループ A に属するか B に属するかは、それぞれに対応する平方距離

$$D_A^2 = \frac{1}{1-\rho^2}\left[\frac{\left(x_1 - \mu_{1A}\right)^2}{\sigma_1^{(2)}} + \frac{\left(x_2 - \mu_{2A}\right)^2}{\sigma_2^{(2)}} - 2\rho \frac{\left(x_1 - \mu_{1A}\right)\left(x_2 - \mu_{2A}\right)}{\sqrt{\sigma_1^{(2)}\sigma_2^{(2)}}} \right] \tag{43}$$

$$D_B^2 = \frac{1}{1-\rho^2}\left[\frac{\left(x_1 - \mu_{1B}\right)^2}{\sigma_1^{(2)}} + \frac{\left(x_2 - \mu_{2B}\right)^2}{\sigma_2^{(2)}} - 2\rho \frac{\left(x_1 - \mu_{1B}\right)\left(x_2 - \mu_{2B}\right)}{\sqrt{\sigma_1^{(2)}\sigma_2^{(2)}}} \right] \tag{44}$$

を評価し、その小さいほうのグループに属する、と判断すればよかった。

その大小を判断する代わりに

$$\varsigma = D_B^2 - D_A^2 \tag{45}$$

を評価し

$$\begin{cases} \varsigma > 0 & \rightarrow Group\ A \\ \varsigma < 0 & \rightarrow Group\ B \end{cases} \tag{46}$$

とすればいい。このςのことを判別関数と呼ぶ。ここでは、この判別関数ςの表式を求める。

$$\begin{aligned} \varsigma &= D_B^2 - D_A^2 \\ &= \frac{1}{1-\rho^2}\left[\frac{(x_1-\mu_{1B})^2}{\sigma_1^{(2)}} + \frac{(x_2-\mu_{2B})^2}{\sigma_2^{(2)}} - 2\rho\frac{(x_1-\mu_{1B})(x_2-\mu_{2B})}{\sqrt{\sigma_1^{(2)}\sigma_2^{(2)}}}\right] \\ &\quad -\frac{1}{1-\rho^2}\left[\frac{(x_1-\mu_{1A})^2}{\sigma_1^{(2)}} + \frac{(x_2-\mu_{2A})^2}{\sigma_2^{(2)}} - 2\rho\frac{(x_1-\mu_{1A})(x_2-\mu_{2A})}{\sqrt{\sigma_1^{(2)}\sigma_2^{(2)}}}\right] \\ &= \frac{1}{1-\rho^2}\left[\frac{(x_1-\mu_{1B})^2-(x_1-\mu_{1A})^2}{\sigma_1^{(2)}} + \frac{(x_2-\mu_{2B})^2-(x_2-\mu_{2A})^2}{\sigma_2^{(2)}} - 2\rho\frac{(x_1-\mu_{1B})(x_2-\mu_{2B})-(x_1-\mu_{1A})(x_2-\mu_{2A})}{\sqrt{\sigma_1^{(2)}\sigma_2^{(2)}}}\right] \end{aligned} \tag{47}$$

となる。

ここで、括弧の中の第 1 項の分子を考える。

$$\begin{aligned} (x_1-\mu_{1B})^2-(x_1-\mu_{1A})^2 &= x_1^2 - 2\mu_{1B}x_1 + \mu_{1B}^2 - x_1^2 + 2\mu_{1A}x_1 - \mu_{1A}^2 \\ &= 2x_1(\mu_{1A}-\mu_{1B}) - (\mu_{1A}-\mu_{1B})(\mu_{1A}+\mu_{1B}) \\ &= \left[2x_1-(\mu_{1A}+\mu_{1B})\right](\mu_{1A}-\mu_{1B}) \\ &= 2\left(x_1-\frac{\mu_{1A}+\mu_{1B}}{2}\right)(\mu_{1A}-\mu_{1B}) \\ &= 2(x_1-\mu_1)(\mu_{1A}-\mu_{1B}) \\ &= 2(x_1-\mu_1)d_1 \end{aligned} \tag{48}$$

ただし、

$$\mu_1 = \frac{\mu_{1A}+\mu_{1B}}{2} \tag{49}$$

$$d_1 = \mu_{1A} - \mu_{1B} \tag{50}$$

である。

同様に、括弧の中の第 2 項の分子は

$$(x_2-\mu_{2B})^2-(x_2-\mu_{2A})^2 = 2(x_2-\mu_2)d_2 \tag{51}$$

である。ただし、

$$\mu_2 = \frac{\mu_{2A}+\mu_{2B}}{2} \tag{52}$$

$$d_2 = \mu_{2A} - \mu_{2B} \tag{53}$$

である。

次に、括弧の中の第 3 項の分子を考える。

$$
\begin{aligned}
&(x_1-\mu_{1B})(x_2-\mu_{2B})-(x_1-\mu_{1A})(x_2-\mu_{2A})\\
&=x_1x_2-\mu_{1B}x_2-\mu_{2B}x_1+\mu_{1B}\mu_{2B}-x_1x_2+\mu_{1A}x_2+\mu_{2A}x_1-\mu_{1A}\mu_{2A}\\
&=(\mu_{2A}-\mu_{2B})x_1+(\mu_{1A}-\mu_{1B})x_2-\mu_{1A}\mu_{2A}+\mu_{1B}\mu_{2B}\\
&=d_2x_1+d_1x_2-\mu_{1A}\mu_{2A}+\mu_{1B}\mu_{2B}
\end{aligned}
\tag{54}
$$

ここで、無理やり μ_1,μ_2 を使う。すなわち

$$
\begin{aligned}
&(x_1-\mu_{1B})(x_2-\mu_{2B})-(x_1-\mu_{1A})(x_2-\mu_{2A})\\
&=d_2x_1+d_1x_2-\mu_{1A}\mu_{2A}+\mu_{1B}\mu_{2B}\\
&=d_2(x_1-\mu_1)+d_2\mu_1+d_1(x_2-\mu_2)+d_1\mu_2-\mu_{1A}\mu_{2A}+\mu_{1B}\mu_{2B}\\
&=d_2(x_1-\mu_1)+d_1(x_2-\mu_2)+d_2\mu_1+d_1\mu_2-\mu_{1A}\mu_{2A}+\mu_{1B}\mu_{2B}
\end{aligned}
\tag{55}
$$

ここで、

$$
\begin{aligned}
&d_2\mu_1+d_1\mu_2\\
&=(\mu_{2A}-\mu_{2B})\frac{\mu_{1A}+\mu_{1B}}{2}+(\mu_{1A}-\mu_{1B})\frac{\mu_{2A}+\mu_{2B}}{2}\\
&=\frac{\mu_{2A}\mu_{1A}+\mu_{2A}\mu_{1B}-\mu_{2B}\mu_{1A}-\mu_{2B}\mu_{1B}+\mu_{1A}\mu_{2A}+\mu_{1A}\mu_{2B}-\mu_{1B}\mu_{2A}-\mu_{1B}\mu_{2B}}{2}\\
&=\mu_{1A}\mu_{2A}-\mu_{1B}\mu_{2B}
\end{aligned}
\tag{56}
$$

より、第 3 項の分子は

$$
\begin{aligned}
&(x_1-\mu_{1B})(x_2-\mu_{2B})-(x_1-\mu_{1A})(x_2-\mu_{2A})\\
&=d_2(x_1-\mu_1)+d_1(x_2-\mu_2)
\end{aligned}
\tag{57}
$$

となる。したがって、

$$
\begin{aligned}
\varsigma&=D_B^2-D_A^2\\
&=\frac{1}{1-\rho^2}\left[\frac{2(x_1-\mu_1)d_1}{\sigma_1^{(2)}}+\frac{2(x_2-\mu_2)d_2}{\sigma_2^{(2)}}-2\rho\frac{d_2(x_1-\mu_1)+d_1(x_2-\mu_2)}{\sqrt{\sigma_1^{(2)}\sigma_2^{(2)}}}\right]\\
&=\frac{2}{1-\rho^2}\left[\left(\frac{d_1}{\sigma_1^{(2)}}-\frac{\rho d_2}{\sqrt{\sigma_1^{(2)}\sigma_2^{(2)}}}\right)(x_1-\mu_1)+\left(\frac{d_2}{\sigma_2^{(2)}}-\frac{\rho d_1}{\sqrt{\sigma_1^{(2)}\sigma_2^{(2)}}}\right)(x_2-\mu_2)\right]
\end{aligned}
\tag{58}
$$

となる。

この表式をベクトルおよび行列で表現する。

まず、ベクトル

$$
\boldsymbol{x}=\begin{pmatrix}x_1\\x_2\end{pmatrix}
\tag{59}
$$

$$\boldsymbol{d} = \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} = \begin{pmatrix} \mu_{1A} - \mu_{1B} \\ \mu_{2A} - \mu_{2B} \end{pmatrix} \tag{60}$$

を導入する。さらに共分散行列は

$$\Sigma = \begin{pmatrix} {\sigma_{11}}^{(2)} & {\sigma_{12}}^{(2)} \\ {\sigma_{21}}^{(2)} & {\sigma_{22}}^{(2)} \end{pmatrix} \tag{61}$$

この共分散行列は相関係数を利用して

$$\begin{aligned} \Sigma &= \begin{pmatrix} {\sigma_{11}}^{(2)} & {\sigma_{12}}^{(2)} \\ {\sigma_{21}}^{(2)} & {\sigma_{22}}^{(2)} \end{pmatrix} \\ &= \begin{pmatrix} {\sigma_{11}}^{(2)} & \dfrac{{\sigma_{12}}^{(2)}}{\sqrt{{\sigma_{11}}^{(2)}{\sigma_{22}}^{(2)}}}\sqrt{{\sigma_{11}}^{(2)}{\sigma_{22}}^{(2)}} \\ \dfrac{{\sigma_{21}}^{(2)}}{\sqrt{{\sigma_{11}}^{(2)}{\sigma_{22}}^{(2)}}}\sqrt{{\sigma_{11}}^{(2)}{\sigma_{22}}^{(2)}} & {\sigma_{22}}^{(2)} \end{pmatrix} \\ &= \begin{pmatrix} {\sigma_{11}}^{(2)} & \rho\sqrt{{\sigma_{11}}^{(2)}{\sigma_{22}}^{(2)}} \\ \rho\sqrt{{\sigma_{11}}^{(2)}{\sigma_{22}}^{(2)}} & {\sigma_{22}}^{(2)} \end{pmatrix} \end{aligned} \tag{62}$$

とも表現される。この共分散行列の逆行列は

$$\begin{aligned} \Sigma^{-1} &= \begin{pmatrix} {\sigma_{11}}^{(2)} & \rho\sqrt{{\sigma_{11}}^{(2)}{\sigma_{22}}^{(2)}} \\ \rho\sqrt{{\sigma_{11}}^{(2)}{\sigma_{22}}^{(2)}} & {\sigma_{22}}^{(2)} \end{pmatrix}^{-1} \\ &= \frac{1}{\begin{vmatrix} {\sigma_{11}}^{(2)} & \rho\rho\sqrt{{\sigma_{11}}^{(2)}{\sigma_{22}}^{(2)}} \\ \rho\rho\sqrt{{\sigma_{11}}^{(2)}{\sigma_{22}}^{(2)}} & {\sigma_{22}}^{(2)} \end{vmatrix}} \begin{pmatrix} {\sigma_{22}}^{(2)} & -\rho\sqrt{{\sigma_{11}}^{(2)}{\sigma_{22}}^{(2)}} \\ -\rho\sqrt{{\sigma_{11}}^{(2)}{\sigma_{22}}^{(2)}} & {\sigma_{11}}^{(2)} \end{pmatrix} \\ &= \frac{1}{{\sigma_{11}}^{(2)}{\sigma_{22}}^{(2)} - \rho^2{\sigma_{11}}^{(2)}{\sigma_{22}}^{(2)}} \begin{pmatrix} {\sigma_{22}}^{(2)} & -\rho\sqrt{{\sigma_{11}}^{(2)}{\sigma_{22}}^{(2)}} \\ -\rho\sqrt{{\sigma_{11}}^{(2)}{\sigma_{22}}^{(2)}} & {\sigma_{11}}^{(2)} \end{pmatrix} \\ &= \frac{1}{1-\rho^2} \begin{pmatrix} \dfrac{1}{{\sigma_{11}}^{(2)}} & -\rho\dfrac{1}{\sqrt{{\sigma_{11}}^{(2)}{\sigma_{22}}^{(2)}}} \\ -\rho\dfrac{1}{\sqrt{{\sigma_{11}}^{(2)}{\sigma_{22}}^{(2)}}} & \dfrac{1}{{\sigma_{22}}^{(2)}} \end{pmatrix} \end{aligned} \tag{63}$$

となる。

すると

$$
\begin{aligned}
\varsigma &= D_B^2 - D_A^2 \\
&= \frac{2}{1-\rho^2}\left[\left(\frac{d_1}{\sigma_1^{(2)}} - \frac{\rho d_2}{\sqrt{\sigma_1^{(2)}\sigma_2^{(2)}}}\right)(x_1-\mu_1) + \left(\frac{d_2}{\sigma_2^{(2)}} - \frac{\rho d_1}{\sqrt{\sigma_1^{(2)}\sigma_2^{(2)}}}\right)(x_2-\mu_2)\right] \\
&= 2(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1}\boldsymbol{d}
\end{aligned}
\tag{64}
$$

となる。これは

$$
\begin{aligned}
& 2(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1}\boldsymbol{d} \\
&= 2(x_1-\mu_1 \quad x_2-\mu_2)\begin{pmatrix} \frac{1}{\sigma_{11}^{(2)}} & -\rho\frac{1}{\sigma_{11}\sigma_{22}} \\ -\rho\frac{1}{\sigma_{11}\sigma_{22}} & \frac{1}{\sigma_{22}^{(2)}} \end{pmatrix}\begin{pmatrix} d_1 \\ d_2 \end{pmatrix} \\
&= \frac{2}{1-\rho^2}(x_1-\mu_1 \quad x_2-\mu_2)\begin{pmatrix} \frac{d_1}{\sigma_{11}^{(2)}} - \rho\frac{d_2}{\sigma_{11}\sigma_{22}} \\ -\rho\frac{d_1}{\sigma_{11}\sigma_{22}} + \frac{d_2}{\sigma_{22}^{(2)}} \end{pmatrix} \\
&= \frac{2}{1-\rho^2}\left[\left(\frac{d_1}{\sigma_{11}^{(2)}} - \rho\frac{d_2}{\sigma_{11}\sigma_{22}}\right)(x_1-\mu_1) + \left(\frac{d_2}{\sigma_{22}^{(2)}} - \rho\frac{d_1}{\sigma_{11}\sigma_{22}}\right)(x_2-\mu_2)\right] \\
&= \varsigma
\end{aligned}
\tag{65}
$$

と確かめられる。

このςの表式の中で、第1項以外はデータによらない。そこで、それを$\boldsymbol{a}$で表す。これは、判別する環境の良さを表す。すなわち

$$
\begin{aligned}
\varsigma &= 2(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1}\boldsymbol{d} \\
&= 2(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{a}
\end{aligned}
\tag{66}
$$

である。すなわち

$$
\boldsymbol{a} = \Sigma^{-1}\boldsymbol{d} \tag{67}
$$

である。

8.6.2. 多変数の判別関数

前節では2変数の判別関数を扱ったが、そのベクトルおよび行列表記は多変数への自然な拡張になっている。すなわち、多変数の判別関数は

$$
\begin{aligned}
\varsigma &= 2(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1}\boldsymbol{d} \\
&= 2(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{a}
\end{aligned}
\tag{68}
$$

となる。ただし

$$\boldsymbol{a} = \Sigma^{-1}\boldsymbol{d} \tag{69}$$

である。

ここで、変数の数が m 個ある場合は

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix} \tag{70}$$

$$\begin{aligned} \boldsymbol{\mu} &= \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{pmatrix} \\ &= \begin{pmatrix} \dfrac{\mu_{1A} + \mu_{1B}}{2} \\ \dfrac{\mu_{2A} + \mu_{2B}}{2} \\ \vdots \\ \dfrac{\mu_{mA} + \mu_{mB}}{2} \end{pmatrix} \end{aligned} \tag{71}$$

$$\Sigma = \begin{pmatrix} \sigma_{11}{}^{(2)} & \sigma_{12}{}^{(2)} & \cdots & \sigma_{1m}{}^{(2)} \\ \sigma_{21}{}^{(2)} & \sigma_{22}{}^{(2)} & \cdots & \sigma_{2m}{}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1}{}^{(2)} & \sigma_{m2}{}^{(2)} & \cdots & \sigma_{mm}{}^{(2)} \end{pmatrix} \tag{72}$$

$$\begin{aligned} \boldsymbol{d} &= \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_m \end{pmatrix} \\ &= \begin{pmatrix} \mu_{1A} - \mu_{1B} \\ \mu_{2A} - \mu_{2B} \\ \vdots \\ \mu_{mA} - \mu_{mB} \end{pmatrix} \end{aligned} \tag{73}$$

である。さらに

$$\begin{aligned}\boldsymbol{a} &= \Sigma^{-1}\boldsymbol{d} \\ &= \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix}\end{aligned} \tag{74}$$

である。このςの符号を同定することでどちらのグループに属するか判断できる。

8.7. 判別誤差

判別誤差の議論をする。

8.8. 多変数の判別効率

ここでは、判別効率を評価する。グループ数はmとする。グループ内平方和を評価する。それは

$$Q_{ij} = \sum_{k=1}^{n_A}\left(x_{kiA} - \mu_{iA}\right)\left(x_{kjA} - \mu_{jA}\right) + \sum_{k=1}^{n_B}\left(x_{kiB} - \mu_{iB}\right)\left(x_{kjB} - \mu_{jB}\right) \tag{75}$$

で定義される。この自由度は二つのグループの平均をとっているのでQ_{ij}の自由度は

$$\begin{aligned} f &= n_A - 1 + n_B - 1 \\ &= n - 2 \end{aligned} \tag{76}$$

である。

ここで、

$$\begin{aligned} D_m^2 &= f\sum_{i=1}^{m}\sum_{j=1}^{m} Q^{ij} d_i d_j \\ &= \boldsymbol{d}^T M^{-1}\boldsymbol{d} \end{aligned} \tag{77}$$

を定義する。ここで

$$M = \begin{pmatrix} Q_{11} & Q_{12} & \cdots & Q_{1m} \\ Q_{21} & Q_{22} & \cdots & Q_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ Q_{m1} & Q_{m2} & \cdots & Q_{mm} \end{pmatrix} \tag{78}$$

$$\begin{aligned} M^{-1} &= \begin{pmatrix} Q_{11} & Q_{12} & \cdots & Q_{1m} \\ Q_{21} & Q_{22} & \cdots & Q_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ Q_{m1} & Q_m & \cdots & Q_{mm} \end{pmatrix}^{-1} \\ &= \begin{pmatrix} Q^{11} & Q^{12} & \cdots & Q^{1m} \\ Q^{21} & Q^{22} & \cdots & Q^{2m} \\ \vdots & \vdots & \ddots & \vdots \\ Q^{m1} & Q^{m2} & \cdots & Q^{mm} \end{pmatrix} \end{aligned} \tag{79}$$

である。この D_m^2 は説明変数の数が m 個の場合に、判別する際のその有意差を表す。これが大きいほど、判別には有効である。これは、新たなデータとは無関係な量であり、新たなデータがきた場合、それに対する環境の判別する能力を示している。

判別効率はデータの種類が多くなればなるほど単純に大きくなる。すなわち、それだけから判断すると、判別する項目は有効であろうがなかろうが多いほどいい、ということになってしまう。判別項目を増やすことによる判別効率の増加と関連する判別精度と関連する分散と、誤差と関連する分散の比から、変数の有効性を評価し、その選択をしなければならない。それを次節で行う。

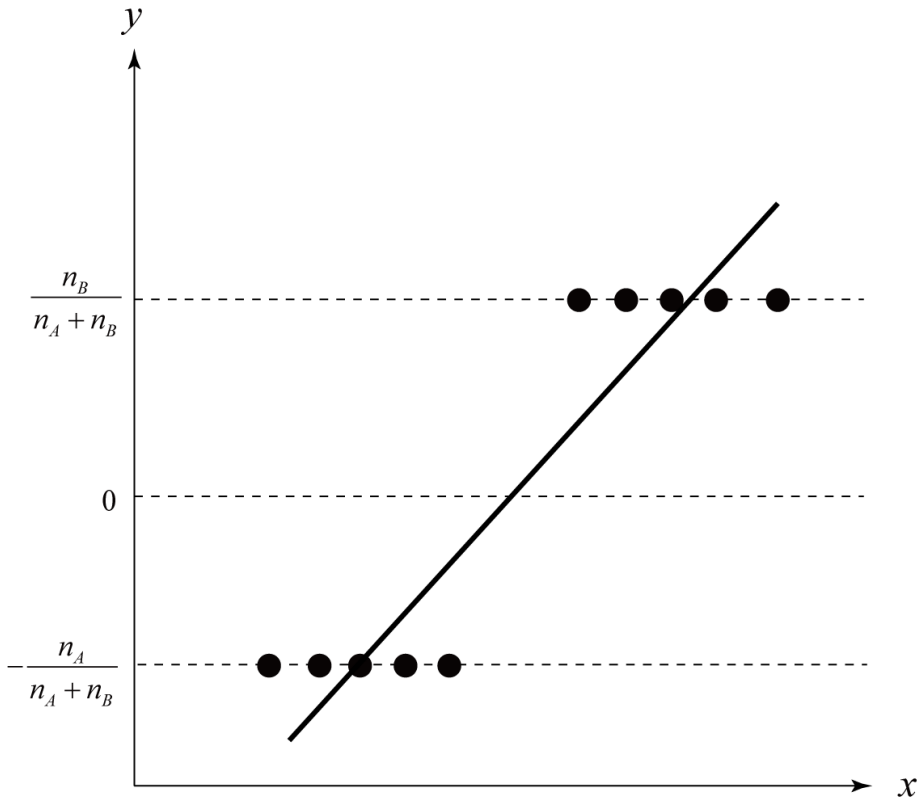


図 1　y に便宜的に数値を与えた回帰直線

8.9. 判別効率と回帰分散の関係

以上で判別分析の主な解析項目は扱った。説明変数は全て有効である、という仮定を暗にこれまでは仮定していた。この説明変数の妥当性をここでは、議論する。

説明変数の妥当性は重回帰分析で議論してきた。したがって、この判別効率を重回帰分析から導出できれば、同様な解析にもっていけると思われる。

判別解析の中では、グループ A,B が目的変数になっていた。つまり、グループ A,B は数

値ではない。一方、回帰分析では目的変数は数値である。これらをつなぐために判別解析の中のグループ A,B を数値で表現する。これは、別の数値であれば任意であるが、利用しやすいように平均を 0 とする。そこでグループ A,B に対する数値をそれぞれ y_A、y_B とし、それぞれ

$$y_A = \frac{n_B}{n_A + n_B} \tag{80}$$

$$y_B = -\frac{n_A}{n_A + n_B} \tag{81}$$

とする。ここでは、狙い通りその平均が 0 になるように設定している。グループ A に属するデータ数は n_A、グループ B に属するデータ数は n_B としている。するとトータルのデータ数 n は

$$n = n_A + n_B \tag{82}$$

である。また、グループ A,B の値は以上の二つの値に固定されている。この状態で y は

$$y_k = \mu_y + b_1\left(x_{k1} - \mu_1\right) + b_2\left(x_{k2} - \mu_2\right) + \cdots + b_p\left(x_{km} - \mu_m\right) + e_k \tag{83}$$

とする。これに対する回帰直線は

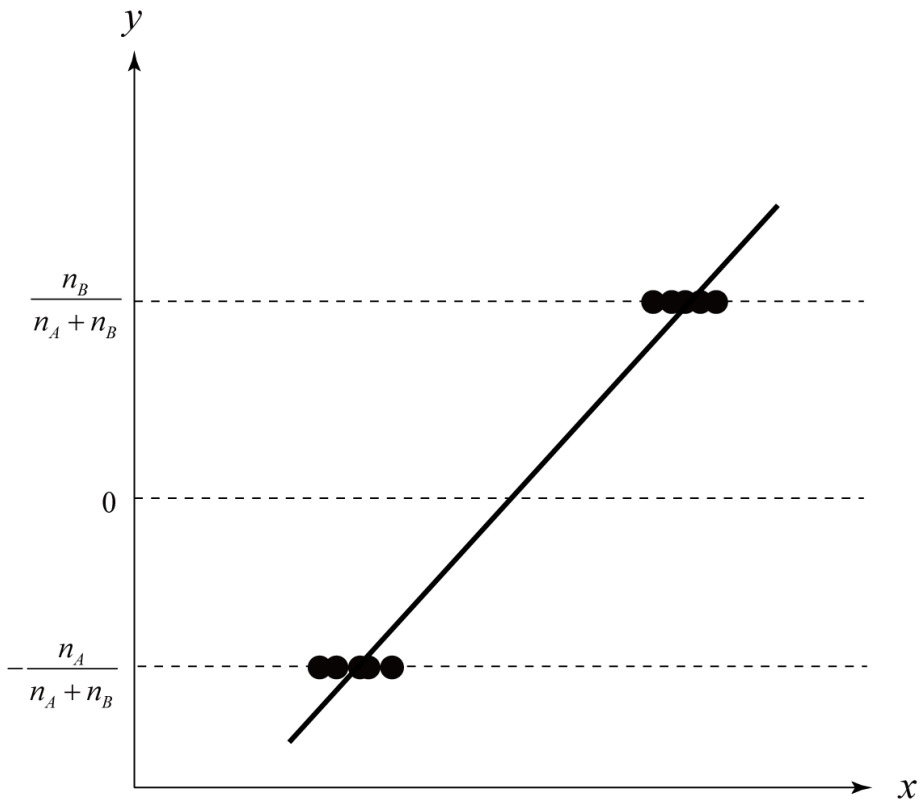
$$Y_k = \mu_y + b_1\left(x_{k1} - \mu_1\right) + b_2\left(x_{k2} - \mu_2\right) + \cdots + b_m\left(x_{km} - \mu_m\right) \tag{84}$$

となる。この誤差 e_k の平方和

$$\sum_{k=1}^{n} {e_k}^2 = \sum_{k=1}^{n} \left(y_k - Y_k\right)^2 \tag{85}$$

の平方和を最小にする問題を考える。

この場合も数に示すように、その変数が判別に有効であればこの誤差に関する平方和は小さくなるし、有効でなければ誤差に関する平方和は大きくなる。



(a)

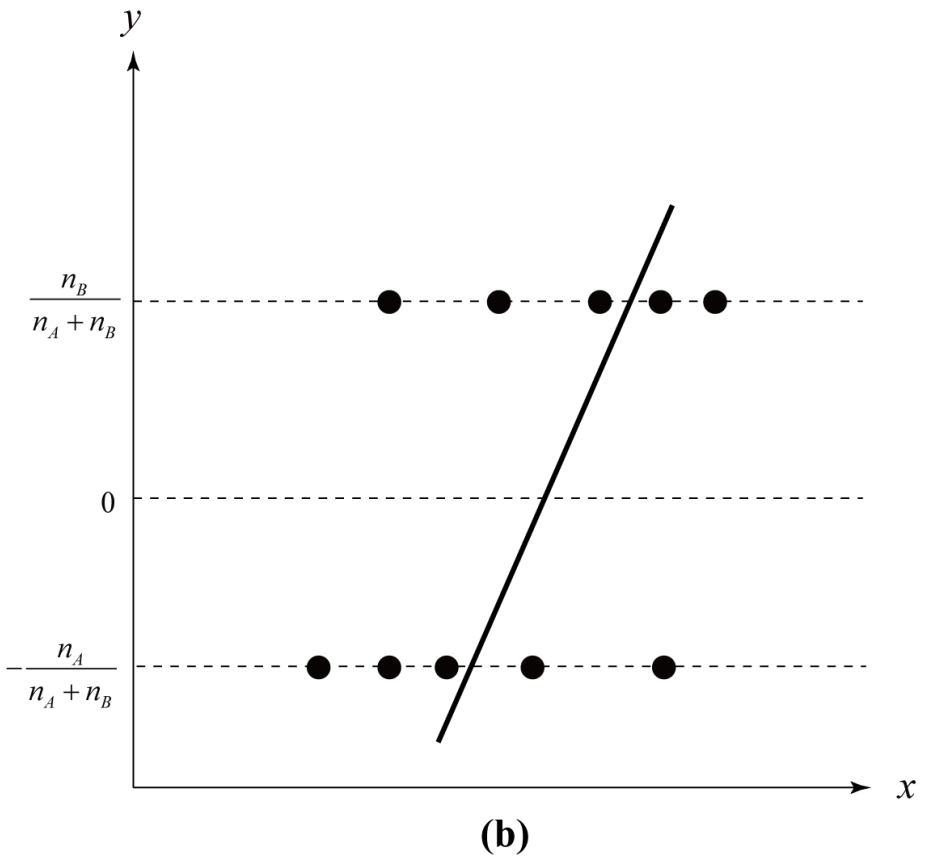


図 2　y に便宜的に数値を与えた回帰直線。.(a)　変数が判別に有効な場合。　(b)　変数が判別に有効でない場合

まずグループ内の平均値と関連した平方和を考える。上の実際のデータはグループを意識せずに x_{ki} と置いていたが、グループ A に属する場合は x_{kiA}、グループ B に属する場合は x_{kiB} とする。すると

$$Q_{ij} = \sum_{k=1}^{n_A}\left(x_{kiA}-\mu_{iA}\right)\left(x_{kjA}-\mu_{jA}\right)+\sum_{k=1}^{n_B}\left(x_{kiB}-\mu_{iB}\right)\left(x_{kjB}-\mu_{jB}\right) \tag{86}$$

$$\begin{aligned} Q_{iy} &= \sum_{k=1}^{n_A}\left(x_{kiA}-\mu_{iA}\right)\left(y_{kA}-\mu_{yA}\right)+\sum_{k=1}^{n_B}\left(x_{kiB}-\mu_{iB}\right)\left(y_{kB}-\mu_{yB}\right) \\ &= 0 \end{aligned} \tag{87}$$

$$\begin{aligned} Q_{yy} &= \sum_{k=1}^{n_A}\left(y_{kA}-\mu_{yA}\right)^2+\sum_{k=1}^{n_B}\left(y_{kB}-\mu_{yB}\right)^2 \\ &= 0 \end{aligned} \tag{88}$$

となる。ここで

$$y_{kA} = \mu_{yA} = \frac{n_B}{n_A+n_B} \tag{89}$$

$$y_{kB} = \mu_{yB} = -\frac{n_A}{n_A + n_B} \tag{90}$$

を利用している。

グループでなく、トータルの平方和を考え、群内の平方和との関連を調べる。トータルの平方和は添え字 Tot をつける。

$$\begin{aligned}
Q_{ij}^{Tot} &= \sum_{k=1}^{n_A}\left(x_{kiA}-\mu_i\right)\left(x_{kjA}-\mu_j\right) \\
&\quad+\sum_{k=1}^{n_B}\left(x_{kiB}-\mu_i\right)\left(x_{kjB}-\mu_j\right) \\
&= \sum_{k=1}^{n_A}\left(x_{kiA}-\mu_{iA}+\mu_{iA}-\mu_i\right)\left(x_{kjA}-\mu_{jA}+\mu_{jA}-\mu_j\right) \\
&\quad+\sum_{k=1}^{n_B}\left(x_{kiB}-\mu_{iB}+\mu_{iB}-\mu_i\right)\left(x_{kjB}-\mu_{jB}+\mu_{jB}-\mu_j\right) \\
&= \sum_{k=1}^{n_A}\left(x_{kiA}-\mu_{iA}\right)\left(x_{kjA}-\mu_{jA}\right)+\sum_{k=1}^{n_B}\left(x_{kiB}-\mu_{iB}\right)\left(x_{kjB}-\mu_{jB}\right) \\
&\quad+\left(\mu_{jA}-\mu_j\right)\sum_{k=1}^{n_A}\left(x_{kiA}-\mu_{iA}\right)+\left(\mu_{iA}-\mu_i\right)\sum_{k=1}^{n_A}\left(x_{kjA}-\mu_{jA}\right)+n_A\left(\mu_{iA}-\mu_i\right)\left(\mu_{jA}-\mu_j\right) \\
&\quad+\left(\mu_{jB}-\mu_j\right)\sum_{k=1}^{n_B}\left(x_{kiB}-\mu_{iB}\right)+\left(\mu_{iB}-\mu_i\right)\sum_{k=1}^{n_B}\left(x_{kjB}-\mu_{jB}\right)+n_B\left(\mu_{iB}-\mu_i\right)\left(\mu_{jB}-\mu_j\right) \\
&= Q_{ij}+n_A\left(\mu_{iA}-\mu_i\right)\left(\mu_{jA}-\mu_j\right)+n_B\left(\mu_{iB}-\mu_i\right)\left(\mu_{jB}-\mu_j\right)
\end{aligned} \tag{91}$$

となる。最後の二つの項の和はさらに以下のように変形される。

$$\begin{aligned}
&n_A\left(\mu_{iA}-\mu_i\right)\left(\mu_{jA}-\mu_j\right)+n_B\left(\mu_{iB}-\mu_i\right)\left(\mu_{jB}-\mu_j\right) \\
&= n_A\left(\mu_{iA}-\frac{n_A\mu_{iA}+n_B\mu_{iB}}{n_A+n_B}\right)\left(\mu_{jA}-\frac{n_A\mu_{jA}+n_B\mu_{jB}}{n_A+n_B}\right) \\
&\quad+n_B\left(\mu_{iB}-\frac{n_A\mu_{iA}+n_B\mu_{iB}}{n_A+n_B}\right)\left(\mu_{jB}-\frac{n_A\mu_{jA}+n_B\mu_{jB}}{n_A+n_B}\right) \\
&= \frac{n_A{n_B}^2}{\left(n_A+n_B\right)^2}\left(\mu_{iA}-\mu_{iB}\right)\left(\mu_{jA}-\mu_{jB}\right)+\frac{{n_A}^2 n_B}{\left(n_A+n_B\right)^2}\left(\mu_{iA}-\mu_{iB}\right)\left(\mu_{jA}-\mu_{jB}\right) \\
&= \frac{n_A n_B}{n_A+n_B}d_i d_j
\end{aligned} \tag{92}$$

よって、

$$Q_{ij}^{Tot} = Q_{ij}+\frac{n_A n_B}{n_A+n_B}d_i d_j \tag{93}$$

となる。

次に Q_{iy} を考える。

$$
\begin{aligned}
Q_{iy}^{Tot} &= \sum_{k=1}^{n_A}\left(x_{kiA}-\mu_i\right)\left(y_{kA}-\mu_y\right)+\sum_{k=1}^{n_B}\left(x_{kiB}-\mu_i\right)\left(y_{kB}-\mu_y\right)\\
&= \sum_{k=1}^{n_A}\left(x_{kiA}-\mu_{iA}+\mu_{iA}-\mu_i\right)\left(y_{kA}-\mu_{yA}+\mu_{yA}-\mu_y\right)\\
&\quad+\sum_{k=1}^{n_B}\left(x_{kiB}-\mu_{iB}+\mu_{iB}-\mu_i\right)\left(y_{kB}-\mu_{yB}+\mu_{yB}-\mu_y\right)\\
&= \sum_{k=1}^{n_A}\left(x_{kiA}-\mu_{iA}\right)\left(y_{kA}-\mu_{yA}\right)+\sum_{k=1}^{n_B}\left(x_{kiB}-\mu_{iBi}\right)\left(y_{kB}-\mu_{yB}\right)\\
&\quad+\left(\mu_{yA}-\mu_y\right)\sum_{k=1}^{n_A}\left(x_{kiA}-\mu_{iA}\right)+\left(\mu_{iA}-\mu_i\right)\sum_{k=1}^{n_A}\left(y_{kA}-\mu_{yA}\right)+n_A\left(\mu_{iA}-\mu_i\right)\left(\mu_{yA}-\mu_y\right)\\
&\quad+\left(\mu_{yB}-\mu_y\right)\sum_{k=1}^{n_B}\left(x_{kiB}-\mu_{iB}\right)+\left(\mu_{iB}-\mu_i\right)\sum_{k=1}^{n_B}\left(y_{kB}-\mu_{yB}\right)+n_B\left(\mu_{iB}-\mu_i\right)\left(\mu_{yB}-\mu_y\right)\\
&= Q_{iy}+n_A\left(\mu_{iA}-\mu_i\right)\left(\mu_{yA}-\mu_y\right)+n_B\left(\mu_{iB}-\mu_i\right)\left(\mu_{yB}-\mu_y\right)\\
&= n_A\left(\mu_{iA}-\mu_i\right)\mu_{yA}+n_B\left(\mu_{iB}-\mu_i\right)\mu_{yB}\\
&= n_A\left(\mu_{iA}-\frac{n_A\mu_{iA}+n_B\mu_{iB}}{n_A+n_B}\right)\frac{n_B}{n_A+n_B}+n_B\left(\mu_{iB}-\frac{n_A\mu_{iA}+n_B\mu_{iB}}{n_A+n_B}\right)\left(-\frac{n_A}{n_A+n_B}\right)\\
&= \frac{n_A{n_B}^2}{n_A+n_B}\left(\mu_{iA}-\mu_{iB}\right)+\frac{{n_A}^2n_B}{n_A+n_B}\left(\mu_{iA}-\mu_{iB}\right)\\
&= \frac{n_An_B}{n_A+n_B}d_i
\end{aligned}
\tag{94}
$$

最後にQ_{yy}を考える。

$$
\begin{aligned}
Q_{yy}^{Tot} &= \sum_{k=1}^{n_A}\left(y_{kA}-\mu_y\right)^2+\sum_{k=1}^{n_B}\left(y_{kB}-\mu_y\right)^2\\
&= n_A{y_A}^2+n_B{y_B}^2\\
&= n_A\left(\frac{n_B}{n_A+n_B}\right)^2+n_B\left(-\frac{n_A}{n_A+n_B}\right)^2\\
&= \frac{n_An_B}{n_A+n_B}
\end{aligned}
\tag{95}
$$

以上を纏めると

$$Q_{ij}^{Tot} = Q_{ij}+wd_id_j \tag{96}$$

$$Q_{iy}^{Tot} = wd_i \tag{97}$$

$$Q_{yy}^{Tot} = w \tag{98}$$

となる。ただし

$$w = \frac{n_A n_B}{n_A + n_B} \tag{99}$$

となる。よって、回帰曲線のb_iを求める方程式は重回帰分析から

$$Q_{11}{}^{Tot} b_1 + Q_{12}{}^{Tot} b_2 + \cdots + Q_{1m}{}^{Tot} b_m = wd_1 \tag{100}$$

$$Q_{21}{}^{Tot} b_1 + Q_{22}{}^{Tot} b_2 + \cdots + Q_{2m}{}^{Tot} b_m = wd_2 \tag{101}$$

・・・

$$Q_{m1}{}^{Tot} b_1 + Q_{m2}{}^{Tot} b_2 + \cdots + Q_{mm}{}^{Tot} b_m = wd_m \tag{102}$$

となる。これから係数b_iが求まる。

これをグループ内の平方数に焼き直すと

$$\left(Q_{11} + wd_1 d_1\right) b_1 + \left(Q_{12} + wd_1 d_2\right) b_2 + \cdots + \left(Q_{1m} + wd_1 d_m\right) b_m = wd_1 \tag{103}$$

$$\left(Q_{21} + wd_2 d_1\right) b_1 + \left(Q_{22} + wd_2 d_2\right) b_2 + \cdots + \left(Q_{2m} + wd_2 d_m\right) b_m = wd_2 \tag{104}$$

・・・

$$\left(Q_{m1} + wd_m d_1\right) b_1 + \left(Q_{m2} + wd_m d_2\right) b_2 + \cdots + \left(Q_{mm} + wd_m d_m\right) b_m = wd_m \tag{105}$$

となる。これを整理して書き直すと

$$Q_{11} b_1 + Q_{12} b_2 + \cdots + Q_{1m} b_m = w\left(1 - \sum_{l=1}^{m} b_l d_l\right) d_1 \tag{106}$$

$$Q_{21} b_1 + Q_{22} b_2 + \cdots + Q_{2m} b_m = w\left(1 - \sum_{l=1}^{m} b_l d_l\right) d_2 \tag{107}$$

・・・

$$Q_{m1} b_1 + Q_{m2} b_2 + \cdots + Q_{pp} b_m = w\left(1 - \sum_{l=1}^{m} b_l d_l\right) d_m \tag{108}$$

となる。

b_iは

$$\begin{aligned} b_i &= \sum_{j=1}^{m} Q^{ij}{}_{Tot} wd_j \\ &= \sum_{j=1}^{m} Q^{ij} \left(1 - \sum_{l=1}^{m} b_l d_l\right) wd_j \\ &= w\left(1 - \sum_{l=1}^{m} b_l d_l\right) \sum_{j=1}^{m} Q^{ij} d_j \end{aligned} \tag{109}$$

と求まる。

これに対応する回帰平方和 S_m^2 は、重回帰分析の結果から（Appemdix）

$$
\begin{aligned}
S_m^2 &= \sum_{k=1}^{n}\left(Y_k - \mu_y\right)^2 \\
&= \sum_{k=1}^{n} {Y_k}^2 \\
&= \sum_{i=1}^{m} b_i {Q_{iy}}^{Tot} \\
&= \sum_{i=1}^{m} b_i w d_i \\
&= w\sum_{i=1}^{m} w\left(1 - \sum_{l=1}^{m} b_l d_l\right)\sum_{j=1}^{m} Q^{ij} d_j d_i \\
&= w^2\left(1 - \sum_{l=1}^{m} b_l d_l\right)\sum_{i=1}^{m}\sum_{j=1}^{m} d_i Q^{ij} d_j \\
&= w^2\sum_{i=1}^{m}\sum_{j=1}^{m} d_i Q^{ij} d_j - \left(w\sum_{l=1}^{m} b_l d_l\right) w\sum_{i=1}^{m}\sum_{j=1}^{m} d_i Q^{ij} d_j \\
&= w^2\sum_{i=1}^{m}\sum_{j=1}^{m} d_i Q^{ij} d_j - S_m^2 w\sum_{i=1}^{m}\sum_{j=1}^{m} d_i Q^{ij} d_j
\end{aligned}
\tag{110}
$$

となる。よって回帰平方和は

$$
S_m^2 = \frac{w^2\sum_{i=1}^{m}\sum_{j=1}^{m} d_i Q^{ij} d_j}{1 + w\sum_{i=1}^{m}\sum_{j=1}^{m} d_i Q^{ij} d_j} \tag{111}
$$

となる。これに判別効率

$$
D_m^2 = f\sum_{i=1}^{m}\sum_{j=1}^{m} d_i Q^{ij} d_j \tag{112}
$$

を代入すると

$$
S_m^2 = \frac{w^2\dfrac{D_m^2}{f}}{1 + w\dfrac{D_m^2}{f}} \tag{113}
$$

と書ける。これは変形すると

$$
\begin{aligned}
S_m^2 &= \frac{w\dfrac{D_m^2}{f}}{1+w\dfrac{D_m^2}{f}}w \\
&= \frac{w\dfrac{D_m^2}{f}}{1+w\dfrac{D_m^2}{f}}S_{yy}^2
\end{aligned}
\tag{114}
$$

と表現できる。これにより、回帰平方和は判別効率と関連付けられた。あとは、回帰分析の変数選択のプロセスを踏んでいけばいい。

$m+1$ 個の説明変数の場合の回帰平方和は

$$
S_{m+1}^2 = \frac{w^2\dfrac{D_{m+1}^2}{f}}{1+w\dfrac{D_{m+1}^2}{f}} \tag{115}
$$

となる。

y の全体平方和は w であるから、変数を $m+1$ 個にした場合の誤差に関する平方和は

$$
\begin{aligned}
S_e^2 &= w - S_{m+1}^2 \\
&= w - \frac{w^2\dfrac{D_{m+1}^2}{f}}{1+w\dfrac{D_{m+1}^2}{f}} \\
&= \frac{w}{1+w\dfrac{D_{m+1}^2}{f}}
\end{aligned}
\tag{116}
$$

となる。これが変数を増やすことによる回帰平方和の増分より小さければ、その変数は判別に有効である、とみなすことができる。

S_{m+1}^2 の自由度は $m+1$ 個の係数からなるから $m+1$、S_m^2 の自由度は m 個の係数からなる m である。よって、$S_{p+1}^2 - S_p^2$ の自由度は

$$
(m+1) - m = 1 \tag{117}
$$

である。

変数 $m+1$ 個の変数の誤差と関連する S_e^2 は

$$
S_{yy}^2 = S_{m+1}^2 + S_e^2 \tag{118}
$$

の関係から求めることができる。S_{yy}^2 は n 個のデータの中で全体の平均を利用しているから

その自由度は$n-1$になる。したがって、S_e^2の自由度は

$$(n-1)-(m+1)=n-2-m \tag{119}$$

である。

以上より変数を 1 個追加した場合の分散比は

$$\begin{aligned}
F &= \frac{\dfrac{S_{m+1}^2 - S_m^2}{1}}{\dfrac{S_e^2}{n-2-m}} \\
&= (n-2-m)\frac{S_{p+1}^2 - S_p^2}{S_e^2} \\
&= (f-m)\frac{\dfrac{w^2\dfrac{D_{m+1}^2}{f}}{1+w\dfrac{D_{m+1}^2}{f}} - \dfrac{w^2\dfrac{D_m^2}{f}}{1+w\dfrac{D_m^2}{f}}}{\dfrac{w}{1+w\dfrac{D_{m+1}^2}{f}}} \\
&= (f-m)\frac{\dfrac{w^2\dfrac{D_{m+1}^2}{f}\left(1+w\dfrac{D_m^2}{f}\right) - w^2\dfrac{D_m^2}{f}\left(1+w\dfrac{D_{m+1}^2}{f}\right)}{\left(1+w\dfrac{D_{m+1}^2}{f}\right)\left(1+w\dfrac{D_m^2}{f}\right)}}{\dfrac{w}{1+w\dfrac{D_{m+1}^2}{f}}} \\
&= (f-m)\frac{D_{m+1}^2 - D_m^2}{\dfrac{f}{w}+D_m^2}
\end{aligned} \tag{120}$$

となる。これと$F[1,f-m;P]$との比較をし、このFが$F[1,f-m;P]$より大きければ変数は判別に有効であると判断できる。

この$F[1,f-m;P]$は数値的に評価しなくてはならない。数値的に評価した結果を図に示す。ここでは、$F[1,n_2;P]$として計算している。結果はn_2に依存するがn_2が大きければ、つまりデータ数が多ければほぼ一定とみなすことができる。したがって、$n_2=1000$の場合の$F[1,1000;P]$を図に示す。これを近似的に用いることもできる。

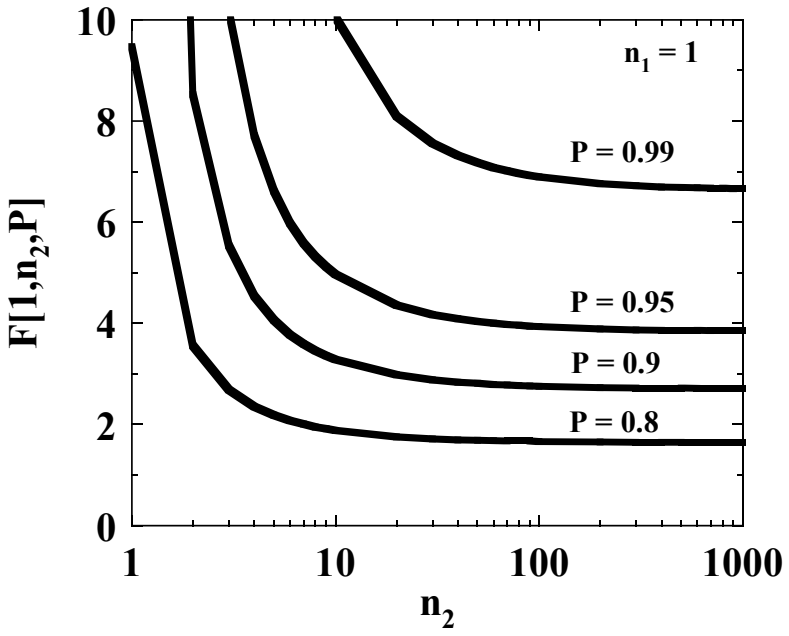


図 3　F 値の n_2 依存性

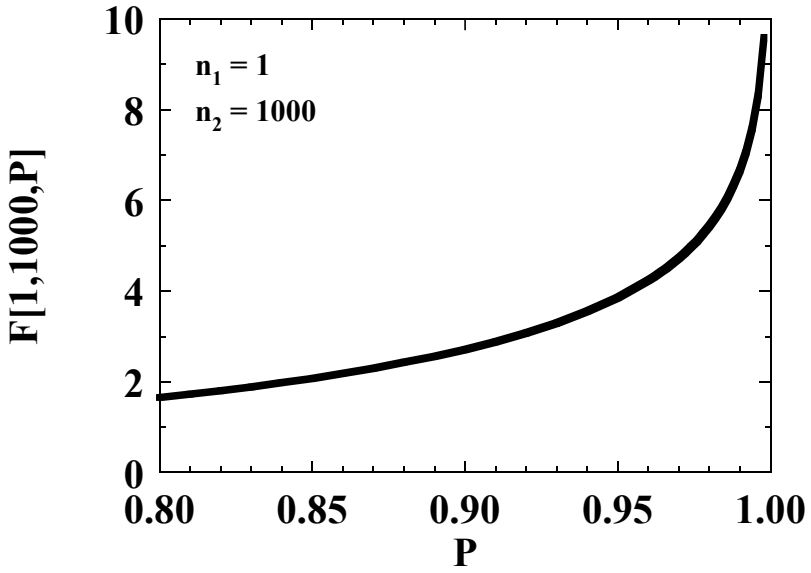


図 4　F 値の P 依存性

別の章で議論した AIC と関連づける試みもなされている。m 変数と $m+1$ 変数を用いた AIC の差を A_m と置いて

$$\begin{aligned} A_m &= AIC[m+1] - AIC[m] \\ &= -n\ln\left(1+\frac{F}{f-m}\right)+2 \end{aligned} \tag{121}$$

となる。この導出の詳細は不明である。これが負であれば、追加した変数は判別に有効である、と判断される。上の式の対数の第 2 項は一般に小さいので、A_m は

$$\begin{aligned} A_m &= -n\ln\left(1+\frac{F}{f-m}\right)+2 \\ &\approx -n\frac{F}{f-m}+2 \\ &= -\frac{n}{f-m}F+2 \end{aligned} \tag{122}$$

と近似される。$n \gg m$ としてこの近似をさらにすすめると

$$\begin{aligned} A_m &\approx -\frac{n}{f-m}F+2 \\ &= -\frac{n}{n-2-m}F+2 \\ &\approx -F+2 \end{aligned} \tag{123}$$

となり、先の $F[1, f-m; P]$ を 2 と近似して評価した場合と一致する。これはデータ数が十分大きい場合の $P=0.85$ に相当する。

8.10. 判別効率を利用した変数選択

判別を決断する場合、その変数が有効であるかどうかを判断する必要がある。

ここでは、二つのブループ A、B を判断する際の変数の有効性を議論する。

まず、二つのグループの平均の差を評価し、それからベクトルを構成する。すなわち、

$$\boldsymbol{d} = \begin{pmatrix} \mu_{1A}-\mu_{1B} \\ \mu_{2A}-\mu_{2B} \\ \vdots \\ \mu_{mA}-\mu_{mB} \end{pmatrix} \tag{124}$$

そして、判別効率平方距離

$$D_m^2 = \boldsymbol{d}^T M^{-1}\boldsymbol{d} \tag{125}$$

を構成する。ただし

$$M = \frac{1}{f}\left(Q_{ij}\right) \tag{126}$$

であり、

$$Q_{ij} = \sum_{k=1}^{n_A}\left(x_{kiA}-\mu_{iA}\right)\left(x_{kjA}-\mu_{jA}\right) + \sum_{k=1}^{n_B}\left(x_{kiB}-\mu_{iB}\right)\left(x_{kjB}-\mu_{jB}\right) \tag{127}$$

である。

これが、判別効率と関連する距離である。これが大きいほど、判別は明確にできる、と判断できる。

この判別効率の式を使って以下の手順で変数の評価をする。

まず、変数 1 個からスタートする。それぞれの変数で D_1^2 を評価する。その中で最大のものを与える変数を選択する。その変数に相当する D_1^2 を D_{1*}^2 とする。

次に、2 個目の変数を $m-1$ 個の中から一つ選び、それと最初の変数を利用する。つまり、変数の組み合わせは $m-1$ 個ある。それぞれの組み合わせで D_2^2 を構成する。この各 D_2^2 に対して

$$F = (f-1)\frac{D_2^2 - D_{1*}^2}{f\dfrac{n_A+n_B}{n_A n_B} + D_{1*}^2} \geq F[1, f-1; P] \tag{128}$$

を満足するか確かめる。満足するものがなければ終了。あれば、その中で最大の F を与える

変数を選択する。それに相当する D_2^2 を D_{2*}^2 とする。

前のステップで満足するものあった場合、このステップに進む。次に、3 個目の変数を $m-2$ 個の中から一つ選び、それと最初の 2 つの変数を利用する。つまり、変数の組み合わせは $m-2$ 個ある。それぞれの組み合わせで D_3^2 を構成する。この各 D_3^2 に対して

$$F=(f-2)\frac{D_3^2-D_{2*}^2}{f\dfrac{n_A+n_B}{n_An_B}+D_{2*}^2}\geq F[1,f-2;P] \tag{129}$$

を満足するか確かめる。満足するものがなければ終了。あれば、その中で最大の F を与える変数を選択する。

以上プロセスを満足するものがなくなるか、全ての変数に対して評価するか、を行う。

以上のプロセス遂行後に残った変数が採用すべき変数となる。

8.11. 多くのグループの場合への拡張

ここでは、グループ A、B の二つを考えた。一般には $G_1,G_2,\cdots,G_l$ の l 個のグループを考えることができる。この場合、

$$D_{G_1}^2=\left(\boldsymbol{x}-\boldsymbol{\mu}_{G_1}\right)^T\Sigma^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}_{G_1}\right) \tag{130}$$

$$D_{G_2}^2=\left(\boldsymbol{x}-\boldsymbol{\mu}_{G_2}\right)^T\Sigma^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}_{G_2}\right) \tag{131}$$

・・・

$$D_{G_l}^2=\left(\boldsymbol{x}-\boldsymbol{\mu}_{G_l}\right)^T\Sigma^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}_{G_l}\right) \tag{132}$$

として最小の平方距離のグループに属すると判断すればいい。ただし、

$$\boldsymbol{x}=\begin{pmatrix}x_1\\x_2\\\vdots\\x_m\end{pmatrix} \tag{133}$$

$$\boldsymbol{\mu}_{G_i}=\begin{pmatrix}\mu_{1G_i}\\\mu_{2G_i}\\\vdots\\\mu_{mG_i}\end{pmatrix} \tag{134}$$

である。

8.12. 多くのグループの場合の変数評価

判別効率を利用した変数選択方法は二つのグループを仮定していた。この方法は 3 つ以上のグループには利用できない。

しかし、3 つ以上のグループの場合でも、二つの組み合わせを選択し、それに対して変数の評価を行う。そして、すべての組み合わせについいておこない、全ての組み合わせについて不要な変数を取り除く、という操作をすればいい。

たとえば A,B,C の 3 グループであれば、$(A,B),(A,C),(B,C)$ の組み合わせ、A,B,C,D であれば $(A,B),(A,C),(A,D),(B,C),(B,D),(C,D)$ の組み合わせを検討する。

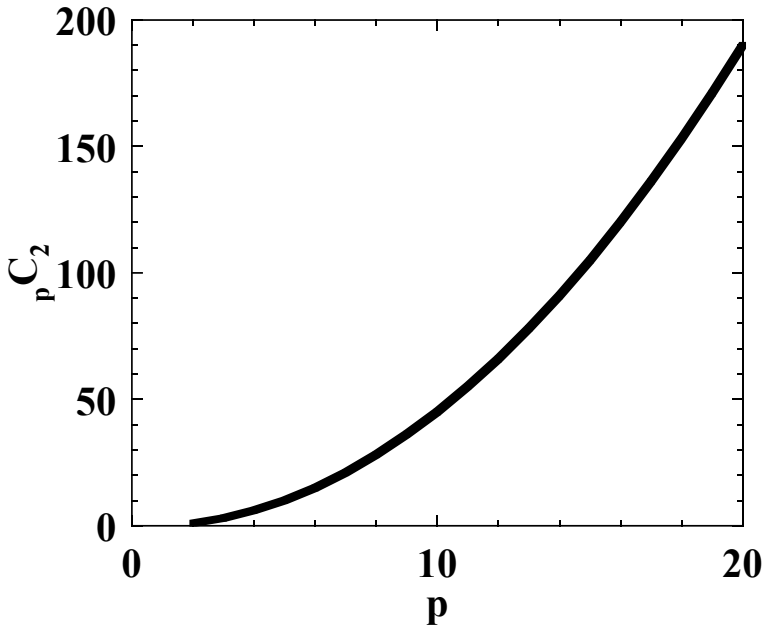


図 5　場合の数のグループ数依存性

8.13. 分散分析による変数評価

変数の選択の際には目的変数は A,B の二種類であった。前節のように目的変数が 2 種類よりも多い場合は二つのグループを選択するが、その組み合わせの数はグループの数が大きくなると急激に増大する。p グループの場合は $_pC_2$ 通りとなる。この場合、説明変数間の相関が強くなければ、以下に示す分析で簡便に評価することができる。

この場合は分散分析を行うことができる。ただしこの場合は変数間の相関は無視する。具体的な作業は以下の簡単に記述する。解析の詳細は分散分析の章を参照されたい。

ここでは、A,B,C の 3 種類のグループを考える。さらに多い場合でも解析方法は同じである。検査の種類は m 種類であるとする。すなわちデータとしては $(x_1,x_2,\cdots,x_m)$ が常に存在する。ここでは検査データ間の相関は無視するため、一つの検査項目のみを考えれいい。そこで以後は一つの検査データのみを考えるとする。

一つの種類の各データはどれかのグループに必ず属する。例えばグループ A に属するデータは x_{kA} というふうに表記する。データの数を n とし、各グループに属するグループの数をそれぞれ n_A, n_B, n_C とする。すると

$$n = n_A + n_B + n_C \tag{135}$$

となる。

各グループの検査の平均値は

$$\mu_A = \frac{1}{n_A}\sum_{k=1}^{n_A} x_{kA} \tag{136}$$

$$\mu_B = \frac{1}{n_B}\sum_{k=1}^{n_B} x_{kB} \tag{137}$$

$$\mu_C = \frac{1}{n_C}\sum_{k=1}^{n_C} x_{kC} \tag{138}$$

となる。

全体の平均は μ は

$$\begin{aligned}\mu &= \frac{1}{n_A + n_B + n_C}\left(\sum_{k=1}^{n_A} x_{kA} + \sum_{k=1}^{n_B} x_{kB} + \sum_{k=1}^{n_C} x_{kC}\right) \\ &= \frac{n_A\mu_A + n_B\mu_B + n_C\mu_C}{n_A + n_B + n_C}\end{aligned} \tag{139}$$

となる。

各水準平均の全体平均に対する偏差の二乗にデータ数を掛けた分散 $S_{ex}^{(2)}$ を考える。すなわち

$$S_{ex}^{(2)} = \frac{n_A(\mu_A - \mu)^2 + n_B(\mu_B - \mu)^2 + n_C(\mu_C - \mu)^2}{n_A + n_B + n_C} \tag{140}$$

を得る。この自由度 ϕ_{ex} を考える。これは μ_A, μ_B, μ_C の三つであるが、それぞれは全体平均と結びつけられるから三つのうち二つが決まれば残りの一つが決まる。よってグループ数を p とすると（ここでは $p=3$）

$$\phi_{ex} = p - 1 \tag{141}$$

となる。したがって、対応する不偏分散 $s_{ex}^{(2)}$ は

$$s_{ex}^{(2)} = \frac{n}{\phi_{ex}} S_{ex}^{(2)} \tag{142}$$

となる。

次にグループ内の分散を考える。各グループのデータの各水準の平均に対する偏差の二

乗を足し合わせた分散 $S_{in}^{(2)}$ は

$$S_{in}^{(2)} = \frac{\sum_{i=1}^{n_A}\left(x_{iA}-\mu_A\right)^2 + \sum_{i=1}^{n_B}\left(x_{iB}-\mu_B\right)^2 + \sum_{i=1}^{n_C}\left(x_{iC}-\mu_C\right)^2}{n} \tag{143}$$

となる。各水準はそれぞれ n_A, n_B, n_C 個のデータがあるが、それぞれの平均からの偏差をとってあるから、自由度はデータ数からグループ数 p だけ減る。つまり

$$\phi_{in} = n - p \tag{144}$$

となる。したがって、対応する不偏分散 $s_{in}^{(2)}$ は

$$s_{in}^{(2)} = \frac{n}{\phi_{in}} S_{in}^{(2)} \tag{145}$$

となる。この比

$$F = \frac{s_{ex}^{(2)}}{s_{in}^{(2)}} \tag{146}$$

は自由度 $\left(\phi_{ex}, \phi_{in}\right)$ の F 分布に従う。F の P 点 $F\left[\phi_{ex}, \phi_{in}, P\right]$ を評価し $F > F\left[\phi_{ex}, \phi_{in}, P\right]$ であるか評価する。もしもそれが成り立つならば、その変数は判別に有効であると判断し、あｈン別分析に利用する。それが成り立たなければ利用しない。

以上の分析をすべての検査項目について行い、判別分析に利用する項目を選択する。

8.14. 金融融資の場合の例

ここでは金融の融資の例をとり、これまでの解析を具体的に問題に適用する。

8.14.1. 教師データ構造および基本パラメータ

リスク x_1、金額 x_2、貸付期間 x_3、利率 x_4 とする。これらは解析において説明変数となる。各メンバーを i で表現し、それぞれの項目のデータはリスク x_{i1}、金額 x_{i2}、貸付期間 x_{i3}、利率 x_{i4} と表現される。また、それに対して融資可であったかの結果もあるとする。これは目的変数であり、これは融資可か不可のカテゴリーデータになる。融資可のグループを A,不可のグループを B とする。以上のように判別分析で扱うのは、数値データからなる複数の説明変数とカテゴリーデータからなる 1 個の目的変数のデータ構造である。この一連のデータのことを融資結果がわかっていることから教師データを呼ぶ。

ここでリスクはお客の信用度を総合評価したものである。職業、年収、勤続年数、家族構成、持ち家の 5 項目をそれぞれ 5 点満点で信用度を評価し、その総合点をリスクとする。したがって、リスクは 0 から 25 の数値になる。ここではそれぞれの項目をどのように評価するのかの詳細にはここでは触れない。

リスク、金額、貸付期間、利率はそれぞれ別の観点からの評価である。したがって、項目間で数値そのものを単純に比較することはできない。そこで、それぞれを規格化してから解析をすすめる。すなわち、分散行列ではなく、相関行列を用いる。解析方法は同じである。

融資可の場合のデータ数n_A、不可の場合の数n_B、トータルのデータ数$n=n_A+n_B$とする。

どちら判定になったかを考えない場合の解析をまずすすめていく。

各項目の平均は

$$\mu_1=\frac{1}{n}\sum_{i=1}^{n}x_{i1} \tag{147}$$

$$\mu_2=\frac{1}{n}\sum_{i=1}^{n}x_{i2} \tag{148}$$

$$\mu_3=\frac{1}{n}\sum_{i=1}^{n}x_{i3} \tag{149}$$

$$\mu_4=\frac{1}{n}\sum_{i=1}^{n}x_{i4} \tag{150}$$

で与えられる。標準偏差は

$$\sigma_1\approx S_1=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{i1}-\mu_1\right)^2} \tag{151}$$

$$\sigma_2\approx S_2=\sqrt{\sum_{i=1}^{n}\left(x_{i2}-\mu_2\right)^2} \tag{152}$$

$$\sigma_3\approx S_3=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{i3}-\mu_3\right)^2} \tag{153}$$

$$\sigma_4\approx S_4=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{i4}-\mu_4\right)^2} \tag{154}$$

で与えられる。ここでは、標本数は十分多いとし、標本分散を母集団の分散として用いている。

これをもとに変数を以下のように規格化する。

$$z_{i1}=\frac{x_{i1}-\mu_1}{\sigma_1} \tag{155}$$

$$z_{i2}=\frac{x_{i2}-\mu_2}{\sigma_2} \tag{156}$$

$$z_{i3}=\frac{x_{i3}-\mu_3}{\sigma_3} \tag{157}$$

$$z_{i4}=\frac{x_{i4}-\mu_4}{\sigma_4} \tag{158}$$

これから各説明変数間の関連をあらわす相関行列

$$R=\begin{pmatrix} r_{11} & r_{12} & r_{13} & r_{14} \\ r_{21} & r_{22} & r_{23} & r_{24} \\ r_{31} & r_{32} & r_{33} & r_{34} \\ r_{41} & r_{42} & r_{43} & r_{44} \end{pmatrix} \tag{159}$$

を得る。ただし

$$r_{ij}=\frac{1}{n}\sum_{k=1}^{n} z_{ki}z_{kj} \tag{160}$$

である。この逆行列を

$$R^{-1}=\begin{pmatrix} r^{11} & r^{12} & r^{13} & r^{14} \\ r^{21} & r^{22} & r^{23} & r^{24} \\ r^{31} & r^{32} & r^{33} & r^{34} \\ r^{41} & r^{42} & r^{43} & r^{44} \end{pmatrix} \tag{161}$$

と表記する。逆行列の求め方についてはここではふれない。逆行列は求められるものとして議論をすすめていく。

つぎに融資可（グループ A）、不可(グループ B)の場合を取り入れたデータ解析をしていく。

グループ A、B であることを明記したデータをリスク z_{i1A}、z_{i1B}、金額 z_{i2A}、z_{i3B}、貸付期間 z_{i3A}、z_{i3B} 利率 z_{i4A}、z_{i4B} と表現する。それぞれの平均は

$$\mu_{1A}=\frac{1}{n}\sum_{i=1}^{n_A} z_{i1A} \tag{162}$$

$$\mu_{1B}=\frac{1}{n}\sum_{i=1}^{n_B} z_{i1B} \tag{163}$$

$$\mu_{2A}=\frac{1}{n}\sum_{i=1}^{n_A} z_{i2A} \tag{164}$$

$$\mu_{2B}=\frac{1}{n}\sum_{i=1}^{n_B} z_{i2B} \tag{165}$$

$$\mu_{3A}=\frac{1}{n}\sum_{i=1}^{n_A} z_{i3A} \tag{166}$$

$$\mu_{3B}=\frac{1}{n}\sum_{i=1}^{n_B} z_{i3B} \tag{167}$$

$$\mu_{4A}=\frac{1}{n}\sum_{i=1}^{n_A} z_{i4A} \tag{168}$$

$$\mu_{4B}=\frac{1}{n}\sum_{i=1}^{n_B} z_{i4B} \tag{169}$$

となる。

8.14.2. 教師データの判別

教師データは判別がはっきりしている。このデータを使い、この判別方法がデータに対してどの程度の精度、または合致度であるかを評価できる。

データ i とグループ A の近さは以下に示すマハラノビスの距離の 2 乗で評価される。

$$D_{iA}^2 = \begin{pmatrix} z_{i1}-\mu_{1A} & z_{i2}-\mu_{2A} & z_{i3}-\mu_{3A} & z_{i4}-\mu_{4A} \end{pmatrix} R^{-1} \begin{pmatrix} z_{i1}-\mu_{1A} \\ z_{i2}-\mu_{2A} \\ z_{i3}-\mu_{3A} \\ z_{i4}-\mu_{4A} \end{pmatrix} \tag{170}$$

同様にこのデータとグループ B の近さは

$$D_{iB}^2 = \begin{pmatrix} z_{i1}-\mu_{1B} & z_{i2}-\mu_{2B} & z_{i3}-\mu_{3B} & z_{i4}-\mu_{4B} \end{pmatrix} R^{-1} \begin{pmatrix} z_{i1}-\mu_{1B} \\ z_{i2}-\mu_{2B} \\ z_{i3}-\mu_{3B} \\ z_{i4}-\mu_{4B} \end{pmatrix} \tag{171}$$

となる。

データが A 、B どちらのグループに属するかは以下のように判別される。

$$\begin{cases} D_{iA}^2 < D_{iB}^2 & \to A \\ D_{iA}^2 > D_{iB}^2 & \to B \end{cases} \tag{172}$$

各データ i はどちらのグループであったかがデータとしてあるから、それ合っているかどうか比較することができる。それを合致率と呼ぶ。

8.14.3. 任意データの判定

リスク x_1 、金額 x_2 、貸付期間 x_3 、利率 x_4 が A 、B どちらのグループに属するか判別したい。

まずデータを以下のように規格化する。

$$u_1 = \frac{x_1 - \mu_1}{\sigma_1} \tag{173}$$

$$u_2 = \frac{x_2 - \mu_2}{\sigma_2} \tag{174}$$

$$u_3 = \frac{x_3 - \mu_3}{\sigma_3} \tag{175}$$

$$u_4 = \frac{x_4 - \mu_4}{\sigma_4} \tag{176}$$

このデータとグループ A の近さは以下に示すマハラノビスの距離の 2 乗で評価される。

$$D_A^2 = \begin{pmatrix} u_1 - \mu_{1A} & u_2 - \mu_{2A} & u_3 - \mu_{3A} & u_4 - \mu_{4A} \end{pmatrix} R^{-1} \begin{pmatrix} u_1 - \mu_{1A} \\ u_2 - \mu_{2A} \\ u_3 - \mu_{3A} \\ u_4 - \mu_{4A} \end{pmatrix} \tag{177}$$

同様にこのデータとグループ B の近さは

$$D_B^2 = \begin{pmatrix} u_1 - \mu_{1B} & u_2 - \mu_{2B} & u_3 - \mu_{3B} & u_4 - \mu_{4B} \end{pmatrix} R^{-1} \begin{pmatrix} u_1 - \mu_{1B} \\ u_2 - \mu_{2B} \\ u_3 - \mu_{3B} \\ u_4 - \mu_{4B} \end{pmatrix} \tag{178}$$

となる。

データが A 、B どちらのグループに属するかは以下のように判別される。

$$\begin{cases} D_A^2 < D_B^2 & \to A \\ D_A^2 > D_B^2 & \to B \end{cases} \tag{179}$$

8.14.4. 条件を提示したい場合

これまでの解析結果から、新たなデータの組に対して、判別結果を得ることができる。そうではなくて、お客にどのような条件を提示できるかを示すにはどうすればいいかを考える。

これまでの、結果に基づいてお客にリスク、金額、貸付期間、利率のどの点を提示すればいいのか、という問題にいきあたる。

我々が表で示すことができるのは、二つの項目までである。したがって、二つの項目をのぞいた項目は、その値を指定できるものとする。その状態で残りの二つの項目を系統的にサーチすればいい。

上の例でいうとお客のリスクと金額が x_{10} 、金額 x_{20} とわかっているものとする。このような客にどのような貸付期間 x_3 、と利率 x_4 を提示すればいいのか、は提示できる。

8.15. まとめ

この章のまとめを行う。

あるグループとそのデータの距離は以下で与えられる。

$$d = \sqrt{D^2}$$

ここで、 D は

$$D^2 = \left(\mathbf{x} - \boldsymbol{\mu}\right)^T \Sigma^{-1} \left(\mathbf{x} - \boldsymbol{\mu}\right)$$

ただし、

$$\Sigma = \begin{pmatrix} \sigma_{11}^{(2)} & \sigma_{12}^{(2)} & \cdots & \sigma_{1p}^{(2)} \\ \sigma_{21}^{(2)} & \sigma_{22}^{(2)} & \cdots & \sigma_{2p}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1}^{(2)} & \sigma_{p2}^{(2)} & \cdots & \sigma_{pp}^{(2)} \end{pmatrix}$$

$$\mathbf{x} - \boldsymbol{\mu} = \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \\ \vdots \\ x_p - \mu_p \end{pmatrix}$$

である。この距離が一番短いグループにそのデータは属すると判断する。

もとの変数を規格化しているとすると

$$D^2 = \left(\mathbf{z} - \boldsymbol{\mu}_z\right)^T R^{-1} \left(\mathbf{z} - \boldsymbol{\mu}_z\right)$$

となる。ただし、

$$R = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{pmatrix}$$

$$\mathbf{z} - \boldsymbol{\mu}_z = \begin{pmatrix} z_1 - \mu_{z_1} \\ z_2 - \mu_{z_2} \\ \vdots \\ z_p - \mu_{z_p} \end{pmatrix}$$

である。

評価する項目の重みを考慮するばあいは

$$D^2 = \begin{pmatrix} w_1\left(z_1 - \mu_{z_1}\right) & w_2\left(z_2 - \mu_{z_2}\right) & \cdots & w_p\left(z_p - \mu_{z_0}\right) \end{pmatrix} R^{-1} \begin{pmatrix} w_1\left(z_1 - \mu_{z_1}\right) \\ w_2\left(z_2 - \mu_{z_2}\right) \\ \vdots \\ w_p\left(z_p - \mu_{z_0}\right) \end{pmatrix}$$

となる。μ_{z_i} と r_{ij} は既存のデータから評価できる。w_i は既存のデータによく合うように定める。