8 デザイン行列とコーディング

この章は一般化線型モデルの準備であり，共変量データの行列表現（デザイン行列）について説明する．デザイン行列のコーディングの数値例がいくつか示される．

キーワード 交互作用，高次の項，ダミー変数，デザイン行列

8.1 デザイン行列

一般化線型モデルは，アウトカム Y とその確率分布を規定する共変量 $X_1, X_2, \ldots, X_p$ の関係を表現したものである．具体的には，Y の条件付期待値が，リンク関数 $g(x)$ を通じて

$$g[\mathrm{E}(Y_i|\boldsymbol{X}_i)] = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1}$$

という構造を仮定する．共変量の指定の仕方が違えば，根本的に異なるモデルを当てはめることになる．だから第一に，N 人の対象者から得られたデータがあるとき，ベクトルでどのように表記し，プログラム上でどのようにコーディングするかを知っておかなければならない．

結果変数とパラメータをそれぞれ N 次元または p 次元の縦ベクトル

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{pmatrix}$$

で表す．そして共変量はデザイン行列と呼ばれる $N \times p$ 行列 $\boldsymbol{X}$ で表す．すると冒頭のモデルは，N 人の対象者について

$$\begin{pmatrix} g[\mathrm{E}(Y_1|\boldsymbol{X}_1)] \\ \vdots \\ g[\mathrm{E}(Y_N|\boldsymbol{X}_N)] \end{pmatrix} = \boldsymbol{X}\boldsymbol{\beta}$$

という関係を仮定していることになる．これは

$$\begin{pmatrix} g[\mathrm{E}(Y_1|\boldsymbol{X}_1)] \\ \vdots \\ g[\mathrm{E}(Y_N|\boldsymbol{X}_N)] \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \cdots + \beta_{p-1} X_{1,p-1} \\ \vdots \\ \beta_0 + \beta_1 X_{N1} + \beta_2 X_{N2} + \cdots + \beta_{p-1} X_{N,p-1} \end{pmatrix}$$

をベクトル表記したものである．

8.2 連続データの扱い

もっとも単純なケースは，共変量が連続データ X_1 のみのモデルであり，これを単回帰という．リンク関数として $g(x) = x$ が用いられる（これを恒等関数という）．この単回帰のモデルは，対象者 i で代表させて

$$\mathrm{E}(Y_i|X_{i1}) = \beta_0 + \beta_1 X_{i1}$$

と書くこともあるし，N 人すべてを

$$\boldsymbol{X}_i = (1 \quad X_{i1})$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

を用いて

$$\begin{pmatrix} \mathrm{E}(Y_1|X_{11}) \\ \vdots \\ \mathrm{E}(Y_N|X_{N1}) \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1\boldsymbol{\beta} \\ \vdots \\ \boldsymbol{X}_N\boldsymbol{\beta} \end{pmatrix} = \boldsymbol{X}\boldsymbol{\beta}$$

と表すこともできる．

共変量が連続データのとき，デザイン行列のコーディングが単位に依存することに注意しよう．それが煩わしいときは，X_1 に対応する共変量の平均が 0 になるように

$$\boldsymbol{X}_i = (1 \quad X_{i1} - \overline{X}_1)$$

というデザイン行列を用いたり，標準偏差が 1 になるように

$$\boldsymbol{X}_i = (1 \quad X_{i1}/\mathrm{SD})$$

とコーディングしたりする．$\overline{X}_1$ と SD は，それぞれ X_1 の平均と標準偏差で

ある．

もちろん Y と X_1 の関係が 1 次関数かどうかはわからない．仮に 2 次関数だとしたら単回帰ではなく重回帰

$$\mathrm{E}(Y_i|X_{i1}) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2$$

となる．この式は X_1 の 2 次式だが，パラメータについては 1 次式という点に注意してほしい．一般化線型モデルとは，パラメータについて線型（linear）という意味であって，共変量の高次の項を含んでいて構わない．

$p-1$ 種類の共変量 $X_1, X_2, \ldots, X_{p-1}$ がすべて連続データのとき，古典的な重回帰分析は

$$\mathrm{E}(Y_i|\boldsymbol{X}_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} = \boldsymbol{X}_i \boldsymbol{\beta}$$

というような表記になる．デザイン行列 $\boldsymbol{X}$ は切片項を含めた $N \times p$ 行列となるだろう．X_1 と X_2 の 2 次の項について考えてみよう．このとき，X_1^2 と X_2^2 の項を追加した

$$\mathrm{E}\,(Y_i|\boldsymbol{X}_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + \beta_3 X_{i2} + \beta_4 X_{i2}^2 + \cdots + \beta_{p+1} X_{i,p-1}$$

という回帰式だけではなく，両者の積 $X_1\,X_2$ を含む

$$\mathrm{E}\,(Y_i|\boldsymbol{X}_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + \beta_3 X_{i2} + \beta_4 X_{i2}^2 + \beta_5 X_{i1} X_{i2} + \cdots + \beta_{p+2} X_{i,p-1}$$

を考えることができる．この項を交互作用（interaction）や統計的交互作用（statistical interaction）という．

8.3 分類データの扱いとダミー変数

共変量が分類データを含むときは，0 または 1 の値をとるダミー変数を用いて，コーディングする必要がある．もっとも簡単な例として，コントロール群 N_0 人のアウトカム $Y_1, Y_2, \ldots, Y_{N0}$ と試験治療群 $N - N_0$ 人のアウトカム $Y_{N0+1}, Y_{N0+2}, \ldots, Y_N$ の平均を比較する状況を考えよう．リンク関数は $g(x) = x$ とする．2 群の平均をそれぞれ推定したいときには

$$\boldsymbol{X} = (\boldsymbol{X}_1 \quad \boldsymbol{X}_2) = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu_0 \\ \mu_1 \end{pmatrix}$$

というデザイン行列となる．コントロール群の平均は

$$\mathrm{E}(Y_i | X_{i1} = 1, X_{i2} = 0) = \mu_0,$$

試験治療群の平均は

$$\mathrm{E}(Y_i | X_{i1} = 0, X_{i2} = 1) = \mu_1$$

と表される．デザイン行列のコーディングによっては，2 群間の差を直接推定することもできる．このとき，デザイン行列とパラメータは

$$\boldsymbol{X} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ \vdots & 0 \\ \vdots & 1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu_0 \\ \beta \end{pmatrix}$$

となる．この場合，β は試験治療群の平均がコントロール群に比べてどれくらい差があるかを表している．コントロール群の平均は

$$\mathrm{E}(Y_i | X_{i1} = 1, X_{i2} = 0) = \mu_0,$$

試験治療群の平均は

$$\mathrm{E}\left(Y_i | X_{i1} = 1, X_{i2} = 1\right) = \mu_0 + \beta$$

と表される．このように，ダミー変数は，0 でコーディングしたカテゴリーが比較の基準になる．この場合の X_{i1} は切片項に相当する．もし，2 群全体の平均を推定したければ

$$\boldsymbol{X} = \begin{pmatrix} 1 & -1/2 \\ \vdots & \vdots \\ \vdots & -1/2 \\ \vdots & 1/2 \\ \vdots & \vdots \\ 1 & 1/2 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \beta \end{pmatrix}$$

とコーディングすればよい．μ は 2 群全体の平均に対応するパラメータになる．このように，デザイン行列には無数のコーディングの仕方があって，それに応じてパラメータの解釈が異なる．

さて，デザイン行列に，切片項，コントロール群，試験治療群という 3 変数を含めたらどうなるだろうか．

$$\boldsymbol{X} = \begin{pmatrix} 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ \vdots & 0 & 1 \\ \vdots & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

2 群しかないのに 3 つのパラメータを推定することはできない．これは，9 章で述べるデザイン行列の列ベクトルが一次従属のときどう対処するかという問題である．