多変量回帰モデルにおける説明変数の選択と予測区間の構成

Input selection and prediction intervals for multivariate regression model

数学専攻　今井沙也可

IMAI Sayaka

1 はじめに

回帰モデルは, 現象の結果と, それに影響を与えると考えられる複数の要因とを結びつけることで, 現象の解釈や予測を行うためのモデルである. 特に, 結果を表す変数である目的変数がただ一つの場合である重回帰モデルを中心に, これまでに様々な研究がなされている. 一方で, 目的変数が複数である多変量回帰モデルは, 重回帰モデルとして個々の目的変数に対するパラメータ推定を行っても同じ点推定量が得られることもあり, 重回帰モデルと比べて多くの研究がなされているとは言い難い.

回帰モデルの基本的な目的はパラメータの点推定や区間推定である. しかし, 説明変数の値を固定したときの目的変数の値の予測も非常に重要である. 例えば重回帰モデルにおいては, 誤差の正規性の仮定の下で, 目的変数を区間で予測した予測区間を求めることもできる. 多変量回帰モデルにおいても同様の考え方で, 多変量の目的変数の値を領域で予測した予測領域 (あるいは予測楕円) を求めることができ, それを用いて各目的変数ごとの予測区間を求めることも可能である (Johnson and Wichern, 2008). こうした区間予測により, 誤差によるばらつきを考慮に入れた予測が可能となる.

回帰モデルにおいて, パラメータ推定や予測の精度も重要な問題である. 多くの説明変数を用いると, 観測値に対する過適合となり, 将来のデータに対する予測や, パラメータ推定の精度が落ちてしまうという問題が生じる (小西, 2010). そこで, 説明変数の選択が重要となる. AIC などの情報量規準を用いた変数選択法は, これまでに多くの研究や応用がなされてきた.

一方で, L_1 正則化に基づく変数選択法が近年脚光を浴びてきている. Lasso (Tibshirani, 1996) に代表される L_1 正則化法は, パラメータの縮小推定と変数選択を同時に行うことができるという特徴を持つ. さらに, 説明変数が高次元の場合などにも有用であるため, ゲノム解析や機械学習など様々な分野で活用されるようになってきた. L_1 正則化法は非常に有用であるが, 推定量が解析的に陽な形で表現できず, 凸 2 次計画問題として計算機を用いて解くことになる. したがって, 推定量の標本分布の導出が難しく, 先に述べた予測領域や予測区間を得ることも難しい.

2 多変量回帰モデルと予測区間

目的変数が複数ある場合に, p 個の説明変数から予測することを考える. いま, q 個の目的変数 $\boldsymbol{Y} = (Y_1, \ldots, Y_q)^T$ および p 個の説明変数 $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ に関し, n 組の観測値 $(\boldsymbol{Y}_i, \boldsymbol{X}_i)$ が観測されたとする. 説明変数 $\boldsymbol{X}$ は確率変数ではなく所与の値であるとする. このとき, 以下のようなモデルを多変量回帰モデルという:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}.$$

ここで, $\boldsymbol{Y}$ は $n \times q$ の目的変数の行列, $\boldsymbol{X}$ は $n \times (p+1)$ の計画行列, $\boldsymbol{B}$ は $(p+1) \times q$ のパラメータ行列, $\boldsymbol{E}$ は $n \times q$ の誤差行列を表す. 誤差項 $\boldsymbol{E}$ は以下の仮定をおく:

$$\begin{aligned} \boldsymbol{E} &= \left(\boldsymbol{E}_{(1)}, \ldots, \boldsymbol{E}_{(m)}\right), \\ E(\boldsymbol{E}_{(i)}) &= \boldsymbol{0}, \qquad\qquad (i, k = 1, 2, \ldots, m) \\ \mathrm{Cov}(\boldsymbol{E}_{(i)}, \boldsymbol{E}_{(k)}) &= \sigma_{ik}\boldsymbol{I}, \quad \boldsymbol{\Sigma} = (\sigma_{ij}), \\ \boldsymbol{E} &\sim N\left(\boldsymbol{0}, \sigma_{ik}\boldsymbol{I}\right). \end{aligned}$$

パラメータ $\boldsymbol{B}$, Σ の点推定は最小二乗法や最尤法によって行う. $\boldsymbol{B}$ の最小二乗推定量と $\boldsymbol{\Sigma}$ の最尤推定量はそれぞれ

$$\begin{aligned} \hat{\boldsymbol{B}} &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}, \\ \hat{\boldsymbol{\Sigma}} &= \frac{1}{n}\left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\right)^T\left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\right). \end{aligned}$$

となる. ただし, $\boldsymbol{X}^T\boldsymbol{X}$ の逆行列は存在すると仮定している.

多変量回帰モデルにおける予測区間を考える. 誤差 $\boldsymbol{E}$ が正規分布に従うとき, 個々の目的変数 Y_{0j} に対する $100\,(1-\alpha)$ %同時予測区間は

$$\left[\boldsymbol{x}_0^T\hat{\boldsymbol{B}}_{(j)} - G(\alpha), \boldsymbol{x}_0^T\hat{\boldsymbol{B}}_{(j)} + G(\alpha)\right] (j = 1, 2, \ldots, m)$$

と書ける. ここで,

$$G(\alpha) = \sqrt{\frac{q\,(n-p-1)}{n-p-q}F_{q,n-p-q}\,(\alpha)}\sqrt{\left\{1 + \boldsymbol{x}_0^T\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{x}_0\right\}\left(\frac{n}{n-p-1}\right)\hat{\sigma}_{jj}}$$

である.

3 多変量回帰モデルにおける変数選択

多変量回帰モデルでの変数選択を行うために, Group Lasso を紹介する. Group Lasso(Yuan and Lin,2006) は, 重回帰モデルにおいて説明変数がいくつかのグループに分かれている場合に, 各グループ内の変数をまとめて選択あるいは除外することのできる方法である. いま, p 個の説明変数が J 個のグループに分かれているとし, 各グループ内の説明変数の個数を $p_j, j = 1, 2, \ldots, J$ とする. このとき, j 番目のグループに属する $n \times p_j$ 計画行列を $\boldsymbol{X}_j$ としたとき, 回帰モデルは以下のように表すことができる:

$$\boldsymbol{y} = \sum_{j=1}^{J}\boldsymbol{X}_j\boldsymbol{\beta}_j + \boldsymbol{\epsilon}$$

ここで, $\boldsymbol{\beta}_j$ は p_j 次元係数パラメータベクトルである. なお, 各説明変数は標準化, 目的変数は中心化されているとする.

このとき, Group Lasso は次の目的関数の最小化により定式化される.

$$\left(\boldsymbol{y} - \sum_{j=1}^{J}\boldsymbol{X}_j\boldsymbol{\beta}_j\right)^T\left(\boldsymbol{y} - \sum_{j=1}^{J}\boldsymbol{X}_j\boldsymbol{\beta}_j\right) + \lambda\sum_{j=1}^{J}\sqrt{p_j}\,||\boldsymbol{\beta}_{\boldsymbol{j}}||_2$$

ここで, $\lambda(>0)$ は正則化パラメータであり, $||\cdot||_2$ はユークリッドノルム $\left(||\boldsymbol{\beta}||_2 = \left(\boldsymbol{\beta}^T\boldsymbol{\beta}\right)^{1/2}\right)$ を表す. Group Lasso においても推定量を解析的に陽な形で表現できないため, 数値的な方法で解くことになる. Group Lasso を多変量回帰モデルで用いる場合, Lasso 推定量の場合と同様に各行列をベクトル表現すればよい. このとき, パラメータ行列 $\boldsymbol{B}$ の各行をひとつのグループと考えることで, 多変量回帰モデル全体における各説明変数の選択および除外を行うことができるようになる (Kim and Xing, 2012).

4 ブートストラップ法による予測区間の構成

説明変数 X は所与の値であるとする. 誤差の実現値と考えられる残差 $\hat{\boldsymbol{E}}$ を用いて, 誤差分布 F を推定することを考える. $\boldsymbol{B}$ の Group Lasso 推定量を $\hat{\boldsymbol{B}}$ とし, チューニングパラメータ λ は既知である (あるいはクロスバリデーション法などにより決定されている) とする. 分散を修正した残差

$$r_{ij} = \frac{\hat{\epsilon}_{ij}}{\sqrt{1-\boldsymbol{x}_{ij}\left(\boldsymbol{X}_i^T\boldsymbol{X}_i\right)^{-1}\boldsymbol{x}_{ij}^T}} - \frac{1}{n}\sum_{s=1}^{n}\frac{\hat{\epsilon}_{sj}}{\sqrt{1-\boldsymbol{x}_{sj}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{x}_{sj}^T}}$$

のリサンプリングを考える. すなわち, 修正した残差の復元抽出を無作為に行い, ブートストラップ残差

$$\boldsymbol{R}^* = \begin{pmatrix} r_{11}^* & \cdots & r_{1q}^* \\ \vdots & \ddots & \vdots \\ r_{n1}^* & \cdots & r_{nq}^* \end{pmatrix}$$

を生成する. 生成されたブートストラップ残差 $\boldsymbol{R}^*$ と予測値 $\hat{\boldsymbol{Y}}$ を用いて, 目的変数のブートストラップ標本

$$\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{B}} + \boldsymbol{R}^*$$

を作成することができる. そこで, 観測データの説明変数 $\boldsymbol{X}$ とブートストラップ標本の目的変数 $\boldsymbol{Y}^*$ を用いることで, 新たな $\boldsymbol{B}$ の推定量 $\hat{\boldsymbol{B}}^*$ を得ることができる. 残差を B 回無作為復元抽出し, $\boldsymbol{B}$ の推定量

$$\hat{\boldsymbol{B}}_1^*, \ldots, \hat{\boldsymbol{B}}_1^*$$

を得ることで, $\boldsymbol{B}$ の分布を推定することができる. さらに, $\hat{\boldsymbol{B}}_1^*, \ldots, \hat{\boldsymbol{B}}_1^*$ を用いて, ブートストラップ予測値は

$$\hat{\boldsymbol{Y}}^* = \boldsymbol{X}\hat{\boldsymbol{B}}^*$$

となることから, ブートストラップ予測誤差

$$\boldsymbol{E}^* = \hat{\boldsymbol{Y}}^* - \hat{\boldsymbol{Y}} - \boldsymbol{r}^*$$

を作成することができる. したがって, ブートストラップ予測誤差 $\boldsymbol{E}^*$ の分布を G^* とし, 上側 (100α) %点を $G^*(\alpha)$ とすれば, Y_{ij} の予測区間は,

$$\left[\hat{Y}_{ij} - G^*(1-\alpha), \hat{Y}_{ij} + G^*(1-\alpha)\right]$$

を計算することで推定できる.

参考文献

[1] Chu C. H., Tsai Y. T., Wang C. L. and, Kwok T. H. (2010). "Exemplar-based statistical model for semantic parametric design of human body". Computers in Industry. 61:541-549.

[2] 川野秀一, 廣瀬慧, 立石正平, 小西貞則 (2010). "回帰モデリングと L_1 型正則化法の最近の展開". 日本統計学会誌. vol.39, no.2, pp.211-242.

[3] Kim, S., Xing, E, P. (2012). "Tree-guided group lasso for multi-response regression with structured sparsity, with an application to EQTL mapping". The Annals of Applied Statistics. vol.6, no.3, pp.1095-1117.

[4] 小西貞則 (2010). "多変量解析入門". 岩波書店.

[5] Richard, A., Johnson, Dean, W., Wichern.(2008). "Applied Multivariate Statistical Analysis". Pearson.

[6] Sandrine, Juan., Frdric, Lantz.(2001). "Application of bootstrap techniques in econometrics:the example of cost estimation in the automotive industry". .

[7] 佐和隆光 (1979). "回帰分析". 朝倉書店.

[8] Simiia, T., Tikka, J. (2007). "Input selection and shrinkage in multiresponse linear regression". Computational Statistics and Data Analysis. vol.52, pp.406-422.

[9] 竹村彰通 (1991). "多変量推測統計の基礎". 共立出版.

[10] Tibshirani, R. (1996). "Regression shirinkage and selection via the lasso". J. Roy. Statist. Soc. Ser. B, vol.58, 267-288.

[11] 汪金芳, 桜井裕仁 (2011). "ブートストラップ入門". 共立出版.

[12] 汪金芳, 田栗正章, 手塚集, 樺島祥介, 上田修功 (2003). "計算統計 I". 岩波書店.

[13] Yuan, M. and Lin, Y. (2006). "Model selection and estimation in regression with grouped variables". J. Roy. Statist. Soc. Ser. B, vol.68, 49-67.

[14] Wu, C. F. J. (1986). "Jackknife, bootstrap and other resampling methods in regression analysis (with discussion)". Annals of Statistics, vol.9, 1218-1228.