数理解析研究所講究録 1439 巻 2005 年 241-246

クロスバリデーションによる比例ハザードモデルの選択

北里大学大学院 薬学研究科
高橋 史朗

1 はじめに

がん臨床研究の第Ⅲ相試験の主たる目的は，処置間の生存期間の違いを検証することである．さらに，生存に寄与する予後因子を同定し，ハイリスク集団の予測生存期間などの予測モデルを構築することである．多くの臨床論文では，Cox の比例ハザードモデルを当てはめ，パラメータの推定値，その信頼区間および Wald タイプの p 値を示している．しかしながら，モデルの当てはまりの良さ，予測精度などに関しては，当然一言も言及されていないのが現状である．モデルの当てはまりや比例ハザード性の検討については，数多くの方法が提案されている．たとえば，Martingale 残差，Martingale 残差の累積和に基づく方法などである．しかし，予測モデルに関する研究は十分とはいえない状況であろう．

予測誤差は，将来得られる観測値をモデルがどれだけよく予測できるかを図る指標である．クロスバリデーションは，回帰分析，判別分析やクラスタリングなどで予測誤差を推定する標準的な方法として広く知られている．そこで本論では，興味のあるデータが右側無情報打ち切り生存時間データの場合へクロスバリデーションを拡張し，Cox の比例ハザードモデルのモデル選択方法について考える．

2 生存時間解析

2.1 記号など

症例 i において，生存時間 $T_i(i=1,\ldots,n)$ は独立に同一の分布 F に従い，打ち切り時間 $C_i(i=1,\ldots,n)$ は独立に同一の分布 G に従うと仮定する．T_i と C_i は独立と仮定する．追跡時間 $U_i=\min(T_i,C_i)$，打ち切り指示変数 $\delta_i=1(T_i\leq C_i)$ と定義する．症例 i の共変量 Z_i とすると，$(U_i,\delta_i,Z_i)(i=1,\ldots,n)$ が観察されるとする．さらに，リスクセット指示変数を $Y_i(t)=1(U_i\geq t)$ で表すものとする．

比例ハザードモデル $\lambda(t|Z_i)=\lambda_0(t)\exp(Z_i^t\beta)$ の仮定のもと，部分尤度スコア関数は次のようになる．

$$\sum_{i=1}^{n}\delta_i\left[Z_i-\sum_{j=1}^{n}\frac{Y_j(t)Z_l exp(Z_l^t\beta)}{\sum_k Y_k(t)\exp(Z_k^t\beta)}\right]=0$$

この推定方程式から得られる推定量 $\hat{\beta}$ は，漸近的に一致推定量であり，漸近正規性を持つことが知られている．

2.2 提案されている予測精度を測る指標

Schemper and Henderson(2000) は，モデルによって説明される変動の割合を特徴付ける R^2 を比例ハザードモデルへ拡張した統計量を提案した．彼らは，計数過程 $N_i^*(t) = 1(T_i \leq t)$，生存関数の期待値 $S(t) = E[1 - N_i^*(t)]$ と条件付期待値 $S(t|Z_i) = E[1 - N_i^*|Z_i]$ を用いる．時点 t を固定したとき，$N_i^*(t)$ は2値データなので，周辺分散 $S(t)(1-S(t))$ と条件付分散 $S(t|Z_i)(1-S(t|Z_i))$ を用い，共変量 Z_i によって説明される変動の割合を特徴付けた．特に,$(0,\tau)$ の範囲の時間平均変動として，

$$\begin{aligned} D(\tau) &= \frac{\int_0^\tau S(t)[1-S(t)]f(t)dt}{\int_0^\tau f(t)dt} \\ D_Z(\tau) &= \frac{\int_0^\tau S(t|Z_i)[1-S(t|Z_i)]f(t)dt}{\int_0^\tau f(t)dt} \end{aligned}$$

を定義した．そして，

$$V(\tau) = \frac{D(\tau) - D_Z(\tau)}{D(\tau)}. \tag{1}$$

O'Quigley and Xu(2001) は，共変量と生存時間の役回りを変更し，生存時間によって説明される共変量のばらつきの割合を提案した．

3 クロスバリデーション

3.1 線型モデルにおけるクロスバリデーション

3.1.1 モデルと予測誤差

線型モデル

$$y = X\beta + e \tag{2}$$

を考える．ここで，y は n 次元応答変数ベクトル，x は 既知の $p \times n$ 次元デザイン行列，β は p 次元の未知パラメータベクトル，そして e は 平均0，共分散行列 $\sigma^2 I$ の n 次元誤差ベクトルである．β のいくつかの要素がゼロであるかもしれないので，より簡単なモデル

$$y = X_\alpha \beta_\alpha + e \tag{3}$$

を M_α であらわすことにする．ここで，x_α および β_α は，それぞれ x および β の部分集合を表す．式3において，考えうる異なるモデルの個数は $2^p - 1$ 個である．もし真のモデル，すなわち β の各要素が0であるか否かを知りえるとするならば，M_α は

- クラス1：真の β のゼロでない要素が，β_α に少なくとも1つ以上含まれていない．
- クラス2：真の β のゼロでないすべての要素が，β_α に含まれている．

に分類できる．明らかに，クラス1に含まれるモデルは正しくなく，そしてクラス2に含まれるモデルは非効率なモデルであろう．最適なモデルを $M_\star$ で表すとすると，$M_\star$ はクラス2のより簡単なモデルとなる．

モデル M_α のもとで，β_α の最小二乗推定量は，

$$\hat{\beta}_\alpha = (X_\alpha^t X_\alpha)^{-1} X_\alpha y.$$

なお，$rank(X_\alpha^t X_\alpha) = d_\alpha$ とする．新たなデータ $(Z_{(i)}, x_{(i)})(i = 1, \ldots,)$ が得られたとする．モデル M_α を用いたとき，二乗予測誤差の平均は

$$m^{-1}\sum_{i=1}^{m}(z_{(i)} - x_{(i)\alpha}^t\hat{\beta})$$

である．ただし，$\hat{\beta}_\alpha$ は以前のデータから算出された最小二乗推定量である．$y_{(\cdot)}$ を与えたもとでの，二乗予測誤差の条件付期待値は，

$$\begin{aligned} E\left[m^{-1}\sum_{i=1}^{m}(z_{(i)} - x_{i\alpha}^t\hat{\beta})\middle| y\right]^2 &= E\left[m^{-1}\sum_{i=1}^{m}(z_{(i)} - x_i\beta + x_i\beta x_{i\alpha}^t\hat{\beta})\middle| y\right]^2 \\ &= \sigma^2 + \frac{1}{m}\sum_i\left(x_i^t\beta - x_{i\alpha}^t\hat{\beta}_\alpha\right)^2. \end{aligned}$$

よって，二乗予測誤差の期待値は，

$$\Gamma_{\alpha,m} = E\left[m^{-1}\sum_{i=1}^{m}(z_i - x_{i\alpha}^t\hat{\beta})\right] = \sigma^2 + m^{-1}d_\alpha\sigma^2 + \Delta_{\alpha,m}. \tag{4}$$

ただし，

$$\begin{aligned} \Delta_{\alpha,m} &= m^{-1}\beta^t X(I_m - P_\alpha)X\beta, \\ P_\alpha &= X_\alpha(X_\alpha^t X_\alpha)^{-1}X_\alpha^t. \end{aligned}$$

$\Gamma_{\alpha,m}$ は，新たな観測値のバラツキと選択されたモデルと推定値の誤差を反映する項 $m^{-1}d_\alpha\sigma^2\Delta_{\alpha,m}$ に分離できる．M_α がクラス 2 のモデルであれば，$X\beta = X_\alpha\beta_\alpha$ であり，

$$\Gamma_{\alpha,m} = \sigma^2 + m^{-1}d_\alpha\sigma^2. \tag{5}$$

一方，M_α がクラス 1 のモデルであれば，P_α が X の部分集合 X_α への射影行列であるため，任意の m に対して $\Delta_{\alpha,m} > 0$.

3.1.2 線型モデルにおけるクロスバリデーション

二乗予測誤差を算出するために，新たな実験を実施することが困難な場合がある．そこで，得られたデータを 2 つに区分する．すなわち，予測誤差を算出するためのデータセット $\{(y_i, x_i), i \in S\}$ と，モデル M_α を構築するデータセット $\{(y_i, x_i), i \in S^c\}$ に区分する．それぞれのデータ数を n_v と n_c で表すものとする．このとき，二乗予測誤差の平均値は，

$$n_v^{-1}||y_s - \hat{y}_{\alpha,S^c}||^2. \tag{6}$$

ただし，$||a|| = (a^t a)^{1/2}$, y_s は $\{(y_i, x_i), i \in S\}$ の n_v 次元ベクトル，y_{α,S^c} はモデル M_α およびモデル構築データ $\{(y_i, x_i), iin S^c\}$ から得られた y_s の予測値である．この式は，以下のとおり書き直すことができる．

$$n_v^{-1}||y_s - \hat{y}_{\alpha,S^c}||^2 = n_v^{-1}\left|\left|(I_{n_v} - Q_{\alpha,s})^{-1}\left(y_s - X_{\alpha,S}\hat{\beta}_\alpha\right)\right|\right|^2 \tag{7}$$

ここで，$X_{\alpha,S}$ は モデル M_α のもとでのバリデーションデータ X_α の部分集合である $(n_v \times d_\alpha)$ 次元デザイン行列であり，$Q_{\alpha,S} = X_{\alpha,S}\left(X_\alpha^t X_\alpha\right)^{-1} X_{\alpha,S}^t$, $\hat{\beta}_\alpha$ は全データ (n) を用いた場合の最小二乗推定量である．

それぞれのモデル M_α に対して，$\Gamma_{\alpha,n}$ のクロスバリデーション推定量は，大きさ n_v の集合 S 上の式 (7) の平均として得られる．クロスバリデーションにより選択されるモデルは，すべての M_α に対するクロスバリデーション推定量のうちそれ最小とするモデルである．Shao(1993) は，以下のことを証明した．

- leave-one-out 法 $(n_v = 1)$
  M_α がクラス 2 に属し，$M_\alpha \neq M_\star$ であったとしても，

$$P(\Gamma_{\alpha,n}^{CV1} < \Gamma_{\star,n}^{CV1}) \neq 0 \tag{8}$$

- leave-n_v-out 法 $(1 < n_v < n)$
  $n_v/n \to 1$, $n_c = n - n_v \to \infty$, および適切な正則条件のもとで，

$$\lim_{n\to\infty} \Pr\left(M_\star \text{が選択}\right) = 1 \tag{9}$$

3.2 比例ハザードモデルにおけるクロスバリデーション

3.2.1 提案する推定量

打ち切りデータがない場合，線型線形モデルにおける損失関数の自然な拡張は，

$$\Gamma = n_v^{-1} \sum_{i=1}^{n_v} [y_i - E(Y_i|Z_i)]^2 \tag{10}$$

ただし，

$$\begin{aligned} E(Y_i|Z_i) &= E\left[\exp(-\hat{\Lambda}(T, Z_i))\right] \\ \hat{\Lambda}(t, Z_i) &= \sum_{j=1}^{n_c} \int_0^t \frac{dN_j(s)}{\sum_{l=1}^{n_c} Y_l(s) \exp\left(\hat{\beta}(Z_l - Z_i)\right)} \end{aligned}$$

打ち切りデータが存在する場合には，直感的に打ち切りデータの不確実性を表す重み W_i を用いた加重平均

$$\Gamma = \left(\sum W_i\right)^{-1} \sum_{i=1}^{n_v} W_i \left[y_i - E(Y_i|Z_i)\right]^2 \tag{11}$$

とすればよいであろう．

重み W_i として，IPCW(Robins & Rotnitzky) を考える．IPCW は，欠測メカニズムが missing at random であるとき，平均および分散の構造を仮定したセミパラメトリックモデルのパラメータを最適に推定する方法として提案された．

従属変数ベクトル $Y_i = (Y_{i1}, \ldots, Y_{iT})^T$，共変量行列 $X_i = (X_{i0}^T, \ldots, X_{iT}^T)^T$ (X_{i0}^t は投与前の共変量)，およびその他の時間依存共変量 $V_{it}(t = 0, \ldots, T)$ とする．症例 i の時点 t において，Y_i および V_i が観察された場合には $R_{it} = 1$, その他の場合には $R_{it} = 0$ をとる指示変数を考える．また，$W_{it} = (V_{it}^T, Y_i t)^T (t = 0, \ldots, T)$, $\bar{W}_{it} = (W_{i0}^T, \ldots, W_{it}^T)$ とする．次の欠測メカニズムを考える．

$$\Pr(R_{it} = 1|R_{i(t-1)} = 1, \bar{W}_{it}, Y_i) = \Pr(R_{it}|R_{i(t-1)}, \bar{W}_{it})$$

これは，現時点までの情報 $\bar{W}_{it}$ を与えたとき，欠測は現時点および将来のデータに依存しない，すなわちMARを仮定していることと同じである．周辺モデル

$$E(Y_{it}|X_i) = g_t(X_i.\beta_0),$$

および観測確率モデル

$$\bar{\lambda}_{it} = \Pr(R_{it} = 1|R_{i(t-1)} = 1, \bar{W}_{it}) = \bar{\lambda}(\alpha_0)$$

を仮定する．ただし，β_0 および α_0 は未知パラメータベクトル，$g_t(\cdot,\cdot)$ および $\bar{\lambda}_{it}(\cdot)$ は既知の関数である．部分尤度

$$L(\alpha) = \prod_i \prod_t \left[\bar{\lambda}_{it}(\alpha)^{R_{it}} \left(1 - \bar{\lambda}_{it}(\alpha)\right)^{1-R_{it}}\right]^{R_{i(t-1)}}$$

を最大にする α （以下，$\hat{\alpha}$ と表す）を与えたとき，推定方程式

$$U(\beta, \hat{\alpha}) = \sum_i D_i(\beta)\Delta_i(\hat{\alpha})\epsilon_i(\beta) = 0$$

を解く方法を提案した．ただし，$\Delta_i(\alpha)$ は，対角要素 $\pi_{it}(\alpha)^{-1}R_{it} = \frac{R_{it}}{\lambda_{i1}(\alpha)\times\ldots\times\lambda_{it}(\alpha)}$ をもつ対角行列，$\epsilon_{it}(\beta) = Y_{it} - g_t(X_i, \beta)$ である．適当な正則条件のもとで，推定方程式は，$n^{1/2}(\hat{\beta} - \beta_0)$ が漸近的に平均ゼロの正規分布に従うような解 $\hat{\beta}$ をもつ．

IPCWの考え方をCoxの比例ハザードモデルにおける予測誤差の重みへ適応すると，

$$W_i = \frac{\delta_i}{\hat{G}(T_i)}. \tag{12}$$

ただし，$G(\cdot)$ はモデル構築データにおける打ち切りデータのKaplan-Meier推定量である．以上より，提案する予測誤差の推定は，

$$\Gamma = \frac{1}{\sum_i \delta_i} \sum_{i=1}^{n_v} \frac{\delta_i}{\hat{G}(T_i)} \left[y_i - E(Y_i|Z_i)\right]^2. \tag{13}$$

4 考察

クロスバリデーションは，回帰分析，判別分析やクラスタリングにおける予測誤差を推定する方法としてしばしば用いられている．本論では，線型回帰モデルにおけるクロスバリデーションの自然な拡張として，Coxの比例ハザードモデルにおける予測誤差の推定法を提案した．しかしながら，提案した統計量の一致性，Shaoが示したような正しいモデルを選択するための条件等に関する検討は，一切なされていない．今後，重みの選択などを含めてこれらの理論的正当性を明らかにし，シミュレーションを通して小標本での性能等を評価していかなければならない．

生存時間解析の領域以外の臨床データに目を向ける．非線型混合モデルの解析（PKデータの解析など）において，モデル構築データと検証データを半分に分割したクロスバリデーションがしばしば用いられているが，理論的正当性は十分でないように思われる．今後，非線型混合モデル等の予測誤差についても検討していく必要があるだろう．

参考文献

[1] Efron, B and Tibshirani, R. (1993). An Introduction to the Bootstrap. Chapman & Hall.

[2] O'Quigley, J. and Xu, R. (2001). Expalined variation in proportional hazards regression. In handbook of statistics in clinical oncology. J. Crowley (ed.), 397-409, New York: Marcel Dekker.

[3] Robins, J. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. JASA 90, 122-129.

[4] Schemper, M. and Henderson, R. (2000). Predictive accuracy and explained variation in Cox regression. Biometrics 56, 249-255.

[5] Shao, J. (1993). Linear Model Selection by Cross-Validation. JASA 88, 486-494.