生存時間解析における「比較」のデザインと解析

杉本知之・下川敏雄・後藤昌司 (2001)

0　序に代えて

実質科学では，その当該分野の現象の理解は，実験と観察の相互研究により深められる．統計的接近法は，実験研究では，デザインする方法論を与え，よく管理されたデータを生成し評価することを可能にする．一方で，観察研究では，いくつかの大きなデータ集合などから，ノイズを排除し統計的に意味のあるシグナルを抽出するか，もしくは有用な統計モデルを構築することによって，当該の現場に有用な観察的視点を与えることが望まれる．ただし，生存時間研究の分野では，データの対象が生物 (人) の生死を扱う特殊性のため，こういった実験研究と観察研究において，他の分野にはあまり存在しない注意が必要である．とくに，限られた実験研究の結果は，より大きく併合された結果により評価されること，併合されて (後ろ向きでも) 数多く吟味されることなどが望まれる．しかし，こういった方法論の進展には，当該分野のさらなる進歩も期待されるが，統計的接近法においても，まだ多く議論の余地を残し，今後にかけてさらに整備されることが必要である．本報告では，このような視点から，生存時間解析における統計的接近法の代表的なツールを概略し，それらの問題点といくつかの発展を議論する．

1　ハザード・モデルとその応用

一般的な比例ハザード・モデル

$$\lambda(t; f(z)) = \lambda_0(t) \exp(f(\beta, z(t))) \tag{1}$$

を考える．ここに，$\lambda_0(\cdot)$ は任意の潜在基礎ハザード関数，β はパラメータ，z は時変型もしくは固定変量とする．通常，$f(\beta, z(t)) = \beta z$ もしくは $f(\beta, z(t)) = \beta z(t)$ とおかれ，このとき，パラメータの推測に関する対数部分 (プロフィール) 尤度 (Cox,1972,1975:Breslow,1972) は

$$pl_n(\beta; z) = n^{-1} \sum_{i=1}^{n} \delta_i \left[\beta z_i(t_i) - \log \left\{ n^{-1} \sum_{j \in \mathscr{R}_i} \exp(\beta z_j(t_i)) \right\} \right] \tag{2}$$

である．ここに，$t_i, \delta_i, R_i (i = 1, \cdots, n)$ は，それぞれ観測時点，中途打ち切り指標，リスク集合である．これより得られる最尤推定値は一致性とオーダ $\sqrt{n}$ の漸近正規性を満たす (Tsiatis,1981:Andersen& Gill,1982)．ここで，$pl_n(\beta; z)$ の β に関するスコア関数を $U_n(\beta)$，2 次微分にマイナスをつけたものを $I_n(\beta)$ とおく．

1.1　順位検定

Fleming-Harrington (1991) のクラスは，$W_n(t) = S_n^p(t)[1 - S_n(t)]^q$ により与えられる重み関数のクラスである．ここに，$S_n(t)$ は併合標本の Kaplan-Meier 推定値である．このクラスからの 2 標本順位検定統計量は

$$\mathscr{Z} = \sum_{i=1}^{k} W_n(t_i) \left\{ d_{i1} - d_i \frac{Y_{i1}}{Y_i} \right\} \Bigg/ \sum_{i=1}^{k} W_n(t_i)^2 d_i \frac{Y_{i1}}{Y_i} \left(1 - \frac{Y_{i1}}{Y_i} \right) \left(\frac{Y_i - d_i}{Y_i - 1} \right) \sim N(0, 1) \tag{3}$$

である．ここに，$Y_i = |\mathscr{R}_i|$，$Y_{i1} = |\mathscr{R}_i \cap$ 標本 $1|$，d_i は t_i での死亡個体数，$d_{i1} = d_i \cap$ 標本 1 である．また，個体 i が標本 1，標本 2 に属するとき，それぞれ $g_i = 1$，$g_i = 0$ とする．$z_i(t)$ を $g_i W_n(t)$ によりおきかえれば，同順位なしのときに，$\mathscr{Z} = U_n(0) \Big/ \sqrt{I_n(0)/n}$ が成りたつ (同順位ありのときには，分母に補正項が追加される)．

次に，(2) 式において $z_i(t) = g_i W_0(t), \hat{z}_i(t) = g_i W_n(t)$ とする．ここに，$W_0(t)$ は $W_n(t)$ の真の値である．通常では，z が観測できないため，推測の尤度として $pl_n(\beta; z)$ の代わりに $pl_n(\beta; \hat{z})$ を用いる．写像

$\phi : D_\phi \subset (D[0,\tau_e]|\beta,\delta_i,t_i,\mathscr{R}_i; i=1,\cdots,n) \mapsto \mathrm{R}$ を，$\phi(z) = \log pl_n(\beta;z)$ と定義する．このとき，いくつかのコンパクト微分とその連鎖性により，有界な β に対して

$$\phi(\hat{z}) = \phi(z) + n^{-1/2}\beta \sum \tilde{z}(t_i)\delta_i[g_i - \sum_{j\in\mathscr{R}_i} g_j \mathrm{e}^{\beta g_j z(t_i)} / \sum_{j\in\mathscr{R}_i} \mathrm{e}^{\beta g_j z(t_i)}] + o(n^{-1/2}) \tag{4}$$

を得る (Sugimoto & Goto,2001)．ここに，$\tilde{z} = n^{1/2}(\hat{z} - z)$ である．上式がこのモデルの推測の漸近的な基本公式になる．

1.2 ノンパラメトリック回帰

ここでは，z を多次元にわたる固定共変量とする．生存時間解析におけるノンパラメトリックな回帰接近法は，(2) を推測の尤度とし，$f(\beta,z)$ をデータに柔軟にあてはめ，理解しやすい関数で近似する方式である．とくに，非線形で交互作用の強い高次元のデータに対して威力を発揮する再帰分割手法の CART 接近法と MARS 接近法を，主要な議題とする (Breiman *et.al.*,1984:Friedman,1991:松原 他,1991:LeBlanc& Crowley,1992,1999)．CART 接近法では，定数関数近似に基づき

$$f(\beta,z) = \sum_{m=1}^{M} \hat{\beta}_m \prod_{k=1}^{K_m} H[s_{km}(z_{v(k,m)} - t_{km})] \tag{5}$$

の形式をとる．ここに，M : モデルの個数，$H[\cdot] : [\cdot] > 0$ のとき 1, $[\cdot] < 0$ のとき 0，K_m : 基底関数の因子数，$v(k,m)$: 共変量番号，t_{km} : 節点，s_{km} :0 か 1，をそれぞれとる．最終モデルの選定には，いくつかの方式が考えられる．交差確認法に基づく LeBlanc& Crowley(1992)，併合を行う松原 他 (1992) がある．計算の高速性と解釈の良さの利点を与えるのが後者である．杉本・松原・後藤 (2001) では，両者の利点の折衷方式をとりいれた．

一方で，MARS 接近法では ($v(k,m)$ に関する制約はいく人かの研究者により除去されつつある)

$$f(\beta,z) = \sum_{m=1}^{M} \hat{\beta}_m \prod_{k=1}^{K_m} [s_{km}(z_{v(k,m)} - t_{km})]_+ \tag{6}$$

の形式をとる．ここに，$[\cdot]_+$ は (1 次) 打ち切りべき基底関数である．(5) の連続への拡張なので，よりの良い予測を与えると予想される．

1.3 中央生存時間

中央生存時間における比較の問題で，漸近的性質に基づく簡便な検定方式は，小標本においてその妥当性はあまり調査されていない．そのため，ある正確方式に基づく方式を提案し，両者を比較する．

参考文献

Andersen, P. K. & Gill, R. D. (1982). Cox's regression model for counting processes:a large sample study. *Ann. Statist.*,**10**,1100-1120. /Breslow, N. E. (1972). Contribution to discussion of paper by D.R.Cox. *J. Roy. Statist. Soc.*, **B34**,216-217. /Breiman, L., Freidman, J. H.,Olshen, R. A., Stone, C. J. (1984). *Classification And Regression Trees.* Wadsworth. /Cox, D. R. (1972). Regression models and life tables. *J. Roy. Statist. Soc.*, **B34**, 187-202. /Cox, D. R. (1975). Partial likelihood. *Biometrika*, **62**, 269-276. /Friedman, J. H. (1991). Multivariate adaptive regression spline. *Ann.Statist.*,**19**, 1-141. /Fleming, T. R. & Harrington, D. P. (1991). *Counting Processes and Survival Analysis.* Wiley. /LeBlanc, M. & Crowley, J. (1992). Relative risk trees for censored survival data. *Biometrics*,**48**,411-425. /LeBlanc,M. & Crowley, J. (1999). Adaptive regression splines in the Cox model. *Biometrics*, **55**, 204-213. /Sugimoto,T. & Goto,M (2001). (Submitting). /Tsiatis, A. A. (1981). A large sample study of Cox's regression model. *Ann. Statist.*, **9**, 93-108. /後藤昌司・松原義弘 (1982)．比例ハザードモデルとその周辺．応用統計学, **11(1)**, 1-26. /松原義弘・渡辺秀章・後藤昌司 (1992)．データ解析過程における樹木表現の諸法．日本分類学会シンポジウム予稿集．/杉本・松原・後藤 (2001).(準備中).