論文の内容の要旨

論文題目　区間打ち切りを含む多イベント生存時間データに対するセミパラメトリックモデル

指導教官　大橋 靖雄 教授

東京大学大学院医学系研究科

平成 14 年 4 月進学

保健学博士後期課程

氏名　吉村 健一

I はじめに

多イベント生存時間データは、それぞれの観察対象が複数のイベントを経験する可能性がある状況や、対象が何らかの人工的或いは自然なクラスターを形成する状況で観察され、イベント間に相関を有することがその特徴である。Wei, Lin & Weissfeld による周辺モデル（WLW モデル）は Cox 比例ハザードモデルの多イベントの状況への 1 拡張であり、全対象の全イベントの生存時間が正確に測定されるか右側打ち切りを受けるかの何れかである状況のみに適用可能である。単イベントの状況と同様に、多イベントの状況においても生存時間データに区間打ち切りを含む可能性がある。区間打ち切りは興味のある確率変数の正確な値を知ることができず、その値が或る区間内に存在することのみが分かる状況で生じる。医学領域ではイベントの発生が定期的に行われる検査によって確認される状況が多い為、区間打ち切りが生じる状況は多い。区間打ち切りを含む場合には、一般的に対象が予め計画された検査機会を逃すこと等によってイベントの発生順位を一意に決められず、その場合には部分尤度に基づくセミパラメトリックな推測法を適用できない。この場面で実際に最も頻用されるアプローチは、最初にイベント発生を検出した日をイベント発生日として補完（time to first detection (TFD)補完）する方法である。補完を一旦行えば通常の統計解析法が全て適用可能となる一方で、正しい補完法を用いない限りは推定値にバイアスを含む。Satten（1996）は単イベントの状況に限り、潜在的な真の生存時間の順位に関する周辺尤度に基づくセミパラメトリックモデルによる推測法を区間打ち切りデータに対して提案した。このモデルの特徴は、観察された区間打ち切りデータに矛盾しない順位のサンプリングとそれに伴う確率的近似法により母数推定を行うこと、Cox 比例ハザードモデルと同様に基準ハザード関数の特定が必要無いこと、区間打ち切りを含まない状況では Cox のモデルに帰着することにある。この方法は計算機に強く依存した演算を要する反面、区間内からの生存時間のサンプリングを中心とした比較的単純なアルゴリズムに基づき、応用上の観

点からも魅力的な方法である。一般的な区間打ち切りを含む多イベント生存時間データに対して適用可能なセミパラメトリックモデルは未だ提案されておらず、Satten(1996)のアプローチを拡張する事は非常に魅力的である。

本論文では、順位に関する周辺尤度に基づく順位のサンプリングを通じて、多イベント生存時間データに対する周辺モデルを区間打ち切りを含む状況へと拡張したセミパラメトリックモデルを提案し、シミュレーションによりその性能を評価する。また、このモデルを循環器疫学コホート研究で実際に観察された糖尿病罹患と高血圧罹患に関する区間打ち切り生存時間データへ適用する。

II 方法

時間 t における対象 i の種類 j のイベントに特異的な周辺ハザード関数 $\lambda_{ij}(t)$ を以下で与える。

$$\lambda_{ij}(t) = \lambda_{j0}(t)\exp\left(\mathbf{x}_{ij}\boldsymbol{\beta}_j\right).$$

$\lambda_{j0}(t)$ は種類 j のイベントに関する時間 t における基準ハザード関数、$\mathbf{x}_{ij}$ は対象 i の種類 j のイベントに関する $1\times p_j$ 共変量ベクトル、$p_j \times 1$ 行列 $\boldsymbol{\beta}_j$ は種類 j のイベントに特異的な周辺対数比例ハザード母数ベクトルである。p_j は種類 j のイベントの共変量数を表す。このモデルは Cox の比例ハザードモデルと同様に $\lambda_{j0}(t)$ の特定は何ら必要とせず、更に WLW モデルと同様にイベント間相関構造の明示的な特定も必要としない。種類 j のイベントに関して、全対象から観察された区間打ち切りを含む生存時間に矛盾しない発生順序のベクトル R_j により構成される有限集合 $\mathcal{R}_j$ を考え、尤度 $\mathcal{L}_j$ をこの全要素に関して足し合わせた以下の式で与える。

$$\mathcal{L}_j \propto \sum_{R_j\in\mathcal{R}_j}\Pr\left(R_j \mid \boldsymbol{\beta}_j, \mathbf{x}_j\right).$$

$S_j(d)$ を種類 j のイベントの d 番目の発生時点におけるリスク集合、D_j は種類 j のイベントの総観察数とすると、$\Pr\left(R_j \mid \boldsymbol{\beta}_j, \mathbf{x}_j\right)$ は基準ハザードを局外母数とした以下の式で与えられる。

$$\Pr\left(R_j \mid \boldsymbol{\beta}_j, \mathbf{x}_j\right) = \sum_{d=1}^{D_j}\frac{\exp\left(\mathbf{x}_{dj}\boldsymbol{\beta}_j\right)}{\sum_{i\in S_j(d)}\exp\left(\mathbf{x}_{ij}\boldsymbol{\beta}_j\right)}.$$

周辺スコア関数と情報量行列をこの尤度から適切に導き、これらに基づく確率的近似法によって興味のある母数、周辺スコア関数、情報量行列の推定を行うと共に後述する多イベントの状況下での統計的推測を可能とするためにスコア残差の推定も同様に確率的近時法を用いて行う。この確率的近似法は、以下の条件付き確率に従った順位のサンプリングを、比例ハザード族に属する任意の作業分布を利用した

作業生存時間のサンプリングに基づいて行うことで実現される。

$$\frac{\Pr\left(R_j \mid \beta_j, \mathbf{x}_j\right)}{\sum_{R_j \in \mathfrak{R}_j} \Pr\left(R_j \mid \beta_j, \mathbf{x}_j\right)}$$

多イベントの状況での統計的推測はロバスト分散推定量に基づいて行う。これにより、提案したモデルを用いれば共変量の効果に関する柔軟な仮説に対する同時検定や推定が可能となり、研究者の興味に沿った柔軟な統計的推測が可能である。

Ⅲ シミュレーションによる評価

提案したモデルの性能を、TFD 補完を伴い WLW モデルを用いる方法を比較対照とし、バイアス及び信頼区間の被覆確率を基準としてシミュレーションにより評価した。ここでは片群 50 例の 2 群比較の設定の下、2 種類のイベントに関する検査が来院時に同時に行われる状況において、群間で来院スケジュールに対称性を有する場合と有さない場合の 2 通りの区間打ち切りの発生機序、帰無仮説及び対立仮説双方の下での値を想定した。その結果、提案したモデルはイベント間の相関の有無、及び区間打ち切りの発生機序と共変量が独立であるかに依らず、バイアスは僅かで、かつ信頼区間による被覆確率も名義上の値に近かった。その一方で、TFD 補完を伴う方法は、特に区間打ち切りの発生機序と共変量が独立でない状況や対立仮説の下でバイアスが大きく、ロバストな推定が行えなかった。

Ⅳ 循環器疫学コホートデータへの適用

提案したモデルを、循環器疫学コホート研究として有名な久山町研究で観察された糖尿病と高血圧の 2 種類のイベントに関する区間打ち切り生存時間データに実際に適用した。本論文では、2 型糖尿病は空腹時血糖 126mg/dl 以上であることと糖尿病治療を受けることの少なくともどちらか一方の条件を満たすこと、高血圧は拡張期血圧が 90mmHg 以上であること、収縮期血圧が 140mmHg 以上であること、高血圧治療を受けることの少なくともどれか 1 つの条件を満たすことを定義として用いた。何れのイベントも定期健診時の検査に依存し、区間打ち切りを受けた。対象は 1988 年に設定されたコホートの内、年齢が 40 歳以上であること、1988 年時点までに 2 型糖尿病と高血圧のどちらにも罹患していないことを全て満たす者とし、1990 年から 2000 年までの大規模及び小規模検診において測定されたものを追跡データとして用いた。1988 年時の久山町の総人口 7,612 の内、上の条件を満たす対象数は 1,467 人（男 566 人、平均年齢 55.8 歳）であった。層化調整因子には性別及び年齢階級（40 歳以上 50 歳未満／50 歳以上 60 歳未

満／60 歳以上)、共変量には 1988 年時点の肥満(BMI≧25kg/m^2)、高トリグリセライド血症(トリグリセライド≧150mg/dl)、現在の飲酒習慣及び喫煙習慣の有無を用いた。その結果、糖尿病罹患に対する肥満のハザード比は 2.08、95%信頼区間(1.24, 3.47)、現在の喫煙状況のハザード比は 1.73、95%信頼区間(0.95, 3.14)、高血圧罹患に対する肥満のハザード比は 1.76、95%信頼区間(1.43, 2.17)、高トリグリセライド血症のハザード比は 1.61、95%信頼区間(1.24, 2.09)であった。各共変量の 2 つのイベントに対する同時検定の結果、肥満では p 値<0.001、高トリグリセライド血症では p 値=0.002 となった。

V 考察

本研究で提案したモデルは、WLW モデルと同様に基準ハザード関数及び相関構造を特定する必要がない。基準ハザード関数を特定しないことは Cox 比例ハザードモデルと同様に適用の場面においては非常に有用な特徴となる。また、相関構造の特定を必要としない事から柔軟なイベント間相関構造に対応可能であると共に、提案したモデルは周辺ハザード関数が正しく特定される限り、母数の一致推定量と共にロバスト分散推定量による共分散行列の一致推定量が得られる。本論文のシミュレーションの結果から、提案したモデルを用いた場合には区間打ち切りを含む状況に対してもバイアスが僅かでロバストな推定が可能であり、仮説検定を行う際の実際の水準も名義水準に近いことが期待できる。共通性の仮定の下ではより高い精度を持つ共通母数を推定できるため、メタボリックシンドロームの様に複数のイベントの総体として定義されるものに関するリスク要因を探索する場面で有用な特徴となりうる。メタボリックシンドロームは予防医学的観点から今後も重要な標的であり、提案したモデルはこの分野での検討に大きな貢献を行うことが可能であると考える。

VI 結論

1. 順位に関する周辺尤度に基づく順位のサンプリングを通じて、多イベント生存時間データに対する周辺モデルを区間打ち切りを含む状況へと拡張したセミパラメトリックモデルを提案した。
2. シミュレーションによりその性能を評価した結果、区間打ち切りの発生機序と共変量が独立でない状況においても、ロバストな母数推定が可能であった。
3. 提案したモデルを循環器疫学コホート研究で実際に観察された糖尿病罹患と高血圧罹患に関する区間打ち切りを生存時間データに適用した結果、共変量の効果に関する柔軟な仮説に対する検定や推定が実際に可能であり、糖尿病罹患と肥満及び現在の喫煙習慣、高血圧罹患と肥満及び高トリグリセライド血症の間にそれぞれ関連を有する事が示唆された。