

連続変数で表される事件の接合関数を用いた 生存分析

福元 健太郎[†]

(受付 2019 年 6 月 3 日；改訂 12 月 11 日；採択 12 月 16 日)

要 旨

ある事件がいつ起きるか(時間変数)に影響する要因を調べるには、生存分析が使われることが多い。しかしそれだけでなく、どんな事件が起きるか(事件変数)にも関心がある場合がある。しかも往々にして、時間変数と事件変数は統計的に独立ではない。もし事件変数が名義変数であれば従属的競合リスク・モデルが使われる。本稿は、事件変数が連続変数である場合を扱う、接合関数を利用したモデルを提案する。そしてモンテ・カルロ・シミュレーションにより、時間変数と事件変数を別々に分析すると(打ち切りがある場合は特に)バイアスが生じるので、本稿の連続変数事件接合関数生存分析モデルを用いて両者を同時に分析する必要があることを示す。最後に、戦後英国の選挙の相対的タイミングと与党の得票率との関係を検討した先行研究のデータを再分析する。選挙が遅いと与党に有利ではあるが、選挙を早くしても必ずしも与党に不利になる訳ではない、という非対称な従属性が明らかになる。

キーワード：生存時間分析，生存解析，タイミング，従属性，選挙，イギリス。

1. はじめに

ある事件がいつ起きるか(時間変数)に影響する要因を調べるには、生存分析が使われることが多い。それだけでなく、どんな事件が起きるか(事件変数)にも関心がある場合がある。しかも往々にして、時間変数と事件変数は統計的に独立ではない。例えば、選挙がいつ行われるかと、選挙結果がどうなるかは、後述するように密接に関連する。しかし、本稿で明らかにするように、時間変数と事件変数を別々に分析するとバイアスが生じるので、両者を同時に分析する必要がある。もし事件変数が名義変数であれば(例えば選挙事由が、議会解散か、内閣不信任か、任期満了か)、従属的競合リスク・モデルが使われる(Fukumoto, 2009; Gordon, 2002)。しかし事件変数が連続変数や順序変数で表される場合の分析手法はこれまでなかった。Fukumoto (2015)で事件が順序変数である場合(例えば、与党過半数か、過半数を得た政党無しか、野党第 1 党過半数か)を扱っているので、本稿は事件が連続変数である場合(例えば与党の得票率)を扱うモデルとして、連続変数事件接合関数生存分析(Copula-based Continuous Event History Analysis, 以下 CCEHA)モデルを提案する¹⁾。

本稿は時間変数と事件変数の従属性をモデル化するにあたり、接合関数(後述)を利用する。接合関数の利点の 1 つは、時間変数や事件変数の周辺分布が何であってよいというモジュール性である。もう 1 つは、時間変数と事件変数の従属性は、既存の二変量分布を超えて、実

[†] 学習院大学 法学部：〒171-8588 東京都豊島区目白 1-5-1

多様なものを表現できる, という柔軟性である. しかも, 従属性は対称的でなくてもよい. 例えば, 本稿が明らかにするように, 選挙が遅いと与党に有利ではあるが, 選挙を早くしても与党に不利になる訳ではない, という非対称な従属性をも扱える.

本稿は次のように構成される. 先ず, CCEHA モデルを提示する. 次いで, モンテ・カルロ・シミュレーションを行う. 時間変数と事件変数に従属性がある時に(かつ場合によっては標本に打ち切りがある時に), 時間モデルと事件モデルを別々に推定すると, 事件モデルの推定にバイアスが生じるが, CCEHA モデルはそれに対処し得る. その後, Smith (2004) が分析した戦後英国の選挙データを CCEHA モデルで再分析する. CCEHA モデルは, 元の分析よりも多くの情報を, 同じデータから引き出す. 最後にモデルの長所と今後の課題を展望する.

2. モデル

2.1 設定

2.1.1 記法

ある観測単位(例えば議会)が時点 $t_0 = 0$ から始まって時点 $T > 0$ まで存続し, そこで事件 E (例えば選挙における与党得票率)を起こして終わるとする. ここで T も E も確率変数かつ連続変数かつ潜在変数であり, それぞれの周辺分布の累積分布関数を

$$T \sim F_T(t | \theta_T), \quad E \sim F_E(e | \theta_E)$$

と表す. θ_T と θ_E はそれぞれの媒介変数ベクトルである. これらの分位変数は当然に標準一様分布 $\mathcal{U}$ に従う.

$$U_T \equiv F_T(T | \theta_T) \sim \mathcal{U}, \quad U_E \equiv F_E(E | \theta_E) \sim \mathcal{U}.$$

T と E の同時分布の累積分布関数は, これら分位変数の関数 $C(\cdot, \cdot)$ で表すことができる.

$$(T, E) \sim F_{TE}(t, e | \theta) \equiv C(u_T, u_E | \theta_C).$$

ここで θ_C は T と E の従属性を決める媒介変数であり, $\theta \equiv (\theta'_T, \theta'_E, \theta'_C)'$ とする. この関数 $C(\cdot, \cdot)$ を接合関数と呼ぶ²⁾. シュクラールの定理により, $C(\cdot, \cdot)$ は一意に定まる. ここで $F_T(\cdot)$ と $F_E(\cdot)$ は 1 回微分可能, $F_{TE}(\cdot, \cdot)$ は(よって $C(\cdot, \cdot)$ も) 2 回微分可能であると想定し, 確率密度関数を次のように表記する.

$$\begin{aligned} \frac{\partial F_T(t | \theta_T)}{\partial t} &\equiv f_T(t | \theta_T), & \frac{\partial F_E(e | \theta_E)}{\partial e} &\equiv f_E(e | \theta_E), \\ \frac{\partial^2 F_{TE}(t, e | \theta)}{\partial t \partial e} &\equiv f_{TE}(t, e | \theta), & \frac{\partial^2 C(u_T, u_E | \theta_C)}{\partial u_T \partial u_E} &\equiv c(u_T, u_E | \theta_C). \end{aligned}$$

2.1.2 接合関数

本稿は 5 種類の接合関数を検討する. 最も基本的なのは正規接合関数である.

$$C_G(u_T, u_E | \theta_{C(G)}) \equiv \Phi_2\{\Phi^{-1}(u_T), \Phi^{-1}(u_E) | \theta_{C(G)}\}.$$

ここで $\Phi(\cdot)$ と $\Phi_2(\cdot, \cdot)$ はそれぞれ 1 変量と 2 変量の標準正規分布の累積分布関数である. $-1 < \theta_{C(G)} < 1$ は 2 変量の相関係数である. 図 1 はいくつかの接合関数の密度を等高線及び濃淡で示したものである. いずれもケンドール相関係数が, 後出の実データの値(約 0.325)に等しくなっている. 図 1 の左上図が正規接合関数であり, 従属性はやや強い($\theta_{C(G)} \equiv 0.488$).

次はファーマー・グンベル・モルゲンシュテルン(FGM)接合関数である.

$$C_{FGM}(u_T, u_E | \theta_{C(FGM)}) \equiv u_T u_E \{1 + \theta_{C(FGM)}(1 - u_T)(1 - u_E)\}.$$

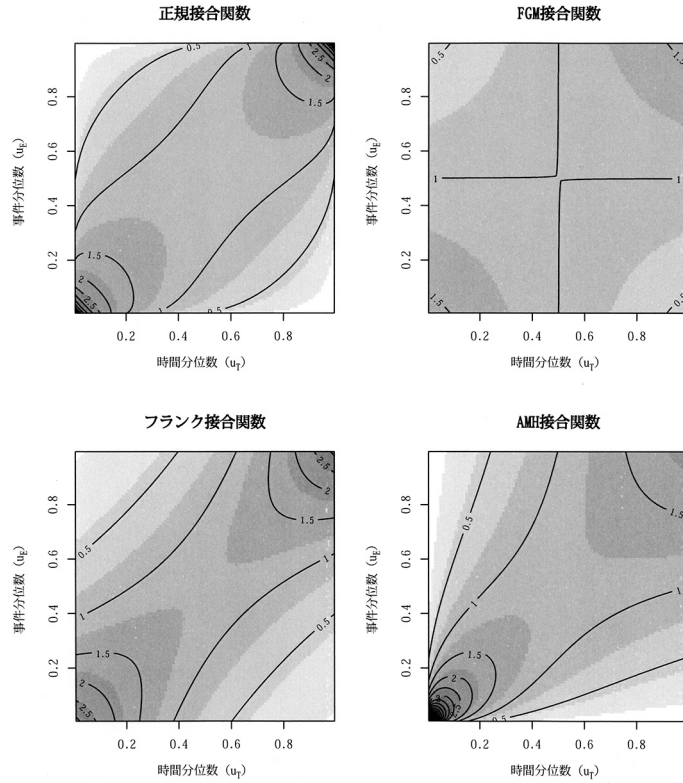


図 1. 接合関数の密度の例.

ここで $-1 < \theta_{C(FGM)} < 1$ である．従属性は弱い(図 1, 右上図, $\theta_{C(FGM)} \doteq 1.462$).

以下では, アルキメデス接合関数と呼ばれる接合関数の種類を紹介する．これはある生成関数 $\varphi(\cdot)$ に対して, 次のように定義される．

$$C_A\{u_T, u_E \mid \varphi(\cdot)\} \equiv \varphi^{-1}\{\varphi(u_T) + \varphi(u_E)\}.$$

その一例はフランク接合関数で, その生成関数は

$$\varphi_F(z \mid \theta_{C(F)}) \equiv -\log \left\{ \frac{\exp(-z\theta_{C(F)}) - 1}{\exp(-\theta_{C(F)}) - 1} \right\}$$

である($\theta_{C(F)}$ が取り得る値の範囲に制約はない)．従属性はやや弱い(図 1, 左下図, $\theta_{C(F)} \doteq 3.205$).

もう 1 つはクレイトン接合関数で, その生成関数は

$$\varphi_K(z \mid \theta_{C(K)}) \equiv (z^{-\theta_{C(K)}} - 1)\theta_{C(K)}^{-1}$$

である．ここで $\theta_{C(K)} > 0$ とする．この接合関数は順相関しか扱えず, 図 2 の左上図 ($\theta_{C(K)} \doteq 0.962$) の通り, 従属性が左裾(左上図の左下隅)で強く右裾(左上図の右上隅)で弱い場合に適切である³⁾．しかしクレイトン接合関数の関連接合関数を用いれば, その他の従属性も扱える．関連接合関数の 1 つは生存接合関数であり, 一般に $C(u_T, u_E)$ に対して

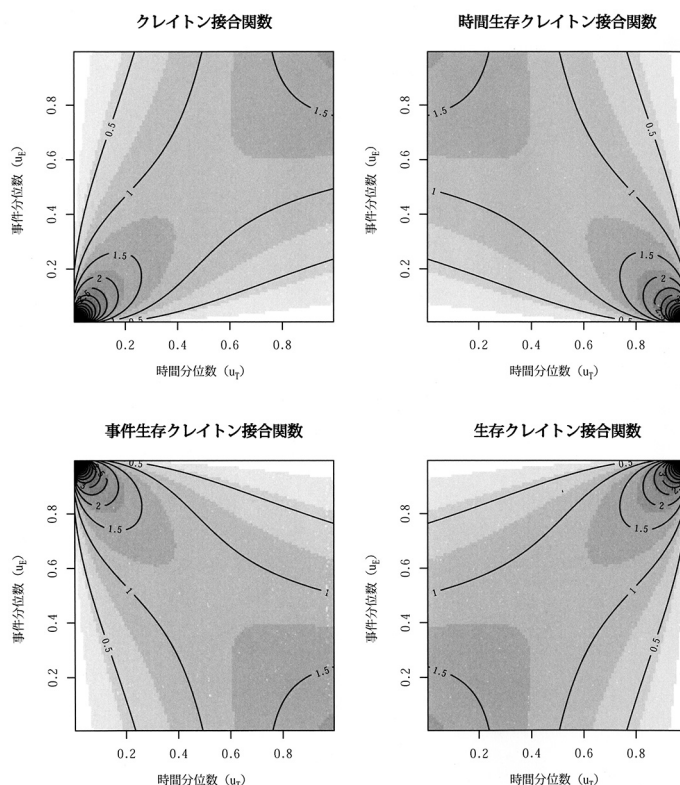


図 2. クレイトン接合関数とその関連接合関数の密度の例.

$\overline{C}^{(TE)}(u_T, u_E) \equiv u_T + u_E - 1 + C(1 - u_T, 1 - u_E)$ と定義される. クレイトン接合関数の生存接合関数の従属性は, 右裾で強く左裾で弱くなる (図 2, 右下図). 別の関連接合関数として $\overline{C}^{(T)}(u_T, u_E) \equiv u_E - C(1 - u_T, u_E)$ がある (本稿では時間生存接合関数と呼ぶ). クレイトン接合関数の場合 (図 2, 右上図), 従属性は事件変数の左裾と時間変数の右裾 (右上図の右下隅) で強く, 事件変数の右裾と時間変数の左裾 (右上図の左上隅) で弱いという, 逆相関となる. さらに別の関連接合関数として $\overline{C}^{(E)}(u_T, u_E) \equiv u_T - C(u_T, 1 - u_E)$ も考えられる (本稿では事件生存接合関数と呼ぶ). クレイトン接合関数の場合 (図 2, 左下図), 従属性は事件変数の右裾と時間変数の左裾 (左下図の左上隅) で強く, 事件変数の左裾と時間変数の右裾 (左下図の右下隅) で弱い⁴⁾.

最後にアリ・ミハイル・ハク (AMH) 接合関数は生成関数として

$$\varphi_{AMH}(z \mid \theta_{C(AMH)}) \equiv \log \left\{ \frac{1 - \theta_{C(AMH)}(1 - z)}{z} \right\}$$

を用いて導出される (図 1, 右下図, $\theta_{C(AMH)} \doteq 0.987$). ここで $-1 < \theta_{C(AMH)} < 1$ である. AMH 接合関数についても 3 つの関連接合関数を検討する.

2.2 推定

ここからは観測単位 i の添字を明記する. 打ち切り時点を $\bar{t}_i$ とする.

2.2.1 事件を観測する場合の尤度

もし $T_i = t_i \leq \bar{t}_i$ であれば、時点 $T_i = t_i$ において事件 $E_i = e_i$ を観測する．そうすると i の尤度は $T_i = t_i$ と $E_i = e_i$ で評価した同時分布密度に比例するから

$$\begin{aligned}\mathcal{L}_i(\boldsymbol{\theta} \mid t_i, e_i) &\propto p(T_i = t_i, E_i = e_i \mid \boldsymbol{\theta}) \\ &= f_{TE}(t_i, e_i \mid \boldsymbol{\theta}) \\ &= \left. \frac{\partial^2 F_{TE}(t, e \mid \boldsymbol{\theta})}{\partial t \partial e} \right|_{t=t_i, e=e_i} \\ &= \left. \frac{\partial u_T}{\partial t} \right|_{t=t_i} \left. \frac{\partial u_E}{\partial e} \right|_{e=e_i} \left. \frac{\partial^2 C(u_T, u_E \mid \theta_C)}{\partial u_T \partial u_E} \right|_{u_T=u_{T,i}, u_E=u_{E,i}} \\ &= f_T(t_i \mid \boldsymbol{\theta}_T) f_E(e_i \mid \boldsymbol{\theta}_E) c(u_{T,i}, u_{E,i} \mid \theta_C)\end{aligned}$$

が導かれる．これは、2つの周辺分布の密度と接合関数の密度の積であり、本モデルのモジュール性をよく表している．すなわち、時間の周辺分布 ($f_T(\cdot)$) 及び・又は事件の周辺分布 ($f_E(\cdot)$) 及び・又は両者の従属性 ($c(\cdot, \cdot)$) を、相互に無関係に変えることができる．例えば、 $f_T(\cdot)$ として対数正規分布、ワイブル分布のどちらを選んでもよいし、そのいずれの場合でも、 $f_E(\cdot)$ として正規分布、 t 分布のどちらを選んでもよいし、以上4通りのどの場合でも、 $c(\cdot, \cdot)$ として正規、FGM、フランク、クレイトン、AMHなどの接合関数から何を選んでもよい．ここに挙げたメニューだけでも、 $2 \times 2 \times 5 = 20$ 通りのモデルがある．

2.2.2 存続が打ち切られる場合の尤度

もし $T_i > \bar{t}_i$ であれば、打ち切り時点 $\bar{t}_i$ までに事件 E_i もそれが起きる時点 T_i も観測しない．これを存続が打ち切られると言う．便宜上、 $t_i = \bar{t}_i, e_i = \phi$ と表記しよう．この場合 i の尤度は打ち切り時点 $\bar{t}_i$ 以降のいずれかの時点 T_i で何らかの事件 E_i が起こる確率に比例する．

$$\begin{aligned}\mathcal{L}_i(\boldsymbol{\theta} \mid t_i, e_i) &\propto \Pr(T_i > \bar{t}_i \mid \boldsymbol{\theta}) \\ &= \int_{\bar{t}_i}^{\infty} \int_{-\infty}^{\infty} f_{TE}(t, e \mid \boldsymbol{\theta}) dt de \\ &= 1 - F_T(\bar{t}_i \mid \boldsymbol{\theta}_T).\end{aligned}$$

2.2.3 最尤法

全観測数を N とすれば、全尤度は

$$\mathcal{L}(\boldsymbol{\theta} \mid \mathbf{t}, \mathbf{e}) = \prod_{i=1}^N \mathcal{L}_i(\boldsymbol{\theta} \mid t_i, e_i)$$

である(ここで $\mathbf{t} \equiv (t_1, t_2, \dots, t_N)'$ 、 $\mathbf{e}$ も同様)．この値を最大化する $\boldsymbol{\theta}$ を推定値 $\hat{\boldsymbol{\theta}}$ とする．このように本稿は、時間モデルと事件モデルを同時に推定する CCEHA モデルを提唱する．

3. シミュレーション

時間変数と事件変数の間に従属性があると、時間モデルと事件モデルを別々に推定することがバイアスをもたらす．そのことを示すため、以下のようにモンテ・カルロ・シミュレーションを行う．

3.1 モデルの特定化

3.1.1 時間モデル

$\theta_T = (\alpha_T, \beta'_T)'$ とし, 説明変数ベクトル $\mathbf{x}_T$ を用いた線形予測値を $\mu_T = \mathbf{x}'_T \beta_T$ で表す. 2 つの分布を考える. 1 つは, 対数正規分布である.

$$F_{T(LN)}(t | \theta_T, \mathbf{x}_T) \equiv \Phi \left\{ \frac{\log(t) - \mu_T}{\exp(\alpha_T)} \right\}.$$

もう 1 つはワイブル分布である.

$$F_{T(W)}(t | \theta_T, \mathbf{x}_T) \equiv 1 - \exp\{-\exp(\mu_T)t^{\exp(\alpha_T)}\}.$$

$\mathbf{x}_T$ は定数項 $x_{T0} = 1$ と 2 つの共変量 x_{T1}, x_{T2} から成るとし, 媒介変数の値を次のように定める⁵⁾.

$$\mathbf{x}_T = (x_{T0}, x_{T1}, x_{T2})', \quad \beta_T = (\beta_{T0}, \beta_{T1}, \beta_{T2})' = (-1, 1, -1)', \quad \alpha_T = 0.$$

3.1.2 事件モデル

$\theta_E = (\alpha_E, \beta'_E)'$ とし, 説明変数ベクトル $\mathbf{x}_E$ を用いた線形予測値を $\mu_E = \mathbf{x}'_E \beta_E$ で表す. E は正規分布に従うとする.

$$F_{E(N)}(e | \theta_E, \mathbf{x}_E) \equiv \Phi \left\{ \frac{e - \mu_E}{\exp(\alpha_E)} \right\}.$$

$\mathbf{x}_E$ は定数項 $x_{E0} = 1$ と 2 つの共変量 x_{E1}, x_{E2} から成るとし, 媒介変数の値を次のように定める.

$$\mathbf{x}_E = (x_{E0}, x_{E1}, x_{E2})', \quad \beta_E = (\beta_{E0}, \beta_{E1}, \beta_{E2})' = (1, -1, 1)', \quad \alpha_E = 2.$$

3.1.3 従属性

時間変数と事件変数の従属性をモデル化するものとして 2 つの接合関数を検討する. 1 つは正規接合関数である. $\theta_{C(G)}$ の値として, 0.762(順相関), 0(無相関), -0.762(逆相関)の 3 つを考える. もう 1 つはクレイトン接合関数である. $\theta_{C(K)} = 1$ とする.

3.2 データ発生

標準一様分布 U から 3 つの乱数 $u_T, \bar{u}, \underline{u}$ を発生させる. 次に以下のような変数を作る.

$$t = F_T^{-1}(u_T), \quad \bar{t} = F_T^{-1}(\bar{u}^{\exp(\psi)}), \quad \underline{t} = F_T^{-1}(\underline{u}^{\exp(4)}).$$

ここで打ち切り調整媒介変数 ψ は 0(半分の観測が打ち切り)もしくは -10(ほとんど打ち切り無し)である. $\underline{t}$ は左側切断時点を表す. もし $t \leq \underline{t}$ であれば, この存続はそもそも観測されない. $t > \bar{t}$ の場合, この存続は打ち切られる. 以上に該当しない, つまり $\underline{t} < t \leq \bar{t}$ であれば, 次の段階に進んで, 事件変数 e を発生させる.

一般に t に条件づけられた E の累積分布関数 $F_{E|T}(e | t)$ は導出するのが往々にして難しいが,

$$F_{E|T}(e | t) = \frac{\partial C(u_T, u_E)}{\partial u_T} \equiv C_{E|T}(u_E | u_T)$$

が成り立つため, 接合関数の微分さえ導出できれば容易である. 正規接合関数の場合は

$$C_{E|T}(u_E | u_T) = \Phi \left\{ \frac{\Phi^{-1}(u_E) - \theta_{C(G)}\Phi^{-1}(u_T)}{\sqrt{1 - \theta_{C(G)}^2}} \right\}$$

であり⁶⁾、クレイトン接合関数の場合は

$$C_{E|T}(u_E | u_T) = (u_T^{-\theta_{C(K)}} + u_E^{-\theta_{C(K)}} - 1)^{-\frac{1}{\theta_{C(K)}} - 1} u_T^{-\theta_{C(K)} - 1}$$

である。以上を踏まえ、乱数 $u_{E|T}$ を標準一様分布 $\mathcal{U}$ から発生させて、まず $u_E = C_{E|T}^{-1}(u_{E|T} | u_T)$ を計算し、ついで $e = F_E^{-1}(u_E)$ を作る⁷⁾。

3.3 推定

以上のように CCEHA モデルに基づいて作成したデータを、本稿はまず CCEHA モデルで推定する⁸⁾。CCEHA モデルは真のデータ発生過程を想定しているから、当然に推定の性能は良い。関心があるのは、時間モデルと事件モデルを別々に推定するとどうなるかである。

まず時間モデルの媒介変数を最尤法で推定する(ここでは生存分析になる)。観測単位 i の事件が観測された場合の尤度は

$$\mathcal{L}_{T,i}(\theta_T | t_i, \mathbf{x}_{T,i}) \propto \frac{f_T(t_i | \theta_T, \mathbf{x}_{T,i})}{1 - F_T(\underline{t}_i | \theta_T, \mathbf{x}_{T,i})}$$

である。存続が打ち切られた場合の尤度は

$$\mathcal{L}_{T,i}(\theta_T | t_i, \mathbf{x}_{T,i}) \propto \frac{1 - F_t(t_i | \theta_T, \mathbf{x}_{T,i})}{1 - F_T(\underline{t}_i | \theta_T, \mathbf{x}_{T,i})}$$

である。推定値 $\hat{\theta}_T$ は $\prod_i \mathcal{L}_{T,i}(\theta_T | t_i, \mathbf{x}_{T,i})$ を最大化するものとして求められる。

次に、事件が観測された観測単位のみを用いて、以下の尤度を最大化するものとして、事件モデルの媒介変数を推定する(ここでは線形回帰分析になる。以下、単純事件モデルと呼ぶ)。

$$\prod_{i: e_i \neq \phi} \mathcal{L}_{E,i}(\theta_E | e_i, \mathbf{x}_{E,i}) \propto \prod_{i: e_i \neq \phi} f_E(e_i | \theta_E, \mathbf{x}_{E,i}).$$

さらに、事件モデルがなるべく CCEHA モデルと対等になるように、時間モデルの残差を説明変数に加えることも考える。まず、時間モデルの推定値に基づく時間変数の中央値を $\hat{t}_i$ と表記する。具体的には、時間変数が対数正規分布に従う場合、

$$\hat{t}_i \equiv \exp(\hat{\mu}_{T,i}),$$

ワイブル分布に従う場合、

$$\hat{t}_i \equiv \left\{ -\frac{\log(1 - 0.5)}{\exp(\hat{\mu}_{T,i})} \right\}^{\exp(-\hat{\alpha}_T)}$$

となる。ここで、対数化した時間変数の実現値から中央値を引いた差を、時間モデルの残差と定義する。

$$\hat{e}_{T,i} \equiv \log(t_i) - \log(\hat{t}_i).$$

これを共変量に加えたものを誤差修正事件モデルとする。

$$\mathbf{x}_{EC,i} = (x_{E0,i}, x_{E1,i}, x_{E2,i}, \hat{e}_{T,i})', \quad \boldsymbol{\beta}_{EC} = (\beta_{E0}, \beta_{E1}, \beta_{E2}, \beta_{TE})'.$$

以上のように時間モデルと事件モデルを別々に推定する手順を分離推定、これに対して CCEHA モデルによる推定を同時推定と呼ぼう。

3.4 シミュレーションの手順

観測数は $N = 1,000$ である。共変量 $x_{T1,i}, x_{T2,i}, x_{E1,i}, x_{E2,i}$ は、シミュレーション全体を通じて 1 回だけ、標準正規分布から乱数発生させた。時間変数の分布(対数正規かワイブル

か), 接合関数(正規かクレイトンか), 従属性媒介変数(正規接合関数の場合に $\theta_{C(G)}$ が正か負か 0 か, クレイトン接合関数の場合は常に $\theta_{C(K)} = 1$), 打ち切り調整媒介変数(打ち切り有り ($\psi = 0$) か無し ($\psi = -10$) か) を選ぶ.

1 回のルーティンで, t と e から成るデータセットを無作為に発生させる. 同時推定と分離推定を行い, 推定値を保存する. ヘシアン行列が特異で計算できない(すなわち推定値が収束しない)場合は, そのデータセットを廃棄する. このルーティンを 1,000 組の推定値が得られるまで繰り返す⁹⁾.

3.5 結果

時間モデルの係数 β_T は, 分離推定による推定値も不偏であり, かつ CCEHA による推定値と同じくらい効率的であったので, 以下では省略する. 事件モデルの係数 β_E については, 分離推定による推定値は, 時間変数と事件変数の従属性を考慮していないために, CCEHA による推定値とそのままでは比較できない. そこで本稿では, $x_{T1} = x_{T2} = x_{E1} = x_{E2} = 0$ の場合の, u_T の各値に対する E の中央値

$$\hat{e}(u_T, \hat{\theta}) \equiv F_{E|T}^{-1}(0.5 \mid u_T, \hat{\theta}_E, x_E, \hat{\theta}_C)$$

について, 両モデルを比べる. 計算方法は 3.2 節と同じ要領で $u_{E|T} = 0.5$ とすればよい.

図 3 は時間変数が対数正規分布に従い, 正規接合関数が用いられた場合である. 6 つの図から成る. 左列は打ち切りがない場合 ($\psi = -10$), 右列は半数が打ち切られた場合 ($\psi = 0$) である. 上段, 中段, 下段はそれぞれ接合関数が順相関 ($\theta_{C(G)} > 0$), 無相関 ($\theta_{C(G)} = 0$), 逆相関 ($\theta_{C(G)} < 0$) を表す場合である. 全ての図で, 横軸は t , 縦軸は $\hat{e}(u_T, \cdot)$ を表す¹⁰⁾. 太点線と太実線はそれぞれ CCEHA モデルと誤差修正事件モデルによる 1,000 組の $\hat{\theta}$ に基づく $\hat{e}(u_T, \hat{\theta})$ の平均値を表す. 斜線部と陰影部は同様に 1,000 組の $\hat{e}(u_T, \hat{\theta})$ の 2.5% 及び 97.5% 分位数に挟まれた区間(以下「95% 区間」と呼ぶ)である. 真の値 $\hat{e}(u_T, \theta)$ は, CCEHA モデルによる $\hat{e}(u_T, \hat{\theta})$ の平均値(太点線)に重なるため, 改めて図示はしていない. CCEHA モデルが真のデータ発生過程であることを考えれば, これは驚くに当たらない. 単純事件モデルは, 時間モデルの残差 $\hat{e}_T$ を考慮していないので, 1,000 組の $\hat{\theta}$ に基づく $\hat{\mu}_E$ の平均値は水平な細実線で表されている. 水平な細点線は, 事件変数の周辺分布の真の平均値 $\mu_E = \beta_{E0} = 1$ を表す. 垂直な細点線は t の真の中央値 $F_T^{-1}(0.5 \mid \theta_T, x_T)$ を表す.

まず, 時間変数と事件変数の間に順相関があり, 標本の一部が打ち切られている場合(右上図), 単純事件モデルによる推定値の平均値(水平な細実線)は(真の値を表す水平な細点線と比べて)下向バイアスがかかり, 符号が真実とは逆で負になる. その理由は以下の通りである. 順相関があれば, 時間変数が大きいほど, 事件変数は大きくなるが, 打ち切られることも多くなる. よって, 潜在的な事件変数は大きいほど観測されにくく, 観測された事件変数の周辺分布は下向きに歪む. 従って, 観測された事件変数の平均値は, 潜在的な(真の)事件変数の平均値より低くなる. 同様に, 逆相関でかつ打ち切りがあれば, 単純事件モデルによる推定値は上向バイアスがかかる(右下図). 相関がないか, 打ち切りがなければ, 単純事件モデルは不偏である. こうした事情はヘックマン選択モデルと似ている.

また CCEHA モデル(太点線・斜線部)と誤差修正事件モデル(太実線・陰影部)はほとんど同じ結果である(そのため太点線は太実線に覆われて見えない). 両者は今の設定では同じモデルだからである. 時間変数と事件変数の間に順相関がある場合(上段), 時間変数が大きいほど, 事件変数は大きくなる. 逆相関の場合(下段), 時間変数が大きいほど, 事件変数は小さくなる. 相関がないと(中段), 時間変数が長くても, 事件変数の平均値は変わらないが, 95% 区間は広くなる. 事件変数の 95% 区間は打ち切りがある方(右列)がない場合(左列)よりも広くなる.

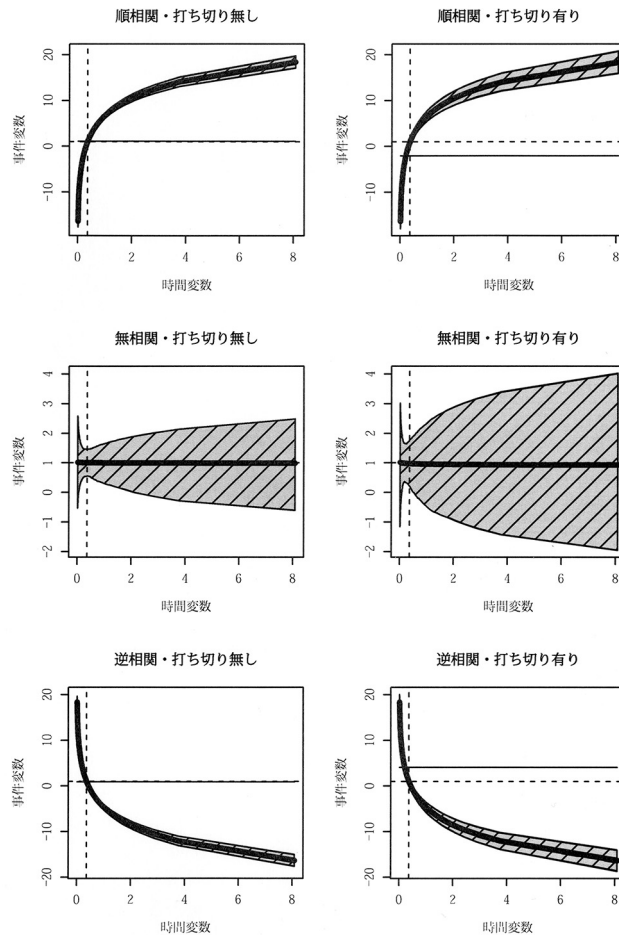


図 3. 正規接合関数と対数正規分布する時間変数を使った場合のモンテ・カルロ・シミュレーション：CCEHA モデル(太点線・斜線部)，誤差修正事件モデル(太実線・陰影部)，単純事件モデル(水平な細実線)，事件変数の周辺分布の中央値(水平な細点線)，時間変数の中央値(垂直な細点線)。

何故なら，事件変数が観測される標本の大きさは，前者の方が後者より小さいからである．興味深いことに，分離推定(陰影部)は同時推定(斜線部)と同じくらい効率的である．

もっとも，誤差修正事件モデルが CCEHA モデルと代替可能であるのは，時間変数が対数正規分布で正規接合関数が当てはまるという，これまで検討してきた極めて特殊な状況下に限られる．図 4 は時間変数の分布が対数正規からワイブルに変わった場合を示している(正規接合関数はそのまま)．すると，時間変数と事件変数の間に正(または負)の相関があれば(上段または下段)，打ち切りが有っても(左列)無くても(右列)，誤差修正事件モデルの推定値(太実線・陰影部)さえ(太点線・斜線部と比べて)下向(または上向)バイアスがかかる．もっとも，時間変数と事件変数が互いに独立であれば(図 4 の中段)，誤差修正事件モデルの推定値は不偏であり(太点線と太実線が重なる)，95% 区間(陰影部)は CCEHA モデルのそれ(斜線部)よりも狭い．なお，時間変数と事件変数の間に従属性があり，かつ打ち切られた標本があれば(図 4 の

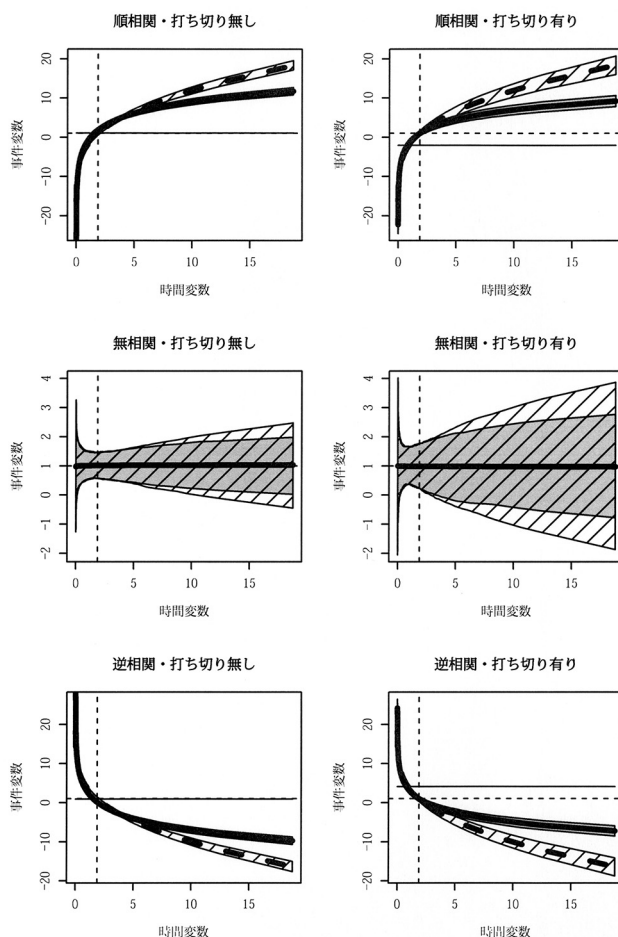


図 4. 正規接合関数とワイブル分布する時間変数を使った場合のモンテ・カルロ・シミュレーション：CCEHA モデル(太点線・斜線部)，誤差修正事件モデル(太実線・陰影部)，単純事件モデル(水平な細実線)，事件変数の周辺分布の中央値(水平な細点線)，時間変数の中央値(垂直な細点線)。

右上図と右下図)，単純事件モデルの推定値にバイアスがあるのは，図 3 の場合と同様である。

最後に，正規接合関数をクレイトン接合関数に変えたらどうだろうか．時間変数は再び対数正規分布するとしよう．図 5 に結果を示す(クレイトン接合関数の場合，時間変数と事件変数の間の相関は正しかない)．誤差修正事件モデルの推定値は先ほどとは逆に上向バイアスがかかる．これは次のような理由による．クレイトン接合関数の場合，時間変数と事件変数がともに小さい場合は順相関だが，両変数とも大きいとほとんど無相関になる(図 2 の左上図)．しかし誤差修正事件モデルは時間変数と事件変数がともに小さい場合の順相関を，両変数とも大きい場合にまで誤って外挿してしまう．そのため時間変数が大きい場合の上向バイアスを生んでしまう．打ち切られた標本があれば(右図)，単純事件モデルの推定値にバイアスがあるのは，図 3 の場合と同様である．

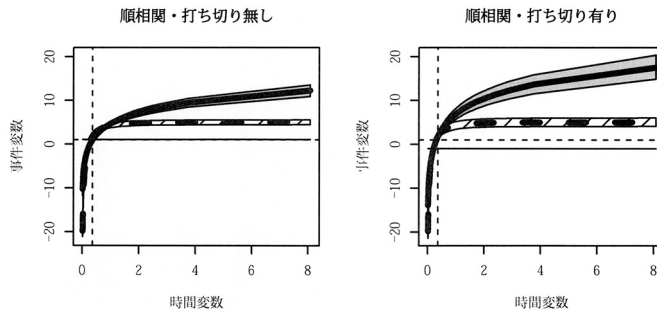


図 5. クレイトン接合関数と対数正規分布する時間変数を使った場合のモンテ・カルロ・シミュレーション：CCEHA モデル(太点線・斜線部)，誤差修正事件モデル(太実線・陰影部)，単純事件モデル(水平な細実線)，事件変数の周辺分布の中央値(水平な細点線)，時間変数の中央値(垂直な細点線)。

4. 実例

4.1 モデルの特定化

本節は CCEHA モデルを用いて Smith (2004) のデータを再分析する¹¹⁾。彼は、1945 年から 2001 年までの英国における選挙のタイミングと結果の関係を分析している。彼の議論によれば、与党はこれから経済状況が悪化することが予想される場合、その時になってから選挙をすると負けるので、先手を打って早めに議会を解散して選挙に持ち込む。しかし有権者もそれがわかるので、結局、予想よりも早く選挙が行われる場合、与党の得票率は落ちる、と主張している。

まず Smith (2004, p. 104, モデル 3.3) は、ワイブル分布を用いた時間モデルにより、選挙のタイミングを検討している。個々の議会は、直前の選挙から始まり、首相が選挙を行うことを発表するという「事件」により終わる¹²⁾。後述する日次時間変数変数を考慮するため、議会を日単位に分割したものを観測単位とする。直前の選挙からの日数が時間変数 (T) である。総観測数は 19,742 (日) であるが、議会の数は 16 であり、そのうち 2 つは打ち切られている¹³⁾。よって、議会開始日以外の $19,742 - 16 = 19,726$ 日は左側切断されており、事件が起きていない $19,742 - 14 = 19,728$ 日は右側打ち切りされている。説明変数 (x_T) は、残任期間(残任日数を 365 で除した商)、残任期間の二乗項、与野党支持率差(直近の月次世論調査における、与党へ投票する意向を持つ人の百分率から、主要野党へ投票する意向を持つ人の百分率を引いた差)、与野党議席差(与党議席数から主要野党議席数を引いた差)である (Smith, 2004, p. 96)¹⁴⁾。

次に Smith (2004, p. 141, モデル 4.5) は、正規分布を用いた(誤差修正ではなく単純)事件モデルにより、選挙の結果を分析している。事件変数 (E) は、与党支持率の変化(「選挙で実際に与党に投票した有権者の百分率から、選挙が発表される直前の世論調査で与党に投票する意向を表明した有権者の百分率を」引いた差 (Smith, 2004, p. 130)) である。選挙の相対的なタイミングを説明変数に含めるため、著者は累積危険率比と彼が呼ぶ値を計算している。これは「選挙が発表される前 30 日間の累積危険率を、それに先立つ 5 ヶ月間の累積危険率で除した比」である (Smith, 2004, p. 140)。すなわち、直前の選挙から s 日目の、先ほどの時間モデルから推定された危険率を $\hat{h}_s$ と記し、直前の選挙から t 日目の累積危険率比を $\sum_{s=t-30}^t \hat{h}_s / \sum_{s=t-183}^{t-31} \hat{h}_s$ と定義する。危険率が上がり始めて 1 ヶ月も経たない頃だと、分子は既に大きいけど分母はまだ小さいので、累積危険率比は大きな値となる。よって相対的に早い選挙の指標になると主張し

ている．この他の説明変数(x_E)は，時間モデルの説明変数(x_T)から残任期間の二乗項を除き，新首相(同じ議会で過去 100 日以内に首相が替わったか否かのダミー変数)，保守党ダミー変数を加えたものである(Smith, 2004, p. 96)．

CCEHA モデルを上述の Smith (2004) のデータに適用する．周辺分布は Smith (2004) の分析と同じで，時間変数はワイブル分布，事件変数は正規分布に従う．時間変数と事件変数の接合関数には生存 AMH 接合関数を用いる(この接合関数を選んだ理由は後述する)．説明変数は 2 組を検討する．1 組では，Smith (2004) の分析と同じ説明変数を用いる．もう 1 組では，累積危険率比を 2 つの理由から外す．1 つには，こうしたアドホックな尺度を拵えなくても，CCEHA モデルは相対的タイミングが事件変数に及ぼす影響を既に考慮に入れているからである．もう 1 つの理由は，累積危険率比が実のところ何を意味しているのか，よくわからないからである．例えば，もし累積危険率の「指数」の比であれば，生存率の比の逆数という意味を持つが，ここではそうっていない．何故過去 1 ヶ月とそれに先立つ 5 ヶ月間を選んだのかも恣意的である．そして何より，そもそも累積危険率比が相対的タイミングを測定しているのか疑わしい．確かに危険率が上がり始めた直後の累積危険率比は高い．しかし，危険率が上がってかなり経った後だけでなく，危険率が上がる前も，累積危険率比は低くなる(前者は分子も分母も大きくなるため，後者は分子も分母も小さくなるため)．すると累積危険率比が小さいからといって，相対的なタイミングが遅いとは限らず，むしろあまりにも早すぎることもあり得る．

4.2 結果

本稿は分離推定により元の分析結果をほぼ再現することに成功した(表 1 の左 2 列(生存分析)と表 2 の左 2 列(単純な線形回帰))¹⁵⁾．CCEHA は同時推定だが，結果は時間モデル(表 1)と事件モデル(表 2)に分けて報告している(いずれも中 2 列が累積危険率比を制御した場合，右 2 列が制御しなかった場合)．時間モデルは，分離推定と同時推定で係数の推定値はほとんど同じである¹⁶⁾．事件モデルは，分離推定によれば残任期間や与野党議席率差の係数が有意でないが，同時推定によると有意になる，といった違いが見られる．

本稿で重要なのは，時間変数と事件変数の従属性である．一方で分離推定によれば，累積危険率比の係数が有意に負なので，選挙の相対的なタイミングが早いほど与党の得票率が減る，あるいは逆にタイミングが遅いと与党の得票率が増える，という単調な従属性が示唆される．他方，同時推定により得られた従属性の媒介変数の推定値 $\hat{\theta}_C = 0.987$ に基づいて生存 AMH 接合関数の密度を図示したのが，図 6 の左図である¹⁷⁾．さらに図 6 の右図で，推定された分位数である $\hat{u}_T$ と $\hat{u}_E$ の散布図を示す．ここで $\hat{u}_T$ は選挙の相対的なタイミングを表す．右上隅に(重なって見えにくい)4 つの観測が集中していることは，左図の右上隅で密度が高くなって

表 1. 英国戦後総選挙のタイミングの分析．

モデル 累積危険率比	分離推定		同時推定 有り		同時推定 無し	
	点推定値	標準誤差	点推定値	標準誤差	点推定値	標準誤差
与野党支持率差	0.130	0.033	0.131	0.029	0.138	0.028
与野党議席率差	-0.017	0.006	-0.014	0.005	-0.016	0.005
残任期間	-4.647	1.027	-4.462	1.025	-4.826	1.005
残任期間 ²	1.215	0.401	1.203	0.413	1.417	0.368
定数項	-3.590	5.204	-4.476	4.276	-6.345	4.249
$\log(\alpha_T)$	1.748	0.541	1.793	0.465	1.998	0.371

表 2. 英国戦後総選挙の結果(与党支持率の変化)の分析.

モデル 累積危険率比	分離推定 有り		同時推定 有り		同時推定 無し	
	点推定値	標準誤差	点推定値	標準誤差	点推定値	標準誤差
残任期間	0.307	0.748	1.309	0.642	2.478	0.869
累積危険率比	-4.212	1.419	-4.232	1.033		
与野党支持率差	-0.250	0.048	-0.302	0.037	-0.340	0.057
与野党議席率差	0.006	0.011	0.021	0.009	0.031	0.013
保守党	2.117	1.374	1.541	0.994	2.456	1.586
新首相	-2.054	1.706	0.134	1.566	1.463	2.150
定数項	1.597	3.122	-1.435	2.532	-7.097	3.171
α_E	0.402	0.189	0.358	0.182	0.693	0.179

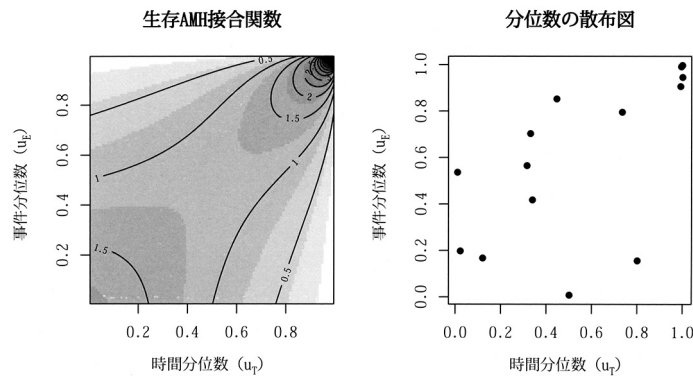


図 6. 英国戦後総選挙の時間変数と事件変数.

いることと呼応している. 逆に左下隅は, それほど密度が高くない(左図)観測も集中していない(右図). ここからわかることは, 選挙までの時間がかかる(時間変数が大きくなる)ほど与党に有利になる(事件変数が大きくなる)が, 逆に選挙を早めに行うからといって必ずしも与党に不利に働くわけではないという, 非対称な従属性が見られることである. 言い方を変えれば, 忍耐強い与党は報われると言えるが, せっかちな与党が罰されるとは限らない. こうしたニュアンスに富んだ選挙のタイミングと結果の関係は CCEHA モデルなしでは明らかにし得ないだろう.

最後に, 生存 AMH 接合関数を用いた理由を述べておこう. 表 3 は, 様々な接合関数を用いた CCEHA モデルの赤池情報量規準を比べたものである(左列が累積危険率比を制御した場合, 右列が制御しなかった場合). 最後の行は分離推定で時間モデルと単純事件モデルの赤池情報量規準値を合計した値である. 事件生存クレイトン接合関数を用いた場合の赤池情報量規準が最小であるが, 実は最適化が失敗している(推定が収束していない). よって本稿は 2 番目に赤池情報量規準が小さい生存 AMH 接合関数を選んだ. これは分離推定の赤池情報量規準の総和(最終行)よりも小さい. 従って, 赤池情報量規準の観点からも, CCEHA による同時推定は分離推定よりも望ましい.

表 3. 様々な接合関数を用いた場合の CCEHA の赤池情報量規準 (最終行のみ分離推定の合計値)。

接合関数		累積危険率比	
		有り	無し
正規		98.2	103.3
FGM		98.0	103.1
フランク		98.1	102.9
クレイトン	標準	99.9	104.7
	時間生存	99.9	104.7
	事件生存	85.3	88.9
	生存	98.4	101.2
AMH	標準	99.4	102.8
	時間生存	98.4	103.8
	事件生存	98.4	102.9
	生存	94.6	101.3
分離推定		97.9	102.7

5. おわりに

本稿で提示した，連続変数事件接合関数生存分析モデルを適用できる研究課題は枚挙に暇がない．疾病が発見されるのが遅いほど，疾病は進行しているだろうか？ストライキが長引くほど，高い賃上げが獲得できるだろうか？政党が野に下っている時間が長いほど，閣僚ポストが少なくとも入閣するだろうか？

これまで，2 変数の同時分布をモデル化する際は，2 変量正規分布や 2 変量ワイブル分布のように，限られた周辺分布に対して，特定の従属性しか扱うことができなかった．あるいは生存分析の文脈では，共通脆弱性などを導入すると，周辺分布が変わってしまうし，脆弱性の分布はもとの周辺分布と共役である必要があった．それに対して接合関数の利点の 1 つは，各変数の周辺分布とは無関係に，変数間の従属性を扱えるという点にある．つまり，各変数の周辺分布が，例えば指数関数族のような何らかの特定のクラスに属さなくてもよいし，互いに共役分布である必要もない (Fukumoto, 2020)．従って，周辺分布や同時分布を，統計的便宜ではなく，研究対象の実態に合わせて選ぶことができるのである．

接合関数のもう 1 つの利点は，非対称な従属性を表現できる，ということである．本稿でも示したように，一方の変数が増えると他方の変数も増えるからと言って，一方の変数が減れば他方の変数も減るとは限らない．そうした微妙な従属性をも，接合関数は扱える．

接合関数の適用事例は，海外に目を向ければ，政治学に限っても皆無ではないが (Braumoeller et al., 2018; Chiba et al., 2015a, 2015b)，まだ広く使われているとは言い難い．本稿が社会科学に対し，接合関数の利点を伝え，広めることに貢献できれば幸いである．

注.

- ¹⁾ Fukumoto (2020) は複数の存続の従属性を軸に生存分析全般を概観する中に本稿のモデルも位置づけている．
- ²⁾ 接合関数全般の説明は，例えば Nelsen (2006) や Trivedi and Zimmer (2007) を参照．
- ³⁾ 本稿で相関とは，スピアマン順位相関やケンドール順位相関を指すことにする．詳しくは例えば Trivedi and Zimmer (2007, pp. 16, 22–23) を参照．

- 4) なお最初に紹介した3つの接合関数($C_G, C_{FGM}, C_A\{\cdot, \cdot | \varphi_F(\cdot)\}$)は、点対称なので生存接合関数に一致する($\overline{C}^{(T^E)}(\cdot, \cdot | \theta_C) = C_G(\cdot, \cdot | \theta_C)$). その他の関連接合関数も媒介変数の正負を逆にすれば一致する($\overline{C}^{(T)}(\cdot, \cdot | \theta_C) = \overline{C}^{(E)}(\cdot, \cdot | \theta_C) = C_G(\cdot, \cdot | -\theta_C)$). 従って関連接合関数を別途検討する必要はない.
- 5) $\alpha_T = 0$ なので指数分布に帰着するが、後に実例でワイブル分布を用いるので、ここでワイブル分布を導入しておく.
- 6) なおここでは E の周辺分布は正規分布なので $\Phi^{-1}(u_E) = (e - \mu_E) / \exp(\alpha_E)$ である.
- 7) この乱数発生方法については Nelsen (2006, p. 41) を参照. なお Trivedi and Zimmer (2007, p. 113) は必ずしも正確ではないように見受けられる.
- 8) 左側切断があるので, $\mathcal{L}_i$ を $1 - F_T(\underline{t}_i | \theta_T, \mathbf{x}_{T,i})$ で除した商を, 観測単位 i の尤度とした.
- 9) 本稿における分析は R Core Team (2019, Version 3.5.3) を用いた. なお本稿の分析を再現するファイル一式を著者のサイト <https://www-cc.gakushuin.ac.jp/~e982440/research/replication.htm> で公開する.
- 10) t は媒介変数の真の値 θ_T を用いて $F_T^{-1}(u_T | \theta_T, \mathbf{x}_T)$ から計算している.
- 11) 元のデータと統計ソフト STATA のコードは, 2008 年 2 月 19 日に <http://www.nyu.edu/gsas/dept/politics/faculty/smith/electiontimingwebpage.html> からダウンロードした. これらを公開し使用を許諾している著者の Smith 氏に感謝する.
- 12) 厳密には, この後に議会が解散され, さらにその後に選挙が行われる. なお 5 年の任期が満了したことはない.
- 13) 1979 年に終わる議会は内閣不信任可決によるため (Smith, 2004, p. 93). 2001 年から始まる議会は観測期間終了のため.
- 14) 残任期間は任期の 5 年から生存時間 ($t/365$) を引いた差であるから, ワイブル分布を修正したベースライン・ハザードを想定していることになる. なお Smith (2004, p. 96) には与野党議席「率」差と書かれているが, データは与野党議席「数」差である.
- 15) 但し, 元の分析結果 (Smith, 2004, p. 104, モデル 3.3, p. 141, モデル 4.5) と比べて本稿の再分析の方が, 時間モデルでは定数項の点推定値がやや大きくなっており, 事件モデルでは標準誤差がおしなべて小さくなっている. また Smith (2004, p. 104, モデル 3.3) の p は本稿の α_T に等しいが, 表 1 では実際に推定している $\log(\alpha_T)$ を報告している.
- 16) 定数項の点推定値は異なるが, 全て標準誤差が大きく, いずれにしても大きな問題ではないだろう.
- 17) 累積危険率比を制御した場合である. 実際には $\log\{(1 + \theta_C)/(1 - \theta_C)\}$ を推定しており, その点推定値と標準誤差はそれぞれ 5.007 と 2.337 である. 累積危険率比を制御しない場合はそれぞれ 4.174 と 2.154 であるが, 含意はあまり変わらない.

謝 辞

本稿は, “What Happens Depends on When It Happens: Continuous or Ordered Event History Analysis,” paper prepared for the Annual Meeting of American Political Science Association, Toronto, September 3–6, 2009 の主として前半部分をもとにしている (後半部分は Fukumoto (2015) として公刊した). 草稿は, the Annual Summer Meetings of the Society for Political Methodology, University of Michigan, July 9–13, 2008 and Yale University, July 23–25, 2009, 日本行動計量学会 (埼玉大学, 2010 年 9 月 22–25 日), 応用統計学会 (統計数理研究所, 2010 年 5 月 20–21 日), 研究集会「時空間現象データに対する統計科学モデルの構築及び解析に関する組織的研究」 (沖縄県青年会館, 2008 年 11 月 27–28 日), 研究集会「官庁統計データの公開にお

ける諸問題の研究」(統計数理研究所, 2008 年 10 月 23–24 日)で報告された。また本研究は, JSPS 科研費 JP20330023, 安倍能成記念教育基金学術研究助成金を受けており, そのことに謝意を表する。さらにコメントをいただいた, 有賀賢一, 大森裕浩, Frederick Boehmke, Robert J. Franzese Jr., Adam Glynn, Jay Goodliffe, Jude Hays, Gary King, John Londregan の諸氏及び担当編集委員と匿名査読者に感謝する。

参 考 文 献

- Braumoeller, B. F., Marra, G., Radice, R. and Bradshaw, A. E. (2018). Flexible causal inference for political science, *Political Analysis*, **26**(1), 54–71.
- Chiba, D., Martin, L. W. and Stevenson, R. T. (2015a). A copula approach to the problem of selection bias in models of government survival, *Political Analysis*, **23**(1), 42–58.
- Chiba, D., Metternich, N. W. and Ward, M. D. (2015b). Every story has a beginning, middle, and an end (but not always in that order): Predicting duration dynamics in a unified framework, *Political Science Research and Methods*, **3**(3), 515–541.
- Fukumoto, K. (2009). Systematically dependent competing risks and strategic retirement, *American Journal of Political Science*, **53**(3), 740–759.
- Fukumoto, K. (2015). What happens depends on when it happens: Copula-based ordered event history analysis of civil war duration and outcome, *Journal of the American Statistical Association*, **110**(509), 83–92.
- Fukumoto, K. (2020). Duration analysis, *The SAGE Handbook of Research Methods in Political Science and International Relations* (eds. L. Curini and R. J. Franzese Jr.), 659–678, Sage, London.
- Gordon, S. (2002). Stochastic dependence in competing risks, *American Journal of Political Science*, **46**(1), 200–217.
- Nelsen, R. B. (2006). *An Introduction to Copulas*, 2nd ed., Springer, New York.
- R Core Team (2019). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna.
- Smith, A. (2004). *Election Timing*, Cambridge University Press, Cambridge, UK.
- Trivedi, P. K. and Zimmer, D. M. (2007). *Copula Modeling: An Introduction for Practitioners*, now Publishers, Boston.

Copula-based Continuous Event History Analysis

Kentaro Fukumoto

Department of Political Studies, Faculty of Law, Gakushuin University

Scholars are interested not just in when an event happens but also in what kind of event happens. Moreover, the latter can depend on the former. If an event variable is nominal, the dependent competing risks approach is available. The model of the present paper, copula-based continuous event history analysis (CCEHA), takes advantage of copulas to model dependence between time and continuous event variables. Monte Carlo simulation illustrates that separate estimation of time and event models creates bias in the estimate of the event model when time and the event are not independent of each other (and some durations are censored). By contrast, CCEHA does not have this problem. We reanalyzed a dataset on election timing and outcome in the postwar U.K. and found an asymmetric positive dependence: later elections helped the governing party, whereas earlier elections did not necessarily hurt it.