イベントヒストリー分析のためのデータ加工とモデル選択 *

麦山 亮太 †

2016 年 2 月 9 日

目次

* 本稿作成にあたり、永島圭一郎氏、西澤和也氏から助言を得た。記して感謝申し上げる。

† 東京大学大学院人文社会系研究科社会学専門分野修士課程

1 はじめに

1.1 イベントヒストリー分析（生存分析）とは？

ある時点において生じる何らかの質的な（その前後で大きく状態が隔たっているような）変化＝イベントが生起する原因を明らかにする手法（Allison 2014）。例）失業、転職、結婚、移住、犯罪、病気…etc

イベントヒストリー分析（生存分析）は、個体の状態の変化を問題とする時に有用な分析手法で、とくにカテゴリカル変数を多く扱う社会学にとっては重要な分析手法であるといえる。にもかかわらず、日本ではイベントヒストリー分析の普及が十分に進んでいないように思われる。これは、分析のためのデータ加工が非常に煩雑であり、かつそれを体系的に解説した（日本語の）テキストが見当たらないということが大きな理由であると考えている。

そこで今回、イベントヒストリー分析のために必要なデータの形式とその加工の方法、分析の際に考慮すべきことについて、具体的な手順（パーソン・ピリオドデータの作成とモデルの選択）を紹介したい。

1.2 あらかじめのおことわり

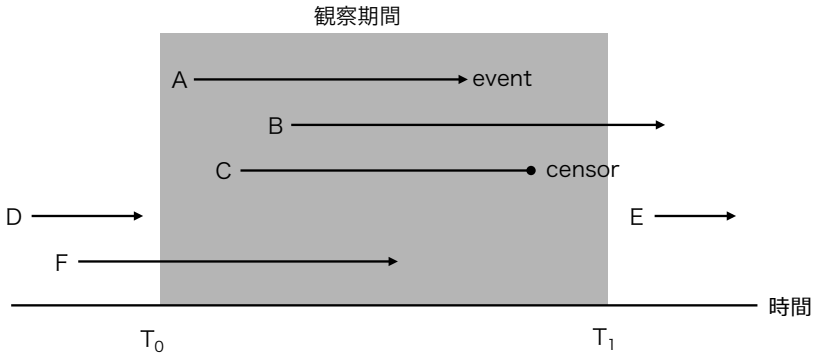
- 基本的に修士論文のための分析に際して身に付けた（楽屋裏的な）ノウハウを中心に構成しています。「イベントヒストリー分析はだいたいどんな分析手法かわかるけど使ったことはない…」くらいのレベルの方を想定しています。イベントヒストリー分析に関する基礎的な知識などは他の教科書など（Yamaguchi 1991; Singer and Willett 2003 = 2014; Andreß et al. 2013; Allison 2014）でフォローしていただければと思います。
- とくに、パーソン・ピリオド（人×時点の数だけ行が存在する）と呼ばれるデータを作成するための方法を紹介します。イベントヒストリー分析に際して使われる別のデータの形式として、episode-splitting（時変の独立変数が変化するごとに行が 1 つ増える）がありますが、個人的には、パーソン・ピリオドデータのほうがより柔軟にデータハンドリングできると思うので、おすすめです。episode-splitting に関心のある方は Blossfeld et al.（2007）などを参照ください。
- 競合リスクモデル（Competed risk model）や多状態モデル（Multi-state model）については扱いません。前者についてはここで紹介した二値の従属変数に関する分析の拡張で対応することができます。後者は Andersen and Keiding（2002）などを参照ください。
- 回顧式調査で収集されたデータを念頭において解説します。パネル調査データでも基本的には同様にできますが、左センサリングの問題があるため分析はやや難しくなると思います。
- シンタックスの紹介は Stata についてのみ行います（SPSS、R、SAS は分からないので…）。日本語で読める標準的な Stata の教科書としては、筒井ほか（2011）や石黒（2014）などがよく読まれていると思います。筒井ほか（2011）はイベントヒストリー分析についても解説しています。

1.3 基本的な用語

- **リスクセット（risk set）**：イベントが生起しうる個人の集合
- **打ち切り（censor）**：観察期間中にイベントが生起せず、観察が打ち切られる場合
- **リスク期間（risk period）**：リスク開始時点からイベントまたは打ち切りが起こるまでの長さ
- **連続時間（continuous time）**と**離散時間（discrete time）**
- **ハザード率（hazard rate）**：イベントの瞬間的な起こりやすさ

- **生存率（survival rate）**：ある時点までにイベントが生起していないケースの比率

図 1　リスクセットの概念



時間を t、イベントが生起する時間を T とすると、ハザード率は以下のように定義される。

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \quad \text{（連続時間）} \tag{1}$$

$$p(t) = \Pr(T = t | T \geq t) \quad \text{（離散時間）} \tag{2}$$

ハザード率は、連続時間の場合は「率 rate」を、離散時間の場合は「確率 probability」を意味する。イベントヒストリー分析において、目的変数となるのはこのハザード率である。

また生存率は以下のように定義される。

$$S(t) = \Pr(T \geq t) \qquad \text{（連続時間・離散時間）} \tag{3}$$

1.4　イベントヒストリー分析を行うときのおおよその手順

(1) 用いるデータを決める
(2) 分析の目的変数を決め、独立変数の検討をつける（時不変のみ or 時変を含む）
(3) パーソン・ピリオドデータを作成する
(4) 生存関数・ハザード関数を描く
(5) モデルを決定する
(6) 独立変数を投入したり除いたりして試行錯誤しながら分析する

2　データの形式と加工

2.1　分析の前に確認すべき事柄

- イベントの生起、イベント生起時点、リスク開始時点を特定できるだけの情報が揃っているか？
  たとえば初婚の分析をするためには、初婚年齢を知る必要がある。転職の分析をするためには、入職した時点、転職した時点を知る必要がある。また時変の独立変数を用いる場合は、それについても変化とそのタイミングに関する情報が必要となる。
- 目的変数は、一回きりのイベント（Nonrepeated event）か、繰り返しイベント（Repeated event）か？
  一回きりのイベントとしては、結婚、死亡など。繰り返しイベントとしては、転職など。後者のほうがより複雑なデータの加工の手間を必要とすることが多い。

- どのような時間の単位（年、月、週など）で変数が測定されているか？
  1 行あたりの時間の長さを揃える必要があるので、測定単位が重要となる。年については、暦年か年齢かという点についても確認が必要。例外がデータに残っている場合もあるので注意（例えば SSM 調査は「原則」1 年を最小の単位としているが、無業期間については 3 か月以上を目安に聴取している。また頻繁に転職している場合もデータに残っている。そのため同じ年に複数の職歴が現れることがある）。

2.2 基本：時変の変数を含まない一回きりのイベントの場合

分析のためには、たとえば、表 1[A] のようなデータから [B] のようなデータを作成する必要がある。実際は時変の変数を用いない場合はパーソン・ピリオドデータを作成する必要は必ずしもないが、後の導入のために役立つので説明する。

表 1　時変の変数を含まない一回きりのイベントの場合（架空例）

[A] パーソン・レベルデータ

id	sex	period	censor
1	1	7	0
2	2	4	1
3	1	3	0

sex: 1 が女性、2 が男性を表す
period: リスクセットに入ってからイベントが生起する、または打ち切られるまでの時間を表す変数
censor: 打ち切りであれば 1、打ち切りでなければ 0 をとる変数

[B] パーソン・ピリオドデータ

id	sex	period	censor	t	event
1	1	7	0	1	0
1	1	7	0	2	0
1	1	7	0	3	0
1	1	7	0	4	0
1	1	7	0	5	0
1	1	7	0	6	0
1	1	7	0	7	1
2	2	4	1	1	0
2	2	4	1	2	0
2	2	4	1	3	0
2	2	4	1	4	0
2	2	3	0	1	0
3	1	3	0	2	0
3	1	3	0	3	1

基本：パーソン・ピリオドデータの作成（Stata）

```
/*各行を period の値だけ複製する。period が 1 以下の値の場合には行は複製されずそのまま*/
expand period
/*t という変数を作成する。_n は 1 から順に 1 ずつ増加していく数列を作成する命令。by id は以下のコマンドを id ごとにそれぞれ行うよう宣言している*/
by id: generate t = _n
/*event という変数を作成。値はすべて 0 をとる。その後値を置き換える。period と t の値が等しく、かつ打ち切りケースでなければ、event の値を 1 にする*/
gen event = 0
replace event = 1 if period == t & censor == 0
```

2.3 応用 1：時変の変数を含む一回きりのイベントの場合

時変の変数を含む一回きりのイベントについて考える。初婚をイベントとする分析を行いたいとする。その際、時変の変数として仕事を用いる。

表 2　時変の変数を含む一回きりのイベントの場合（架空例）

[A] パーソン・レベルデータ

id	sex	job1	jobst1	job2	jobst2	job3	jobst3	job4	jobst4	marage	agenow
1	1	3	22	2	26	9	30	.	.	28	32
2	2	1	20	.	.	.	.	.	.	88	23
3	1	2	18	9	20	5	29	9	33	20	35

[B] パーソン・ピリオドデータ

id	sex	job	age	t	event
1	1	3	22	1	0
1	1	3	23	2	0
1	1	3	24	3	0
1	1	3	25	4	0
1	1	2	26	5	0
1	1	2	27	6	0
1	1	2	28	7	1
2	2	1	20	1	0
2	2	1	21	2	0
2	2	1	22	3	0
2	2	1	23	4	0
3	3	2	18	1	0
3	3	2	19	2	0
3	3	2	20	3	1

job*: *番目の仕事の種類を表す変数。9 は無業を表す。
jobst*: *番目の仕事を始めたときの年齢を表す変数。
marage: 結婚した年齢。88 は非該当。
agenow: 調査時点の年齢。
注）job は、開始年齢を常に優先してとっているが、例外もある。たとえば id3 の 3 行目に注意されたい。元のデータでは、20 歳の時に無業で、かつ結婚したことがわかる。こうした場合、どちらが先であるかをこちらで判断する必要が出てくる。日本ではかつて結婚退職の慣行が広くみられた（今もなおみられる）。これを考えると、無業になってから結婚した、というよりは、結婚して職場を離れたと考えるほうが妥当だろう。したがってここでは、20 歳時まで job1 を続けていたものとみなした。

応用 1：パーソン・ピリオドデータの作成（Stata）

```
/*リスク期間を表す変数を作成。ここでは働き始めてから結婚までの期間とする。結婚していない場合は現在の年齢とする。*/
gen period = marage - jobst1 + 1 if marage ~= 88
replace period = agenow - jobst1 + 1 if marage == 88
replace period = 50 - jobst1 if agenow > 50 & marage == 88 *50歳で打ち切り
/*行をperiodの数だけ複製し（period <= 1の場合は複製しない）、idで昇順に並び替え*/
expand period
sort id
/*リスク暴露期間および年齢を作成*/
by id: gen t = _n
by id: gen age = jobst1 + t - 1
/*年齢からイベント生起時点を特定する*/
gen event = 0
replace event = 1 if marage == age
/*時変の変数jobを作成。当該の仕事の開始年齢を利用し、どの範囲に入っているかどうかでその仕事についているかどうかを判断する*/
gen job = .
forvalues i = 1/4{
 replace job = job‘i’ if jobst‘i’ <= age
 /*ただし結婚時点で無業であった場合、1時点前の仕事を採用する*/
 replace job = job[_n-1] if job == 9 & event == 1
}
```

2.4 応用 2：時変の変数を含む繰り返しイベントの場合

時変の変数を含む繰り返しイベントの場合を考える。無業への移動をイベントとする分析を行いたいとする。無業への移動は当然、同一個人に複数回起こりうるイベントである。データ加工の際に注意すべき点は以下のとおり。

- 個人は、有業のときにのみリスクセットに入る。したがって、無業期間については、リスクセットから除外する必要がある。
- 同一個人が複数回リスクセットに入るため、これらをそれぞれ区別しつつも、同一個人のなかにネストしているという情報は残す。
- イベントが生起したときの年齢に注意。無業期間の開始年齢のときまでは、1 つ前の仕事を続けていたものと考えられる。

表 3　時変の変数を含む繰り返しイベントの場合（架空例）

[A] パーソン・レベルデータ

id	sex	job1	jobst1	job2	jobst2	job3	jobst3	job4	jobst4	marage	agenow
1	1	3	22	2	26	9	30	.	.	28	32
2	2	1	20	.	.	.	.	.	.	88	23
3	1	2	18	9	20	5	29	9	33	20	35

[B] パーソン・ピリオドデータ

id	episode	sex	job	age	marriage	t	event
1	1	1	3	22	0	1	0
1	1	1	3	23	0	2	0
1	1	1	3	24	0	3	0
1	1	1	3	25	0	4	0
1	1	1	2	26	0	5	0
1	1	1	2	27	0	6	0
1	1	1	2	28	1	7	0
1	1	1	2	29	1	8	0
1	1	1	2	30	1	9	1
2	2	2	1	20	0	1	0
2	2	2	1	21	0	2	0
2	2	2	1	22	0	3	0
2	2	2	1	23	0	4	0
3	3	1	2	18	0	1	0
3	3	1	2	19	0	2	0
3	3	1	2	20	1	3	1
3	4	1	5	29	1	1	0
3	4	1	5	30	1	2	0
3	4	1	5	31	1	3	0
3	4	1	5	32	1	4	0
3	4	1	5	33	1	5	1

注）job はつねに開始年齢を優先するが、無業への移動が生起した時点について注意が必要。たとえば id1 は 26 歳から job = 2 であり、30 歳までこれをつづけ、30 歳になったときに無業へと移動した、というふうに判断する。
episode は、リスクセットに入っている「→」の個数をカウントした変数というふうに考えれば良い。
リスク暴露期間 t は、episode ごとに作成する。

応用 2：パーソン・ピリオドデータの作成（Stata）

```
/*リスク期間を表す変数を作成。繰り返しイベントの場合は、観察を打ち切る時点までとる*/
gen period = agenow - jobst1 + 1
expand period
sort id
/*年齢を作成*/
by id: gen age = jobst1 + _n - 1
/*仕事を示す変数を作成*/
gen job = .
forvalues i = 1/4{
 replace job = job‘i’ if jobst‘i’ <= age
}
/*結婚を示す変数を作成*/
gen marriage = 0
replade marriage = 1 if marage <= age
/*無業期間はリスクセットに含まれないので、複数行ある無業期間は最初のものだけ残す*/
drop if job == 9 & job[_n-1] == 9
/*イベントを定義する*/
gen event = 0
replace event = 1 if job == 9
/*イベントが生起した行について、1 つ前の時点の job を採用する*/
by id: replace job = job[_n-1] if event == 1
/*リスクセットに入っている個数で番号を振った変数 episode を作成*/
gen flag = 0
replace flag = 1 if id[_n-1] ~= id[_n] | (event[_n-1] == 1 & id[_n-1] ~= .)
gen episode = 1
replace episode = episode[_n-1] + flag[_n] if id[_n-1] ~= .
/*episode に対してリスク暴露期間を割り当てる*/
sort episode
by episode: gen t = _n
/*作成したデータのチェック*/
browse id episode job age marriage t event
*****************************************
/*時変の変数はいろいろ柔軟に作成できる。たとえば、結婚 1 年前~結婚 1 年後を表す変数を作成する場合を考える*/
gen marimpact = 0
by episode: replace marimpact = 1 if marriage[_n] == 0 & marriage[_n+1] == 1
by episode: replace marimpact = 2 if marimpact[_n-1] == 1
by episode: replace marimpact = 3 if marimpact[_n-1] == 2
/*1:結婚 1 年前 2:結婚 3:結婚 1 年後、をそれぞれ表す。*/
```

3 生存関数・（累積）ハザード関数のプロット

生命表（Life table）：パーソン・ピリオドデータを用いて生命表を作成する際は、各エピソードの一番最後の行だけを使うようにオプションをつける必要がある。

ltable 関数を用いてパーソン・ピリオドデータから生命表を作成（Stata）

```
/*生命表を作成。，以下はオプションで、それぞれ以下のような意味。*/
*tvid(id)：id ごとに、もっとも t の大きな行だけを計算に用いることを宣言する
*noadjust：Stata はデフォルトで修正をかけるので、これをなくす
*by(sex)：sex ごとに別々の生命表を作成
*test：by() で分けたグループに関して、Log-rank 検定をする
*graph：グラフを表示する
/*そのほか、オプションで hazard を指定することで、ハザード率を求めることもできる。*/
ltable t event , tvid(episode) noadjust by(sex) test graph
```

生命表とそのプロットから基底ハザード関数がおおよそどのような形状をしているかを考えることで、よりよいモデルの選択につながる。

きちんとした生存関数を描くためには、st 系の関数を使う必要がある。こちらも、各エピソードの一番最後の行だけを使うようにオプションをつける必要がある。

sts graph 関数を用いて生存関数を作成（Stata）

```
/*各エピソードのうちもっとも t の大きい行を特定する変数を作成する。一回きりのイベントの場合は、
episode の部分を id に置き替え*/ \\
gen last = 0
by episode: replace last = 1 if episode ~= episode[_n+1]
/*データが生存時間データであることを宣言する。繰り返しイベントでない場合は、id() のオプション
は不要*/
stset t if last == 1, id(episode) failure(event == 1)
/*Kaplan-Meier 生存関数を描く*/
sts graph, by(sex)
/*Nelson-Aalen 累積ハザード関数を描く*/
sts graph, cumhaz by(sex)
/*各種検定を行う。一度に複数の検定を行うことはできない*/
sts test sex, logrank
sts test sex, wilcoxon
sts test sex, cox
stset, clear *stset を解除する
```

4 モデルの選択

4.1 離散時間モデル

ロジットモデル（Logit model）：リンク関数としてロジットリンクを指定。

$$p(t) = \frac{\exp\left(\lambda(t) + \sum_{k=0}^{K} \beta_k X_k\right)}{1 + \exp\left(\lambda(t) + \sum_{k=0}^{K} \beta_k X_k\right)} \tag{4}$$

$\lambda(t)$ は基底ハザード（baseline hazard）関数を意味する。基底ハザード関数は自由に関数形を設定できる。
(4) を変形することで以下を得る。

$$\log \frac{p(t)}{1 - p(t)} = \lambda(t) + \sum_{k=0}^{K} \beta_k X_k \tag{5}$$

(5) の左辺の分母は生存率に相当する。ロジットモデルを用いるのが標準的だが、当然プロビットモデルでも同様に推定できる。

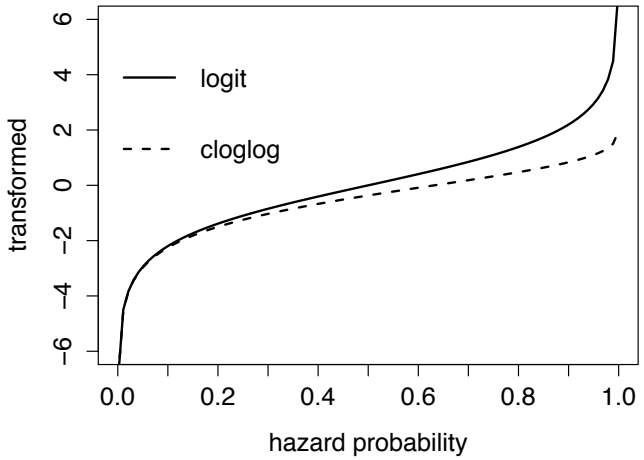
補対数対数モデル（Complementary log-log model）：リンク関数として補対数対数リンクを指定[*1]。

$$p(t) = 1 - \exp\left[-\exp(\lambda(t) + \sum_{k=0}^{K} \beta_k X_k)\right] \tag{6}$$

基底ハザード関数 $\lambda(t)$ は自由に設定して良い。(6) を変形することで以下を得る。

$$\log[-\log(1 - p(t))] = \lambda(t) + \sum_{k=0}^{K} \beta_k X_k \tag{7}$$

図 2　logit リンクと cloglog リンクの比較



ロジットモデルと補対数対数モデルの違い：比例オッズの仮定 vs 比例ハザードの仮定
(4) の場合、$\exp(\beta_k)$ は独立変数 1 単位の変化に対するハザード確率の「オッズ」の変化量（オッズ比が何倍に

[*1] 補対数対数モデルを用いた分析の実例として、Jacob and Kleinert(2014)。この論文では後述するランダム効果を組み込んだモデルを用いている。これ以降、各モデルを用いた論文を紹介していくが、筆者の読んでいる論文は社会階層・労働市場研究に偏っているので注意。

なるか）を表すのに対して、(6) の場合、$\exp(\beta_k)$ は独立変数 1 単位の変化に対するハザード確率の変化量（ハザード比が何倍になるか）を表すという点で、解釈がより直接的である[*2]。その意味で、補対数対数モデルは、後に述べる連続時間モデルとより近いモデルといえる。他方で、補対数対数モデルの場合、リンク関数は非対称であるので、係数の正負を逆転して解釈することはできない（佐々木 2009）。

図 2 に示したように、ハザード確率がかなり高い（0.2∼0.25 くらいを超える）現象を扱う場合、ロジットモデルと補対数対数モデルのズレは大きくなってくる。こうした場合、ロジットモデルを用いるほうが安全と思われる[*3]。

離散時間モデルの適用（Stata）

```
/*表 3[B] のデータに対してイベントヒストリー分析を行うものとする*/
/*ロジットモデルの適用。基底ハザード関数を特定化（ここでは時点ダミーとする）し、個人をクラスタとするロバスト標準誤差を算出*/
logit event marriage i.t , vce(cluster id)
glm event marriage i.t , family(binomial) link(logit) vce(cluster id)
/*補対数対数モデルの適用*/
cloglog event marriage i.t, vce(cluster id)
glm event marriage i.t, fam(binomial) link(cloglog) vce(cluster id)
estat sum *走らせたモデルで用いた変数について記述統計量を出力
```

4.2 連続時間モデル

連続時間モデルの基本形は以下のように表すことができる。

$$h(t) = \lambda(t) \exp\left(\sum_{k=0}^{K} \beta_k X_k\right) \tag{8}$$

基底ハザード関数 $\lambda(t)$ の形状にどのような分布関数を仮定するかによって、さまざまなモデルを組むことができる。これをパラメトリック・モデルという。色々な種類があるが、ここでは以下の 4 つのみ取り上げる[*4]。

指数モデル（exponential model）

$$\lambda(t) = b \tag{9}$$

ゴンペルツ・モデル（Gompertz model）

$$\lambda(t) = b \exp(ct) \tag{10}$$

ワイブル・モデル（Weibull model）

$$\lambda(t) = ba^b t^{b-1} \tag{11}$$

対数ロジスティックモデル（Log-logistic model）

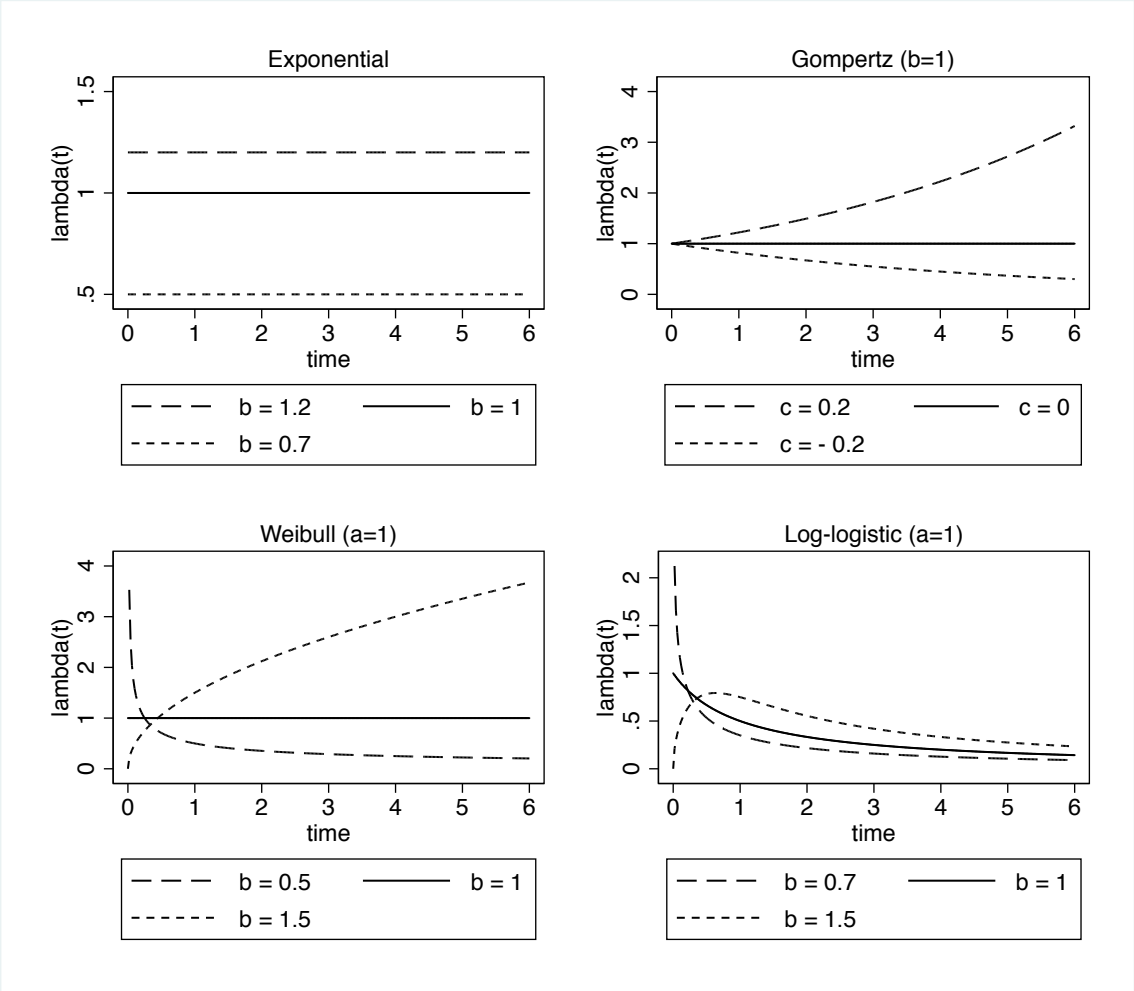
$$\lambda(t) = (ba^b t^{b-1})/(a - (at)^b) \tag{12}$$

[*2] ただしいずれのモデルに関しても、独立変数の限界効果（partial effect）は他の独立変数の値によって変化するので、OLS 回帰分析と比べると解釈はやや複雑になる。せっかく Stata を使っているのであれば、margins コマンドなどを用いて具体的に独立変数の値によってどの程度ハザード確率が変化するかどうかを確かめることを推奨する。margins コマンドの使い方については筒井（2011）などが参考になる。

[*3] ただし社会学ではこうした現象は必ずしも多くないように思うので、基本的にはどちらを選んでもさほど問題はないだろう。参考までに、修士論文で行った無業への移動に関する分析では、女性についてのハザード確率は最大で 0.11 程度であった。

[*4] ゴンペルツモデル、ワイブルモデル、対数ロジスティックモデルを用いた分析の例として、それぞれ順に Hachen (1992)、Petersen et al.（2000）、Breen（1992）。

図 3　パラメトリック・モデルの基底ハザード関数の形状



ただし実際もっとも使用頻度の高いのは、基底ハザード関数に対して上記のような確率分布を仮定せず、適当な区間に区切って基底ハザードを近似する以下のモデルである。

区分別定率モデル（Piecewise-constant model）：適当な時間で区切った時間のダミー変数（ステップ関数）を投入することで基底ハザードの形状を表現する[*5]。ハザード率を見て場当たり的に基底ハザード関数の形状を決める方法といえる[*6]。

上記のように基底ハザード関数の形状を特定化するモデルとは別に、基底ハザード関数の形状を特定化しないで推定することも可能。これをセミパラメトリック・モデルといい、以下が代表的。

Cox 比例ハザードモデル（Cox proportional hazard model）：Cox 回帰モデルなどとも言われる[*7]。部分最尤法を用いて、(8) の基底ハザード関数の形状を特定化することなく、共変量の情報だけを用いて係数を

[*5] 区分別定率モデルを使った分析例として、Gash（2008）。

[*6] 時間に関するダミー変数を投入しない場合を指数モデル（exponential model ＝基底ハザードが時間によらず一定とするモデル）といい、連続時間のイベントヒストリー分析のなかではもっとも基本的なモデルとして取り上げられる。

[*7] Cox 比例ハザードモデルを用いた分析例として Ishida et al.（2002）。

推定する。ハザード関数の形状が未知であっても共変量の係数を求めることができるが、比例ハザード性の仮定に違反する場合、推定値にバイアスが生じるのに対して有効な対策が取れない。

比例ハザード性の仮定：ハザード比が時間によらず一定であるとする仮定。

$$\frac{h_2(t)}{h_1(t)} = \exp\left[\sum_{k=1}^{K} \beta_k (X_{k1} - X_{k2})\right] \tag{13}$$

比例ハザード性の逸脱については、ハザード性への違反が考えられる独立変数と時間との交互作用項などを投入することで一定の解決を図ることができる。
比例ハザード性の仮定に反するかどうかを確認する方法（シェーンフェルド残差のプロット）もあるが、それよりも比例ハザード性の仮定が理論的に妥当かどうかが重要である。例えば、学歴が初婚イベントに与える影響について、リスク期間の初期では低学歴のほうが結婚ハザードが高いが、リスク期間の後期ではむしろ高学歴のほうが結婚ハザードが高いという関連があるとき、比例ハザード性を逸脱している。これは理論的にもありそうな話。もちろん、他の変数（職業や雇用形態に関する変数）を統制すればこの関係がなくなるのであれば、比例ハザード性の仮定には違反していない[*8]。

連続時間モデルの適用（Stata）

```
/*推定を行うにあたり、用いるデータが生存時間データであることを宣言する*/
stset t, failure(event == 1) id(episode)
/*区分別定率モデルを推定。t に関する独立変数を投入しない場合は指数モデルに等しい*/
*dist：ハザード率の分布関数を宣言
*nohr：nohr をつけない場合は、exp(β) の値が出力される
streg i.marriage i.sex i.t, dist(exp) nohr vce(cluster id)
/*区分別定率モデルを推定するためのモジュール stpiece を使う場合*/
ssc install stpiece *stpiece モジュールをインストールする
stpiece, tp(0(1)7) nohr
/*そのほかのパラメトリックモデル*/
streg i.marriage i.sex, dist(gomperz) nohr
streg i.marriage i.sex, dist(weibull) nohr
streg i.marriage i.sex, dist(loglogistic) nohr
/*Cox 比例ハザードモデル*/
streg i.marriage i.sex, dist(exp) nohr vce(cluster id)
```

4.3 連続時間モデル vs 離散時間モデル

離散時間ロジットモデルでも、基底ハザード関数さまざまな関数型を仮定することで、連続時間のときに紹介したような多様なハザードを表現することが可能。もっとも大きな違いは、ハザードの性質が「率（rate）」であるか「確率（probability）」であるかという点にある。

連続時間モデルを用いることができるかどうかは、タイ（tie）（＝同じ時間にイベントが生起すること）がどれくらいあるかに依存している。連続時間の場合、タイが多いと推定値にバイアスが生じると言われている（Singer and Willet 2003 = 2014）。日本の社会学では離散時間ロジットモデルが主流であるが、それは年単

[*8] このように、独立変数について比例ハザード性が成り立たないことを（間接的に）示している典型例として佐々木（2012）が挙げられる。

位より細かい測定単位をとっている社会調査データが少ないことから生じているのではないかというのが個人的な見立て。

4.4 他の変数による基底ハザード関数の近似

リスク暴露期間 t を用いて基底ハザード関数を近似する必要は必ずしもなく、同じく時間を表現した、より理解しやすい変数で近似する方法もある。例えば初婚のイベントヒストリー分析を行う場合は、リスク暴露期間よりも年齢を用いるほうが分かりやすいだろう。

5 その他、あまり取り上げられないトピック

5.1 依存性の問題

(1) 状態依存性（State Dependence）：従属変数（ハザード率）が、それ以前の状態の影響を受けること。以下に 2 つの例を挙げる。

- 以前にイベントを経験したことが、ハザード率に影響をおよぼす[*9]（失業するとその後再び失業しやすくなる、一度犯罪に走ると再度犯罪に至りやすくなる、etc）
- イベントが生起しない期間が長くなることが、ハザード率に影響をおよぼす（企業に勤め続けると転職しにくくなる、交際期間が長くなると交際を解消しづらくなる、etc）

とくに後者を、**持続依存性（Duration dependence）**という[*10]。持続依存性は以下の 2 つを区別して捉えることが有効である（Yamaguchi 1987）。

- ある独立変数の状態が持続していることが、ハザード率に対して効果をもつ。→時変の独立変数を投入することで調整することが可能
- ある従属変数の状態が持続していること（そのもの）が、ハザード率に対して効果をもつ。→基底ハザード関数の形状によって調整することが可能

(2) 率依存性（Rate dependence）：従属変数（ハザード率）が、それ以前のハザード率の高さ（低さ）の影響を受けること。特に時変の変数を考慮する際に問題となる。例えば、結婚することが昇進に対しておよぼす影響について分析を行うとする（結婚プレミアム・ペナルティの議論）。ここで結婚の係数が正であったとしても、これを単純に結婚の効果と解釈することはできない。なぜなら、こうした関係が、昇進のハザードが高い（将来に見込みのある）人々が結婚に至りやすい、というセレクションによって生じているかもしれないからだ。このような場合、結婚確率と昇進のハザード双方に影響をおよぼすとみられる交絡変数を統制することが有効である（普通の回帰分析と同じ発想）。

月並みだが、イベントヒストリー分析を行う際に、イベントの生起に対してどのような理論の下でモデルを構築するかをよく考える必要があるだろう[*11]。

[*9] こうした状態依存性について、Heckman は見せかけの（Spurious）状態依存性と真の（True）状態依存性を区別することが有効であると述べる（Heckman 1980, 1981）。見せかけの状態依存性とは、状態そのものの間に相関関係がなくとも、当該の状態を経験する確率に影響をおよぼす変数の値が個人によって異なっている。言い換えれば、「個人の異質性」によって生じる擬似相関であるということを意味する。真の状態依存性は、このような見せかけの状態依存性を除いた上で残るものとして捉えられる。

[*10] 非正規雇用の持続依存性を検討した論文の例として Gebel（2009）。

[*11] 余談だが、社会科学において経路依存性（Path dependency）という語がよく使われるが、これはとくに統計的な意味を考慮せず、単に昔の状態が今まで続く、というような意味で使われがちである。過去の状態が現在まで続いている、という経験的な現象が観察された時、ここで挙げた状態依存性、持続依存性、率依存性のように、要因を分解して考えることには一定の意味があるだろう。

5.2 個人の異質性の問題

個人の異質性（Unobserved heterogeneity, Frailty）：ハザード率の違いが観察可能な変数だけでは捉えられないときには、推定にバイアスが生じる。

リスクセットのなかにハザード率の異なる下位集団が存在しているとき、時間が経過するにつれて、ハザード率の高い集団はリスクセットから抜けていく。こうしたセレクションによって、リスク期間の後ろの方では、ハザード率の低い集団がリスクセットに占める割合が増加していく。この結果、負の持続依存性を過大に評価し、正の持続依存性を過少に評価するバイアスが生じる。

以上の個人の異質性にともなう問題は、パネルデータ分析などと同様に、個人効果を考慮したモデルを組むことで対処が可能である[*12]。

表 4　イベントヒストリー分析における固定効果・ランダム効果モデルの使い分け

	時不変の独立変数の効果に関心が	
イベントの種類	ない	ある
1 回きりのイベント	ランダム効果モデル	ランダム効果モデル
繰り返しイベント	固定効果モデル	ランダム効果モデル

例えば、ランダム効果離散時間ロジットモデルは以下のように表せる。

$$\log \frac{p(t)}{1-p(t)} = \lambda(t) + \sum_{k=0}^{K} \beta_k X_k + u, \quad u \sim Normal(0, \sigma) \tag{14}$$

繰り返しイベントの場合は、個人ダミーを投入することで、固定効果モデルを推定することもできる。ただしこの場合、イベントを 1 回しか経験していない個人は分析から除外される。しかし 1 回きりのイベントの場合には固定効果モデルを用いることは難しく、未だ標準的な方法は確立されていない[*13]。

[*12] 個人効果を考慮することで、結果の解釈は異なるものになる。通常のイベントヒストリー分析の場合は、独立変数の係数は個人「間」の違いを表すのに対して、個人の異質性を考慮する場合には、個人「内」の違いに近い解釈をすることになる。この点はパネルデータ分析とよく似ている。

[*13] Allison らは 1 回きりのイベントで固定効果推定を行うことのできる方法（case-time control method）を開発したが（Allison and Christakis 2006; Allison 2009）、欠点も多く経験的な研究への適用例は少ない。

イベントヒストリー分析におけるランダム効果・固定効果モデル（Stata）

```
/*離散時間の場合*/
xtset id t
/*ランダム効果モデルを適用*/
xtlogit event marriage, re
xtcloglog event marriage, re
/*固定効果モデルを適用。xtcloglog で固定効果モデルを適用する関数は Stata には実装されていないので、個人ダミーを投入して推定する*/
xtlogit event marriage, fe
xtcloglog event marriage i.id
**************************
/*連続時間の場合*/
stset t, failure(event == 1) id(episode)
/*ランダム効果モデル（Shared-frailty model)*/
streg i.marriage i.sex, dist(exp) shared(id) nohr
/*固定効果モデル*/
streg i.marriage i.sex i.id, dist(exp) strata(id)
*stcox, stpiece などでも同様のオプションをつけることで推定できる。
```

Shared-frailty model は、個人の異質性はあれど、それらが同一の確率分布にしたがうハザード関数を共有していると仮定するが、別の確率分布にしたがう異質な集団からなると想定するモデル（Unshared-frailty model）を考えることもできる。いわゆる混合分布モデル（Finite mixture model）の枠組み。そしてモデルはさらに複雑に…

5.3 時変の独立変数の効果をモデリングする

縦断的な分析において時変の独立変数を用いるとき、それがどのような時間的スパンで目的変数に影響を及ぼすかを考える必要がある（Allison 1994）。とくに、カテゴリカルな独立変数の場合はより注意が必要である。時間が細かく測定されている場合には、ラグつき独立変数を投入するというのが 1 つの手である。

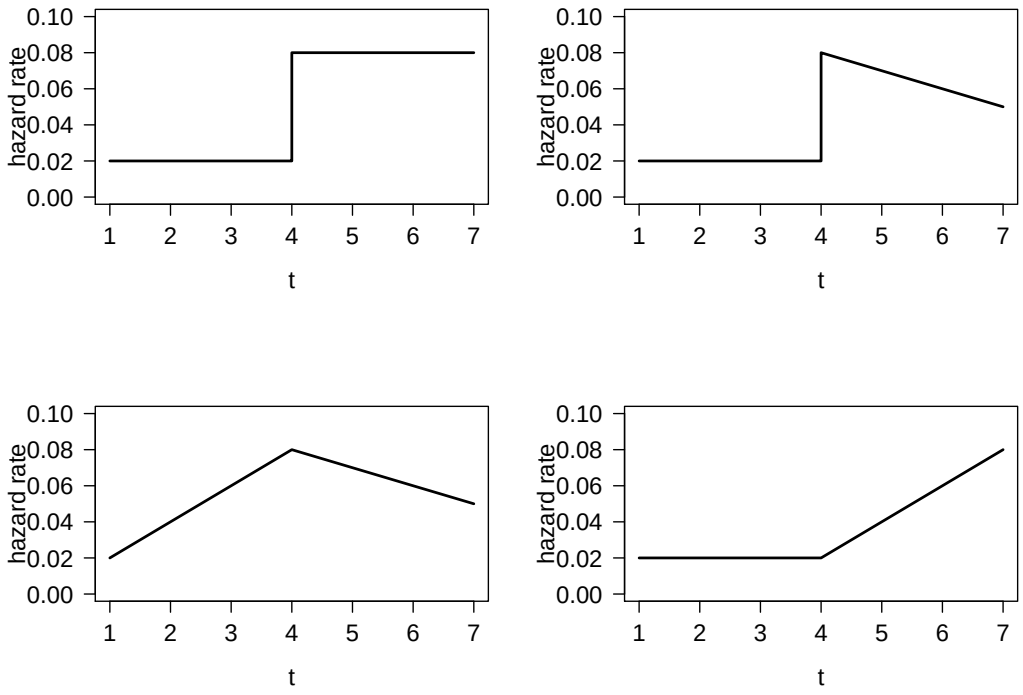
関連する議論として、Impact function の話をしておきたい。例えば、子どもをもつことが女性の離職に与える効果を明らかするイベントヒストリー分析を行う。以下のようなモデルを考える。時間が 1 年単位で離散的に測定されているものとする。$f(\cdot)$ はリンク関数を意味する。D_t $(t = -1, 0, 1, 2, 3)$ は、子どもをもつ 1 年前、子どもをもった年、その 1 年後、2 年後、3 年後を表すダミー変数である。

$$f[h(t)] = \alpha_{-1}D_{-1} + \alpha_0 D_0 + \alpha_1 D_1 + \alpha_2 D_2 + \alpha_3 D_3 + \sum_{k=0}^{K} \beta_k X_k \tag{15}$$

つまり、単純に子どもをもつかどうかがハザード率に対して影響をもつ、と考えるのではなく、子どもをもつ 1 年前（妊娠）～ 出産 ～ 育児期のそれぞれで異なる影響があり、3 歳を超えると影響を持たなくなる、といったようなモデルを考えることができる[*14]。そのほかにも、Impact function の部分は想定する理論に応じて適当な関数型を作成してよい[*15]。図 4 には、独立変数の変化が従属変数に与える効果に関する例を図示した。

[*14] このように直近の変化の影響に着目した分析例として、Kalmijn and Luijkx（2005）。
[*15] イベントヒストリー分析ではないが、Impact function を用いた分析例として Schmelzer（2012）。

図 4　Impact function の例（$t = 4$ のときに独立変数の変化が起こった場合）



6　まとめ

イベントヒストリー分析の手順は以下のように整理される（再掲）。

(1) 用いるデータを決める
(2) 分析の目的変数を決め、独立変数の検討をつける（時不変のみ or 時変を含む）
(3) パーソン・ピリオドデータを作成する
(4) 生存関数・ハザード関数を描く
(5) モデルを決定する
(6) 独立変数を投入したり除いたりして試行錯誤しながら分析する

イベントヒストリー分析はデータの加工にややクセがあるが、慣れてしまえばそこまで難しくはない。また、柔軟にモデルを組むことによって、興味深い知見を得ることも可能である。さらに、最近徐々に流行しつつあるパネルデータ分析との接続をするときにも、イベントヒストリー分析を通して得た考え方やデータの加工の方法は非常に役に立つ。本資料を利用することで、イベントヒストリー分析のハードルが少しでも低くなることを願っている（一緒にイベントヒストリー分析やりましょう！！）。

参考文献

Allison, Paul D., 1994, “Using Panel Data to Estimate the Effects of Events,” *Sociological Methods & Research,* 23(2): 174–199.

Allison, Paul D., 2009, *Fixed Effects Regression Models,* Thousand Oaks: Sage.

Allison, Paul D., 2014, *Event History and Survival Analysis (Second Edition),* Thousand Oaks: Sage.

Allison, Paul D. and Nicholas A. Christakis, 2006, "Fixed-Effects Methods for the Analysis of Non-repeated Events," *Sociological Methodology,* 36(1): 155–72.

Andersen, Per Kragh and Niels Keiding, 2002, "Multi-State Models for Event History Analysis," *Statistical Methods in Medical Research,* 11(2): 91–115.

Andreß, Hans-Jürgen, Katrin Golsch, and Alexander W. Schmidt, 2013, *Applied Panel Data Analysis for Economic and Social Surveys,* Springer.

Blossfeld, Hans-Peter, Katrin Golsch, and Gøtz Rohwar, 2007, *Event History Analysis with Stata,* New york: Psychology Press.

Breen, Richard, 1992, "Job Changing and Job Loss in the Irish Youth Labour-Market: A Test of a General Model," *European Sociological Review,* 8(2): 113–25.

Gash, Vanessa, 2008, "Bridge or Trap? Temporary Workers' Transitions to Unemployment and to the Standard Employment Contract," *European Sociological Review,* 24(5): 651–68.

Gebel, Michael, 2009, "Fixed-Term Contracts at Labour Market Entry in West Germany: Implications for Job Search and First Job Quality," *European Sociological Review,* 25(6): 661–75.

Hachen, David S., 1992, "Industrial Characteristics and Job Mobility Rates," *American Sociological Review,* 57(1): 39–55.

Heckman, James J. and George J. Borjas, 1980, "Does Unemployment Cause Future Unemployment? Definitions, Questions and Answers from a Continuous Time Model of Heterogeneity and State Dependence," *Economica,* 47:247 – 83.

Heckman, James J., 1981, "Heterogeneity and State Dependence," Sherwin Rosen eds., *Studies in Labor Markets,* University of Chicago Press; 91–140.

Ishida, Hiroshi, Kuo Hsien Su, and Seymour Spilerman, 2002, "Models of Career Advancement in Organizations," *European Sociological Review,* 18(2): 179–98.

石黒格，2014，『改訂 Stata による社会調査データの分析——入門から応用まで』北大路書房.

Jacob, Marita and Corinna Kleinert, 2014, "Marriage, Gender, and Class: The Effects of Partner Resources on Unemployment Exit in Germany," *Social Forces,* 92(3):839 – 71.

Kalmijn, Matthijs and Ruud Luijkx, 2005, "Has the Reciprocal Relationship between Employment and Marriage Changed for Men? An Analysis of the Life Histories of Men Born in The Netherlands between 1930 and 1970," *Population Studies,* 59(2): 211–31.

Petersen, Trond, Ishak Saporta, and Marc-David L. Seidel, 2000, "Offering a Job: Meritocracy and Social Networks," *American Journal of Sociology,* 106(3): 763–816.

佐々木尚之，2009，「JGSS 統計分析セミナー——イベントヒストリー分析の適用例」『JGSS で見た日本人の意識と行動：日本版 General Social Surveys 研究論文集』8: 91–105.

佐々木尚之，2012，「不確実な時代の結婚——JGSS ライフコース調査による潜在的稼得力の影響の検証」『家族社会学研究』24(2):152–64.

Schmelzer, Paul, 2012, "The Consequences of Job Mobility for Future Earnings in Early Working Life in Germany: Placing Indirect and Direct Job Mobility into Institutional Context," *European Sociological Review,* 28(1): 82–95.

Singer, Judith D. and John B. Willett, 2003, *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence,* Oxford University Press.（＝菅原ますみ監訳，2012，『縦断データの分析 I：変化についてのマルチレベルモデリング』朝倉書店・同，2014，『縦断データの分析 II：イベント生起のモデリング』朝倉書店.）

筒井淳也，2011，「Stata の予測値コマンド (3)」社会学者の研究メモ（2016 年 1 月 8 日最終閲覧，`http://d.hatena.ne.jp/jtsutsui/20110809/1341498218`）.

筒井淳也・平井裕之・水落正明・秋吉美都・坂元和靖・福田亘孝，2011，『Stata で計量経済学入門（第 2 版）』ミネルヴァ書房.

Yamaguchi, Kazuo, 1987, “Event-History Analysis : Its Contributions to Modeling and Causal Inference,” *Sociological Theory & Methods,* 2(1): 61–82.

Yamaguchi, Kazuo, 1991, *Event History Analysis,* Thousand Oaks: Sage.