第 3 章

離散時間ハザードモデル

3.1 イベントヒストリー分析の概要

パネルデータに対する主要な分析手法の 1 つとして、イベントヒストリー分析がある。イベントヒストリー分析とは、あるイベントの発生パターンとその要因に関する分析手法の総称である。別名，生存分析（survival analysis)、ハザード分析（hazard analysis)、期間分析（duration analysis)、failure-time analysis ともいわれる。

イベントヒストリー分析では、リスク人口（population at risk）におけるイベント発生確率である「ハザード率（hazard rate)」を分析の対象とする。リスク人口とは、イベントを経験する可能性がある人口を指す。例えば、離婚をイベントとして分析を行う場合、離婚のリスク人口は有配偶の男女であり、未婚者や死別者、離別者はリスク人口に含まれない。ハザード率は、より正確には「時間 t に至るまでの期間に、当該イベントが起こらなかったという条件のもとでの、時間 t におけるイベント発生の瞬間確率（instantaneous rate)」（津谷 2002, p. 429 右段, ll. 49-52）を指し、以下のように表わされる。

$$h(t) = \lim_{\Delta t \to 0} [P(t + \Delta t > T \geq t | T \geq t)/\Delta t] \quad \cdots (1)$$

ハザードとは、英語で「危険」を意味する言葉であるが、これはハザード率の概念が死亡を分析対象とすることの多い生物統計において発展したことに由来している。通常、リスク人口におけるイベント発生確率は、イベント発生のリスク開始時点からの「時間」によって異なる。また、イベント発生確率が時間の経過とともにどのようなパターンを示すのかも、対象となる集団・人口によって異なる。イベントヒストリー分析は、このハザード率を時間の関数として特定し、それが単数あるいは複数の説明要因によってどのように変化するのかを明らかにする多変量回帰分析である。モデルのパラメータは、最尤法（maximum likelihood method）もしくは部分尤度法（partial likelihood method）によって推定される。時間の関数として表わされるハザード率は、ベースライン・ハザード（baseline hazard）と呼ばれ、モデルの他の要因を統制した場合におけるイベント発生確率の基本的なパターンを表わす。

また、モデルにおける説明変数は共変量（covariate）と呼ばれる。共変量には時間によって値が変化する変数と、そうでないものがある。前者を「時間依存性共変量（time-varying covariate)」

といい、年齢や配偶関係、職業、あるいは学歴といった変数がこれにあたる。

一方、後者を「時間独立共変量（time constant covariate）」と呼ぶ。性別や生年月日、出身地などがこれにあたる。時間依存性共変量を用いることができるのは、時間の概念をもつイベントヒストリー分析ならではの利点である。

イベントヒストリー分析において重要な概念にセンサリングがある。観察対象となるイベントのリスク期間について、その終了時点が明らかではない場合をセンサリングという。このうち、観察期間中にイベントが生起しないケースを右センサリング（right-censoring）といい、観察期間前にイベントが生起しているケースを左センサリング（left-censoring）という[*1]（Guo 1993, Allison 1995）。左センサリングについては、イベントヒストリー分析をはじめ、他の統計分析においても対処することができない。しかし、右センサリングについては、イベントヒストリー分析では、イベントが生起しなかった時点までの情報を分析に反映して、リスク人口全体を対象とした分析を行うことができる。また、パラメーター推定についても、モデルにおいて右センサリングがイベントの生起ハザード率と独立に発生している（無相関である）と仮定できる場合、バイアスのない値を算出することができる（Allison 1995）。これを無作為センサリングの仮定（random censoring assumption）という。

3.2 離散時間モデルの概要

イベントヒストリー分析にはいくつかのモデルがある。本稿において解説するのは、イベントヒストリー分析のうち、時間の測定単位が連続的（際限なく細かい）とはみなせず、離散的（序数的）である場合に利用される分析手法である離散時間ロジットモデルならびに離散時間 complementary log-log モデル（以下、離散時間 CLL モデルと略す）（Allison 1982）である。

3.2.1 離散時間ロジットモデルの概要

離散時間ロジットモデルのモデル式は以下によって表される。

$$\ln[P_t/(1\text{-}P_t)]=a_t+b_1X_1+b_2X_2(t)+\ldots+b_kX_k(t) \quad \cdots (2)$$

P_t：ハザード確率、a_t：時間変数、b_k：共変量 X_k の回帰係数、X_k：共変量 k

(2) 式より分かるように、離散時間ロジットモデルは、各リスク時点でのハザード確率 P_t のロジット[*2]を被説明変数とする回帰モデルである。ここでいうハザード確率とは、時間 t までにイベントが発生していないという条件の下で、時間 t+1 までの期間にイベントが発生する確率を意味する。前述のハザード率は、時間の区切りが無視できるほど小さい場合に定義される確率密度

*1 社会科学において、左センサリングの定義は曖昧であり、リスク期間の開始時点が不明な場合を左センサリングという場合もある（Guo 1993, Allison 1995）。本稿では Guo（1993）に倣い、そのようなケースは左打ち切り（left-truncation）として左センサリングとは区別する。

*2 ロジットとはオッズを自然対数化した値をいう。オッズとは、イベントが生起しない確率（1-P）に対するイベント生起確率（P）の比ことを指し、P / (1-P) として表される。

(probability density) であり、ここでいうハザード確率とは厳密には異なるものであることに留意されたい。(2) 式はロジットモデル（ロジスティック回帰分析）と類似しており、回帰係数を指数化してハザード確率のオッズ比として解釈することができる。ただし、ロジットモデルでは確率 P を扱うのに対して、離散時間ロジットモデルでは、ハザード確率 P_t を用いる。また、離散時間ロジットモデルでは、定数 a や共変量 X がリスク期間中に変化することを許容している点も通常のロジットモデルとは異なる。回帰係数 b_k は、共変量 X_k がハザード確率のロジットに与える効果を意味している。ただし、離散時間ロジットモデルでは、回帰係数 b_k は共変量 X_k のリスク期間を通じた平均的な効果を表していることに留意する必要がある[*3]。また、時間変数 a_t は、ハザード確率のベースライン・対数オッズ (baseline log odds) である。ベースライン・対数オッズは、すべての共変量 X が 0 であった場合におけるハザード確率のロジットの時間推移を表しており、時間経過にともなうイベントの基本的な発生パターンを表す。

3.2.2 離散時間 complementary log-log モデルの概要

一方、離散時間 CLL モデルのモデル式は以下によって表される。

$$\ln[-\ln(1-P_t)]=a_t+b_1X_1+b_2X_2(t)+\ldots+b_kX_k(t) \quad \cdots (3)$$

P_t：ハザード確率、a_t：時間変数、b_k：共変量 X_k の回帰係数、X_k：共変量 k

(3) 式においては、左辺における Pt の扱いにおいて (2) 式とは異なる。これは離散時間ロジットモデルがハザード確率のオッズを従属変数とするモデルであるのに対し、離散時間 CLL モデルでは連続時間において仮定されるハザード確率そのものを従属変数とするモデルであるためである。これについて解説すると、連続時間を仮定するモデルでは、ハザード率 λ と累積生存確率 S(t) は以下の式によって表すことができる。

$$\lambda \quad = \quad -\frac{d}{dt}\ln[S(t)] \quad \cdots (4)$$

時点 t から t+1 までの 1 期間における累積生存確率 S(t) は $(1-P_t)$ で表されるため、(3) 式の左辺を指数化した $(-\ln(1-P_t))$ は (4) 式の右辺の近似となり、連続時間におけるハザード率を仮定した値となる。したがって、離散時間 CLL モデルでは回帰係数 b_k を指数化した値である $\exp(b_k)$ は共変量 X_k のハザード確率の比を表す (Allison 1982)。また、時間変数 a_t は、ベースライン・対数ハザード確率 (baseline log hazard probability) となる。

3.2.3 両モデルの違いについて

離散時間ハザードモデルとしては、CLL モデルではなくロジットモデルを用いたものが一般的である。どちらのモデルを用いても推定結果に質的な相違（回帰係数の統計的有意性や影響力の方

[*3] 共変量 X_k と時間変数 a_t の交互作用項をモデルに組み入れることで、係数 β_k がリスク期間を通じて変化することを許容するモデルを構築することが可能である（山口 2002c, 津谷 2002)。

向など）は生じない。しかし、前述のように回帰係数を指数化して得られる値である exp(b) の解釈について相違が生じる。離散時間ロジットモデルの exp(b) は基準カテゴリーに対するハザード確率のオッズ比を表すのに対し、離散時間 CLL モデルによるそれはハザード確率の比を表すという違いがある（Allison 1982）。オッズと確率は、確率が非常に小さい値である場合にはほぼ同じ値を示す。しかし、年を単位としたイベントのハザード確率は場合によっては非常に小さいとは言えず、ハザード確率とハザード・オッズとを同義的に解釈することができない。例えば、あるカテゴリーのハザード・オッズが基準カテゴリーの 3 倍という結果を得ても、それは必ずしもハザード確率が 3 倍であることを意味するのではない。一方、ハザード比は、あるカテゴリーのハザード確率が基準カテゴリーに比べて何倍高いのか（低いのか）、あるいは共変量の一単位の増加によって、ハザード確率が何倍高くなるのか（低くなるのか）を表しているため、ハザード確率の差異についてより直接的な解釈が可能である。近年、CLL モデルは、STATA や SAS などの汎用的な統計分析ソフトによって容易に使用できることから、ここではロジットモデルと CLL モデルの 2 つを用いて結果を比較する。

3.2.4 パネルデータにおける離散時間モデルの利用について

離散時間モデルは、パネルデータと最も親和性が高いイベントヒストリー分析であるといえる。なぜならば、通常個人を対象としたパネル調査では、調査が行われるのは年に 1 回であり、各年における結婚や出産、就業状態等の変化は、調査時点の状態の変化によって測定されることが多いためである。例えば結婚であれば、ある個人が結婚したか否かは、前年の調査で未婚であった人が当年の調査で有配偶であることによって把握されることが多い[*4]。そのため、結婚の生起は t-1 年から t 年の間に起きたことは明らかであっても、具体的にいつ、例えば何月に起きたのかまでは不明である場合が多々ある。このような場合には、イベントの生起時点に関する情報は年単位でしか把握することができず、連続時間を仮定することができない[*5]。したがって、イベント発生月が不明である場合には、ハザード率の近似として、リスク期間別のハザード確率を用いた離散時間モデルを利用することが最も簡便かつ直接的である。

また、離散時間モデルは、通常の統計ソフトに装備されているロジスティック回帰分析ならびに CLL 回帰分析のパッケージを利用できるため、適用が比較的容易であるといえる。むしろ、同モデルの適用において最も中心的な作業は、人-期間別データの作成である。以下では、Stata を用いた人-期間別データの作成方法について解説する。

---

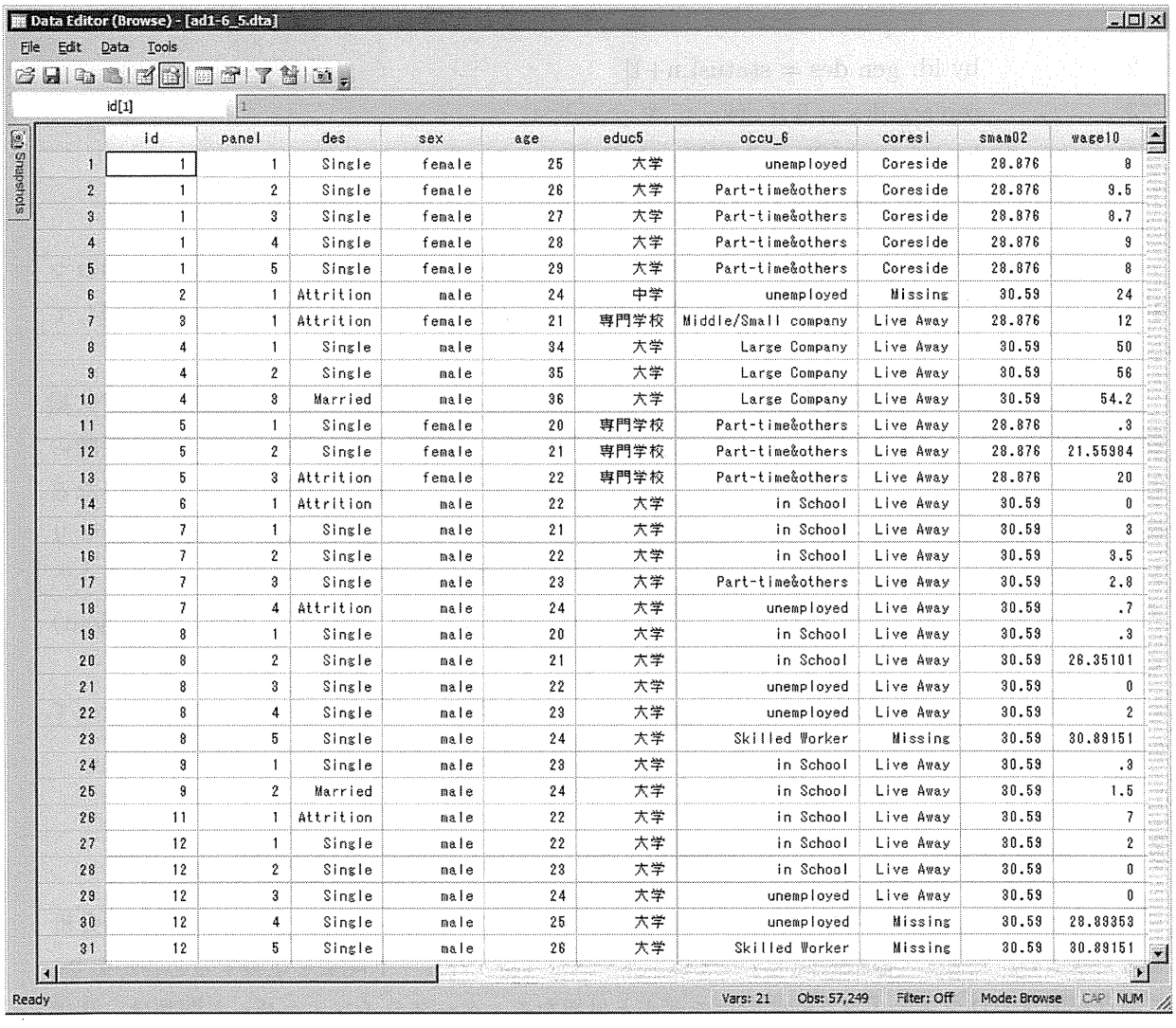
*4 なかにはイベント発生時点に関する質問を追加して、結婚や出産などのイベントについて、月単位でその生起時点を把握しているパネル調査もある。本稿で用いる「21 世紀成年者縦断調査」もそうした調査の 1 つである。イベント発生月に関する情報があるパネルデータでは、月を時間単位とした連続時間モデルの適用が可能である。ただし、その場合、連続時間モデル用にデータを再構築する必要が生じるなど、後に述べる離散時間モデル用のデータ作成に比べて作業が煩雑となる。一方で、時間依存性共変量も月単位で測定されている場合においては、共変量とイベントの生起順序を厳密に区別できるため、連続時間モデルの利用に利点がある。中間的な方法としては、離散時間モデルを利用しつつも、月単位の情報を用いて、時間依存性共変量とイベントの生起順序を区別し、各時点における共変量の値に反映させるという方法もある。

*5 特に、連続時間を仮定したイベントヒストリーモデルとしてよく用いられる Cox 回帰においては、同一時点において複数のイベント生起が観察されるような場合にはパラメーター推定にバイアスが生じることが知られており、その利用には注意が必要である（Allison 1995）。

3.3 人-期間別データの作成方法

離散時間モデルの適用においては、はじめに、リスク開始からイベントが発生するか、もしくはセンサリングとなった時点までの人-期間別データ（person-period data）を作成する。次に、この人-期間別データに対して、イベントが生起するか否かのダミー変数を従属変数とする通常のロジスティック回帰分析（ロジット分析ともいう）もしくは complementary log-log 回帰分析を行う。

図 1 人-期間別データの例



Data Editor (Browse) - [ad1-6_5.dta]

File Edit Data Tools

id[1] 1

	id	panel	des	sex	age	educ5	occu_6	coresi	smam02	wage10
1	1	1	Single	female	25	大学	unemployed	Coreside	28.876	8
2	1	2	Single	female	26	大学	Part-time&others	Coreside	28.876	9.5
3	1	3	Single	female	27	大学	Part-time&others	Coreside	28.876	8.7
4	1	4	Single	female	28	大学	Part-time&others	Coreside	28.876	9
5	1	5	Single	female	29	大学	Part-time&others	Coreside	28.876	8
6	2	1	Attrition	male	24	中学	unemployed	Missing	30.59	24
7	3	1	Attrition	female	21	専門学校	Middle/Small company	Live Away	28.876	12
8	4	1	Single	male	34	大学	Large Company	Live Away	30.59	50
9	4	2	Single	male	35	大学	Large Company	Live Away	30.59	56
10	4	3	Married	male	36	大学	Large Company	Live Away	30.59	54.2
11	5	1	Single	female	20	専門学校	Part-time&others	Live Away	28.876	.3
12	5	2	Single	female	21	専門学校	Part-time&others	Live Away	28.876	21.55984
13	5	3	Attrition	female	22	専門学校	Part-time&others	Live Away	28.876	20
14	6	1	Attrition	male	22	大学	in School	Live Away	30.59	0
15	7	1	Single	male	21	大学	in School	Live Away	30.59	3
16	7	2	Single	male	22	大学	in School	Live Away	30.59	3.5
17	7	3	Single	male	23	大学	Part-time&others	Live Away	30.59	2.8
18	7	4	Attrition	male	24	大学	unemployed	Live Away	30.59	.7
19	8	1	Single	male	20	大学	in School	Live Away	30.59	.3
20	8	2	Single	male	21	大学	in School	Live Away	30.59	26.35101
21	8	3	Single	male	22	大学	unemployed	Live Away	30.59	0
22	8	4	Single	male	23	大学	unemployed	Live Away	30.59	2
23	8	5	Single	male	24	大学	Skilled Worker	Missing	30.59	30.89151
24	9	1	Single	male	23	大学	in School	Live Away	30.59	.3
25	9	2	Married	male	24	大学	in School	Live Away	30.59	1.5
26	11	1	Attrition	male	22	大学	in School	Live Away	30.59	7
27	12	1	Single	male	22	大学	in School	Live Away	30.59	2
28	12	2	Single	male	23	大学	in School	Live Away	30.59	0
29	12	3	Single	male	24	大学	unemployed	Live Away	30.59	0
30	12	4	Single	male	25	大学	unemployed	Missing	30.59	28.89353
31	12	5	Single	male	26	大学	Skilled Worker	Missing	30.59	30.89151

Ready Vars: 21 Obs: 57,249 Filter: Off Mode: Browse CAP NUM

図 1 は、Stata のデータウィンドウよりコピーした結婚分析における人-期間別データの画面である。図 1 ではデータの値ラベルを表示しているが、実際には各ラベルには数値データが入力されている。id 変数は個人を識別する番号であり、panel 変数は独立変数が測定された調査回を示している。des は panel の翌年の調査回における結婚の生起状況を示しており、結婚が生起していなければ「Single」、結婚が生起していれば「Married」、調査から脱落していた場合は「Attrition」と

表示されている。先に述べたように、上記のデータは、リスク開始（この場合、調査の開始時点）からイベント（この場合、結婚）が発生するか、もしくはセンサリング（脱落か未婚のまま第6回調査を向かえた場合）となった時点までの人-期間別データとなっている。また、ここでは、各レコードに対して、panelの翌年の調査におけるイベント生起の状況が示されている点に注意されたい。これは、モデルにおいては独立変数の従属変数に対する時間的先行を確保するため、独立変数はすべて前回調査によって得た値を用いているためである。つまり、t-1年における個人の属性・状態によって、t-1年からt年までに生起した結婚に対する因果推論を行うわけである。このような同一ID内におけるラグを付けるStataのコマンドは以下である。

```
1    sort id panel
2    by id: gen des = status[_n+1]
3    replace des = 9 if des==. & panel<4
4    la def des 0"Single" 1"Married" 9"Attrition"
5    la val des des
```

statusという変数が当年調査における配偶関係（無配偶、有配偶）を表しており、2行目のコマンドによって、同一ID内において、翌年調査における配偶関係を当年調査のレコードに付帯している。

この操作によって、最終調査回（ここでは第6回調査）のデータレコードではdesがすべて欠損値となるのでデータからは削除している。また上記の操作とは逆に、前年調査における独立変数の値を翌年調査のレコードに付帯するという作業も考えられるが、この場合は説明変数の候補となる複数の変数にラグをつけなければならないために作業が煩雑である。また、脱落による右センサリングが生じた場合には、上記の操作に比べてデータレコードが1レコード少なくなってしまうため、サンプルの持つ情報を最大限分析に反映するという観点において、統計的に効率的（efficient）ではない。

図1ではidとSEXは時間独立共変量であるが、それ以外の変数は時間依存性共変量となっている。

3.4 離散時間ロジットモデルの分析プログラムと出力例

人-期間別データを作成できれば、あとは通常のロジスティック回帰分析もしくはCLL回帰分析を行えばよい。ただし、離散時間モデルでは、ベースライン・ハザード関数を定義する必要があり、この点が通常のロジスティック回帰分析モデルとは異なる。ベースライン・ハザード関数は、リスク時間（duration）を表す変数にどのような操作化を行うかによって異なる。ダミー変数を用いたステップ関数、2次関数や自然対数、WeibulやGomperzなどのパラメトリックな時間分布を仮定したものなどが考えられる。ベースラインハザードの形状が実際のデータに対して当てはまりが悪いと、共変量のパラメータ推定にもバイアスが生じる。そのため、できるだけ適合性の高いハザード関数を選択することが重要である。また、モデルの節約性（parsimoniousness）の観点から、できるだけ少ないパラメーターでこれを表現できることが理想である。ここではスプライン関数を用

いる方法（Panis 1994）について解説する。

スプライン関数ではリスク期間をいくつかの区間に分けて、各区間内における対数ハザード確率（あるいは対数ハザードオッズ）が同じ傾きをもって線形に増加あるいは減少することを仮定するモデルである。対数ハザード確率の増減の傾きは、同一区間内では一定であるが、異なる区間においては異なる傾きをもつことができる。そのため、比較的少ないパラメーターで自由度の高いベースライン・ハザード関数を設定することができる。最近の研究では Raymo（2003）などにおいて用いられている。Stata によるスプライン関数のコマンドは以下である。

```
1        gen age2 = age - 20
2        lspline age2 f 4 9
```

ここでは 1 行目において、年齢の実数を表す変数 age からサンプルの最低年齢である 20 歳を引いた age2 をまず作成した。その後、2 行目のコマンドにより、age2 に対するスプライン変数を作成した。age2 を用いることにより、モデルの切片（定数）の値は、すべての独立変数が 0 であった場合の 20 歳の女性のハザード確率を表す。age をそのまま用いてスプライン変数を作成しても共変量のパラメーターについては全く同じ値を得る。しかし、切片の値はすべての独立変数が 0 であった場合の 0 歳の女性のハザード確率を表すものとなってしまうため、非常に小さい値を得る。また、このような値は非現実的な仮想値であり、結果の解釈に混乱を招く恐れがあるので、ここでは 20 歳に centered した age の値をもとにスプライン変数を作成した。

離散時間ロジットモデルならびに離散時間 CLL モデルのコマンド例を以下に示す。

```
1        #delimit;
2        logit des1 f1-f3 b1.panel b2.educ3a b2.occu_6 i.coresi smam02s lnwage
3        i.wagem2 if sex==0
4        ;
5        est store logit
6
7        cloglog des1 f1-f3 b1.panel b2.educ3a b2.occu_6 i.coresi smam02s lnwage
8        i.wagem2 if sex==0
9        ;
10       est store clog
11
12       est tab logit clog, eform star(.10 .05 .01) stats(N ll chi2 df_m) b(%9.4f)
13
14       # delimit cr;
```

1 行目のコマンドは、コマンドが改行してもセミコロン；まで同一コマンドとしてみなすことを指示するコマンドである。なお、14 行目のコマンドはこれを解除することを指示するコマンドである。

2-3行目は通常のロジットモデルのコマンドであるが、ここでは分析に用いるデータが人-期間別データであるため、離散時間ロジットモデルとなる。同様に7-8行目のコマンドが離散時間CLLモデルとなる。なお、共変量の変数名の冒頭についている「b1.」、「b2.」、「i.」などの記号は、これらがダミー変数もしくはカテゴリー変数であることを示している。bについては、後に来る数字がその共変量の基準カテゴリーとなることを指定している。なお、iについては単にその共変量がダミー変数もしくはカテゴリー変数であることを指定しているだけで、基準カテゴリーまでは指定していない。その場合、その共変量の最も小さい値が基準カテゴリーとなる。

5行目と10行目のコマンドは分析結果をメモリー内に保存するコマンドであり、それぞれの分析結果を「logit」、「clog」として保存している。12行目のコマンドでそれらを呼び出し、テーブル形式で表示している。stat()において表示するモデル統計量などを細かく指定できるが、ここでは統計的有意水準を示す星マーク、パーソン-イヤー数、Loglikelihood、カイ2乗値、モデル自由度を表示する。また、推定値は回帰係数bとハザードオッズexp(b)のどちらで表示するのかを選べるが、ここではeformというオプションを追加して、exp(b)で表示することを選択する。また、b(%9.4f)では推定結果を小数点以下4桁まで表示するように指定している。

12行目のコマンドで表示されるアウトプットは以下である。

Variable	l5f	cl5f
f1	1.2139***	1.2115***
f2	1.0908***	1.0870***
f3	0.8869***	0.8902***
panel		
2	1.0742	1.0712
3	1.1470	1.1420
4	1.1651*	1.1601*
5	1.2379**	1.2276**
educ3a		
1	0.9933	0.9896
educ3a		
3	1.0917	1.0888
5	1.2462***	1.2370***
occu_6		
1	0.8597	0.8642
occu_6		
3	1.1893**	1.1818**
4	1.0340	1.0330
5	1.0121	1.0131
6	1.1016	1.1052
7	0.5774***	0.5824***
9	1.2326*	1.2225*
coresi		
1	1.0105	1.0106
2	1.1159	1.1126
3	1.0446	1.0399
smam02s	0.7508***	0.7570***
lnwage	1.2567***	1.2496***
wagem2		
1	0.6696***	0.6772***
2	0.5203***	0.5276***
_cons	0.0078***	0.0079***
N	26843	26843
ll	-5.32e+03	-5.32e+03
chi2	332.8224	332.7596
df_m	24.0000	24.0000

legend: * p<.1; ** p<.05; *** p<.01

3.5 成年者縦断調査への適用例 （結婚）

3.5.1 離散時間ロジットモデルと離散時間 CLL モデルの比較

表 1 女性の初婚のハザード確率に対する離散時間ロジットモデルならびに離散時間 complementary log-log モデルの推定結果

	離散時間ロジット exp(β)		離散時間CLL exp(β)	
年齢スプライン				
20-25歳	1.214	***	1.212	***
25-30歳	1.091	***	1.087	***
30-39歳	0.887	***	0.890	***
年次（対：2002-03年）				
2003-04年	1.074		1.071	
2004-05年	1.147		1.142	
2005-06年	1.165	*	1.160	*
2006-07年	1.238	**	1.228	**
学歴（対：高校卒）				
中学校卒	0.993		0.990	
短大・専門学校卒	1.092		1.089	
大学・大学院卒	1.246	***	1.237	***
職業（対：中小企業雇用）				
大企業雇用	0.860		0.864	
専門・技術職	1.189	**	1.182	**
自営・家従・会社役員	1.034		1.033	
非正規雇用	1.012		1.013	
無職	1.102		1.105	
学生	0.577	***	0.582	***
不明	1.233	*	1.223	*
親との同別居（対：親と別居）				
両親と同居	1.011		1.011	
片親と同居	1.116		1.113	
不明	1.045		1.040	
居住都道府県のSMAM-28	0.751	***	0.757	***
Ln（年間勤労所得）	1.257	***	1.250	***
年間勤労所得不明ダミー	0.670	***	0.677	***
年間勤労所得ゼロダミー	0.520	***	0.528	***
定数	0.008	***	0.008	***
person-year数	26843		26843	
カイ2乗値	332.822		332.760	
自由度	24		24	

* p<.1; ** p<.05; *** p<.01

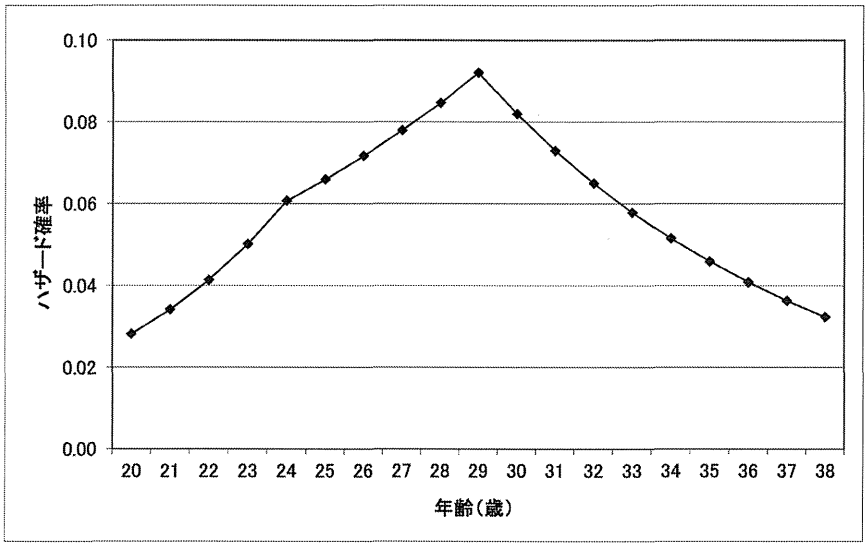
表 1 は、離散時間ロジットモデルならびに離散時間 CLL モデルによる初婚のハザード確率の推定結果を表している。分析結果は回帰係数 b を指数化してえられたハザード・オッズあるいはハザード比によって示した。両モデルの結果は質的にも量的にもほぼ同じ値を示している。このことは年齢別のハザード確率が十分に小さいために、ハザード・オッズとハザード比がほぼ同等に解釈できることを意味する。しかし、この仮定は常に成立するわけではないので、注意が必要である。以下においては、離散時間 CLL モデルの結果を中心に解説する。

exp(b) は、カテゴリー変数については、当該カテゴリーが基準カテゴリーに対して、初婚のハザード確率（ロジットモデルの場合は、ハザード確率のオッズ）が何倍高いのか（あるいは低いの

か）を表す。また、量的変数の場合は共変量1単位当たりの増加による初婚のハザード確率の増加分は、exp(b) を乗数倍した値によって得られる。例えば、2002年の居住都道府県におけるSMAM（静態統計の率から得られた平均婚姻年齢：singulate mean age at marriage）[*6]については、これが1年上昇する毎に、0.757の乗数倍ずつ初婚のハザード確率が減少していくことを意味する。そのため、SMAMが30歳の都道府県（データでは例えば、東京都）出身の女性は、SMAMが28歳（データでは例えば岩手県）の女性に比べて、初婚のハザード確率が43％低い（$=0.757^2-1$）と解釈される。

また、年間勤労所得については、これが不明であったりゼロであった場合には平均値を代入した。したがって、年間勤労所得不詳ダミーや年間勤労所得ゼロダミーは、年間勤労所得が平均値であった場合の初婚ハザード確率を基準カテゴリーとしたハザード比を示している。

図2 初婚のベースライン・ハザード



＊ 年間勤労所得が300万円で、年齢を除く他の共変量がすべて基準カテゴリーである場合。

年齢スプラインの効果は初婚のベースライン・ハザードを示しており、他の共変量がすべてゼロ（あるいは基準カテゴリー）であるとした場合の初婚の年齢別生起パターンを示している。なお、ここでは簡略化のため、ベースラインハザードと他の共変量との交互作用を考慮しない等比ハザードモデルを示している。年間勤労所得が300万円で、他の共変量がすべて基準カテゴリーの値であるケースを仮定すると、切片の値である0.028（$=0.008*1.250^{ln(300)}$）を基点として、20歳から24歳までの間は、初婚のハザード確率が1.212の乗数倍ずつ上昇し、25-29歳の間はこれが1.087の乗数倍ずつ上昇し、30歳以降においては0.890の乗数倍ずつ減少していくことを意味する。これをグラフに表すと図2のようになる。

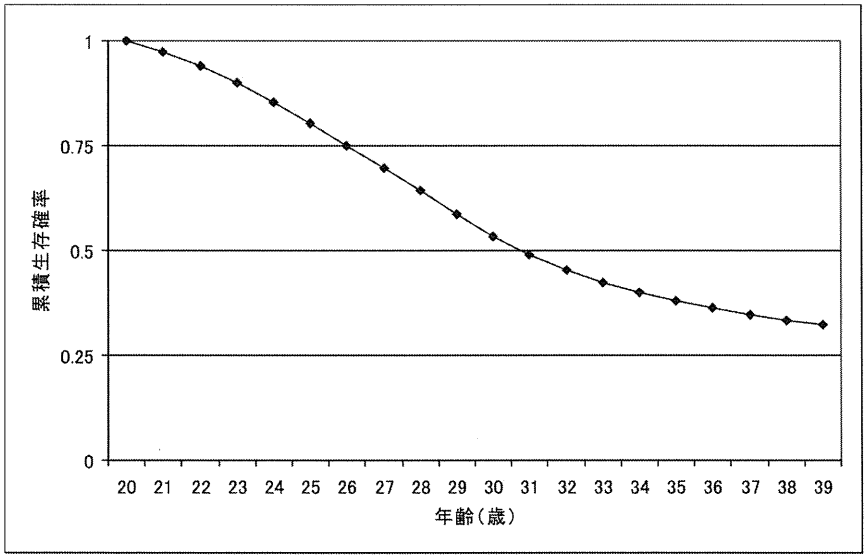
[*6] SMAMは国勢調査の年齢別配偶関係割合から算出されるため、2000年の値と2005年の値を用いて、2002年の値を線形補完した。

さらに、離散時間モデルにおいては、以下の式の利用して、ベースライン・ハザードから累積生存確率を算出することができる。

$$S_t = S_{t-1}[1\text{-}P_t] \quad \cdots (5)$$

図 2 で算出されたベースライン・ハザードに (5) 式を適用して得られた初婚の累積生存確率を図 3 に示す。図 3 における 39 歳時における未婚率は 32 % と推定されており、国立社会保障・人口問題研究所（2006）による将来人口推計における仮定値と比較してもやや高めに推定されているように思われる。パネル調査においては、調査からの脱落が不可避であり、初婚と脱落が無作為に生起しない場合、つまり結婚しやすい女性ほど調査から脱落する確率が高いというような傾向が、共変量を統制した後にも認められる場合には、本章における分析のように脱落を右センサリングとして扱うことには問題がある。なぜならば、本来、結婚の生起としてカウントされるべき女性が、脱落によりこれに寄与しないため、結婚を予測するパラメーターを過小に推定することになるためである。

図 3 初婚の累積生存確率



3.5.2 無作為センサリングの仮定に関する感応分析

この分析で用いたデータにおいて、無作為センサリングが仮定できるのか否かについて、簡単な感応分析 (sensitivity analysis) を行ってみる。ここでは Allison (1995) による感応分析の方法に従って、1) 脱落を初婚として扱う、2) 脱落は最終調査回（第 6 階調査）まで残った後に右センサリングしたものとして扱う、の 2 通りのモデルにより離散時間 CLL 分析を行った。脱落が初婚の生起過程とは無作為に生じると仮定できるならば、2 つのモデルのパラメーター推定値はほぼ同じ値を示すはずである。この分析の結果を表 2 に示す。

モデル 1 は、脱落を初婚として扱ったモデルであり、脱落サンプルは高い初婚ハザードをもつと

表2 初婚のハザード分析における脱落の取り扱いに関する感応分析の結果：離散時間CLLモデル

	モデル1 A=M exp(β)	モデル2 A~= M exp(β)
年齢スプライン		
20-25歳	1.020	1.208 ***
25-30歳	1.051 ***	1.083 ***
30-39歳	0.937 ***	0.894 ***
年次(対:2002-03年)		
2003-04年	1.008	1.202 **
2004-05年	0.911 **	1.470 ***
2005-06年	0.859 ***	1.656 ***
2006-07年	0.907 *	1.756 ***
学歴(対:高校卒)		
中学校卒	1.054	0.982
短大・専門学校卒	1.079 **	1.097
大学・大学院卒	1.149 ***	1.247 ***
職業(対:中小企業雇用)		
大企業雇用	0.999	0.851
専門・技術職	0.994	1.179 *
自営・家従・会社役員	1.119	1.014
非正規雇用	1.071	1.000
無職	1.255 ***	1.068
学生	1.020	0.578 ***
不明	1.202 ***	1.194
親との同別居 (対:親と別居)		
両親と同居	0.518 ***	1.156 *
片親と同居	0.587 ***	1.253 **
不明	0.795 ***	1.080
居住都道府県のSMAM-28	1.043 *	0.734 ***
Ln(年間勤労所得)	1.071 ***	1.252 ***
年間勤労所得不明ダミー	1.167 ***	0.636 ***
年間勤労所得ゼロダミー	0.773 ***	0.533 ***
定数	0.180 ***	0.005 ***
person-year数	26843	32889
カイ2乗値	490.8841	450.9818
自由度	24	24

* p<.1; ** p<.05; *** p<.01

仮定したものである。一方、モデル2は、すべての脱落サンプルは最終調査回までリスクサンプルとして残ったものとして扱ったモデルで、脱落サンプルが低い初婚ハザードをもつことを仮定している。両モデルを比較すると、年齢スプラインや学歴の効果は比較的近い値を示しているが、その他の変数、例えば、年次、職業、親との同別居、SMAMなどについては全く逆の効果を示していることが明らかである。したがって、モデルの共変量を統制した後も、脱落と初婚には何らかの強い相関があることが示唆される。

上記のように、脱落と初婚の生起過程が無作為とは過程できず、脱落を右センサリングとして扱うと、初婚のパラメーター推定に過小バイアスが生じる。この問題については、脱落を初婚の競合リスクイベントとして捉えることで一定の解決を与えることができる。次章においては、パネルデータを用いたイベントヒストリー分析における脱落の影響について考察し、その影響を除去するための分析手法を紹介する。

参考文献

Allison, Paul D., 1982. "Discrete-Time Methods for the Analysis of Event Histories", *Sociological Methodology*, 13: 61-98.

Allison, Paul D., 1995. *Survival Analysis Using The SAS System: A Practical Guide.* Cary: SAS Institute Inc.

Guo, Guang, 1993. "Event-History Analysis for Left-Truncated Data" *Sociological Methodology* 23:217-43.

Panis, Constantijn, 1994. "The Piecewise Liner Spline Transformation." Stata Technical Bulletin, vol. 3, issue 18: 146-149.

Raymo, James M., 2003. "Educational Attainment and the Transition to First Marriage Among Japanese Women" *Demography* 40: 83-103.

津谷典子, 2002, 「イベント・ヒストリー分析」, 日本人口学会編, 『人口大事典』, 428-31 ページ, 培風館。

山口一男, 2002c, 「イベントヒストリー分析 (最終回)」, 『統計』, 2002 年 11 月号, 55-60 ページ。

第 4 章

SURF モデル

4.1 はじめに

近年、わが国でも盛んに用いられるようになった分析手法の 1 つにイベントヒストリー分析 (event-history analysis) がある。パネルデータを用いたイベントヒストリー分析では、最も一般的な方法として、脱落はセンサリング (censoring) (観察打ち切り例) として扱われてきた。しかし、脱落と対象とするイベントとが独立に生起しない場合、このような処置はパラメーターの推定にバイアスをもたらす。

Hill (1997) は、パネルデータを用いた離婚要因に関するイベントヒストリー分析を例に、SURF モデル (Shared Unmeasured Risk Factors Model) (Hill et al. 1993) を適用することによって、この問題に対処できることを示している。SURF モデルとは、McFadden (1981) が多項ロジットモデルの拡張として導いたネステッド・ロジットモデル (nested logit model) を離散時間ハザードモデルに応用したものである。本稿では、SURF モデルの概要について解説するとともに、成年者縦断調査を用いた初婚分析を用いた適用事例について紹介する。

4.2 離散時間ロジットモデルにおける競合イベントの取り扱い

あるイベントの生起によって、他のイベントの生起リスクがなくなる場合、2 つのイベントは競合するイベント (competing events) であるという。例えば、死因別死亡率の分析において、「癌による死亡」と「心臓病による死亡」は相互に競合するイベント (mutually competing events) である。また、相互にではなく、一方のみが他方の競合するイベントとなることもある。例えば、結婚は婚前妊娠にとって競合するイベントである。しかし、婚前妊娠が生起しても、結婚のリスクはなくならないため (むしろ増大する)、婚前妊娠は結婚に競合するイベントではない。

パネル調査における脱落は、あらゆるイベントにとって競合するイベントである。なぜならば、脱落が生じることによって対象とするイベントの生起リスクが観測できなくなるためである。また、イベントの生起によって、少なくともリスク期間における脱落の発生リスクは消失する。そのため、脱落は常に対象とするイベントと相互に競合するイベントであるといえる。

山口 (2002) によれば、離散時間モデルにおける競合するイベントの取り扱いには次の 3 つの方

法がある。1）競合する他のすべてのイベントをその生起時点でセンサリングとして扱う、2）競合するイベントを従属変数とする離散時間多項ロジットモデル（discrete-time multi-nominal logit model）を適用する、そして3）競合するイベントを従属変数とするSURFモデルを行う。以下に山口（2002）を参照しつつ、どのような場合に各方法を使用するべきなのかについて解説する。なお、以下では相互に競合するイベントAとイベントBがあるとする。

競合するイベントを右センサリングとして扱うという第1の方法は、最も一般的に用いられる手法である。しかし、離散時間モデルにおいてこの方法が妥当であるのは、イベントAとイベントBのハザード確率$P_A(t)$と$P_B(t)$の積が無視できるほど小さい場合のみである。連続時間を仮定するモデルにおいては、競合するイベントの同時発生モデルにおいて、各イベントが独立に起こるという条件が成立する場合、競合するイベントをセンサリングとして扱うことが可能である。この条件が成立するには、競合するイベントが2つとも起こらない確率が各イベントの生存確率の積となる必要がある（$S_{A+B}(t) = S_A(t) \times S_B(t)$）。しかし、離散時間モデルにおいては、時点tにおいてイベントAもBも起こらない確率は、1-$P_A(t)$-$P_B(t)$であり、(1- $P_A(t)$)×(1- $P_B(t)$) とはならない。したがって、離散時間モデルでは、競合イベントが独立である時の条件である (1- $P_A(t)$)×(1- $P_B(t)$) に対して、$P_A(t) \times P_B(t)$分だけ誤差が生じることとなる。そのため、イベントAかB、あるいは双方の生起確率が著しく小さく、$P_A(t) \times P_B(t)$が無視できるほど小さい場合に限り、他の競合イベントをセンサリングとして扱うことが妥当となる。

$P_A(t) \times P_B(t)$が無視できるほど小さくない場合、第2の方法である多項ロジットモデルによる競合リスク分析が検討される。この方法では、前項において解説した人-期間別データに対して、多項ロジットモデルを適用し、競合する各イベントのハザード確率の同時推定を行う（Allison 1982）。だたし、多項ロジットモデルではIIA（Independence from Irrelevant Alternatives）の仮定を前提としている。IIAの仮定とは、いかなる2つの確率の比も他の確率の大きさによる影響を受けないことをいう。$P_A(t)$と$P_B(t)$がともに起こらない確率を$P_C(t)$（=1-$P_A(t)$-$P_B(t)$）とすると、IIAが成立するとき、以下の関係が成り立つ。

① $P_A(t)/P_C(t)$が$P_B(t)$に依存しない
② $P_B(t)/P_C(t)$が$P_A(t)$に依存しない

①の関係が成立する時、イベントBが起こらないという条件の下でイベントAの生起確率が、イベントBの生起確率から独立である（AはBから条件付きで独立）。また、②の関係が成立する時、イベントAが起こらないという条件の下でイベントBの生起確率が、イベントAの生起確率から独立である（BはAから条件付きで独立）。IIAが成立する時、条件付きでイベントAとイベントBの決定要因が独立と考えられるため、離散時間多項ロジットモデルを適用することができる。

競合するイベントの条件付き生起確率にIIAが成立するか否かをより直接的に検証し、かつIIAが成り立たない場合でも偏りなくパラメーターを推定する方法が、第3の選択肢であるSURFモデルである。したがって、初婚要因の離散時間モデルにおいては、脱落をセンサリングとして扱う第1の方法とSURFモデルを用いる第3の方法を比較することによって、パラメーター推定にお

けるバイアスの大きさについて検討することが可能となる。また、SURF モデルでは脱落と初婚の非観察要因[*1]に相関があるか否かを統計的に検定することができる（Hill, et. al. 1993）。IIA が成立する場合、この相関は 0 となる。そのため、SURF モデルによる分析を通して、第 2 の方法である離散時間多項ロジットモデルの適用が妥当か否かを検討することもできる。次節では SURF モデルの概要について述べる。

4.3 SURF モデルの概要

SURF モデルとは、McFadden（1981）が多項ロジットモデルの拡張として導いたネステッド・ロジットモデルを Hill 等（1993）が離散時間モデルに応用したものである。その要諦は、競合イベントの同時分析において、各イベントの誤差項に部分的な相関を許容することで、多項ロジットモデルにおける IIA の仮定を緩和することにある。以下に、Hill 等（1993）や山口（2002）を参照しつつ、その概要について述べる。

m 個の競合するイベントがある場合に、個人 i が t 時においてどのイベントを経験するのかは、各イベントの潜在的な生起傾向（state propensity index）によって決定されている。この潜在的な生起傾向は、直接には観察できない連続量（latent variable）で、確率のような固定範囲をもたないとする。こうした条件の下、個人 i の t 時における潜在的なイベント生起傾向 S_{tmi} は以下の (1) 式によって表すことができる。

$$\begin{aligned} S_{t0i} &= \beta_0^{*\prime} X_{t0i} + \epsilon_{t0i} \\ S_{t1i} &= \beta_1^{*\prime} X_{t1i} + \epsilon_{t1i} \\ &\vdots \qquad \vdots \qquad \vdots \qquad (1) \\ S_{tmi} &= \beta_m^{*\prime} X_{tmi} + \epsilon_{tmi} \end{aligned}$$

S_{tmi} は、説明変数の分散によって説明される部分 $\beta_m^{*\prime} X_{tmi}$ と誤差分 ϵ_{tmi} とに分けられる。式 (1) では、個人 i は S_{tmi} が最も高いイベントを経験すると仮定する。離散時間ハザードモデルにおいては、リスク開始時点においてイベント未経験の状態である S_{t0i} の値が最も高いと仮定される。他の潜在的イベント生起傾向 S_{tmi} がこれを超えるまで、いずれのイベントも生起しない。しかし、誤差項による攪乱もしくは共変量 X_{tmi} の値の変化によって、S_{tmi} が S_{t0i} を超えると最も潜在的イベント生起傾向が高いイベントが生起する。

簡略化のため、ここで競合するイベントが 2 つであるとする。誤差項 ϵ_{tmi} に極値分布を仮定すると、これがイベント間で独立である場合に IIA が成立し、離散時間多項ロジットモデルを得る。しかし、ϵ_{t0i} は他の 2 つから独立であるが、ϵ_{t1i} と ϵ_{t2i} の間に相関がある場合、SURF モデルを得る[*2]。

ここで注目すべきは、多項ロジットの成立要件である IIA は、競合イベントの同時分析における

[*1] ハザードモデルにおける観察されない異質性（unobserved heterogeneity）と同義である。
[*2] SURF モデルの数式的展開については、Hill 等（1993）や山口（2002）を参照のこと。

誤差項、すなわち非観察要因が、各イベント間で独立であるときに成立するということである。初婚と脱落について考えてみると、これは非常に強い仮定であるといえる。なぜならば、パネル調査においては、結婚はそれ自体が脱落の要因となるためである（坂本 2006)。女性にとって結婚は転居を伴うことが多い。そのため、結婚直後のサンプル捕捉が困難となる。また、結婚により夫や夫の家族による調査拒否、またそれを忌避することによる本人からの調査拒否などが発生することも報告されている（坂本 2006)。分析において用いる「21世紀成年者縦断調査」においても同様に、少なくとも一定割合の脱落は結婚によって生起しているものと思われる。その結果、結婚と脱落の生起傾向は類似したものとなり、非観察要因についても共通の傾向をもつ可能性が高いのである。

SURFモデルでは、非類似係数（index of dissimilarity）ρ を説明変数の回帰係数と同時に推定する。ϵ_1 と ϵ_2 の相関係数は 1-ρ^2 として表される。したがって、ρ が1の時は競合イベントの非観察要因には相関がない、つまりIIAを仮定できることを意味する。また、ρ の標準誤差もモデルで計算されるため、ρ が1と統計的に有意に異なるのかの検証も行うことができる（山口 2002)。この ρ の解釈を通して、競合するイベントの非独立性の存在やその強さについて検証することができる。また、説明変数の回帰係数は、ρ すなわち、競合するイベント間における非観察要因の相関、を補正した上で得られた値となる。モデルで ρ を統制することは、競合するイベントの生起過程に条件付き独立を留保した状態を統計学的に作り出すことに等しい。そのため、SURFモデルにおける回帰係数は、競合するイベントが起こらなかった場合に、説明変数が当該イベントのハザード確率に与える効果を表す。

4.4 2段階推定によるSURFモデルの適用手順

SURFモデルはHillの開発した独自のソフトウェア（Turbo Pascal compilerとDOS-base PCを使用）[*3]によって最も効率的に推定できるが、ロジットモデルを用いた2段階推定によって、簡便なモデルを適用することができる。ここでは、初婚と脱落を競合イベントとして取り扱う場合を例として、Hill等（1993）や山口（2002）によって示されている2段階推定によるSURFモデルの適用手順を以下に示す。なお、Hill等（1993）による方法と山口（2002）における解説には重要な違いがあり、あまり数学に詳しくない初学者はその適用方法について混乱をきたす恐れがある。両者におけるモデル解説に違いについて、この機会に気づいた点をまとめておいたので、興味がある読者は章末の付録を参照されたい。

① はじめに、通常の離散時間ロジットモデルと同様に人?期間別データを作成する。また、従属変数Y(t)はイベントが生起していなければ0、初婚が生起する場合は1、そして脱落が生起する場合を2となるようにコーディングする。

② ①で作成した人-期間別データより、初婚もしくは脱落を経験したサンプル（Y(t)が1もしくは2のケース）のみを取り出し、結婚対脱落を対比としたロジットモデルを行う。ここでの分析は、結婚か脱落が生起したとして、それが脱落ではなく結婚である確率を推定するモデルとなる。

[*3] このソフトウェアは、http://lib.stat.cmu.edu にて無料で公開されている。

③ ②で得られた回帰係数をもとにして、以下の値を算出する。

$z_1(t) = \log[1 + \exp(-\sum_k b_k x_k(t))]$ ・・・(2)
$z_2(t) = \log[1 + \exp(\sum_k b_k x_k(t))]$ ・・・(3)

この時、$\sum_k b_k x_k(t)$ は②のモデルで得られた予測値を表す。$z_1(t)$ と $z_2(t)$ の値をで作成した人?期間別データの各レコードに対して計算して、変数として付帯する。さらに、このデータに結婚か脱落が生起した場合に1、いずれも生起せずに未婚のままである場合に0をとる新しい変数 Y*(t) を作成して追加する。

④ ③で作成した人?期間別データを用いて、従属変数を Y*(t) とする離散時間ロジット分析を行う。ただし、この時で作成した $z_1(t)$ もしくは $z_2(t)$ の一方を説明変数としてモデルに追加する。$z_1(t)$ を追加した場合には、脱落を経験せずに結婚するというハザード確率の回帰係数 β_{1k} を得る。一方、$z_2(t)$ を追加した場合は、結婚せずに脱落するというハザード確率の回帰係数 β_{2k} を得る。なお、とでは異なる説明変数をもつことも可能である（Hill et al. 1993)。この時、$z_1(t)$ と $z_2(t)$ の回帰係数として算出されるのが ρ の推定値である。ρ は $z_1(t)$ と $z_2(t)$ のどちらを用いても全く同じ値を示し、理論的には0から1までの値をとる。結婚と脱落の観察されない異質性（誤差項）の相関係数は、1-ρ^2 によって与えられる。

⑤ ④で得た分析結果では ρ が0であるという帰無仮説に対するP値が示されている。しかし、ここでは ρ の標準誤差を用いて、ρ が1である、つまり結婚と脱落の相関係数が0であるという帰無仮説を検定するようにP値を計算しなおす必要がある。

4.5　2段階推定によるSURFモデルの利用における留意点

SURFモデルの適用においてはいくつか留意する点がある。第1に、2段階推定によるSURFモデルでは、競合するイベントの非観察要因の相関はリスク期間を通じて一定と仮定されている(Hill et al. 1993)。したがって、非観察要因がリスク期間を通じて、結婚と脱落に異なる影響を与える場合、この仮定が成立しない。例えば、調査の初期においては結婚を契機として脱落するサンプルが多いが、調査回が進むにつれて結婚以外の事由による脱落が増えるという場合には、非観察要因の相関がリスク期間を通じて一定であることを仮定できない。この仮定が成立しない場合、時間依存性共変量のパラメーターや ρ の推定値にバイアスが生じる（Hill et al. 1993)。しかし、非観察要因の相関がリスク期間を通じて変化する場合においても、時間固定共変量の回帰係数についてはバイアスが少なく、比較的安定的に推定されることが示されている（Hill et al. 1993)。また、この仮定が満たされない場合には、ρ の推定値が1に近づく傾向があるため、ρ が1と有意に異ならない場合においても、競合するイベントの非観察要因に相関がある可能性が高いことが指摘されている（Hill et al. 1993)。

さらに、これは2段階推定の場合に限らないが、SURFモデルでは誤差項に負の相関を仮定することができないという制約がある（山口 2002)。例えば、婚前同棲の解消について競合するイベントが結婚と別離である場合、非観察要因（例えば、性格の相性）は結婚に対しては正の効果をも

ち、別離に対しては負の効果をもつことが十分に起こりえる。しかし、ρは理論上、$0 < \rho \leq 1$の範囲の値を取るため、非観察要因の相関係数1-ρ^2は正であることが仮定されている[*4]。したがって、非観察要因の相関が負である競合イベントは、SURFモデルでは分析することができない[*5]。

また、SURFモデルにおける推定上の問題として、2段階推定においては、パラメーター推定値の標準誤差が平均してやや小さめに推定される可能性が指摘されている（山口 2002)。これは1段階目のパラメーター推定値には実際には誤差があるにもかかわらず、2段階推定では定数として扱うことから生じる。しかし、通常はこのバイアスは有意度に影響を与えない程度であるため、それほど問題とはならない（山口 2002)。

最後に、ρが統計的に有意に1と異なる場合（$\rho \neq 1$)、回帰係数β_kを厳密にはオッズ比として解釈することができないという制約がある（山口 2002)。そのため、本稿における分析ではオッズ比ではなく、回帰係数を用いて解釈を行う。

6.6. 2段階推定によるSURFモデルのプログラムと出力例

2段階推定によるSURFモデルは離散時間ハザードモデルの一種であるので、通常の離散時間ハザードモデルと同様に、人-期間別データを作成する必要がある。以下では、前章で用いたのと同じ、結婚をイベントとする人-期間別データを用いてSURFモデルのコマンド例を示すこととする。人-期間別データについては前章を参照されたい。なお、使用するソフトウェアはStataのVersion 12である。

```
1   #delimit;
2   logit des1 f1-f3 b1.panel b2.educ7 b2.occu_6 i.coresi smam02s
3   lnwage i.wagem2 b3.marint if des>0
4   ;
5   # delimit cr;
6   est store m1
7
8   predict z,xb
9   gen z1 = ln(1+exp(-z))
10  gen z2 = ln(1+exp(z))
11
12  #delimit;
13  logit des2 f1-f3 b1.panel b2.educ7 b2.occu_6 i.coresi smam02s
14  lnwage i.wagem2 b3.marint z1
15  ;
16  # delimit cr;
```

[*4] しかし、実際の分析においては、ρの推定値が$0 < \rho \leq 1$の範囲を超えることが頻繁に起こりうる。ρの推定値が統計的に有意で、1より大きいか、0より小さい場合には2段階推定の妥当性に問題があると考えられ、その結果は信頼できない（山口 2002)。

[*5] このような場合は、Vermunt（1997）によって提案されている離散時間多項ロジットモデルに潜在クラスを導入する方法が推奨されている（山口 2002)。

```
17          est store m2
18
19          #delimit;
20          logit des2 f1-f3 b1.panel b2.educ7 b2.occu_6 i.coresi smam02s
21          lnwage i.wagem2 b3.marint z2
22          ;
23          # delimit cr;
24          est store m3
25
26          est tab m1 m2 m3, star(.10 .05 .01) stats(N ll chi2 df_m) b(%9.4f)
```

上記のコマンド例においては、1行目から5行目においては、結婚か脱落が生起したレコードのみに対して、脱落を0、結婚を1とした場合のロジットモデルを行っている。ここで従属変数であるdes1とは、結婚が起きた場合に1、脱落もしくは未婚状態にある場合に0をとるダミー変数である。3行目のif以降のコマンドでは、結婚か脱落が生起したレコードのみに分析サンプルを限定している。desという変数はイベントの生起状態を表す変数で、これが0である時は未婚、1である時が結婚、2である時が脱落を表している。そのため、結婚か脱落が起きたレコードのみを選択している分析に用いている。

6行目のコマンドでは推定結果をあとで呼び出せるように、「m1」という名前でメモリー内に保存している。

8-10行目のコマンドでは、結婚対脱落のロジットモデルの予測値から$z_1(t)$ならびに$z_2(t)$を作成し、変数として保存している。9行目、10行目がそれぞれ（2）式と（3）式に対応している。

12-16行目では、$z_1(t)$を用いた結婚のSURFモデルが、19-23行目では$z_2(t)$を用いた脱落のSURFモデルが行われており、それぞれm2、m3というモデル名で推定結果を保存している。ここで用いられている従属変数のdes2は、結婚か脱落のいずれかが起きた場合を1、未婚状態にある場合を0とするダミー変数である。したがって、両モデルの回帰係数は、$z_1(t)$をモデルに含むか$z_2(t)$をモデルを含むかによって、異なる値を示すことになる。

最後の26行目では、これまで保存した推定結果を呼び出し、テーブル形式で表示している。stat()において表示するモデル統計量などを細かく指定できるが、ここでは統計的有意水準を示す星マーク、パーソン-イヤー数、Loglikelihood、カイ2乗値、モデル自由度を表示する。また、推定値は回帰係数bとハザードオッズ比exp(b)のどちらで表示するのかを選べるが、前述のようSURFモデルにおけるexp(b)は、離散時間ロジットにおけるexp(b)とは同義に解釈できないことから（山口 2002)、ここでは回帰係数bを使用する。exp(b)で表示する場合は、コンマ以降に「eform」とタイプすればよい。また、b(%9.4f)では小数点以下4桁まで表示するように指定している。