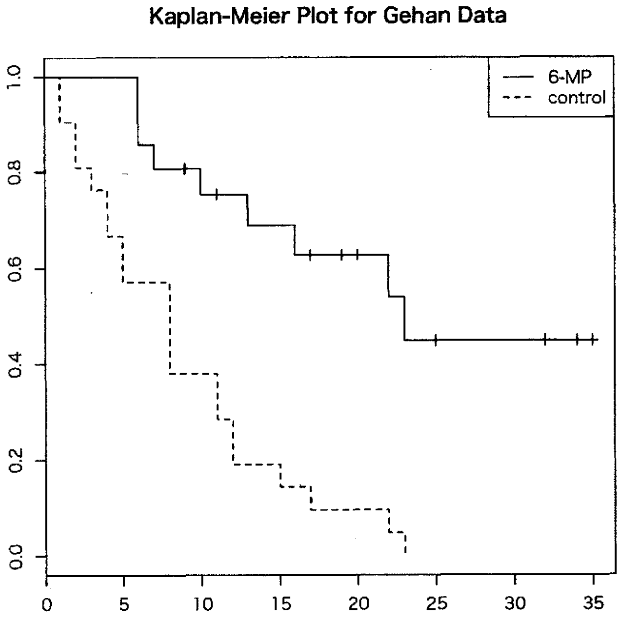
カプラン・マイヤ推定量は、Rで以下のように推定される。

カプラン・マイヤ推定量

```
gehan.sf <- survfit(Surv(time,cens) ~ treat, data= gehan)
print(gehan.sf)
summary(gehan.sf)
plot(gehan.sf, lty = seq(2),
main = "Kaplan-Meier Plot for Gehan Data")
legend("topright", c("6-MP", "control"), lty=seq(2))
```

出力結果

```
> print(gehan.sf)
Call: survfit(formula = Surv(time, cens) ~ treat, data = gehan)

              records n.max n.start events median 0.95LCL 0.95UCL
treat=6-MP         21    21      21      9     23      16      NA
treat=control      21    21      21     21      8       4      12
> summary(gehan.sf)
Call: survfit(formula = Surv(time, cens) ~ treat, data = gehan)

                treat=6-MP
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    6     21       3    0.857  0.0764        0.720        1.000
    7     17       1    0.807  0.0869        0.653        0.996
   10     15       1    0.753  0.0963        0.586        0.968
   13     12       1    0.690  0.1068        0.510        0.935
   16     11       1    0.627  0.1141        0.439        0.896
   22      7       1    0.538  0.1282        0.337        0.858
   23      6       1    0.448  0.1346        0.249        0.807

                treat=control
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     21       2   0.9048  0.0641      0.78754        1.000
    2     19       2   0.8095  0.0857      0.65785        0.996
    3     17       1   0.7619  0.0929      0.59988        0.968
    4     16       2   0.6667  0.1029      0.49268        0.902
    5     14       2   0.5714  0.1080      0.39455        0.828
    8     12       4   0.3810  0.1060      0.22085        0.657
   11      8       2   0.2857  0.0986      0.14529        0.562
   12      6       2   0.1905  0.0857      0.07887        0.460
   15      4       1   0.1429  0.0764      0.05011        0.407
   17      3       1   0.0952  0.0641      0.02549        0.356
   22      2       1   0.0476  0.0465      0.00703        0.322
   23      1       1   0.0000     NaN           NA           NA
```



また、両群に差があるかをログランク検定により検定することができる。

ログランク検定と出力結果

```
> survdiff(Surv(time,cens) ~ treat, data= gehan)
Call:
survdiff(formula = Surv(time, cens) ~ treat, data = gehan)

                N Observed Expected (O-E)^2/E (O-E)^2/V
treat=6-MP     21        9     19.3      5.46      16.8
treat=control  21       21     10.7      9.77      16.8

 Chisq= 16.8  on 1 degrees of freedom, p= 4.17e-05
```

4.3 Cox 比例ハザードモデル

Cox 比例ハザードモデルとは、2 つ以上のグループのイベント発生時間の比較に関心があるとき、両者のハザード関数が比例的な関係にあると仮定し、イベントの発生時間に影響を与えると考えられる共変量を用いてこれを説明するモデルである。

簡単な例を考えてみよう。今、ある年齢層において、男性の死亡に関するハザード関数と女性のそれが比例していると仮定する。女性のハザード関数を基準にとり、$h_0(t)$ と表すことにする。このとき、男性のハザード $h_1(t)$ について、

$$h_1(t) = h_0(t)\exp(\beta)$$

という関係が成立すると考えよう。すると、このパラメータ β を推定することで、両者のハザード関数の差を定量的に評価できることになる。

次に、これに教育水準の高低を加えたいとする。この場合、女性・高学歴を基準ハザードとし、今度は、性別が男性なら 1、女性なら 0 となる変数 z_1、教育水準が低学歴なら 1、高学歴なら 0 となる変数 z_2 を考えると、

$$h(t|z_1, z_2) = h_0(t)\exp(\beta_1 z_1 + \beta_2 z_2)$$

という関係を仮定して、パラメータ β_1、β_2 を推定することにより評価が行えることになる。このようなモデルが Cox 比例ハザードモデルである。より一般的には、

$$h(t|Z) = h_0(t)\exp(\beta' Z)$$

という形となる。

R では coxph という関数を使って比例ハザードモデルのパラメータ推定を行えるようになっている。そこで、Gehan データにおいて、6-MP 処置群を基準ハザード $h_0(t)$ に取り、対照群のハザードが $h_1(t) = h_0(t)\exp(\beta)$ と表されるとした場合のパラメータ推定を行ってみよう。

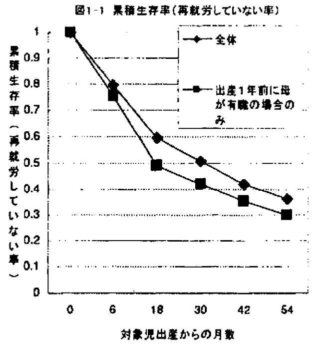
Cox 比例ハザードモデルと出力結果

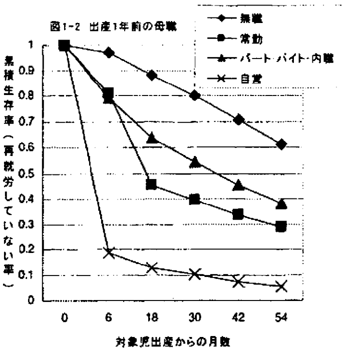
```
gehan.ph <- coxph(Surv(time,cens) ~ treat, data= gehan)
summary(gehan.ph)

Call:
coxph(formula = Surv(time, cens) ~ treat, data = gehan)

  n= 42

                coef exp(coef) se(coef)     z Pr(>|z|)
treatcontrol 1.5721    4.8169   0.4124 3.812 0.000138 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

             exp(coef) exp(-coef) lower .95 upper .95
treatcontrol     4.817     0.2076     2.147     10.81

Rsquare= 0.322   (max possible= 0.988 )
Likelihood ratio test= 16.35  on 1 df,   p=5.261e-05
Wald test            = 14.53  on 1 df,   p=0.0001378
Score (logrank) test = 17.25  on 1 df,   p=3.283e-05
```

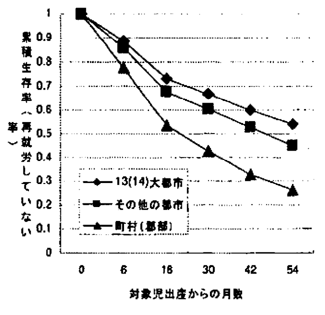
ここで、exp(coef) で示されているのが $\exp(\beta)$ の値に相当する。対照群のハザード関数は、基準ハザードに比べて約 4.8 倍も高いものとなっていると推定される。

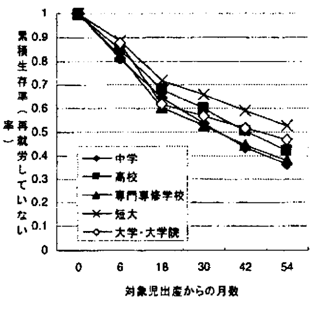
4.4 出生児縦断調査への応用例

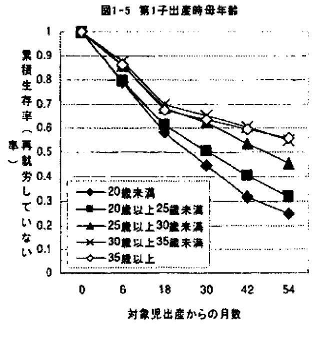
次に示したのは、西野 (2008) による、出生児縦断調査において、イベントの発生を再就労とした生存時間分析を行い、カプラン・マイヤ推定量及びをCox比例ハザードモデルを適用した例である。

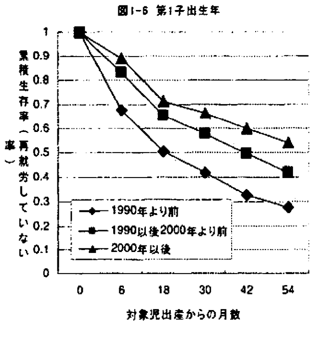












— 126 —

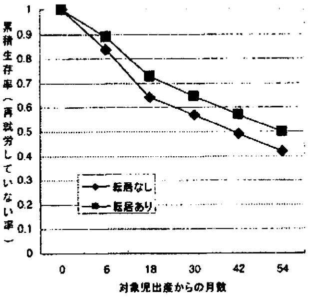


図1-7 この5年間に自治体を越える転居をしたか

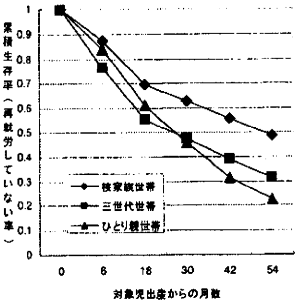


図1-8 再就労時/センサー時の世帯構成

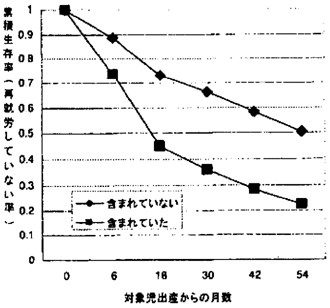


図1-9 再就労時/センサー時のふだんの保育者に
祖父母が含まれていたか

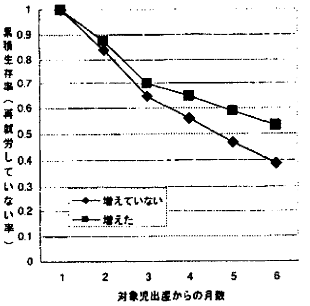


図1-10 この5年間に子どもが増えたか

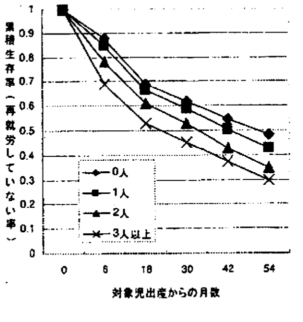


図1-11 きょうだい数（第1回調査時）

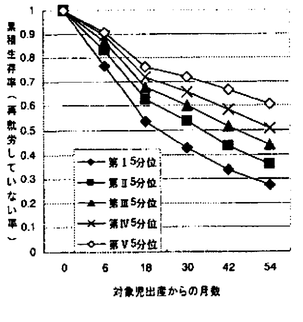


図1-12 父の就労収入額5分位（第1回調査時）

表2 離散時間ロジットモデルによる分析結果

説明変数	全体 係数(β)	全体 ハザード率オッズ比 Exp(β)	13(14)大都市のみ 係数(β)	13(14)大都市のみ ハザード率オッズ比 Exp(β)	その他の都市のみ 係数(β)	その他の都市のみ ハザード率オッズ比 Exp(β)	町村(郡部)のみ 係数(β)	町村(郡部)のみ ハザード率オッズ比 Exp(β)
出産1年前(R=無職) *該当=1 該当せず=0								
常勤	1.605	4.979 ***	1.740	5.695 ***	1.578	4.847 ***	1.377	3.961 ***
パート	1.144	3.139 ***	1.277	3.587 ***	1.121	3.067 ***	0.986	2.681 ***
自営	3.157	23.506 ***	3.511	33.476 ***	3.176	23.962 ***	2.629	13.856 ***
居住地都市規模 (t) *13大都市=1、その他都市=2、町村=3	0.021	1.022 **	—	—	—	—	—	—
母最終学歴(R=高校) *該当=1 該当せず=0								
中学	0.078	1.081	0.265	1.304 *	0.112	1.119	-0.098	0.907
専門専修学校	0.136	1.146 ***	0.075	1.078	0.163	1.178 ***	0.169	1.184 **
短大	-0.166	0.847 ***	-0.249	0.780 ***	-0.145	0.865 ***	-0.052	0.949
四大・大学院	0.031	1.031	0.023	1.023	0.096	1.100 *	0.070	1.072
第1子出産時母年齢	-0.052	0.950 ***	-0.037	0.963 ***	-0.055	0.947 ***	-0.038	0.963 ***
第1子出生年	-0.035	0.966 ***	-0.065	0.937 ***	-0.032	0.969 ***	-0.027	0.974 **
5年間に転居経験あり *該当=1 該当せず=0	0.004	1.004	-0.060	0.942 **	-0.030	0.970 **	0.111	1.117 ***
三世代世帯 (t) *該当=1 該当せず=0	0.039	1.040	-0.305	0.737 ***	-0.039	0.962	0.158	1.171 **
ひとり親世帯 (t) *該当=1 該当せず=0	0.891	2.438 ***	0.726	2.067 ***	0.959	2.608 ***	0.859	2.362 ***
ふだんの保育者に祖父母あり (t) *該当=1 該当せず=0	0.859	2.360 ***	0.861	2.365 ***	0.859	2.361 ***	0.799	2.224 ***
子どもが増えたか (t) *該当=1 該当せず=0	-0.919	0.399 ***	-0.840	0.432 ***	-0.919	0.399 ***	-0.874	0.417 ***
きょうだい数 (t)	0.273	1.314 ***	0.190	1.209 ***	0.294	1.341 ***	0.217	1.242 ***
父の収入額5分位 (t) *第Ⅰ分位=1、第Ⅱ分位=2…	-0.018	0.982 ***	-0.022	0.978 ***	-0.023	0.977 ***	0.002	1.002
パーソンピリオドデータ数	112508		25437		70443		16271	

※(t)は時間依存変数，Rはreference category　　* = p<.05，** = p<.01，*** = p<.001

（表2続き）

説明変数	出産1年前に母が有職の場合のみ 係数（β）	ハザード率オッズ比 Exp(β)	出産1年前に母が有職の場合のみ・13(14)大都市のみ 係数（β）	ハザード率オッズ比 Exp(β)	出産1年前に母が有職の場合のみ・その他の都市のみ 係数（β）	ハザード率オッズ比 Exp(β)	出産1年前に母が有職の場合のみ・町村（郡部）のみ 係数（β）	ハザード率オッズ比 Exp(β)
出産1年前(R=無職) *該当=1 該当せず=0								
常勤	—	—	—	—	—	—	—	—
パート	-0.486	0.615 ***	-0.508	0.602 ***	-0.476	0.622 ***	-0.412	0.662 ***
自営	1.437	4.209 ***	1.638	5.147 ***	1.485	4.415 ***	1.164	3.202 ***
居住地都市規模 (t) *13大都市=1、その他都市=2、町村=3	0.019	1.019	—	—	—	—	—	—
母最終学歴(R=高校) *該当=1 該当せず=0								
中学	0.122	1.130	0.448	1.566 **	0.125	1.133	-0.034	0.967
専門専修学校	0.185	1.203 ***	0.181	1.199 *	0.217	1.243 ***	0.148	1.159 *
短大	-0.091	0.913 **	-0.094	0.910	-0.082	0.922	-0.046	0.955
四大・大学院	0.174	1.190 ***	0.216	1.242 *	0.241	1.273 ***	0.072	1.075
第1子出産時母年齢	-0.016	0.984 ***	0.002	1.002	-0.020	0.980 ***	-0.008	0.992
第1子出生年	-0.054	0.947 ***	-0.078	0.925 ***	-0.049	0.952 ***	-0.042	0.959 **
5年間に転居経験あり *該当=1 該当せず=0	0.003	1.003	-0.082	0.921 **	-0.039	0.961 **	0.109	1.116 ***
三世代世帯 (t) *該当=1 該当せず=0	-0.010	0.990	-0.149	0.861	-0.123	0.884 **	0.103	1.109
ひとり親世帯 (t) *該当=1 該当せず=0	0.530	1.699 ***	0.916	2.499 ***	0.419	1.520 **	0.416	1.516 *
ふだんの保育者に祖父母あり (t) *該当=1 該当せず=0	1.010	2.747 ***	0.903	2.467 ***	1.028	2.796 ***	0.970	2.638 ***
子どもが増えたか (t) *該当=1 該当せず=0	-1.198	0.302 ***	-1.215	0.297 ***	-1.215	0.297 ***	-1.065	0.345 ***
きょうだい数 (t)	0.448	1.565 ***	0.445	1.560 ***	0.487	1.627 ***	0.348	1.417 ***
父の収入額5分位 (t) *第Ⅰ分位=1、第Ⅱ分位=2…	-0.036	0.965 ***	-0.034	0.966 ***	-0.041	0.960 ***	-0.018	0.982 ***
パーソンピリオドデータ数	46140		9595		28538		7899	

※(t)は時間依存変数, Rはreference category　　* = p<.05, ** = p<.01, *** = p<.001

参考文献

[1] Klein, J. P. and M. L. Moeschberger (2009)『生存時間分析』，シュプリンガー・ジャパン株式会社.

[2] 西野淑美 (2008)「出産後再就労のタイミングと促進要因のイベントヒストリー分析」，金子隆一（編）『厚生労働科学研究費「パネル調査（縦断調査）に関する統合的高度統計分析システムの開発研究」平成 19 年度総括研究報告書』, pp.123–132.

[3] 中澤港 (2007)『R による保健医療データ解析演習』，ピアソン・エデュケーション.

第5章

離散時間ハザードモデル

5.1 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、多くの分析において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。

5.2 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）とは仮定できず、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。

5.2.1 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。

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

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、ベースライン・ログオッズ (baseline log odds) である。ベースライン・ログオッズは、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。

5.2.2 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。

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

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。

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

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 のハザード比を表す (Allison 1982)。また、時間変数 a_t は、ベースライン・ログ・ハザード (baseline log hazard) となる。その他の分析結果の解釈は離散時間ロジットモデルと同様である。

[*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)。

5.2.3 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、それをハザード確率に置き換えて解釈することはできない。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、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。

5.2.4 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実際的である。

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

*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。

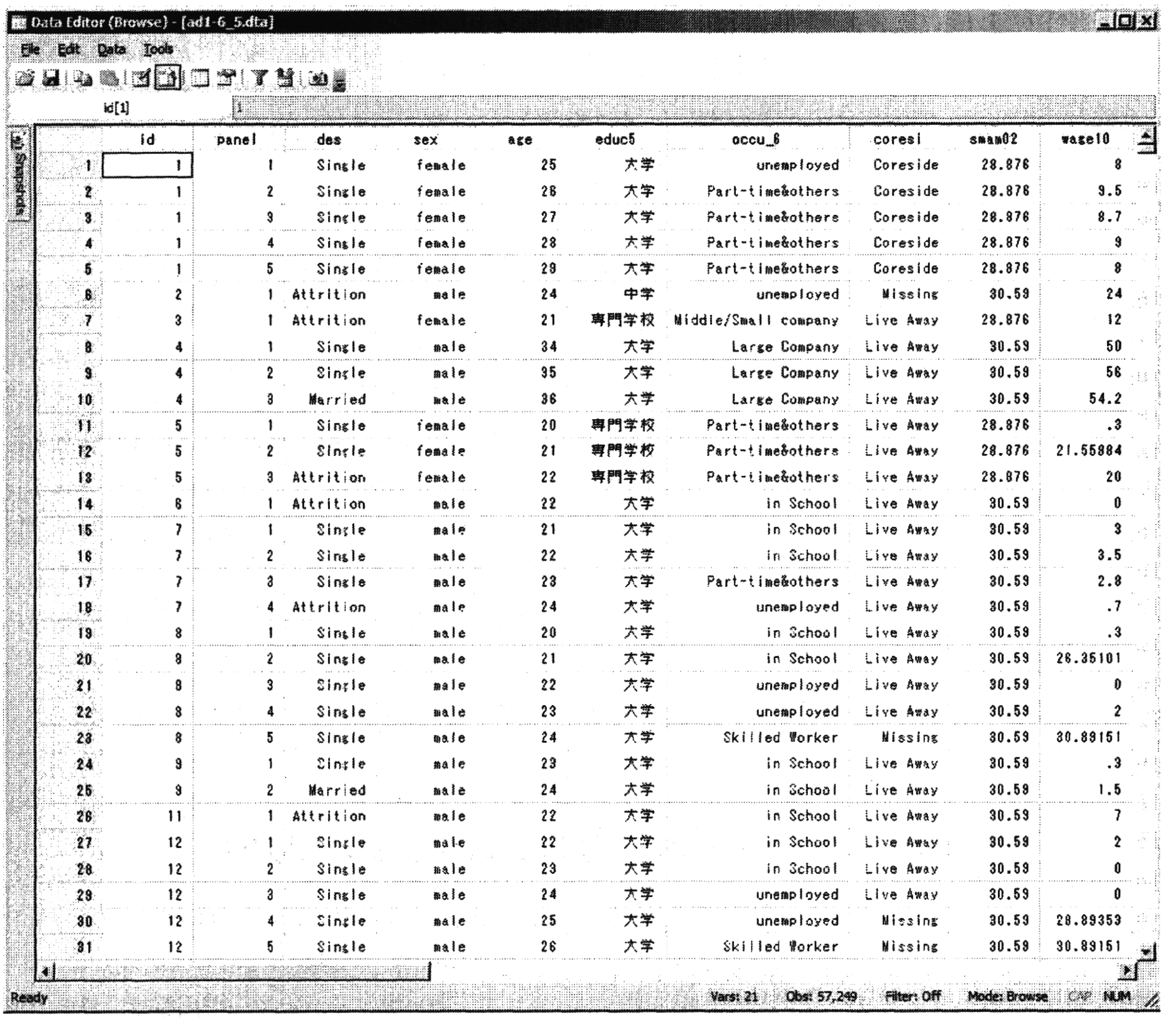
*5 特に、連続時間を仮定したイベントヒストリーモデルとしてよく用いられる Cox 回帰においては、同一時点において複数のイベント生起が観察されるような場合にはパラメーター推定にバイアスが生じることが知られており、その利用には注意が必要である（Allison 1995)。

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。

5.3　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	21.55884
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。

5.4 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、各区間内におけるハザード確率が同じ比率をもって指数関数的に増加あるいは減少することを仮定するモデルである。ハザード確率の増減の比率は、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比率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。

```
#delimit;
logit des1 f1-f3 b1.panel b2.educ3a b2.occu_6 i.coresi smam02s lnwage
i.wagem2 if sex==0
;
est store logit

cloglog des1 f1-f3 b1.panel b2.educ3a b2.occu_6 i.coresi smam02s lnwage
i.wagem2 if sex==0
;
est store clog

est tab logit clog, eform star(.10 .05 .01) stats(N ll chi2 df_m) b(%9.4f)

# 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。

表 1 離散時間ロジットモデルならびに離散時間 complementary log-log モデルの出力例

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

5.5 成年者縦断調査への適用例（結婚）

5.5.1 離散時間ロジットモデルと離散時間 CLL モデルの比較

表 2 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		離散時間CLL	
	exp(β)		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

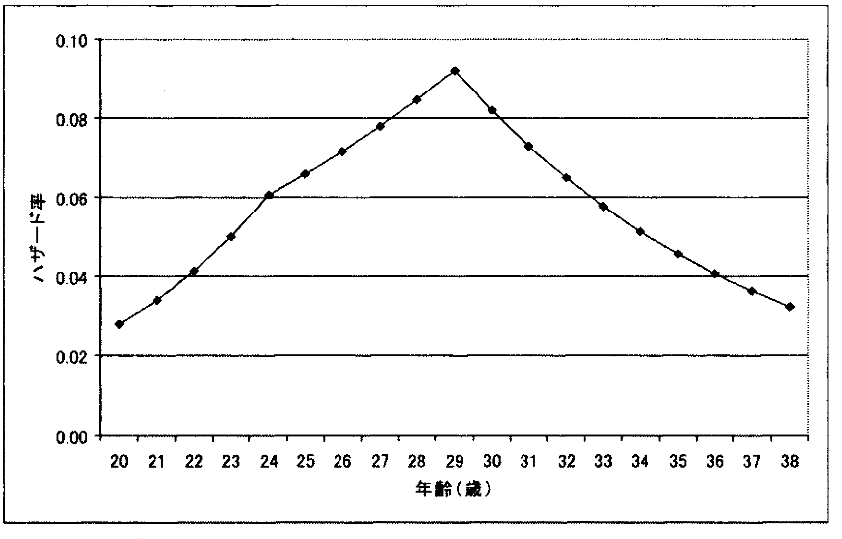
表 2 は、離散時間ロジットモデルならびに離散時間 CLL モデルによる初婚のハザード確率の推定結果を表している。分析結果は回帰係数 b を指数化してえられたハザード・オッズあるいはハザード比によって示した。両モデルの結果は質的にも量的にもほぼ同じ値を示している。このことは年齢別のハザード確率が十分に小さいために、ハザード・オッズとハザード比がほぼ同等に解釈できることを意味する。しかし、この仮定は常に成立するわけではないので、注意が必要である。以下においては、離散時間 CLL モデルの結果を中心に解説する。

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ゼロの場合。

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ゼロ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。