

表7 昭和35年、45年、55年、平成2年および12年の乳幼児身体発育調査による体重の分布(kg)

年・月・日齢	男 子					
	昭和35年	昭和45年	昭和55年	平成2年	平成12年	21世紀パネ ル調査
0年0-6月末まで	7.40	7.80	7.80	7.75	7.79	8.90
0年6-12月末まで	8.80	9.30	9.49	9.39	9.33	9.56
1年0-6月末まで	10.00	10.40	10.50	10.56	10.37	10.65
1年6-12月末まで	11.00	11.60	11.82	11.70	11.43	10.88
2年0-6月末まで	11.60	12.30	12.18	12.33	12.07	12.88
2年6-12月末まで	12.50	13.20	13.27	13.35	13.12	13.16
3年0-6月末まで	13.30	14.10	14.28	14.32	14.13	14.66
3年6-12月末まで	14.20	15.00	15.22	15.28	15.15	15.32
4年0-6月末まで	15.00	15.80	16.12	16.24	16.15	16.83
4年6-12月末まで	15.80	16.60	17.01	17.22	17.27	16.76
5年0-6月末まで	16.60	17.40	17.91	18.27	18.36	-
5年6-12月末まで	17.40	18.20	18.86	19.38	19.48	-
6年0-6月末まで	-	-	19.88	20.60	20.56	-

年・月・日齢	女 子					
	昭和35年	昭和45年	昭和55年	平成2年	平成12年	21世紀パネ ル調査
0年0-6月末まで	6.90	7.30	7.33	7.23	7.18	-
0年6-12月末まで	8.40	8.90	8.91	8.83	8.67	8.62
1年0-6月末まで	9.50	9.90	10.10	9.95	9.86	10.05
1年6-12月末まで	10.40	11.30	11.34	11.09	10.97	10.28
2年0-6月末まで	11.10	11.70	11.89	11.72	11.55	12.38
2年6-12月末まで	12.00	12.60	12.88	12.79	12.58	12.62
3年0-6月末まで	12.90	13.40	13.86	13.83	13.62	14.21
3年6-12月末まで	13.80	14.30	14.82	14.85	14.63	14.90
4年0-6月末まで	14.60	15.20	15.76	15.88	15.73	16.49
4年6-12月末まで	15.40	16.10	16.67	16.92	16.79	16.89
5年0-6月末まで	16.20	17.00	17.55	17.99	17.92	-
5年6-12月末まで	17.00	18.00	18.38	19.11	18.94	-
6年0-6月末まで	-	-	19.15	20.14	20.04	-

表8 昭和35年、45年、55年、平成2年および12年の乳幼児身体発育調査による身長分布(cm)

年・月・日齢	男子					21世紀パネル調査
	昭和35年	昭和45年	昭和55年	平成2年	平成12年	
0年0-6月末まで	65.5	66.7	66.6	67.1	66.8	67.00
0年6-12月末まで	73.1	74.2	74.3	74.3	74.4	74.60
1年0-6月末まで	78.4	80.1	80.1	80.6	80.2	80.01
1年6-12月末まで	83.0	84.9	85.2	85.3	85.5	81.13
2年0-6月末まで	85.0	87.1	87.2	87.4	87.1	88.94
2年6-12月末まで	88.5	90.8	91.1	91.3	91.0	90.21
3年0-6月末まで	91.9	94.4	94.8	95.0	94.7	96.26
3年6-12月末まで	95.0	97.8	98.2	98.6	98.3	98.60
4年0-6月末まで	98.2	101.2	101.5	102.1	101.6	104.09
4年6-12月末まで	101.4	104.3	104.6	105.4	104.9	104.83
5年0-6月末まで	104.4	107.1	107.6	108.6	108.1	-
5年6-12月末まで	107.4	109.6	110.6	111.6	111.4	-
6年0-6月末まで	-	-	113.6	114.5	114.9	-

年・月・日齢	女子					21世紀パネル調査
	昭和35年	昭和45年	昭和55年	平成2年	平成12年	
0年0-6月末まで	64.0	65.2	65.3	65.4	65.4	-
0年6-12月末まで	71.6	73.0	73.0	73.0	72.7	72.00
1年0-6月末まで	77.1	78.7	79.0	79.4	79.1	78.84
1年6-12月末まで	81.4	83.7	84.1	83.9	84.4	79.83
2年0-6月末まで	83.7	86.1	86.3	86.0	86.0	87.76
2年6-12月末まで	87.2	89.5	90.2	90.1	89.9	89.00
3年0-6月末まで	90.7	93.0	93.9	94.0	93.7	95.24
3年6-12月末まで	94.1	96.4	97.5	97.7	97.4	97.64
4年0-6月末まで	97.3	99.8	100.9	101.3	101.0	103.28
4年6-12月末まで	100.4	103.1	104.1	104.7	104.3	104.77
5年0-6月末まで	103.3	106.2	107.1	107.9	107.6	-
5年6-12月末まで	106.3	109.1	109.8	110.9	110.8	-
6年0-6月末まで	-	-	112.2	113.8	113.8	-

表9 体重の成長に関するパネル推定(全体)

Dependent Variable: lnbdywht	Pooling		Between		Random		Fixed	
	Estimated Coefficient	t	Estimated Coefficient	t	Estimated Coefficient	z	Estimated Coefficient	t
survivalday	0.002	1142.57	0.002	332.03	0.002	1312.04	0.002	1272.91
survivalday Sq	-8.64E-07	-682.65	-8.94E-07	-185.5	-8.65E-07	-780.11	-8.64E-07	-754.53
lnkosodate	0.005	10.16	0.002	2.27	0.004	9.46	0.005	8.56
_cons	1.132	1371.13	1.117	567.76	1.131	1368.35	1.131	1439.73
Diagnostics								
Number of observation	175549		175549		175549		175549	
Number of groups			46526		46526		46526	
R-sq within			0.968		0.968		0.968	
between			0.894		0.894		0.894	
overall	0.947		0.947		0.947		0.947	
F test that all $u_i=0$							F(46525, 129020)=2.40	
sigma_u					0.082		0.111	
sigma_e					0.132		0.132	
rho					0.277		0.416	
Breusch-Pagan Lagrangean multiplier test for random effect					chi2(1) = 19850.17 Prob > chi2 = 0.0000			
Hausman Test							chi2(2) = 9.36 Prob > chi2 = 0.0093	

表10 身長の高成長に関するパネル推定(全体)

Dependent Variable: lnbdyght	Pooling		Between		Random		Fixed	
	Estimated Coefficient	t	Estimated Coefficient	t	Estimated Coefficient	z	Estimated Coefficient	t
survivalday	0.001	1312.41	0.001	410.17	0.001	1480.13	0.001	1430.78
survivalday Sq	-3.15E-07	-718.1	-3.29E-07	-211.7	-3.16E-07	-806.28	-3.15E-07	-776.55
lnkosodate	0.001	3.86	0.000	1.1	0.000	2.55	0.000	1.26
_cons	3.901		3.896	6177.05	3.901		3.901	
Diagnostics								
Number of observation	170336		170336		170336		170336	
Number of groups			46459		46459		46459	
R-sq within			0.979		0.979		0.979	
between			0.937		0.937		0.937	
overall	0.967		0.967		0.967		0.967	
F test that all $u_i=0$							F(46458, 123874)=2.20	
sigma_u					0.026		0.037	
sigma_e					0.046		0.046	
rho					0.247		0.401	
Breusch-Pagan Lagrangean multiplier test for random effect					chi2(1) = 15789.23 Prob > chi2 = 0.0000			
Hausman Test							chi2(2) = 8.68 Prob > chi2 = 0.013	

表 11 体重の成長に関するパネル推定(男子)

Dependent Variable: lnbdywht	Pooling		Between		Random		Fixed	
	Estimated Coefficient	t	Estimated Coefficient	t	Estimated Coefficient	z	Estimated Coefficient	t
survivalday	0.002	828.07	0.002	243.95	0.002	937.86	0.002	908.71
survivalday Sq	-8.83E-07	-499.63	-9.17E-07	-137.75	-8.82E-07	-563.28	-8.81E-07	-544.05
lnkosodate	0.004	6.08	0.002	1.41	0.003	4.91	0.003	3.79
_cons	1.145	994	1.130	419.29	1.145	991.44	1.146	1029.7
Diagnostics								
Number of observation	91193		91193		91193		91193	
Number of groups			24185		24185		24185	
R-sq within			0.966		0.967		0.967	
between			0.896		0.896		0.896	
overall	0.947		0.946		0.947		0.947	
F test that all $u_j=0$							F(24184, 67005) = 2.25	
sigma_u							0.079	
sigma_e							0.135	
rho							0.255	
Breusch-Pagan Lagrangean multiplier test for random effect	chi2(1) = 8694.94 Prob > chi2 = 0.0000							
Hausman Test							chi2(2) = 6.29 Prob>chi2 = 0.043	

表 12 身長の成長に関するパネル推定(男子)

Dependent Variable: lnbdyght	Pooling		Between		Random		Fixed	
	Estimated Coefficient	t	Estimated Coefficient	t	Estimated Coefficient	z	Estimated Coefficient	t
survivalday	0.001	950.71	0.001	300.4	0.001	1064.89	0.001	1028.82
survivalday Sq	-3.19E-07	-523.85	-3.37E-07	-156.59	-3.19E-07	-584.22	-3.19E-07	-562.35
lnkosodate	0.000	2	0.000	0.97	0.000	0.61	0.000	-0.81
_cons	3.906	9926.3	3.901	4511.52	3.906	9919.66	3.907	
Diagnostics								
Number of observation	88536		88536		88536		88536	
Number of groups			24148		24007		24007	
R-sq within			0.979		0.979		0.979	
between			0.938		0.938		0.938	
overall	0.967		0.966		0.967		0.967	
F test that all $u_j=0$							F(24147, 64385) = 2.13	
sigma_u							0.026	
sigma_e							0.046	
rho							0.235	
Breusch-Pagan Lagrangean multiplier test for random effect	chi2(1) = 7570.28 Prob > chi2 = 0.0000							
Hausman Test							chi2(2) = 12.43 Prob>chi2 = 0.0020	

表 13 体重の成長に関するパネル推定(女子)

Dependent Variable: lnbdywh	Pooling		Between		Random		Fixed	
	Estimated Coefficient	t	Estimated Coefficient	t	Estimated Coefficient	z	Estimated Coefficient	t
survivalday	0.002	799.5	0.002	231.33	0.002	922.15	0.002	894.71
survivalday Sq	-8.45E-07	-472.44	-8.70E-07	-127.72	-8.45E-07	-542.24	-8.45E-07	-524.46
Inkosodate	0.006	8.37	0.002	1.84	0.006	8.7	0.007	8.55
_cons	1.117	959.15	1.102	392.96	1.115	956.05	1.115	1009.89
Diagnostics								
Number of observation	84356		84356		84356		84356	
Number of groups			22341		22341		22341	
R-sq within			0.969		0.969		0.969	
between			0.896		0.896		0.895	
overall	0.949		0.949		0.949		0.949	
F test that all $u_i=0$							F(22340, 62012) = 2.45	
sigma_u					0.081		0.110	
sigma_e					0.128		0.128	
rho					0.286		0.421	
Breusch-Pagan Lagrangean multiplier test for random effect					chi2(1) = 9888.44 Prob > chi2 = 0.0000			
Hausman Test					chi2(2) = 7.96 Prob > chi2 = 0.019			

表 14 身長に関するパネル推定(女子)

Dependent Variable: lnbdyght	Pooling		Between		Random		Fixed	
	Estimated Coefficient	t	Estimated Coefficient	t	Estimated Coefficient	z	Estimated Coefficient	t
survivalday	0.001	916.83	0.001	286.33	0.001	1030.58	0.001	995.45
survivalday Sq	-3.11E-07	-497.77	-3.22E-07	-146.17	-3.11E-07	-557.12	-3.11E-07	-536.08
Inkosodate	0.001	3.47	0.000	0.67	0.001	3.12	0.001	2.69
_cons	3.896	9684.78	3.891	4324.11	3.895	9655.38	3.895	9899.85
Diagnostics								
Number of observation	81800		81800		81800		81800	
Number of groups			22311		22311		22311	
R-sq within			0.980		0.980		0.980	
between			0.938		0.938		0.938	
overall	0.968		0.968		0.968		0.968	
F test that all $u_i=0$							F(22310, 59486) = 2.15	
sigma_u					0.026		0.037	
sigma_e					0.045		0.045	
rho					0.244		0.399	
Breusch-Pagan Lagrangean multiplier test for random effect					chi2(1) = 7046.58 Prob > chi2 = 0.0000			
Hausman Test					chi2(2) = 0.68 Prob > chi2 = 0.711			

図1 調査回別の身長の高さのヒストグラム

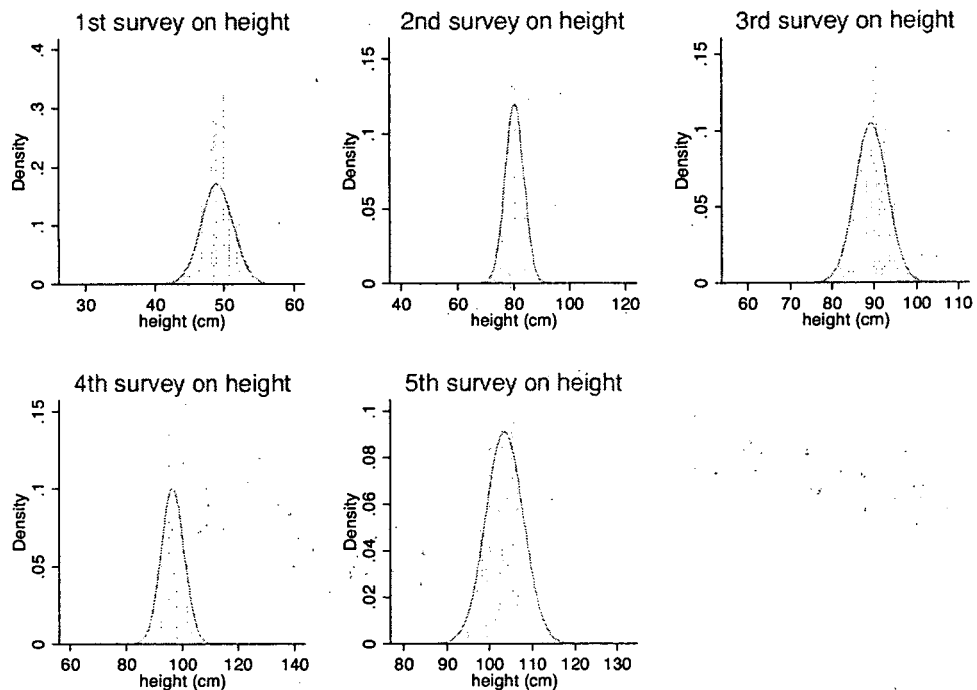


図2 調査回別の体重のヒストグラム

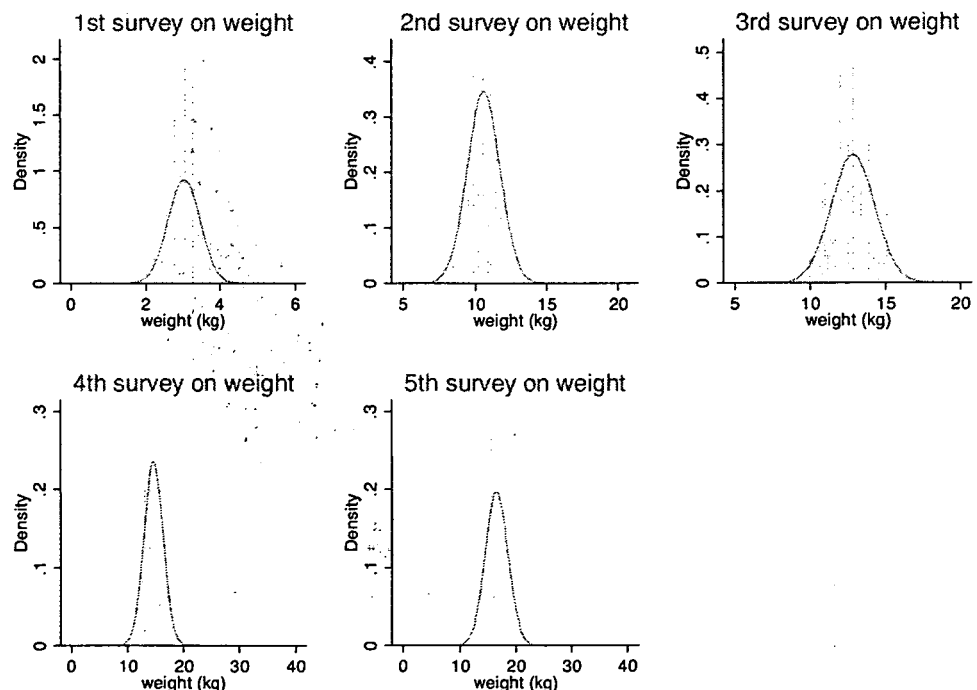


図 3 身長と体重のプロットと統計的推定図

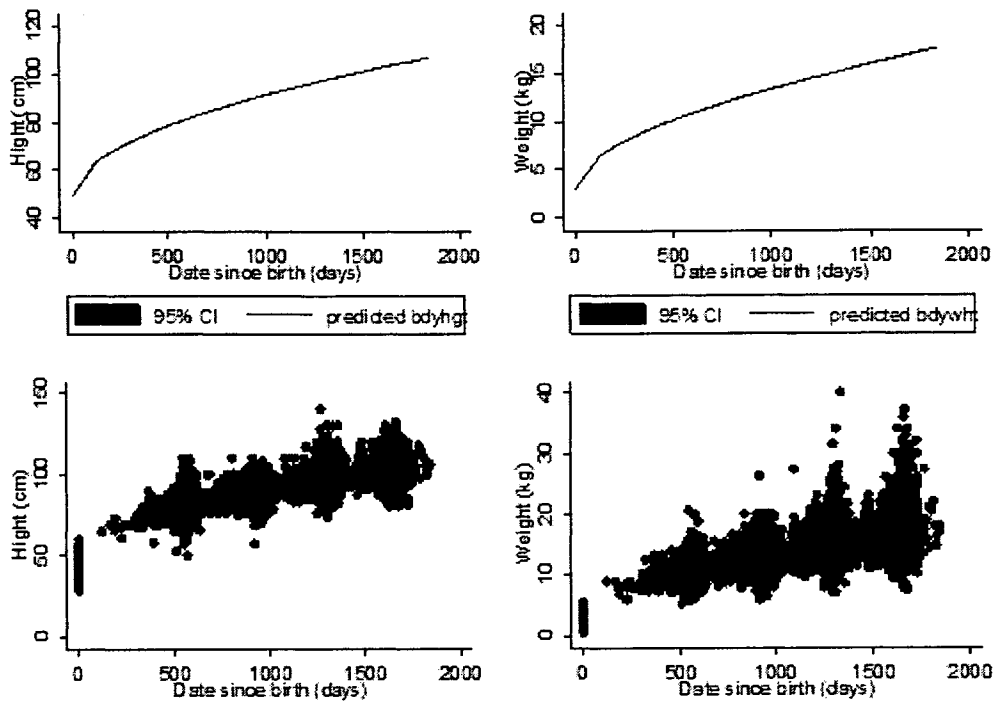


図 4 身長と体重のクロスプロット図

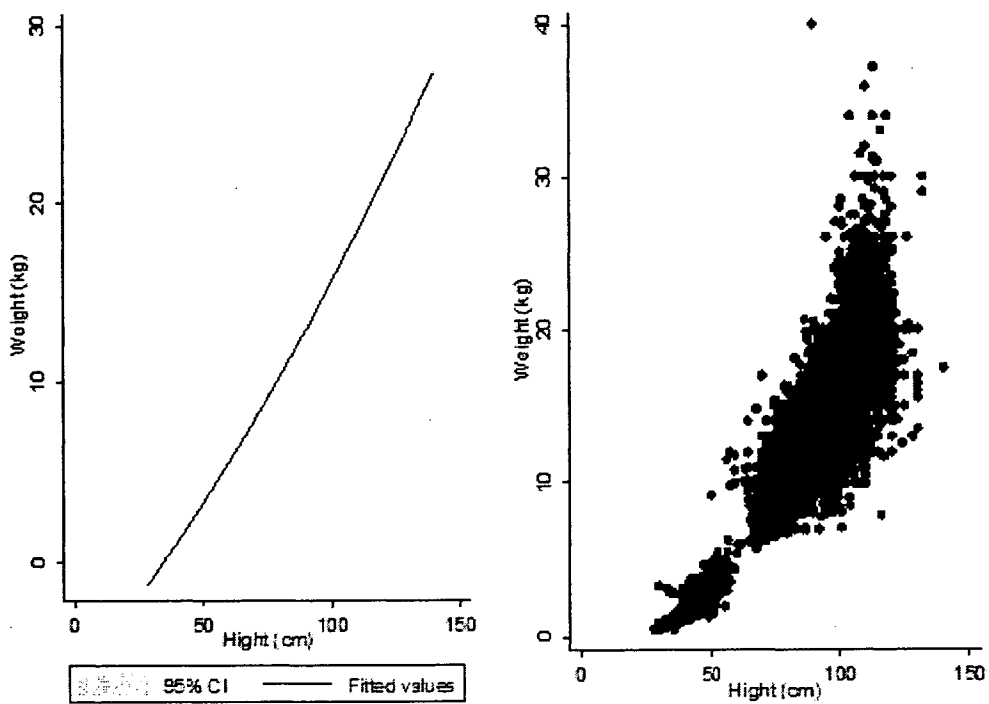


図5 身長と体重の変化のプロット図と統計的推定図

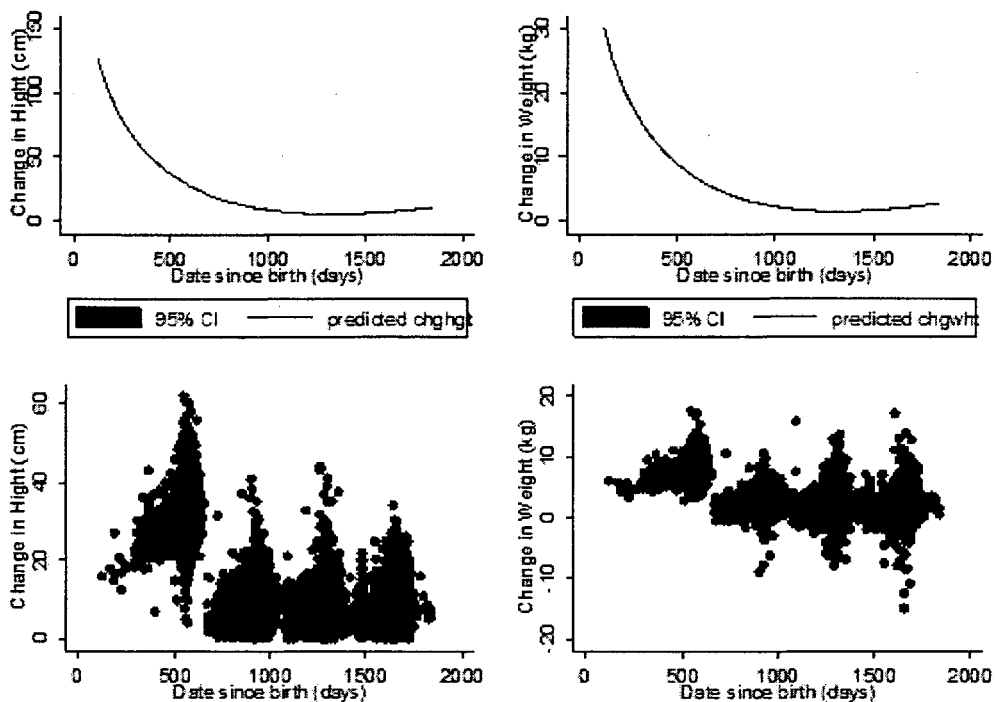
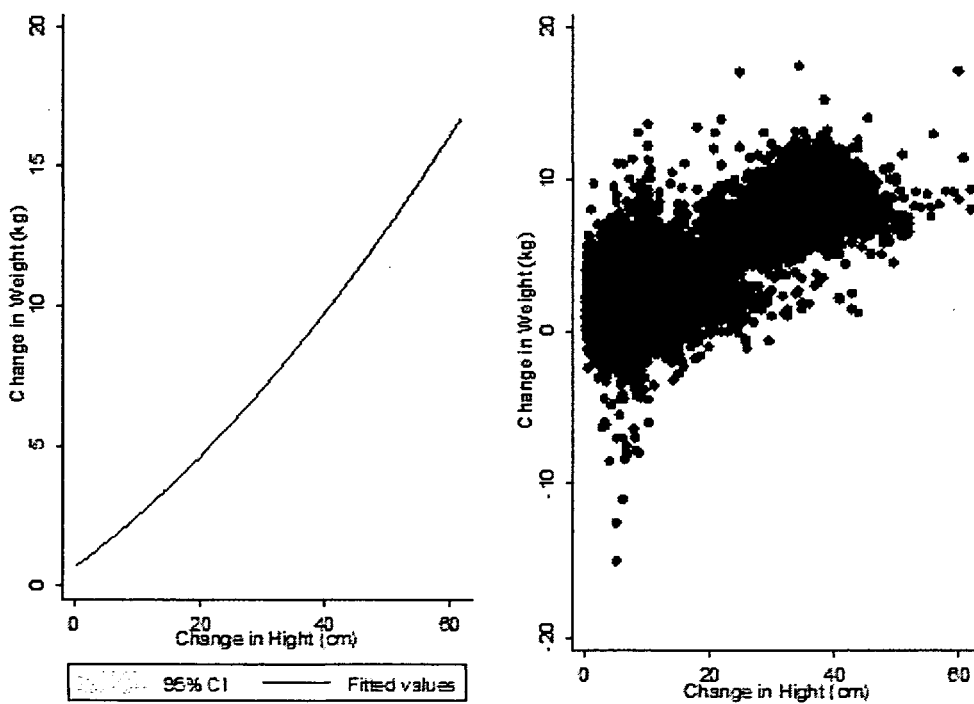


図6 身長と体重の変化のクロスプロット図



6 「21世紀成年者縦断調査」を用いた初婚の要因分析： ネステッド・ロジットモデルによる初婚と脱落の競合ハザード分析

福田 節也

1. はじめに

パネル調査において脱落は不可避の問題である。脱落はサンプル数の減少やサンプル属性の偏向をもたらす。そのため、記述統計やクロス集計といった単純統計の算出においては、脱落の取り扱いが分析者にとって喫緊の課題である¹。一方、多変量解析によるパラメータ推定においては、脱落による影響はそれほど深刻ではないことが報告されている (Fitzgerald et. Al. 1998, Lillard and Panis 1998, Alderman et. Al. 2001, Falaris 2003)。いずれの研究においても、モデルが正しく設定されている限り、脱落は多変量解析による回帰係数の推定には有意な影響を与えないとの結論を得ている。しかし、この結論は多変量解析において脱落による影響を無視してよいということを意味するのではない。使用する分析手法や従属変数によっては、脱落は分析の妥当性を左右する重要な問題となりうる。

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の推定にバイアスをもたらす。本稿では、イベントヒストリー分析の一種である離散時間ロジットモデル (discrete-time logit model) を用いた初婚の要因分析を例として、脱落がパラメータ推定に与える影響について明らかにする。具体的には、1) 脱落をセンサリングとして扱い、通常の離散時間ロジットモデルを行った場合と、2) 脱落を初婚に競合するイベントとして扱い、双方のパラメータを同時推定 (simultaneous estimation) によって求める場合で初婚要因のパラメータを比較する。なお、後者のモデルにおいては、Hill ら (1993) によって提案された離散時間ネステッド・ロジットモデル (discrete-time nested logit model) を適用する²。分析によって、1) 初婚と脱落が独立に生起しているのかを統計的に検定し、両者が非独立であった場合、2) 脱落をセンサリングとして扱うと、初婚要因のパラメータ推定にどの程度バイアスが生じるのかを提示する。

¹ この問題への対処法としては、脱落サンプルの属性を加味したウェイトの使用などが提案されている (Lepkowski and Couper 2002, Wooldridge 2002)。

² Hill ら (1993) は、SURF (Shared Unmeasured Risk Factors) モデルの呼称を用いている。

2. パネルデータを用いた離散時間ロジットモデル

2-1. 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) 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。

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

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³ (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、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)。

2-2. 離散時間ロジットモデル

イベントヒストリー分析にはいくつかのモデルがある。本論文において用いるのは、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 (Allison 1982) である。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。

$$\ln[P_t / (1 - P_t)] = a_t + b_1 X_1(t) + b_2 X_2(t) + \dots + b_k X_k(t) \quad \dots (2)$$

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⁴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 において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 (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、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。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、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

³ 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) と呼ぶことにする。

⁴ 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。

する必要がある⁵。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。

離散時間ロジット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⁶。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ット分析を行う。なお、パラメーターの推定は最尤法によって行われる。

離散時間ロジットモデルは、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⁷。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。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ロジットモデルを利用することが最も簡便かつ実際的である⁸。

⁵ 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)。

⁶ 人・期間別データの詳細については、山口 (2001b) を参照されたい。

⁷ 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、連続時間モデルの適用が可能である。しかし、イベント生起の有無については明らかであっても、発生月に関する情報が欠損している場合には、なんらかの仮定によって、発生月を代入する必要がある。また、後に述べるように本稿では脱落を初婚に競合するイベントとして扱う。脱落をイベントとして扱う場合、調査と調査の間でいつ脱落が発生したのかを論理的に定義することは不可能である。そのため、本稿における分析では、離散時間モデルを適用することが最も合理的な選択となる。

⁸ Allison (1982) は、ロジットモデルの代わりに complementary log-log モデルを使用して同様の分析が可能であることを示している。complementary log-log モデルを使用した離散時間ハザードモデルによって算出される係数 β は、連続時間を仮定する等比ハザードモデルから算出された係数 β の近似値となる。そのため、離散時間ロジットモデルの $\exp(\beta)$ はハザード確率のオッズ比を表すが、離散時間 complementary log-log モデルの $\exp(\beta)$ は、連続時間を仮定する等比ハザードモデルと同様にハザード比として直接的に解釈できるという利点がある (Allison 1995)。近年の統計パッケージには、complementary log-log モデルをサポートするものも多いため、併せてその使用を検討すべきである。

2-3. 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 (2002b) 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離散時間ネステッド・ロジットモデル (以下、ネステッド・ロジットモデル) を行う。以下に山口 (2002b) を参照しつつ、どのような場合に各方法を使用す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いずれ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、 $(1 - P_A(t)) \times (1 - P_B(t))$ 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条件の $(1 - P_A(t)) \times (1 - P_B(t))$ 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)。ただし、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 $P_C(t) (=1-P_A(t) \cdot P_B(t))$ 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 の選択肢であるネステッド・ロジットモデルである。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 の方法とネステッド・ロジットモデルを用いる第 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、ネステッド・ロジットモデルでは脱落と初婚の非観察要因 (shared unmeasured risk factors)⁹ 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、ネステッド・ロジット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。次節ではネステッド・ロジットモデルの概要について述べる。

3. 離散時間ネステッド・ロジットモデルの概要

3-1. 概要

離散時間ネステッド・ロジットモデルとは、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 (2002b) を参照しつつ、その概要について述べる。

⁹ 連続時間を仮定した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。

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。その場合、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 (3) 式によって表すことができる。

$$\begin{aligned}
 S_{t0i} &= \beta_0^* X_{t0i} + \varepsilon_{t0i} \\
 S_{t1i} &= \beta_1^* X_{t1i} + \varepsilon_{t1i} \\
 &\vdots \\
 S_{tmi} &= \beta_m^* X_{tmi} + \varepsilon_{tmi}
 \end{aligned} \tag{3}$$

S_{tmi} は、説明変数の分散によって説明される部分 $\beta_m^* X_{tmi}$ と誤差分 ε_{tmi} とに分けられる。式

(3) 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、ネステッド・ロジットモデルを得る¹⁰。

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 (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。

ネステッド・ロジットモデルでは、非類似係数 (index of dissimilarity) ρ を説明変数の

¹⁰ ネステッド・ロジットモデルの数式的展開については、Hill ら (1993) や山口 (2002b) を参照のこと。

係数と同時に推定する。 ε_1 と ε_2 の相関係数は $1-\rho^2$ 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 2002b)。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、ネステッド・ロジットモデルにおける係数は、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。

3-2. 適用手順

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ら(1993)や山口(2002b)によって示されているネステッド・ロジットモデルの適用手順を以下に示す。

(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は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、初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、初婚か脱落が生起したとして、それが脱落ではなく初婚である確率を推定するモデルとなる。

(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。

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

$$z_2(t) = \log[1 + \exp(\sum_k b_k x_k(t))]$$

この時、 $\sum_k b_k x_k(t)$ は(2)のモデルで得られた各レコードごとの予測値を表す。 $z_1(t)$ と $z_2(t)$ の値を(1)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、各レコードに対して変数として付帯する。さらに、このデータに初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 (3) 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競合イベントへの決定要因 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競合イベントへの決定要因 β_{2k} を得る。なお、(2) と (4) 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 と z_2 の係数として算出されるのが ρ の推定値である。 ρ は z_1 と z_2 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。初婚と脱落の観察されない異質性 (誤差項) の相関係数は、 $1 - \rho^2$ によって与えられる。また、この時 ρ の標準誤差は ρ が 1 である、つまり初婚と脱落の相関係数が 0 であるという帰無仮説を検定している。

3-3. 利用における留意点

ネステッド・ロジット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、このモデルは連続時間を仮定する競合ハザードモデルには適用できない。離散時間モデルによる同時的な競合ハザード分析では、人・期間別データに対する多項ロジット、多項プロビット、あるいは多項 complementary log-log といったモデルを適用する。これらのモデルでは、いずれも回帰式に何らかのパラメトリックな分布をもつ誤差項 ε を仮定している。そのため、競合イベントのそれぞれについて誤差項、すなわち非観察要因の影響力に関する値を個人別に算出し、その相関係数を推定することができる。しかし、通常の連続時間ハザードモデルでは誤差項が仮定されない (Allison 1995)。連続時間モデルでは、誤差項の分布はベースライン・ハザード関数によって部分的に表現されるに留まり、明示的に定義されないためである (Blossfeld and Rowher 2002)。そのため、通常の連続時間モデルでは、競合するイベントごとに誤差項を算出することができず、したがって、その相関係数を推定することもできない¹¹。

次に、ネステッド・ロジットモデルでは、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、初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

¹¹ 連続時間ハザードモデルにおいて、パラメトリックな分布をもつ誤差項を仮定するモデルは、いくつか紹介されている (Blossfeld and Rowher 2002)。なかでも、誤差項の分布についてのパラメトリックな仮定が少ない Heckman と Singer (1984) によるものが代表的である。近年において、Heckman と Singer (1984) のモデルを拡張した分析手法が注目を集めている。例えば、Lillard と Waite (1993) は、誤差項を仮定した連続モデルを用いて、ハザード率の同時回帰モデルを行っている。このモデルでは、依存関係にある複数のハザード率を対象として、観察できない異質性の相関を統制した上で、各ハザード率のパラメーター推定を行うことが可能である。この方法では、対象とするハザード率は競合するイベントに限らず、同時推定を行うことが可能である。そのため、あるイベントが他のイベントに対しても選択バイアス (内生性 Endogeneity) を統制した上で、因果関係を検証できるという利点をもつ (山口 2002a)。しかし、その適用には aML (Lillard and Panis 2003) という特殊な統計ソフトが必要となる。なお、aML は以下のサイトから無料でダウンロードして使用することができる。

<http://www.applied-ml.com/index.html>

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指摘されている(Hill et. al. 1993)。

さらに、ネステッド・ロジットモデルでは、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 2002b)。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 $1 - \rho^2$ は正であることが仮定されている¹²。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、ネステッド・ロジットモデルでは分析することができない¹³。

また、ネステッド・ロジット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 2002b)。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。

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 2002b)。そのため、本稿における分析ではオッズ比ではなく、係数を用いて解釈を行う。離散時間モデルにおける係数は、ハザード確率のログオッズに対する共変量の影響力を表しており、共変量 X_k における1単位の増加はイベント生起のログオッズを β_k 分増加(減少)させると解釈することができる。

4. データと分析対象

分析には、厚生労働省が2002年11月より実施している「21世紀成年者縦断調査」(以後、成年者調査)の第1回から第4回(パネル1からパネル4)までの個票データを使用した。成年者調査は、男女の結婚、出産、就業等の実態及び意識の経年変化の状況を把握することを目的として実施されているパネル調査である。同調査は、全国より無作為に抽出された20・34歳(平成14年10月末日現在)の男女及びその配偶者35,448人を対象としており、第1回調査では対象サンプルの83.4%にあたる29,566人より回答を得ている。調査は毎年11月に行われており、前回調査の回答者を対象として継続的に回答を得ている。

¹² 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 2002b)。

¹³ 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る。

調査方法は調査員による留置き法を基本とし、第1回調査以降に転出した者に対しては郵送法が併用されている¹⁴。成年者調査はわが国における既存のパネル調査の中でも最大規模の調査の1つである。

分析対象は、第1回調査において未婚である女性である¹⁵。未婚女性のうち、調査期間を通じて性別や年齢の回答に不備がある者、ならびに分析に用いる説明変数に外れ値や欠損値がある者を除外したところ、分析の対象となるサンプル数は7,588人であった。

分析では、初婚ならびに脱落を競合するイベントとして取り扱い、第1回調査における未婚女性が結婚するか、調査から脱落するか、未婚のまま第4回調査を向かえるまでをリスク期間とした離散時間ロジットモデルならびに、その応用的分析手法であるネステッド・ロジットモデルによるイベントヒストリー分析を行う。

なお、パネルデータに対するイベントヒストリー分析では、かなりの割合で左打ち切り(left-truncation)が発生することに留意する必要がある。左打ち切りとは、観察期間より前にイベント生起のリスクが開始していることをいう(Guo 1993, Allison 1995)。本調査においては、20-34歳の女性を分析対象としている。女性の法定結婚年齢は16歳であるため、実際には同調査におけるすべての未婚女性が左打ち切りのケースといえる。結婚リスクの高いサンプルは調査開始までの間に結婚してしまう可能性が高いため、左打ち切りのケースはイベント発生リスクの低いケースに限定される可能性が指摘されている(Allison 1995)。しかし、同調査においては第1回調査のサンプルは無作為に抽出されていることから、調査開始時における未婚女性にはこのようなセレクション効果は生じておらず、無作為に選択されていると仮定できる。また、Guo (1993)はイベントのリスク開始時点が明らかである場合、条件付き尤度法(conditional likelihood approach)を用いることで左打ち切りによる偏りなく推定が行えることを示している¹⁶。本分析においては、最も早く初婚が観察される年齢である20歳を初婚リスクの開始時点と仮定して、条件付き尤度法によるパラメーター推定を行う¹⁷。

¹⁴ 留置き法では、調査員があらかじめ配布した調査票に被調査者が自ら記入し、密封したものを後日調査員が回収する留置き法によって回収されている。また、第1回調査以降に転出した者については、厚生労働省から郵送された調査票に被調査者が自ら記入し、郵送により厚生労働省に提出する方法を用いている(厚生労働省大臣官房統計情報部 2005)。

¹⁵ 「21世紀成年者縦断調査」では結婚歴に関する項目がないため、未婚者と既婚の独身者を区別することができない。そのため、本稿では子どものいない独身女性を未婚女性と定義する。「2000年国勢調査」(総務省統計局, 2001)によると、20-34歳の男女のうち、独身者に占める離死別者の割合は男が1.63%、女が4.41%である。本分析では、子どもがいる独身者を分析より除外しているため、分析サンプルにおける既婚者の割合はこれと同程度かそれ以下と推測される。

¹⁶ Guo (1993)の条件付き尤度法による分析手法は、山口(2001a)において、区間調査法によるイベントヒストリー分析として示されている。

¹⁷ ただし、今回の分析では調査開始から3年間に生じた初婚と脱落のみを取り扱う。そのため、各年齢における初婚経験は対応する出生コーホートに限定して生起している。例えば、34歳以降の結婚を経験できるのは、調査開始時に34歳であった1968年出生コーホートの女性のみである。同様にその他の年齢においても、最大で3つの単年コーホートにしか結婚が経験されない。そのため、ハザード確率の年齢効果はコーホート効果と分離することができない。このことは年齢の係数が、普遍的な効果ではなく、コーホートに特有の効果である可能性があることを意味する。

これらの分析を通して、1) 初婚と脱落の生起過程にはどの程度の相関があるのか、2) 脱落を右センサーとして扱い、通常の離散時間ロジットモデルによる分析を行った場合、初婚要因のパラメーター推定にはどの程度のバイアスが生じるのかを明らかにする。

5. 記述統計

5-1. 初婚と脱落の推移

表 1 初婚と脱落の発生頻度

		未婚→			合計
		未婚	結婚	脱落	
パネル1→2	n	6110	316	1162	7588
(2002年-2003年)*	%	80.5	4.2	15.3	100.0
パネル2→3	n	4926	304	880	6110
(2003年-2004年)	%	80.6	5.0	14.4	100.0
パネル3→4	n	4069	283	574	4926
(2004年-2005年)	%	82.6	5.8	11.7	100.0
合計	n	15105	903	2616	18624
(2002年-2005年)	%	81.1	4.9	14.1	100.0
*:各年とも11月時点		X ² (4)=84.02, p<0.01			

表 1 は、各調査間における結婚と脱落の発生頻度を表している。パネル 1 において未婚であった女性のうち、4.2%にあたる 316 人がパネル 2 までの間に結婚している。一方、パネル 2 に回答しなかった脱落ケースは 1,162 人と全体の 15.3%に達している。同様に、パネル 2 の時点で未婚であった女性のうち、パネル 3 までの間に結婚したケースは 304 人 (5.0%) であり、脱落は 880 ケース (14.4%) であった。パネル 3 からパネル 4 においては、結婚が 283 ケース (5.8%)、脱落が 574 ケース (11.7%) となっている。

未婚女性に対する結婚と脱落の発生頻度をみると、調査回が進むにつれて、結婚が増加し、脱落が減少する傾向がみられる。調査回が進む毎に結婚が増加しているのは、より結婚が起きやすい 20 歳代後半へと回答者の年齢がシフトしていることを反映しているものと思われる。また、脱落については、調査に非協力的な回答者は早い段階で脱落するため、調査が回を重ねるにつれて、調査に協力的な回答者が残ることとなる。そのため、あとの調査回ほど脱落が少ないものと思われる。

減少傾向にあるとはいえ、調査からの脱落率が各年で 10%を超えていることに留意する必要がある。結婚と脱落の各生起確率のうち、どちらかあるいは一方が無視できるほど小さい場合には、それぞれのイベントが独立に生起していると仮定して、一方のイベントの生起を他方のイベントのセンサーされたケースとして取り扱うことが可能である (山口 2002b)。しかし、ここでは脱落の生起確率が比較的大きく、また初婚の生起確率も無視できるほどには小さくはない。そのため、脱落をセンサーとして扱い、初婚をイベントとす