4.6 成年者縦断調査への適用例

4.6.1 離散時間ロジットモデルとSURFモデルの比較

初婚ハザード確率に関する離散時間ロジットモデルならびに2段階推定によるSURFモデルの推定結果を表1に示した。表1の1列目は脱落を右センサリングとして扱った通常の離散時間ロジットモデルの結果を示している。第2列目以降は、SURFモデルの推定結果となる。第2列目はSURFモデルの第1段階推定である結婚対脱落のモデルになる。第3列目と第4列目はSURFモデルの第2段階推定の結果で、それぞれ初婚のハザード確率と脱落のハザード確率を推定している。z_1 ならびに z_2 の係数値は、0.611で一致しており、0以上1未満の数値であることから、SURFモデルが成功裏になされたことが分かる。また、この値は統計的有意水準5%で0とは異なり、10 %で1とは異なる値であることから、結婚と脱落の誤差項の間には0.63（=1-0.611^2）の正の相関があると推定される。ここでは、第1列目と第3列目の結果を比較することで、脱落と結婚の生起過程における相関を無視した場合とこれを統制した場合とで結果がどのように異なるのかをみてみよう。

z_1 と z_2 の係数が1である、つまりIIAが仮定できるという帰無仮説は、統計的有意水準10 %でかろうじて棄却できるという水準であるものの、第1列目と第3列目の分析結果の解釈は質的に大きく異なるものとなっている。その理由については、ここでは結婚と脱落の非観察要因の相関に時間的な変化がないものと仮定した分析を行っているが、実際にはこれらの相関に時間による変化があるためと思われる。この仮定からの逸脱は、時間固定変数のパラメーター推定には影響を与えないが、時間依存変数のパラメーター推定には何らかのバイアスがかかっている可能性を示唆する(Hill et al. 1993)。

年次の効果をみると、通常の離散時間ロジットモデルでは2005年以降結婚が生起しやすくなっていることを示しているが、SURFモデルの結果は、これは調査回数が進むごとに脱落が減少することによってもたらされる見せかけの効果であることを示している。脱落が後半の調査で起きにくくなる傾向を統制すると、2002年から2007年までの期間における結婚の期間効果は認められない。学歴の影響についてみると、高卒女性と比べて、短大専門学校卒の女性の初婚ハザードは、離散時間ロジットモデルでは有意に異ならないが、SURFモデルでは1 %水準で有意に高いとの結果を得ている。

また、親との同別居が結婚に与える影響が両者では大きく異なっている。離散時間ロジットモデルでは親との同別居は結婚に対して全く有意な影響を与えていないが、SURFモデルでは両親と同居している女性は、親から独立している女性に比べて結婚のハザード率が有意に低いことが示されている。また、居住都道府県のSMAMの影響は離散時間ロジットモデルでは負の影響を与えており、1 %水準で有意であるが、SURFモデルではその影響が有意とはなっていない。年間勤労所得の影響については、SURFモデルでは弱まっているが有意性は1 %に保たれている。しかし、年間勤労所得不明ダミーの影響がSURFモデルでは有意ではなくなっている。これらの結果は、各共変量の影響についての解釈が、どちらのモデルを使用するかで大きく異なる可能性を示している。

表 1 女性の初婚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 SURF モデルの推定結果

	離散時間ロジットモデル		SURFモデル	
	(1)	(2)	(3)	(4)
	結婚ハザード	結婚(対:脱落)	結婚ハザード	脱落ハザード
	b	bi	b1	b2
年齢スプライン				
20-25歳	0.194 ***	0.180 ***	0.120 **	0.011
25-30歳	0.087 ***	0.054 **	0.076 ***	0.043 ***
30-39歳	−0.120 ***	−0.076 ***	−0.104 ***	−0.057 ***
年次(対:2002-03年)				
2003-04年	0.072	0.097	0.059	0.000
2004-05年	0.137	0.357 ***	0.053	−0.165 ***
2005-06年	0.153 *	0.399 ***	0.001	−0.243 ***
2006-07年	0.213 **	0.419 ***	0.070	−0.186 ***
学歴(対:高校卒)				
中学校卒	−0.007	0.036	0.076	0.054
短大・専門学校卒	0.088	0.142 *	0.147 ***	0.060
大学・大学院卒	0.220 ***	0.323 ***	0.292 ***	0.095 *
職業(対:中小企業雇用)				
大企業雇用	−0.151	−0.266 **	−0.114	0.048
専門・技術職	0.173 **	0.208 *	0.074	−0.053
自営・家従・会社役員	0.033	−0.039	0.105	0.128
非正規雇用	0.012	−0.109	0.031	0.098 **
無職	0.097	−0.218	0.155 *	0.288 ***
学生	−0.549 ***	−0.651 ***	−0.345 *	0.053
不明	0.209 *	0.014	0.217 ***	0.209 ***
親との同別居				
(対:親と別居)				
両親と同居	0.010	0.690 ***	−0.432 ***	−0.854 ***
片親と同居	0.110	0.707 ***	−0.287 *	−0.719 ***
不明	0.044	0.248 *	−0.154 *	−0.305 ***
居住都道府県のSMAM-28	−0.287 ***	−0.450 ***	−0.158	0.117 ***
Ln(年間勤労所得)	0.228 ***	0.186 ***	0.155 ***	0.042
年間勤労所得不明ダミー	−0.401 ***	−0.705 ***	−0.156	0.274 ***
年間勤労所得ゼロダミー	−0.653 ***	−0.559 **	−0.522 ***	−0.181
定数	−4.854 ***	−2.924 ***	−3.327 ***	−1.540 ***
$z1^{*1}$			0.611 *	
$z2^{*1}$				0.611 *
person−year数	26843	4942	26843	26843
カイ2乗値	332.822	471.8969	496.6502	496.6502
自由度	24	24	25	25

* p<.1; ** p<.05; *** p<.01
*1: *z1*および*z2*においては、係数が1と有意に異なるか否かの検定を行い、p値を算出した。

4.6.2 離散時間多項ロジットモデルと SURF モデルの比較

次に、結婚と脱落のハザード確率に対する離散時間多項ロジットモデル（IIA モデル）と SURF モデルの推定結果の比較を表 2 に示した。対象とするイベントと競合するイベントのハザード確率が無視できるほど小さくない場合には、競合イベントをセンサリングとして扱うよりも、多項ロジットモデルによる競合リスクモデルを適用する方が理想的である。しかし、離散時間多項ロジットモデルでは、競合するイベント間で非観察要因に相関がないことを仮定している（IIA の仮定）。ここでは、この IIA の仮定が満たされない場合（z_1 と z_2 が統計的に有意に 1 よりも小さく、0 よりも大きい場合）に分析結果にはどのような違いがもたらされるのかを示す。

表 2 をみると、IIA モデル（離散時間多項ロジットモデル）の結果は、表 1 の離散時間ロジット

モデルに非常に近いものとなっている。そのため、SURFモデルと比較すると、離散時間ロジットモデルの結果と比較した場合とほぼ同じような相違がみられる。ただし、脱落の推定モデルにおいては、IIAモデルとSURFモデルで非常に近い結果が得られている。したがって、やはり離散時間多項ロジットモデルを用いたとしても、結婚と脱落の非観察要因に相関がある場合には、モデルの選択が結婚要因の推定結果の質的解釈に大きな影響を与えることが示唆される。

表2 女性の初婚ハザード確率に対する
離散時間多項ロジットモデル（IIA）ならびにSURFモデルの推定結果

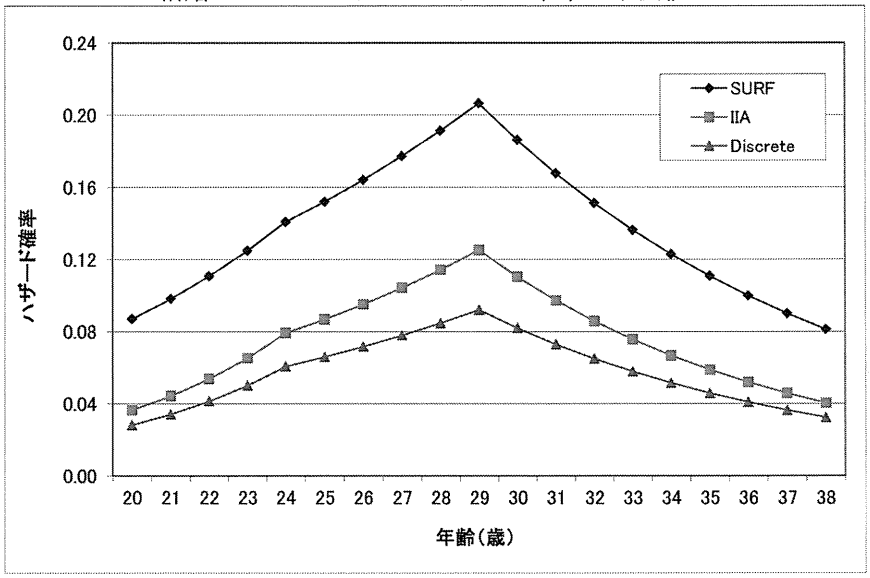
	IIAモデル		SURFモデル	
	(1) 結婚	(2) 脱落	(3) 結婚	(4) 脱落
	β1	β2	b1	b2
年齢スプライン				
20-25歳	0.194 ***	−0.001	0.120 **	0.011
25-30歳	0.091 ***	0.034 ***	0.076 ***	0.043 ***
30-39歳	−0.126 ***	−0.046 ***	−0.104 ***	−0.057 ***
年次(対:2002-03年)				
2003-04年	0.071	−0.001	0.059	0.000
2004-05年	0.111	−0.186 ***	0.053	−0.165 ***
2005-06年	0.114	−0.296 ***	0.001	−0.243 ***
2006-07年	0.180 *	−0.239 ***	0.070	−0.186 ***
学歴(対:高校卒)				
中学校卒	0.004	0.077	0.076	0.054
短大・専門学校卒	0.098	0.076 *	0.147 ***	0.060
大学・大学院卒	0.235 ***	0.114 **	0.292 ***	0.095 *
職業(対:中小企業雇用)				
大企業雇用	−0.144	0.061	−0.114	0.048
専門・技術職	0.160 *	−0.097	0.074	−0.053
自営・家従・会社役員	0.052	0.153	0.105	0.128
非正規雇用	0.025	0.104 *	0.031	0.098 **
無職	0.139	0.298 ***	0.155 *	0.288 ***
学生	−0.540 ***	0.071	−0.345 *	0.053
不明	0.238 **	0.204 **	0.217 ***	0.209 ***
親との同別居 (対:親と別居)				
両親と同居	−0.146	−0.945 ***	−0.432 ***	−0.854 ***
片親と同居	−0.029	−0.798 ***	−0.287 *	−0.719 ***
不明	−0.019	−0.345 ***	−0.154 *	−0.305 ***
居住都道府県のSMAM-28	−0.264 ***	0.154 ***	−0.158	0.117 ***
Ln(年間勤労所得)	0.230 ***	0.013	0.155 ***	0.042
年間勤労所得不明ダミー	−0.346 ***	0.336 ***	−0.156	0.274 ***
年間勤労所得ゼロダミー	−0.671 ***	−0.121	−0.522 ***	−0.181
z1*1			0.611 *	
z2*1				0.611 *
定数	−4.625 ***	−1.446 ***	−3.327 ***	−1.540 ***
person-year数	26843		26843	26843
カイ2乗値	974.5784		496.6502	496.6502
自由度	48		25	25

* p<.1; ** p<.05; *** p<.01
*1: *z1*および*z2*においては、係数が1と有意に異なるか否かの検定を行い、p値を算出した。

4.6.3 モデル予測値における相違

最後に、モデルから推定されるベースライン・ハザードや累積生存確率にどのような違いがみられるのかについて考察する。

図 1 SURF モデル、IIA モデル、離散時間ロジットモデルによる
結婚のベースライン・ハザード確率の予測値 *



* 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。

図 1 は、SURF モデル、IIA モデル、離散時間ロジットモデルによる結婚のハザード確率の予測値を示している。離散時間ロジットモデルでは、結婚のベースライン・ハザード確率が最も低く推定されている。これは、結婚と脱落の生起過程に相関があり、独立とはみなせないにも関わらず、脱落を右センサーとして扱うことによって、回帰係数にバイアスが生じているためである。IIA モデルにおいては、これが若干軽減されているものの、SURF モデルで推定される値よりは依然として低い値を示している。

図 1 を累積生存確率によってみたものが、図 2 である[*6]。各累積生存確率は、Curde Mar が離散時間ロジット、Net Mar（IIA）が IIA モデル、Net Mar（SURF）が SURF モデルに対応している。また、Crude M&A は、結婚か脱落のいずれかが生起するハザード確率をもとに算出された累積生存確率を示す。39 歳時における累積生存確率は、それぞれ 32%、25%、9%、0.4% となっている。

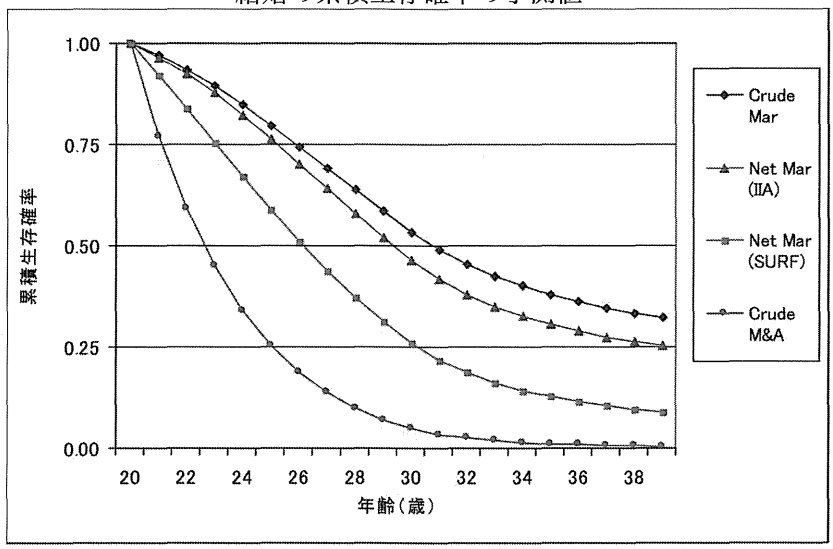
図 2 の中で最も高い生存曲線を描いている Crude Mar は未婚をベースとして結婚を減少要因とした単要因減少の生存確率と考えることができる。一方、最も低い生存曲線を描く Crude M&A は、未婚をベースとして結婚と脱落を減少要因とした多要因減少の生存確率と捉えることができる。また、これらに挟まれた 2 本の生存曲線である Net の生存確率は、脱落という選択肢がないと仮定した場合における結婚の生存確率を示している。Net Mar(SURF) と Net Mar(IIA) で異なる値が得られるのは、脱落した女性が結婚と未婚のどちらに分類されるかについての扱いが、両モデルで異なるためである（Hill et al. 1993）。IIA モデルでは、ほとんどの脱落サンプルを未婚状態とみなしている。なぜならば、IIA モデルにおいては、結婚と脱落は全く異なるイベントである

*6 SURF モデルの生存確率の算出方法は、通常の離散時間モデルとは異なる。詳しくは Hill 等（1993）を参照されたい。

と仮定されており、脱落サンプルが結婚と未婚のどちらに分類されるかは、非脱落サンプルにおける結婚と未婚の割合に応じて決定されている。データでは非脱落サンプルの94%が未婚状態にあるため、脱落のほとんどが未婚と分類されている。そのため、Net Mar(IIA) と Crude Mar は非常に近い値を示している。一方、SURFモデルによる生存曲線は、Crude Mar と Crude M&A の中間くらいに位置している。これは、SURFモデルでは、IIAモデルに比べて、より多くの脱落サンプルが結婚として扱われているためである。これは非観察要因に関する限り、結婚と脱落が類似したイベントであり、モデルでは脱落サンプルの約6割（r=0.63）が脱落しなければ結婚していたと推定されているためである。

また、Crude M&A の生存曲線は、39歳時における生存率がわずかに0.4%であることを示している。このことは、パネル1からパネル6までに観察されたペースで結婚と脱落が生起すると仮定した場合、20歳で調査に参加した未婚女性が39歳まで結婚も脱落も経験せずにいる確率が0.4%しかないことを意味する。

図2 SURFモデル、IIAモデル、離散時間ロジットモデルによる
結婚の累積生存確率の予測値 *



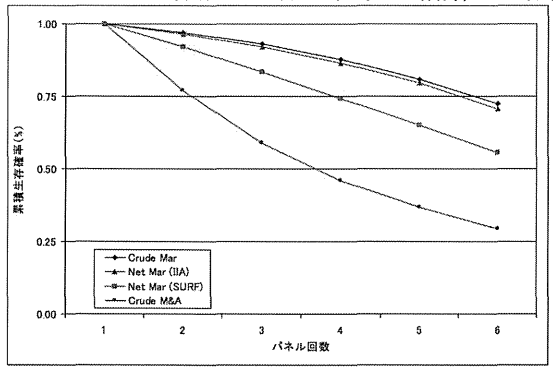
* 年間勤労所得が300万円で、年齢を除く他の共変量がすべて基準カテゴリーである場合。

ところで、これまで行ってきた結婚の分析では、パネル1からパネル6までの5年間の期間観察をあたかもコーホートの行動であるかのように解釈している点において、若干の注意が必要である。図1や図2の結果は、あくまで6年間の観察に基づく仮設コーホートの動きとして理解されるべきであろう。そのため、実際に1968-82年生まれのコーホートにおいて、39歳時における未婚率がSURFモデルで推定されるように9%程度で収まるのか否かについては、本分析から結論を得ることができない。また、本分析では ρ がリスク期間を通じて不変であるとの仮定をおいているが、おそらくその仮定は正しいものではない。そのため、この仮定が満たされないことによるバイアスが、ここでのSURFモデルによる結婚の生存確率にどのような影響を与えているのかが解明されなければならない。そのためには、Hillが公開しているSURFモデル専用のソフトウェア（脚注3を参照のこと）を使用して、ρ がリスク期間を通じて変化することを許容したモデルによ

る推定が必要であろう。

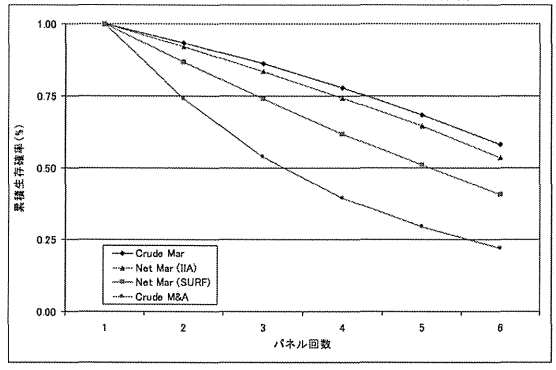
最後に、参考までに図 3-a から図 3-c では調査回数を時間軸として各モデルの累積生存確率を算出した。各年齢において、第 6 回調査までに約 75 %の未婚サンプルが結婚あるいは脱落によって失われていることは注目に値する。また、図 2 で確認されたように、Crude Mar と Net Mar(IIA) との差は総じて小さいものの、Crude Mar と Net Mar(SURF) との間には 16-18% ポイントの非常に大きな差がみられる。したがって、離散時間ロジットモデルや IIA モデルを用いた結婚の分析では、回帰係数の解釈ならびにその予測値について質的・量的双方の観点から注意が必要である。

図 3-a. パネル1時点で20歳の女性の結婚の生存曲線



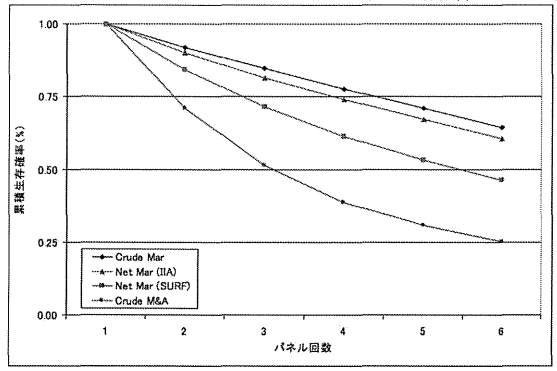
* 年間勤労所得が300万円で、年齢と年次を除く他の共変量がすべてゼロの場合。

図 3-b. パネル1時点で25歳の女性の結婚の生存曲線



* 年間勤労所得が300万円で、年齢と年次を除く他の共変量がすべてゼロの場合。

図 3-c. パネル1時点で30歳の女性の結婚の生存曲線



* 年間勤労所得が300万円で、年齢と年次を除く他の共変量がすべてゼロの場合。

<付記>

Hill 等（1993）の原典においては、単身者（single）が同棲あるいは結婚するか否かの競合リスクモデルを例としている。そこでは結婚ではなく、同棲に至るモデルのパラメーター推定が主な関心として論じられている。そのため、Hill 等の論文における同棲と結婚は、本稿における結婚と脱落にそれぞれ対応する。

Hill 等の解説においては、z_1 の値として本文中にある（4）式、つまり z_2 の値を用いている。そのため、第 2 段階推定式において推定されるパラメーターのうち、第 1 段階推定式にも使用された説明変数の回帰係数は、同棲ではなく、結婚を選択するモデルのパラメーターとなっている（本稿でいうならば、結婚ではなく脱落モデルのパラメーター）。

SURF モデルでは、競合する 2 つのイベントのモデルにおいて、β_{1k}（同棲モデルの回帰係数）と β_{2k}（結婚モデルの回帰係数）の間に以下の式が成り立つ（Hill et al. 1993, p. 270, ll. 19-21）。

$\beta_{1k}=\rho b_k+\beta_{2k}$

$\beta_{2k}=\text{-}\rho b_k+\beta_{1k},$

（いずれも、第 1 段階推定において、同棲を 1、結婚を 0 とした場合。本稿では、結婚を 1、脱落を 0 とした場合。なお、β_{2k} の変換式については筆者による追記。）

そのため、上記の変換式をつかって、結婚モデルの回帰係数から同棲モデルの回帰係数を計算している（Hill et. al. 1993, p. 269, ll. 9-12）。

Hill 等（1993）によると、このような方法で推定された β_{1k} については標準誤差が分析では算出されない。そのため、各共変量の統計的有意性についての検定を行うためには、検定したい変数を②および④から除いて推定したモデルとこれを含めたモデルで Log-likelihood を比較して、尤度比検定（Log-likelihood ratio test）を行う必要があるという。

しかし、この方法は間接的で、モデルに多数の説明変数がある場合などはとても煩雑である。さらに、上記の方法で尤度比検定を行うと、各変数の回帰係数は結婚と脱落のそれぞれのモデルで異なるにもかかわらず、その P 値が全く同じとなってしまうという問題が生じる。説明変数の統計的有意性の検定について、なぜ上記のような記述がされているのか、どのような条件で尤度比検定を行えば結婚と脱落のそれぞれのモデルで正しい P 値が得られるのかについては、論文から理解することができなかった。

一方、山口（2002）では z_1 の値として本文中の（4）式ではなく（3）式を用いている。この場合、分析において直接的に回帰係数 β_{1k} が推定される。したがって分析後に上記のような回帰係数の再計算を行う必要は無い。また、モデルにおいて回帰係数の標準誤差や P 値も直接推定されるため、尤度比検定などを行う必要もない。つまり、Hill 等（1993）によって解説されている（4）式を（3）式に入れ替えることにより、直接的に対象とするイベントのモデル推定値を得ることができるのである（！）。したがって、本稿では Hill 等（1993）によるオリジナルの方法ではなく、山口（2002）により改良された方法を採用した。

なお、第 1 段階推定式に使用せず、第 2 段階推定式においてのみ使用される変数については、結婚と脱落の双方に同じ影響を与えていると仮定される。そのため、結婚と脱落のどちらのモデルに

おいても、その回帰係数は同一となる。これは両モデルで $\beta_{1k}=\beta_{2k}$ の制約を与えているに等しい (Hill et al. 1993, p. 270, footnote 23)。一方、第1段階方程式で用いた変数は、すべて第2段階方程式で用いる必要がある。そうしないと結婚モデルと脱落モデルで ρ の値が異なってしまい、SURFモデルが成功裏に行われない。また、同様の理由により、2段階推定による方法では、各競合イベントについて（例えば、同棲と結婚もしくは結婚と脱落）説明変数が異なるモデルを設定することはできない。

Hillが公開しているSURFモデルのソフトウェア（Turbo Pascal compilerとDOS-base PCを使用）[3]を用いると、ρ が時間とともに変化するモデルや、各競合イベントについてまったく異なるモデルを設定するなどの、より柔軟なモデルを構築することができるようであるが (Hill 1997)、残念ながら、筆者はまだその適用にまでは至っていない。

参考文献

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

Hill, Daniel H., 1997. "Adjusting for Attrition in Event-History Analysis" Sociological Methodology 27: 393-416.

Hill, Daniel H., William G. Axinn, and Arland Thornton, 1993. "Competing Hazards with Shared Unmeasured Risk Factors" *Sociological Methodology* 23： 245-77.

Macfadden, D., 1981. "Economic Models of Probabilistic Choice" in *Structural Analysis of Discrete Data with Econometric Applications*, edited by C. M. Manski and D. McFadden, Cambridge, Mass: MIT Press.

Vermunt, J. K., 1997. *Loglinear Models of Event Histories.* Thousand Oaks, CA: Sage.

坂本和靖, 2006, 「サンプル脱落に関する分析：「消費生活に関するパネル調査」を用いた脱落の規定要因と推計バイアスの検証」, 『日本労働研究雑誌』, 第551号, 55-70ページ。

山口一男, 2002, 「イベントヒストリー分析（14)」, 『統計』, 2002年10月号, 66-71ページ。

第 5 章

固定効果・ランダム効果モデル

本章では、パネルデータ分析の基本的な分析手法である固定効果・ランダム効果モデルについて、統計解析ソフト R による実行例を見ながら統計学的理論を解説し、数値解析例を示すとともに、出生児縦断調査への適用例を紹介する[*1]。

5.1 通常の線形回帰モデル

5.1.1 理論編

いま、N 個の個体を $i = 1, \cdots N$ で表す。変数としては、被説明変数 y_i と、これに対する K 種類の説明変数 $X_i' = \begin{bmatrix} x_{1i} & x_{2i} & \cdots & x_{Ki} \end{bmatrix}$ を考える。このとき、定数項を α、定数項以外の回帰係数を $\beta' = \begin{bmatrix} \beta_1 & \beta_2 & \cdots & \beta_K \end{bmatrix}$、誤差項を u_i として、回帰式は、

$$y_i = \alpha + X_i'\beta + u_i \quad i = 1, \cdots N$$

と書くことができる。

以下、単純化のため、説明変数が一つ、すなわち、$K = 1$ のケースを考える。この場合、上の式は、

$$y_i = \alpha + x_i\beta + u_i \quad i = 1, \cdots N$$

となる。

ここで、残差の平方和が最小になるようにパラメータ α、β を決定するのが最小二乗法 (OLS) である。OLS 推定量は、以下のような仮定の下で、他のいかなる線形推定量よりも分散が小さくなるというよい性質 (BLUE) を持つ (Gauss-Markov の定理)。

1. u の期待値が 0 ($E(u) = 0$)
2. u の分散は均一で、$i \neq j$ について、u_i と u_j は無相関 ($E(uu') = \sigma^2 I$)
3. u は説明変数 x と無相関 ($E(xu') = 0$)

[*1] 本章の内容については、Balatagi (2005)、Croissant and Millo (2008)、北村 (2005)、樋口美雄 [等] (2006) を参考にしている。

α、β の OLS 推定量を $\hat{\alpha}$、$\hat{\beta}$ と書くと、

$$\hat{\beta} = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\alpha} = \bar{y} - \bar{x}\hat{\beta}$$

となる。ただし、$\bar{x}$, $\bar{y}$ は x,y の平均値、σ_x^2, σ_{xy} は x の分散と、x,y の共分散である。また、

$$S_{xx} = \sum_{i=1}^{N}(x_i - \bar{x})^2$$

$$S_{xy} = \sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

である。

5.1.2 R による計算（原始的な方法）

5.1.1 節で見た方法をそのまま用いれば、α、β の推定量を求めることが可能である。ここでは、下に示す仮想的なデータである「データセット A」について、OLS 推定量を求める問題を考える。

データセット A は以下のようなデータであり、個体を識別する ID（Ind）、時間（time）ならびに仮想の変数である X と Y からなるパネルデータとなっている。以下の R による実行例では、"Ydf" と名付けられたデータフレームに格納して取り扱う。

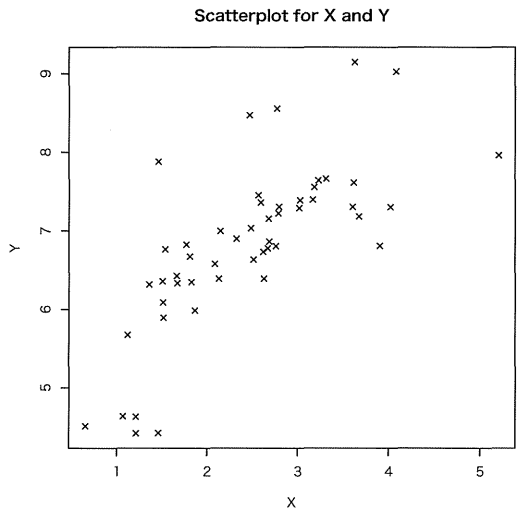
データセット A(データフレーム Ydf) の内容

```
> Ydf
   Ind time        Y        X
1    1    1 6.318302 1.358518
2    1    2 6.765785 1.538198
3    1    3 6.673199 1.812872
4    1    4 7.361044 2.589778
5    1    5 7.305603 2.791172
6    2    1 4.512218 0.652563
7    2    2 4.636834 1.068677
8    2    3 4.630790 1.213461
9    2    4 4.419180 1.215412
10   2    5 4.423203 1.461623
11   3    1 6.633481 2.513192
12   3    2 6.392095 2.623966
13   3    3 6.863346 2.684149
14   3    4 7.303656 4.019452
15   3    5 7.961180 5.209157
16   4    1 5.674991 1.122794
17   4    2 6.089810 1.512576
18   4    3 5.894734 1.518251
19   4    4 5.986789 1.865590
20   4    5 6.809329 3.904122
21   5    1 7.880392 1.463115
22   5    2 8.470850 2.468216
23   5    3 8.556446 2.769034
24   5    4 9.147444 3.630287
25   5    5 9.024920 4.083941
26   6    1 6.348810 1.826399
27   6    2 6.390482 2.129157
28   6    3 6.733463 2.619354
29   6    4 6.777782 2.668968
30   6    5 6.806236 2.758063
31   7    1 6.996998 2.147962
32   7    2 7.032473 2.483352
33   7    3 7.397037 3.164326
34   7    4 7.663331 3.312535
35   7    5 7.613489 3.613653
36   8    1 6.824914 1.773091
37   8    2 7.454234 2.563154
38   8    3 7.390236 3.023997
39   8    4 7.555790 3.179776
40   8    5 7.641978 3.227601
41   9    1 6.358868 1.505719
42   9    2 6.578637 2.088625
43   9    3 7.154565 2.677834
44   9    4 7.222814 2.783386
45   9    5 7.182551 3.674878
46  10    1 6.428335 1.663196
47  10    2 6.335629 1.670820
48  10    3 6.898983 2.324702
49  10    4 7.290997 3.015024
50  10    5 7.305412 3.604515
```

データセットAの変数XとYの関係をプロットすると以下のようになる。

パッケージとプロット

```
plot(Ydf$X, Ydf$Y, type="p", pch=4,
 main = "Scatterplot for X and Y",
 xlab = "X", ylab = "Y")
```



次に、5.1.1節で述べた理論式を直接用いる「原始的な方法」によって、α、βの推定量を求めると以下の通りである。

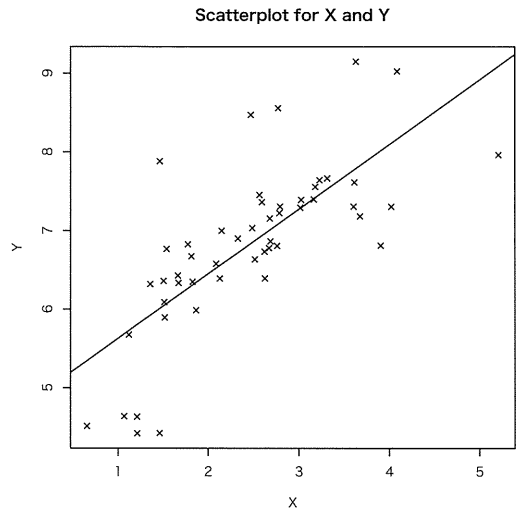
Rによる計算（原始的な方法）

```
V_XY <- cov(cbind(Ydf$X, Ydf$Y))
beta_hat <- V_XY[1,2] / V_XY[1,1]
x_bar <- mean(Ydf$X)
y_bar <- mean(Ydf$Y)
alpha_hat <- y_bar - x_bar * beta_hat
print(c(alpha_hat, beta_hat))

plot(Ydf$X, Ydf$Y, type="p", pch=4,
    main = "Scatterplot for X and Y",
    xlab = "X", ylab = "Y")
abline(alpha_hat, beta_hat)
```

出力結果

```
> print(c(alpha_hat, beta_hat))
[1] 4.807486 0.821806
```



5.1.3 R による計算（関数 lm を利用する方法）

R には線形回帰モデルを推定するための関数 lm が用意されている。これを利用すれば、α、β の推定量のみならず、線形回帰モデルに関する様々な推定量を得ることが可能である。

lm の中には、モデル式といわれる形式で回帰式を記述する。この場合、Ydf というデータフレームの Y を X で説明するという意味になる。説明変数が二つ以上あるときは"+"で結ぶ。結果のサマリーは summary 関数で得られる。なお、ここで得られる OLS 推定量は、同一個体が複数回含まれるデータをすべてプールして推定していることからプーリング推定量（pooling estimates）と呼ばれる。

R による計算（関数 lm を利用する方法）

```
lm.ols <- lm(Y ~ X, data = Ydf)
summary(lm.ols)
plot(Ydf$X, Ydf$Y, type="p", pch=4,
    main = "Scatterplot for X and Y",
    xlab = "X", ylab = "Y")
abline(lm.ols)
```

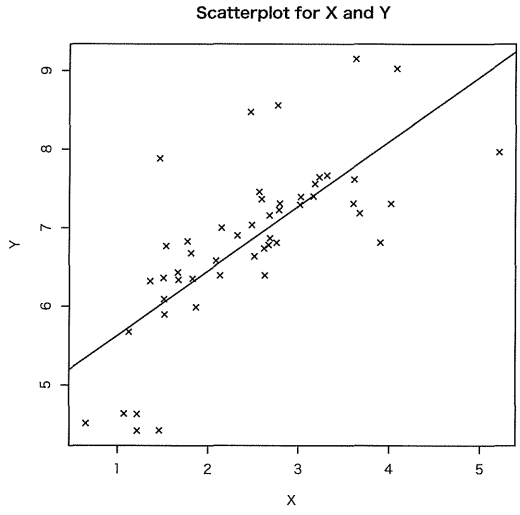
出力結果

```
Call:
lm(formula = Y ~ X, data = Ydf)

Residuals:
     Min       1Q   Median       3Q      Max
-1.58545 -0.26072  0.04754  0.29899  1.87051

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.8075     0.2888   16.65  < 2e-16 ***
X             0.8218     0.1100    7.47 1.40e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7288 on 48 degrees of freedom
Multiple R-squared: 0.5376,Adjusted R-squared: 0.5279
F-statistic:  55.8 on 1 and 48 DF,  p-value: 1.405e-09
```



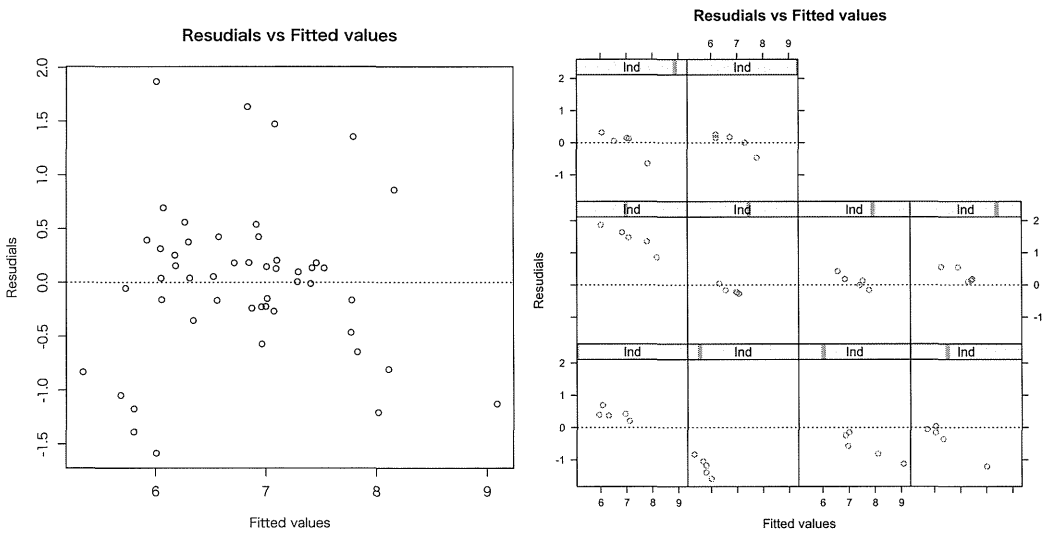
5.2 回帰モデルの残差

データセット A の変数 Y を変数 X で説明する回帰モデルについて、モデルにより推定された値に対する残差をプロットしてみよう。

モデルにより推定された値に対する残差をプロット

```
plot(predict(lm.ols), residuals(lm.ols),
    main = "Resudials vs Fitted values",
    xlab = "Fitted values", ylab = "Resudials")
abline(h=0, lty=3)

library(lattice)
xyplot(residuals(lm.ols) ~ predict(lm.ols) | Ind, data=Ydf,
    panel = function(x,y){
        panel.xyplot(x,y)
        panel.abline(h=0, lty=3)
        },
    main = "Resudials vs Fitted values",
    xlab = "Fitted values", ylab = "Resudials")
```



ここから、各個体別に残差を観察した場合、推定された値が大きくなるほど残差が減少する傾向が見られることがわかる。これは、観察不可能な個体の効果の存在を示唆している。

5.3 パネルデータの表示法

パネルデータは、同一の個体に対して複数時点での観察を行うことから、クロスセクションデータと時系列データの両方を併せ持ったデータであるといえる。したがって、パネルデータに対する回帰分析は、通常のクロスセクションデータや時系列データと異なり、変数に個体番号と時刻の2つの添字を併せ持っている。

いま、N 個の個体を $i = 1, \cdots N$ で表し、時刻を $t = 1, \cdots, T$ で表す。変数としては、被説明変数 y_{it} と、これに対する K 個の説明変数 $X'_{it} = \begin{bmatrix} x_{1,it} & x_{2,it} & \cdots & x_{K,it} \end{bmatrix}$ を考える。このとき、定数項を α、定数項以外の回帰係数を $\beta' = \begin{bmatrix} \beta_1 & \beta_2 & \cdots & \beta_K \end{bmatrix}$、誤差項を u_{it} として、回帰式は、

$$y_{it} = \alpha + X'_{it}\beta + u_{it} \quad i = 1, \cdots N; t = 1, \cdots, T$$

と書くことができる。本章では、一元配置誤差構成要素回帰モデル (One-way Error Component Regression Model) を対象とする。これは、誤差項が

$$u_{it} = \mu_i + \nu_{it}$$

と表されるモデルである。ここで、μ_i は観察不可能な個体の効果、ν_{it} は撹乱項である。

5.4 固定効果モデル

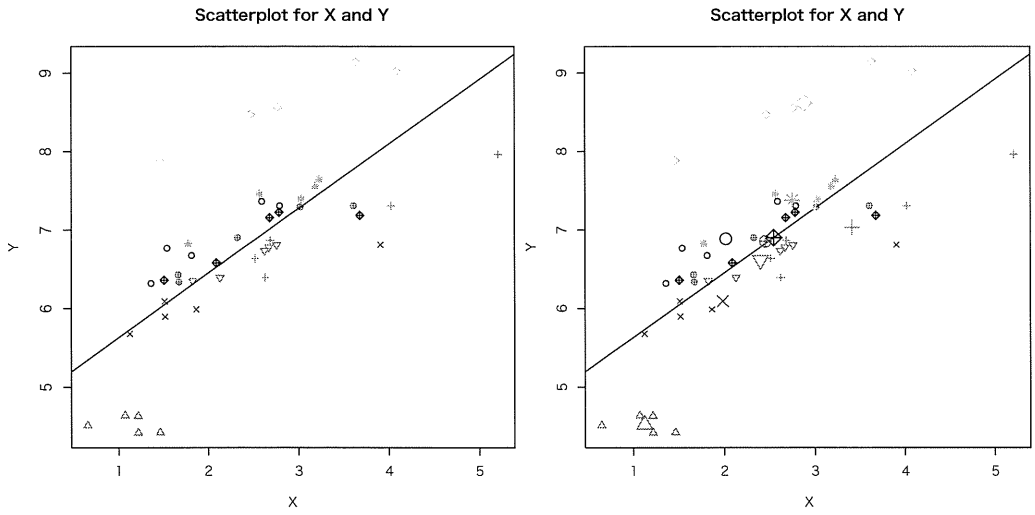
固定効果モデルとは、μ_i: 観察不可能な個体の効果をパラメータとして推定するモデルである。すなわち、傾きは同一だが、個体毎に切片が異なる直線で回帰をするモデルといえる。推定にあたっては、まず、異なる切片の影響を除去するため、X と Y から個体毎の平均値を引き去った変数を考えて、これに回帰を施すことにより傾きを推定する。

この原理を、データセット A を使ってグラフ上で考えてみよう。まず、データセット A の散布図を個体毎に色分けして描き直す。

データセット A の散布図（個体別）

```
plot(Ydf$X, Ydf$Y, type="p", col = Ydf$Ind, pch= Ydf$Ind,
    main = "Scatterplot for X and Y",
    xlab = "X", ylab = "Y")
abline(lm.ols)

# 個体毎の平均値を表示
Yb <- tapply(Ydf$Y,Ydf$Ind,mean)
Xb <- tapply(Ydf$X,Ydf$Ind,mean)
points(Xb,Yb, pch = seq_along(Yb), cex = 2, col = seq_along(Yb), lwd=3)
```

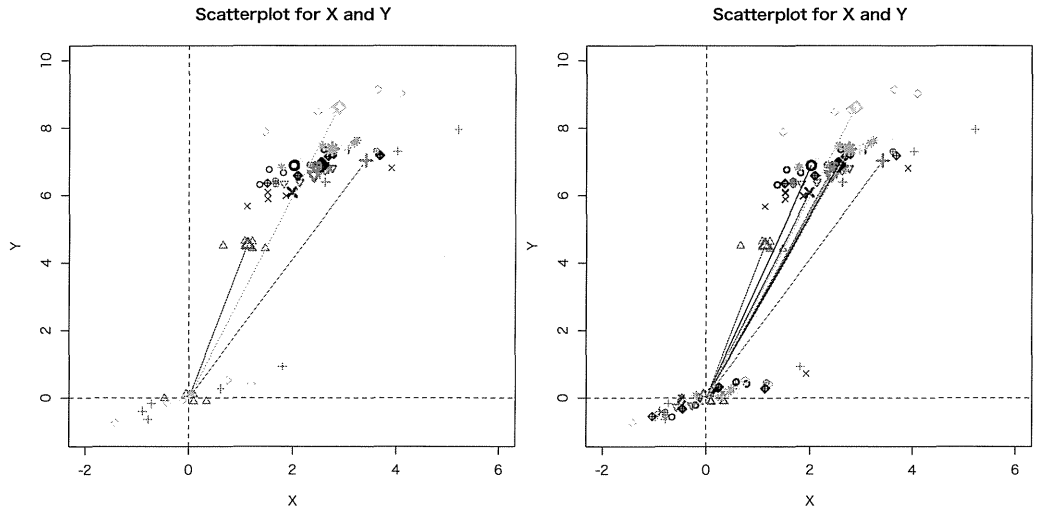


全データに当てはめた OLS の傾きが、同一個体毎に当てはめた OLS の傾きよりもやや大きいように見える。右は、個体毎に 5 時点の X,Y の平均値のポイントをを少し大きめのマーカーで表示したものである。個体毎の異質性を除いた傾きを推定するために、この大きなマーカーを原点に移すように移動する。

個体毎の平均値を引く

```
# 個体毎の平均値を引いたデータを作成
Yw <- Ydf$Y - Yb[Ydf$Ind]
Xw <- Ydf$X - Xb[Ydf$Ind]

plot(Ydf$X, Ydf$Y, col = Ydf$Ind, pch= Ydf$Ind, ylim=c(-1,10), xlim=c(-2,6),
    main = "Scatterplot for X and Y",
    xlab = "X", ylab = "Y")
points(Xw, Yw , col = Ydf$Ind, pch= Ydf$Ind)
points(Xb,Yb, pch = seq_along(Yb), cex = 1.5, col = seq_along(Yb), lwd=3)
abline(h=0, v=0, lty=2, lwd = 0.5)
arrows(Xb, Yb, 0, 0, col = seq_along(Yb), length = 0.1)
```



左の図は、水色・オレンジ・緑で示された個体についての原点への移動の様子を示したものである。点線は X 軸と Y 軸を示すことから、点線の交点が原点を示している。各個体の大きなマーカーから原点へ向かう矢印が、その個体が移動されるベクトルを示しており、原点の周りには移動されたデータが示されている。右の図はこの操作を全ての個体に対して適用した結果である。

このデータに線形回帰モデルを当てはめたときの傾きが固定効果モデルの X の回帰係数となる。

固定効果モデル (原始的な方法)

```
lm.w <- lm(Yw ~ Xw)
summary(lm.w)

plot(Ydf$X, Ydf$Y, col = Ydf$Ind, pch= Ydf$Ind, ylim=c(-1,10), xlim=c(-2,6),
    main = "Scatterplot for X and Y",
    xlab = "X", ylab = "Y")
points(Xw, Yw , col = Ydf$Ind, pch= Ydf$Ind)
points(Xb,Yb, pch = seq_along(Yb), cex = 1.5, col = seq_along(Yb), lwd=3)
abline(h=0, v=0, lty=2, lwd = 0.5)
abline(lm.w)
```

出力結果

```
Call:
lm(formula = Yw ~ Xw)

Residuals:
      Min        1Q    Median        3Q       Max
-0.267974 -0.090150  0.001145  0.100978  0.221316

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept) 2.115e-16  1.919e-02 1.10e-14        1
Xw          4.716e-01  2.673e-02    17.64   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1357 on 48 degrees of freedom
Multiple R-squared: 0.8664,Adjusted R-squared: 0.8636
F-statistic: 311.3 on 1 and 48 DF,  p-value: < 2.2e-16
```