社会調査に基づく、市民属性別の結婚年齢推定の試み
-1985 年の SMM 調査を事例に-

○根岸美知　滝元駿太朗　李皓（静岡大学）

Sample Manuscript for SICE Symposium of Social Systems Section

概要－　人口推定にはコーホート法を始めとする様々な手法があるが，これらの手法は政策を取り入れたシナリオ分析が難しい。我々は政策シナリオが市民のライフイベント確率に影響を与える仮定を前提に，社会調査のアンケートデータに基づくライフイベントの確率推定を行い，その上でその確率に基づく人口推定手法を模索する。本研究は 1985 年の SSM 調査データに対し，生存時間解析モデルの一つである、離散時間ロジットモデルと簡易的な機械学習モデルを用いて，年齢や職業，居住地などの様々の市民属性に基づく結婚確率の推定を試みる。

キーワード:　離散時間ロジットモデル 機械学習 推定

1　はじめに

近年、我が国では生涯未婚率が高まっている。結婚が成立する人とそうでない人の違いを解明することは，結婚する人を増加させることの一助となり，少子高齢化や人口減少といった現代の日本が抱えている人口問題の解決へとつながる。

我々は、市民一人一人の状況を分析することで，結婚が成立する人の特徴を導出できるのではないかと仮定した。

そこで本研究ではSSM調査研究会が10年に1度行う社会階層と社会移動全国調査（SSM調査 1985）の個票データに対して分析を行い，婚姻に至る影響を及ぼしているであろう要因の推定と得られた結果より，推定を行う手法の比較を行う。

2　使用したデータについて

本研究ではSMM調査の1985年のデータを使用した。SMM調査は，1955年以来，10年に一度行われている大規模調査であり，これまで7回の調査を行ってきた。最新の調査は2015年に行われている。SSM調査の1985年のデータのファイルはタブ区切りのファイルとなっていて，その中に男女それぞれ1000人以上の回答データが格納されている他，ラベルファイルや記録コードなどのデータ解釈用のファイルもある。

以下にSSM調査1985年の概要をTable1のようにしてまとめる。

Table1:SSM1985の概要

調査番号	**0762**
寄託者	2015SSM調査管理委員会
調査対象	【男性A票】　20歳～69歳の男性
	【女性票】　20歳～69歳の女性
調査対象の範囲	個人
サンプルサイズ	【男性A票】　標本数：2,030，回収数：1,239　（回収率 61.03%）
	【女性票】　標本数：2,171，回収数：1,474　（回収率 67.89%）
調査時点	1985年11月～1986年2月
調査地域	全国（300地点）
標本抽出	抽出方法：層化2段・等間隔抽出
	抽出台帳：選挙人名簿
調査方法	訪問面接

3 データ分析

3.1 離散時間ロジットモデルを用いた分析

我々は離散時間ロジットモデルを用いて分析を行った。離散時間ロジットモデルはある時点 t までにイベント（今回は結婚）が発生していないという条件のもとで t 時点でイベントが発生する確率（ハザード確率）を予測するモデルであり，以下の式（1）（2）で表される。なお、P（t）はハザード確率、b（t）は時間変数，x_kは共変量，a_kは回帰係数，n は説明変数を表す。

$$P(t) = \frac{1}{1+\exp(-z(t))} \quad \cdots(1)$$

$$z(t) = b(t) + \sum_{k=1}^{n}(a_k x_k(t)) \quad \cdots(2)$$

本研究では，離散時間ロジットモデルを適用するために SMM 調査のデータをパーソン・ピリオドデータに変換し，統計ソフトの R で読み込んだ。次いで，婚姻状態を目的変数，都道府県，本人の最終学歴，個人収入など計 12 個の変数を説明変数とし，男女それぞれの性別で 15-25 才，25-34 才，34-65 才の３つの年齢区分に分け，R の GLM 関数を用いて離散時間ロジットモデルの回帰係数の推定を行い，step 関数で AIC（赤池情報量規準）により上記に挙げた変数の中から変数選択を行い，計算した。

計算結果の一例を以下に示す

Table2:離散時間ロジットモデル（男性 26-34 才）

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.44685	0.74371	-4.635	3.57E-06
エリア近畿	0.02467	0.14251	0.173	0.862578
エリア九州	0.37139	0.16772	2.214	0.026801
エリア甲信越	0.48176	0.22107	2.179	0.029311
エリア四国	0.31613	0.29242	1.081	0.279663
エリア中国	0.4138	0.20517	2.017	0.043713
エリア東海	0.479	0.16618	2.882	0.003946
エリア東北	0.38697	0.1962	1.972	0.048574
エリア北海道	0.55752	0.25742	2.166	0.030329
エリア北陸	1.08793	0.29218	3.723	0.000197
Q5A本人の最終学歴旧制高等小学校	0.46392	0.30859	1.503	0.132753
Q5A本人の最終学歴旧制尋常小学校	0.60727	0.37123	1.636	0.101873
Q5A本人の最終学歴旧制大学（大学院を含む）	-1.06765	0.48832	-2.186	0.028788
Q5A本人の最終学歴旧制中学校・実業学校・師範学校	0.13136	0.33359	0.394	0.693736
Q5A本人の最終学歴新制高校	-0.1603	0.28945	-0.554	0.579708
Q5A本人の最終学歴新制大学（大学院を含む）	-0.23013	0.29864	-0.771	0.44095
Q5A本人の最終学歴新制短大・高専	-0.11207	0.43233	-0.259	0.79546
Q5A本人の最終学歴新制中学校	0.30776	0.30294	1.016	0.30968
Q25B個人収入50	0.23687	0.4707	0.503	0.614802
Q25B個人収入100	-0.02209	0.37956	-0.058	0.953588
Q25B個人収入150	-0.02132	0.34301	-0.062	0.950446
Q25B個人収入200	0.37559	0.33022	1.137	0.255371
Q25B個人収入250	0.40423	0.34256	1.18	0.237986
Q25B個人収入300	0.56312	0.3248	1.734	0.082962
Q25B個人収入350	1.09992	0.33417	3.291	0.000997
Q25B個人収入400	0.77515	0.33104	2.342	0.019204
Q25B個人収入450	0.89216	0.34482	2.587	0.009673
Q25B個人収入500	1.06166	0.33953	3.127	0.001767
Q25B個人収入550	0.89929	0.35316	2.546	0.010885
Q25B個人収入600	1.15878	0.3494	3.317	0.000911
Q25B個人収入650	0.5204	0.40651	1.28	0.200481
Q25B個人収入700	1.06837	0.41719	2.561	0.010442
Q25B個人収入750	2.20448	0.57937	3.805	0.000142
Q25B個人収入800	1.86441	0.50056	3.725	0.000196
Q25B個人収入850	1.03019	0.51924	1.984	0.047252
Q25B個人収入900	1.70441	0.52834	3.226	0.001255
Q25B個人収入950	0.9117	0.98928	0.922	0.356746
Q25B個人収入1000	1.92746	0.53488	3.604	0.000314
Q25B個人収入1050	1.39638	0.61201	2.282	0.022512
Q25B個人収入1100	2.23103	1.06422	2.096	0.036046
Q25B個人収入1150	-12.1143	293.8915	-0.041	0.96712
Q25B個人収入1200	1.37916	0.78969	1.746	0.08073
Q25B個人収入1250	0.37116	1.10694	0.335	0.7374
Q25B個人収入1450	1.85911	1.26933	1.465	0.143019
Q25B個人収入1500	1.69914	0.65566	2.591	0.009556
Q25B個人収入2000	1.85911	1.26933	1.465	0.143019
Q25B個人収入DK.	0.78253	0.31998	2.446	0.014462
現在年齢	0.04074	0.02123	1.919	0.054979

Table2 は男性の 26-34 才のデータであり，最終学歴と個人収入と出身地域と現在年齢が有意のある説明変数であるとして選ばれた。

表より，個人収入が 350 万円をこえてくると z 値が高く，P 値が 0.05 を下回っている傾向がみられ，年収が結婚成立に影響を与えているということが分かる。

出身地域についても同様に見ていくと北陸や東海地方においてこの年代の男性の結婚がよく成立しているということが分かる。

また，Table2 のパラメータを利用し，結婚イベントが発生する確率を計算することもできる。例えば最終学歴が新制大学(大学院含む)で年収が 800 万円の人、出身地域が東海、現在年齢が 30 才の人の場合

$$\frac{1}{1+\exp(-1(-3.44685-0.23013+30\times 0.0474+1.86441+0.47900))}$$

という式で表され，これを計算すると約 0.52 となるので上記の条件にある男性が 30 才で約 52%の確率で結婚すると分かる。また、このデータを用いて予測を行ったところ男性の方は RMSE の値が 5.089759087，女性の方が RMSE の値が 3.693025149 となった。

3.2 機械学習を用いた分析

ID が 1 のデータはテスト用，2,3 のデータはトレーニング用…といった形でデータを 2 個飛ばしにしてテスト用データに振り分けることによりデータをトレーニングデータとテスト用データに 2:1 の比率にして分けた。

結婚時の年齢を目的変数とし, 以下の項目を説明変数とし性別ごとにニューラルネットワークを用いて分析を行った。

性別	説明変数
女性	兄弟数，出生順位，現職.従業上の地位，現職.役職名，最終学歴，満足度.勤め先，満足度.生活全般，父の学歴，父の主な職業.従業上の地位，母の学歴
男性	兄弟数，出生順位，現職.従業上の地位，現職.役職名，最終学歴，満足度.勤め先，満足度.生活全般，父の学歴，父の主な職業.従業上の地位，重要度.高い収入

パラメータの個数を 10 にし分析した結果、女性のデータは中間層のサイズが 1、重み減衰の値が 0.003162278 であるモデル、男性のデータは中間層のサイズが 1、重み減衰の値が 0.003162278 であるモデルが選定された。しかし、Kappa 統計量が女性は 0.09749825、男性は 0.07325322 と非常に小さく、これらのモデルでは結婚時の年齢を適切に説明しているとは言い難い。

また、得られたモデルを用いてテスト用データを用いて予測を行ったところ、女性のデータの予測結果を以下に示す。テスト用データ 492 個のうち、

・未婚を正しく予測したもの：62 個
・未婚だが既婚と予測したもの：8 個
・既婚だが未婚と予測したもの：37 個
・既婚のデータを既婚と予測したもの：370 個
・「DK(不明)」が含まれていたもの：15 個

また、既婚のデータを既婚と予測したものの結婚時の年齢の二乗平均平方根誤差は 3.08412 であった。

この結果を踏まえると今回得られたデータは未婚と予測することが多いため未婚のものを正しく予測できているが、既婚のデータまで未婚と判定してしまった。これは未婚は一つの要素だが既婚の場合結婚時の年齢一つ一つを一つの要素として機械学習を行ったためだと考えられる。

4 おわりに

我々は結婚行動に影響を与える要因を調べるために既存の大規模社会データに対して離散時間ロジットモデルと簡易的な機械学習の二種類のアプローチを行い検証した。機械学習の方では Kappa 統計量の値が低くなり，適切なモデルとは言えないが，生成したモデルを用いて，テストデータの予測を行うことができた。離散時間ロジットモデルの方は，収入や出身地域などでできた計算式を用いて，結婚確率を推定することができた。

両手法ともに今回できなかったことは，今後の研究の課題としていきたい。

最後に本研究を行うにあたり，データを提供してくださった東京大学社会学研究所附属社会調査・データアーカイブ研究センターに感謝の意を表します。

参考文献

1) 「ファクターによる人口動態に関連する行動確率計算による人口推定の試み」李 皓（静岡大学）
2) 1985 年 SSM 調査，1985 | SSJDA Direct（東京大学社会学研究所附属社会調査・データアーカイブ研究センター）