滋賀大学 DS 教育研究センター 研究資料

Report No. 1

Ver: 2021 年 7 月 13 日 (0 時 41 分)

作成：助教　李鍾贊

和歌山県における健康寿命の延伸
「健康長寿日本一わかやま」を目指して
Final

滋賀大学データサイエンス教育研究センター

www.ds.shiga-u.ac.jp

2021 年 7 月 13 日

目 次

第 1 章　本事業の概要, はじめに

1.1　背景

日本は, ここ 25 年の間, 平均寿命の延伸, 死亡率の低下により, 高齢化率が 2016 年において 27%を示しており, 既に「高齢社会 (総人口に対して 65 歳以上の割合が 14%以上)」を過ぎて「超高齢社会」に入っている.[1] こうした現状を考慮すると国・自治体の健康政策も「健康の質」を上げる方向に立案する必要性が求められる. 海外では既に国の保健対策をデータに基づいて行う変革が実施されており (Global Burden of Disease : Generating Evidence, Guiding Policy, 2010), 和歌山県の保健活動にもデータに基づくエビデンスが必要と考えられる.

1.2　目的

和歌山県の健康・医療・介護に関するデータ, 経済状況・ボランティア参加率等の社会環境因子に関わるデータを利活用した現状分析を実施するとともに和歌山県の位置づけや強み・弱みを把握し, 得られた新たな知見を県の施策に反映し, 県民の健康寿命の延伸を図る. 寿命及び健康寿命を用いて統計解析し, 今後, 和歌山県の健康及びヘルスケア産業における政策立案に役に立つ参考資料を示すことを目的とする.

1.3　実施期間

令和 2 年 11 月 1 日 ～ 令和 3 年 3 月 31 日まで

1.4　データ

データは和歌山県が収集した 47 都道府県の公的データを活用することにする. その他経済, データの詳細は後述するが, 文化, など多様なデータを用いる.

1.5　方法

データ分析には主に統計ソフト R 4.0.4 を利用した。提供データの形式に適合する統計手法を取り入れ, 平均寿命や健康寿命との関連を分析する。

影響を与える要因を探るため, 疾病と関連する医学的変数のほか, 社会的変数等を説明変数に取り入れ分析を行う。分析には、説明変数の変数選択、多変量解析よる次元縮約を行い、分析可能なデータとして加工を加えた後に、分析を行う。

1.6　期待効果

本県の現状に関して, 県民及び県外から移住を検討する人に向けて正しい情報発信の資料として活用されることが期待される. ビックデータ時代に, 他県に先駆けて官学連携による健康データを活用

[1] 高齢化率とは総人口に対して 65 歳以上の高齢者人口が占める割合. 世界保健機構 (WHO) や国連の定義によると, 高齢化率が 7 %を超えた社会を「高齢化社会」, 14 %を超えた社会を「高齢社会」, 21 %を超えた社会を「超高齢社会」と定義.

する取り組みは, データに基づく県政を推奨している国の方針とも当てはまるので, 他県のベンチマーク事例になることが期待される.

第 2 章　データと変数

本研究で用いるデータは和歌山県の「和歌山県データ利活用推進センター」が 2018 年度滋賀県の??? 研究で使用したデータに基づき、2021 年時点で収集可能な同様のデータを収集して滋賀大学に csv 形式にて提供されたものである。

全てのデータはインターネットから容易にダウンロードが可能である公的データである。(データの出典は ?? 参考)。

データの次元は 47 都道府県を個体、162 項目を変数とする性別ごとの 2 つのデータセットである。変数の中、平均寿命と健康寿命の 2 つの変数は分析の目的変数となり、残りの 160 変数は説明変数として扱う。本研究ではこれらの 2 種類の寿命変数を以下で寿命変数と呼ぶ。

160 説明変数の項目は両者のデータセットで共通するが、性別によってその数値が異なる変数がある。例えば、平均寿命は、和歌山県の男性の平均寿命は 79.94 歳、和歌山県の男性の平均寿命は 86.47 歳のように性別ごとに異なる数値が得られる性別ごとの情報が分かる変数であるが、表 2.1 の「居住_持ち家比率」変数は性別ごとに調べられ変数ではなく、男性のデータセットも女性のデータセットでも同じ数値が記入されている (例、和歌山県「持ち家比率」は性別に関係なく 73 %と同じ数値)。

本研究では、平均寿命ように性別の区別のある変数を「性別変数」に、「居住_持ち家比率」のように性別の意味の無い変数 (もしくは、データ収集時点で性別分けデータ入手できなかった変数) を「共通変数」に呼ぶことにする。表 2.2 に本研究に用いた性別変数の一覧を、表 2.1 に共通変数の一覧を示す。

Table. 2.1　共通変数 (98 個)

	var_name_Jpn...2	var_name_Jpn...4	var_name_Jpn...6	var_name_Jpn...8
1	受療率_入院_悪性新生物_2017	行政基盤_教育費割合（県財政）	家計_消費支出（一世帯当たり 1 か月）	現金給与総額_2016
2	受療率_入院_心疾患_2017	教育_最終学歴が大学・大学院卒の者の割合	家計_教育費割合（対消費支出）	生鮮肉 (世帯数消費支出)_2014
3	受療率_入院_脳血管疾患_2017	労働_1 次産業就業者比率	家計_教養娯楽費割合（対消費支出）	生鮮肉 (世帯数消費支出)_2015
4	受療率_外来_悪性新生物_2017	労働_2 次産業就業者比率	家計_貯蓄現在高	生鮮肉 (世帯数消費支出)_2016
5	受療率_外来_心疾患_2017	労働_3 次産業就業者比率	家計_スマートフォン所有数量（千世帯当たり）	生鮮肉平均_世帯数消費支出 (2014〜2016)
6	受療率_外来_脳血管疾患_2017	労働_完全失業率	家計_パソコン所有数量（千世帯当たり）	菓子類 (世帯数消費支出)_2014
7	病院数_2019	文化・スポーツ_図書館数（人口 100 万人当たり）	家計_自動車所有数量（千世帯当たり）	菓子類 (世帯数消費支出)_2015
8	診療所数_2019	健康・医療_一般診療所数 (可住地面積 100km^2 当たり)	家計_タブレット端末所有数量（千世帯当たり）	菓子類 (世帯数消費支出)_2016
9	がん治療認定医数_2020	文化・スポーツ_スポーツの行動者率	人口・世帯_高齢単身者世帯の割合	菓子類平均_世帯数消費支出 (2014〜2016)
10	循環器専門医数_2020	文化・スポーツ_旅行・行楽行動者率	高血圧疾患_入院 2014 年	果物 (世帯数消費支出)_2014
11	内視鏡専門医数_2020	居住_持ち家比率	高血圧疾患_外来 2014 年	果物 (世帯数消費支出)_2015
12	書籍購入代金_2019	居住_一戸建住宅比率	糖尿病_入院 2014 年	果物 (世帯数消費支出)_2016
13	人口・世帯_年少人口割合 2020	居住_上水道給水人口比率	糖尿病_外来 2014 年	果物平均_世帯数消費支出 (2014〜2016)
14	人口・世帯_老年人口割合 2020	居住_下水道普及比率	肉類_2014	全国学力・学習状況 (公立学校数)(中学校)_2015
15	人口・世帯_生産年齢人口割合 2020	文化・スポーツ_ボランティア活動行動者率	魚介類_2014	全国学力・学習状況 (公立学校数)(小学生)_2015
16	人口・世帯_粗死亡率 2020	居住_都市公園面積（人口 1 人当たり）	牛乳_2014	う蝕外来総数_2014
17	人口・世帯_共働き世帯割合 2020	居住_都市公園数 (可住地面積 100km^2 当たり)	乳製品_2014	歯周疾患 (歯肉炎) 外来総数_2014
18	自然環境_年平均気温	健康・医療_一般病院数 (可住地面積 100km^2 当たり)	卵_2014	骨の密度障害_2014
19	自然環境_年平均相対湿度	居住_主要道路舗装率	大豆_2014	骨折_2014
20	自然環境_降水量（年間）	居住_市町村舗装率	一定のバリアフリー化率_2018	歯の補てつ_2014
21	自然環境_雪日数（年間）	健康・医療_一般歯科診療所数（人口 10 万人当たり）	高度のバリアフリー化率_2018	アルツハイマー等 (脳血管疾患)_2014
22	経済基盤_県民所得	健康・医療_医療施設に従事する医師数（人口 10 万人当たり）	バリアフリー_手すりがある 2018	ジニ係数総世帯_2014
23	行政基盤_財政力指数	健康・医療_保健師数（人口 10 万人当たり）	バリアフリー_廊下などが車いすで通行可能な幅 2018	収入ジニ係数勤労世帯_2014
24	行政基盤_収支比率	安全_交通事故発生件数（人口 10 万人当たり）	バリアフリー_段差のない屋内 2018	
25	行政基盤_生活保護費割合（県財政）	家計_実収入（一世帯当たり 1 か月）	総実労働時間_2016	

Table. 2.2　性別変数 (62 個)

	var_name_Jpn...2	var_name_Jpn...4	var_name_Jpn...6
1	人口	趣味・娯楽－読書	喫煙率 (計 100 本以上,6ヵ月以上&直近 1ヵ月)(40～74 歳)2014
2	75 歳未満調整死亡率_悪性新生物_2018	自己啓発・訓練－学習・自己啓発・訓練率	20 歳に比べて 10kg 体重増加 (40～74 歳)2014
3	75 歳未満調整死亡率_悪政新生物_2019	自己啓発・訓練－芸術・文化	歩く速度が速い (同年齢と比較)(40～74 歳)2014
4	年齢調整死亡率_心疾患_2015	自己啓発・訓練－英語	飲酒日 1 日当たり 2 合以上飲む割合 (頻度)(40～74 歳)2014
5	年齢調整死亡率_脳血管疾患_2015	自己啓発・訓練－英語以外の外国語	毎日酒を飲む割合 (頻度)(40～74 歳)2014
6	60 歳以上人口_2015	自己啓発・訓練－パソコンなどの情報処理	睡眠休養が十分とれている (40～74 歳)2014
7	学習率_2016	ボランティア活動－安全な生活のための活動	朝食を抜くことが週 3 回ある (40～74 歳)2014
8	読書率_2016	ボランティア活動－自然や環境の活動	夕食後に間食することが週 3 回ある (40～74 歳)2014
9	スポーツ総行動率	ボランティア活動－災害活動	野菜摂取量_2016(20 歳以上平均値 (g/日)
10	スポーツ総行動率-器具を使ったトレーニング	脳血管疾患_年齢調整死亡率 2015	食塩摂取量_2016(20 歳以上平均値 (g/日)
11	スポーツ行動率－ウォーキング	悪性新生物 (胃)_年齢調整死亡率 2015	BMI 平均値_2016(男性 20～69 歳)(女性 40～69 歳)(単位 Kg/m^2)
12	旅行・行楽－旅行・行楽・観光総行動率	悪性新生物 (大腸)_年齢調整死亡率 2015	歩数_2016(20 歳以上平均値 (歩/日)
13	旅行・行楽－旅行率	悪性新生物 (肝及び肝内胆管)_年齢調整死亡率 2015	
14	旅行・行楽－行楽率	悪性新生物 (気管、気管支及び肺)_年齢調整死亡率 2015	
15	旅行・行楽－観光率	悪性新生物 (乳房)_年齢調整死亡率 2015	
16	ボランティア総行動率－総数	悪性新生物 (子宮)_年齢調整死亡率 2015	
17	ボランティア総行動率－まちづくり活動	心疾患_年齢調整死亡率 2015	
18	ボランティア総行動率－国際協力活動	肺炎_年齢調整死亡率 2015	
19	ボランティア総行動率－健康や医療サービスに関係した活動	急性心筋梗塞_年齢調整死亡率 2015	
20	ボランティア総行動率－高齢者を対象とした活動	血圧を下げる薬の使用 (40～64 歳)2014	
21	ボランティア総行動率－障害者を対象とした活動	インシュリン注射、血糖を下げる薬の使用 (40～74 歳)2014	
22	ボランティア総行動率－子供を対象とした活動	コレステロールを下げる薬の使用 (40～74 歳)2014	
23	趣味・娯楽－趣味娯楽総行動率	就寝前の 2 時間以内に夕食 (40～74 歳)2014	
24	趣味・娯楽－園芸・庭いじり・ガーデニング	日常生活において歩行等の身体活動 (1 日 1 時間以上実施)(40～74 歳)2014	
25	趣味・娯楽－スポーツ観覧	軽く汗をかく運動週 2 回 (40～74 歳)2014	

第 3 章　方法

第 2 章では今回の分析に用いるデータのその変数の詳細について紹介した。

この章で分析の方向性と基本設定を紹介するとともに、分析の全体像を示す。第 2 章で紹介したデータをどのように分析に用いて、分析をおこない、最終的第 ??? 章で結論づけた健康に影響を与える因子がどのように導きだしたかについて説明する。

本研究ではモデルの基本設定として性別ごとの寿命変数が目的変数となり、残り 160 変数は説明変数の候補群となる設定である。したがって、説明変数として用いられる変数を目的変数である寿命変数の原因になる可能性を内在していると仮定した上図 3.1 に示すようなモデルを考慮するのが自然的な発想である。

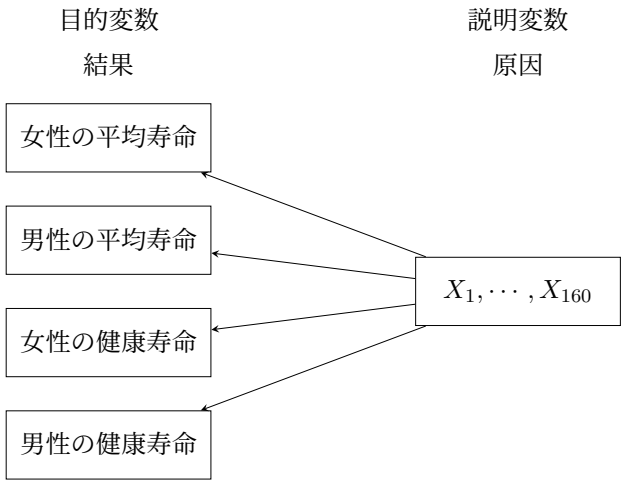

Fig. 3.1　モデリングの基本構造

しかし、本研究で用いられるデータは都道府県がデータの 47 個体となるデータであるため、図 3.1 のように全ての説明変数を一つのモデルに取り入れて 160 個の回帰係数 $\beta_0, \beta_0, \cdots, \beta_{160}$ を推定することが技術的にできない。

以上の理由で、モデリングを進むための前処理として、160 変数の中から比較的に扱いやすい数の説明変数の数を減らす必要がある。この前処理は 2 段階で構成される。

1 段階の前処理では 160 変数の中から一定の基準 (基準については後術）を満たす変数をを選択し, 説明変数の数を減らす。結果として、男性の場合は、10 個の変数が、女性の場合は、17 個の変数がをそれぞれの性別の説明変数として選択された（4 章参照)。2 段階の前処理では 1 段階の前処理で選択された変数に対して、多変量分析の因子分析を用いて 2 つの共通因子を抽出する (第 5 章)。本研究ではこの 2 つの共通因子を説明変数とし, 寿命変数を目的変数に設定して第 6 章以降でモデリングを行う。

言い換えると、以上の前処理が終わった時点で、モデルへの適用が困難であた図 3.1 の構造は、多様なモデルが適用できるような単純な構造となり (図 3.2)、目的変数 (寿命変数) に影響を与える変数を 2 つの共通因子を通して調べることができる。

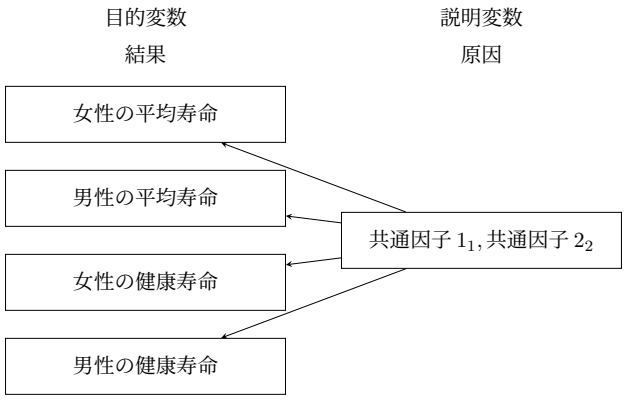

Fig. 3.2　説明変数の共通因子を用いたモデリング

例えば、男性の平均寿命 Y とすると線形回帰モデルは

$$E(Y|F_1, F_2) = \beta_0 + \beta_0 F_1 + \beta_2 F_2 \tag{3.1}$$

のようになり、右辺の $\beta_0, \beta_0,$ を推定し、平均寿命への影響が調べられる。

第 6 章と 7 章では 2 つの共通因子を用いて様々なモデリングを適用し、寿命への影響を分析する。第 6 章では線形回帰モデルを含むより一般化した一般化線形モデルを利用し、寿命変数と 2 つの共通因子との関連性に有無を調べる。

最後に 7 章ではベイズモデルを用い、2 つの共通因子が寿命変数の事後分布にどのような影響を与えるかを分析する。

主に R version 4.0.4 を用いて行った。また、説明変数は標準化処理を行った。

第 4 章　変数選択

第 3 章で述べたように、本研究ではモデリング手法を用いてい寿命と関連のある変数を探索および一般化線形モデルなど統計的確定的なモデルを使い分析すると設定した。第 3 章で述べたように、本研究で用いられるデータは 47 都道府県が個体となるマクロデータであるが、説明変数の数は 160 個と都道府県の数 47 を超えており、モデリングの係数を推定することが困難なため、変数選択の前処理を行う必要がある。

変数選択が必要となる他の理由としては、全ての説明変数が目的変数である寿命変数に統計的に有意な変数とは限らない、また、似たような情報を持つ変数をモデルに同時に取り入れると統計的に正しくモデルが推定できない計算上の問題があるからである。

変数選択の過程において基本的に AIC 基準に基づいて行う。AIC 基準のような定量的な変数選択を使う利点の一つは、目的変数と関連のない説明変数を定量的に選別できる点にある。

4.1　選択された説明変数

この節では AIC により選択された説明変数を紹介する。また、変数選択がどのように行ったのかについてそのアルゴリズムを示す。まず、表 4.1 に変数選択を行った後、最終的に寿命と AIC によりモデルを選択した結果残った変数すリストを示す。男性の説明変数は 18 個、女性は 10 個の説明変数である。

これらの変数は 2 種類の目的変数 (女性の平均寿命、健康寿命、男性の平均寿命、健康寿命) に対して、統計的に有意な説明変数となる候補になるとも言える。

Table. 4.1　変数選択後の変数

	f_var	m_var
1	受療率_外来_脳血管疾患_2017	受療率_入院_心疾患_2017
2	人口・世帯_老年人口割合 2020	自然環境_年平均気温
3	人口・世帯_生産年齢人口割合 2020	健康・医療_保健師数（人口 10 万人当たり）
4	自然環境_年平均気温	家計_貯蓄現在高
5	労働_完全失業率	人口・世帯_高齢単身者世帯の割合
6	居住_都市公園数 (可住地面積 100km^2 当たり)	悪性新生物 (大腸)_年齢調整死亡率 2015
7	高血圧疾患_外来 2014 年	自己啓発・訓練－パソコンなどの情報処理
8	悪性新生物 (大腸)_年齢調整死亡率 2015	一定のバリアフリー化率_2018
9	ボランティア総行動率－総数	自己啓発・訓練－芸術・文化
10	受療率_外来_心疾患_2017	自己啓発・訓練－英語以外の外国語
11	居住_一戸建住宅比率	
12	75 歳未満調整死亡率_悪政新生物_2019	
13	診療所数_2019	
14	バリアフリー_手すりがある 2018	
15	循環器専門医数_2020	
16	家計_スマートフォン所有数量（千世帯当たり）	
17	ボランティア総行動率－高齢者を対象とした活動	

この中、疾患係変数をみると女性の場合、「受療率_外来_脳血管疾患_2017」,「高血圧疾患_外来 2014 年」,「悪性新生物 (大腸)_年齢調整死亡率 2015」が男性の場合は、「受療率_入院_心疾患_2017」,「悪性新生物 (大腸)_年齢調整死亡率 2015」が平均寿命と健康寿命のいずれかと関連があると結果である。これらの変数が正負の方向については、次節で説明する。

4.2 線形回帰モデルからの探索

本研究では第 5 章以降で抽出する因子を説明変数として用いるが、ここでは、AIC 基準で選択された変数を用いて寿命への影響を考えてみる。

4.2.1 平均寿命の線形回帰モデルからの探索

Table. 4.2 女性の線形回帰 (平均寿命)

	var_name_Jpn	estimate	statistic	p.value
2	受療率_外来_脳血管疾患_2017	-0.00	-2.80	0.01
3	人口・世帯_老年人口割合 2020	-0.16	-2.40	0.02
4	人口・世帯_生産年齢人口割合 2020	-0.25	-3.26	0.00
6	労働_完全失業率	-0.25	-2.97	0.01
9	悪性新生物 (大腸)_年齢調整死亡率 2015	-0.09	-3.22	0.00
12	居住_一戸建住宅比率	-0.03	-3.01	0.01
18	ボランティア総行動率 – 高齢者を対象とした活動	0.14	1.95	0.06

Table. 4.3 男性の線形回帰 (平均寿命)

	var_name_Jpn	estimate	statistic	p.value
3	自然環境_年平均気温	0.06	3.17	0.00
4	健康・医療_保健師数（人口 10 万人当たり）	0.01	2.77	0.01
5	家計_貯蓄現在高	0.00	2.37	0.02
6	人口・世帯_高齢単身者世帯の割合	-0.10	-3.83	0.00
7	悪性新生物 (大腸)_年齢調整死亡率 2015	-0.10	-5.43	0.00
8	自己啓発・訓練 – パソコンなどの情報処理	0.08	2.66	0.01
11	自己啓発・訓練 – 英語以外の外国語	0.10	1.70	0.10

4.2.2 健康寿命の線形回帰モデルからの探索

Table. 4.4 女性の線形回帰 (健康寿命)

	var_name_Jpn	estimate	statistic	p.value
2	受療率_外来_脳血管疾患_2017	-0.01	-2.76	0.01
3	人口・世帯_老年人口割合 2020	0.25	2.25	0.03
4	人口・世帯_生産年齢人口割合 2020	0.35	2.75	0.01
5	自然環境_年平均気温	0.29	7.03	0.00
6	労働_完全失業率	-0.29	-2.12	0.04
8	高血圧疾患_外来 2014 年	0.14	4.94	0.00
9	悪性新生物 (大腸)_年齢調整死亡率 2015	0.17	3.61	0.00
12	居住_一戸建住宅比率	0.07	4.51	0.00
13	75 歳未満調整死亡率_悪政新生物_2019	-0.10	-4.41	0.00
14	診療所数_2019	-0.03	-4.60	0.00
15	バリアフリー_手すりがある 2018	-0.00	-4.62	0.00
16	循環器専門医数_2020	0.00	4.37	0.00
17	家計_スマートフォン所有数量（千世帯当たり）	-0.01	-6.02	0.00

Table. 4.5 男性の線形回帰 (健康寿命)

	var_name_Jpn	estimate	statistic	p.value
6	人口・世帯_高齢単身者世帯の割合	-0.16	-3.08	0.00
7	悪性新生物 (大腸)_年齢調整死亡率 2015	-0.06	-1.71	0.10
11	自己啓発・訓練 – 英語以外の外国語	0.20	1.82	0.08

4.3 変数選択のアルゴリズム

この節ではどのような手順で変数選択を行なったのかについて記す。理解を助けるために、図 4.1 にそのアルゴリズムの flow を示す。基本的な考え方は最初の列が 47 を超える説明変数のデータ行列 X_0 から出発して、線形回帰分析が可能になるように、複数のグループに説明変数を分けた後（図 4.1 の赤色の部分)、それぞれの説明変数を用いて回帰モデルと AIC 変数選択を行う（図 4.1 の緑色と青色の部分)。図 4.1 は $\boldsymbol{X}_0$ を 3 つのデータ $\boldsymbol{X}_1$, $\boldsymbol{X}_2$,$\boldsymbol{X}_3$ に分けた例を示したが、本研究で用いられる 160 の変数であると、最低 4 分割が必要となる。手順は同様である。

今回の分析では平均寿命及び健康寿命を目的変数とし、以上の個別の変数を説明変数として使わずに、これらに基づく因子分析により 2 つの因子を抽出してそれらの 2 つの因子を説明変数として用いる。

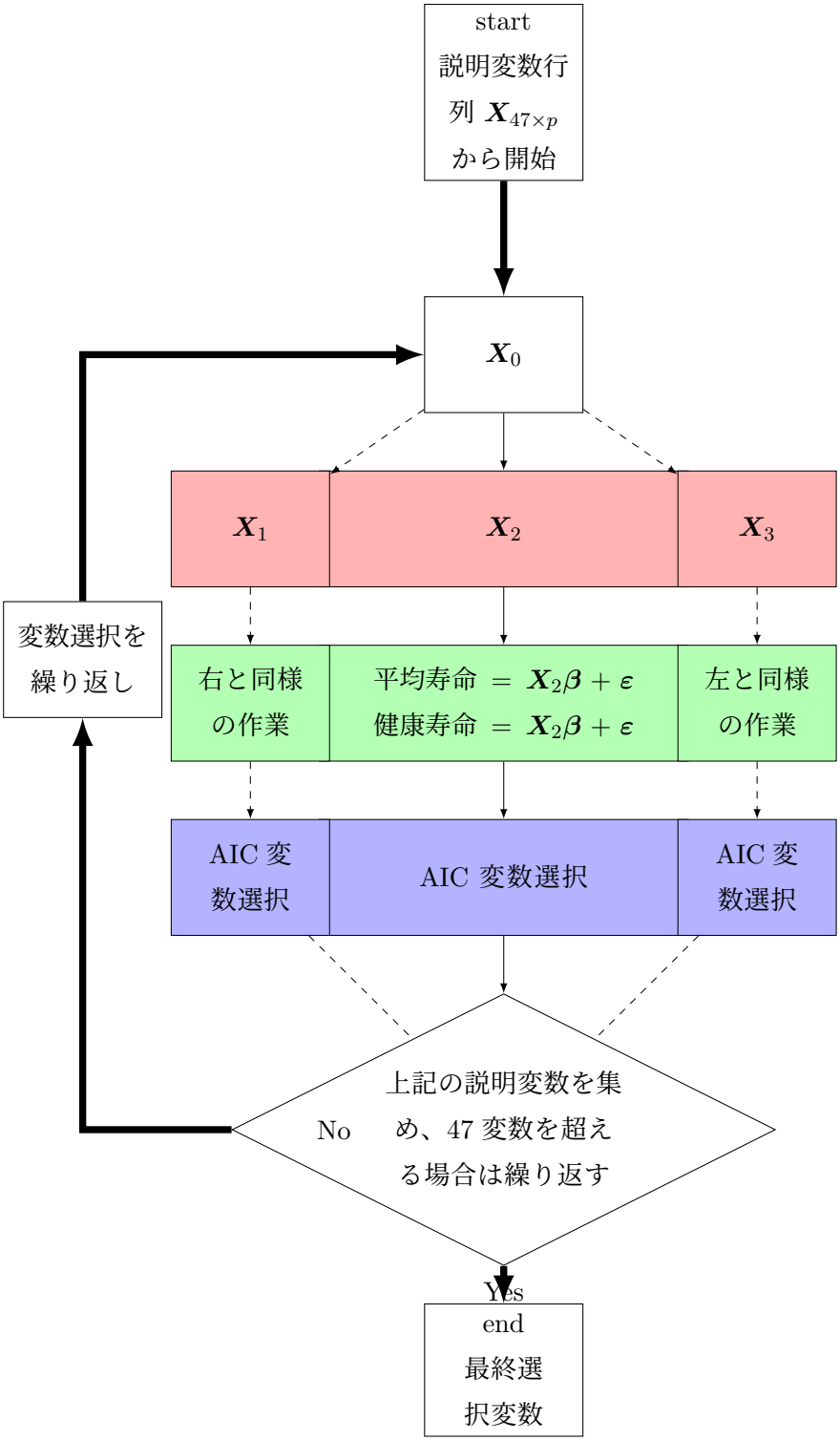

Fig. 4.1　変数選択アルゴリズム

第 5 章　因子分析による次元縮約

5.1　因子分析とは

因子分析とは、多変量データに潜む共通因子を探り出すための使われる多変量解析手法である。

様々な事象（観測変数）を手がかりにして、潜在的に存在する概念（潜在変数）を推定する方法とも言える。本分析で用いられる（観測された）説明変数がこの観測変数に該当し、これらの観測された変数は共通する潜在変数の結果ととして現れたとの考え方である。(図 5.1)

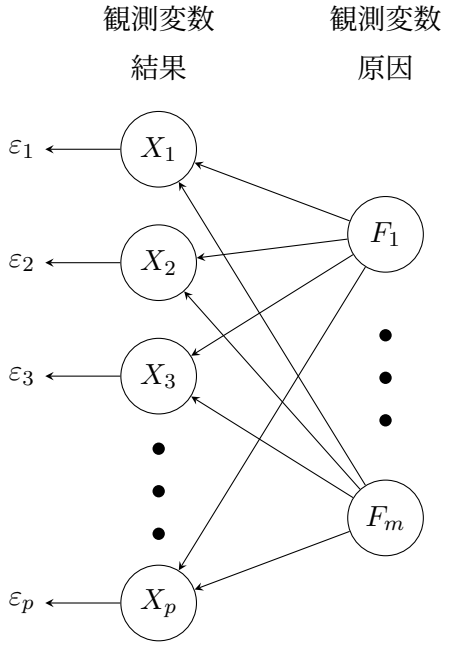

Fig. 5.1　因子分析の概念図

$$
\begin{aligned}
X_1 &= l_{11}F_1 + l_{12}F_2 + \varepsilon && (5.1)\\
&\vdots && (5.2)\\
X_p &= l_{p1}F_1 + l_{p2}F_2 + \varepsilon
\end{aligned}
$$

$$
x = Lf + \varepsilon \tag{5.3}
$$

$$
\begin{aligned}
X_1 &= l_{11}F_1 + l_{12}F_2 + \varepsilon\\
X_2 &= l_{21}F_1 + l_{22}F_2 + \varepsilon
\end{aligned}
$$

これは目には見えず、直接測ることができない「知能」というものが存在し、それが具体的な知能テストや試験などの結果として現れるという考え方を元にしています。

かない、ということになります。このような考え方の視点を広げてみると、消費者の意識、態度は全て潜在的な概念と考えることができ、具体的にモデルを描くと以下の図 1 のようになります。楕円が「悲しい気持ち」という潜在変数で、共通因子といいます。そして四角の X1FF5EX3 までの観測変数が、私達が見ることができる実際の現象と言うことができます。もちろん他の観測変数でも、「悲しい」という心が表出する可能性はあるでしょう。消費者の意識、態度を潜在的な概念と考えたモデル図図 1 また、この図にある e1FF5Ee3 は独自因子といい、それぞれの観測変数に固有の情報を表しています。つまり、それぞれの観測変数が以下のような形で分解されることになります。観測変数　＝　共通因子　＋　独自因子観測変数 = 共通因子 + 独自因子の統計的な概念を図示すると、図 2 のようになります。観測変数 z を共通因子空間 S(F) で説明するとして、z のなかで S(F) で説明できる部分は h となり、できなかった部分が e となります。ここで、e と h は直行するので、z,h,e で直角三角形ができます。このことが、z を h と e に分解する、という意味になり、ピタゴラスの定理が分散分析の根拠になっています。通常、分析をするにあたって関心を持つのは共通因子の方になりますので、単純に「因子」といったときは独自因子ではなく、共通因子を指します。ここで着目していただきたいのが、観測変数の全てを説明するモデルを作るのではなく、他の観測変数の中から共通する部分だけを共通因子空間によって説明し、それだけでは説明できない固有の要素を独自変数として残す、という点です。つまり、より小さな変数で人々の意識を理解する試みが行われており、「枝葉を捨てて、エッセンスを見る」というのが根本的な考え方となっています。この時、それぞれの観測変数に固有の枝が独自因子 e と言うことができます。（朝野熙彦「マーケティング・リサーチ」講談社の第 2 章から引用）観測変数 = 共通因子 + 独自因子の統計的な概念図 2 因子分析結果の読み方

因子分析で得られる指標

因子負荷量

直交解を求めた場合に限りますが各変数と各因子の相関を表します。その場合は因子負荷量は、相関係数なので 0 から ± 1 の値をとります。バリマックス回転が直交解の方法としてよく利用されます。しかし、次の項の共通性の推定により、独自因子の情報は共通因子空間から除かれていることに注意してください。通常、この因子負荷量が高い変数を考慮して、因子の名前をつけます。共通性

各変数が因子空間で表される分散を表しています。0 から 1 の値をとります。これも直交解を求めた場合に限りますが共通性は、各因子負荷量の 2 乗和となります。寄与率

ある因子がどの程度の説明力を持っているか割合を表します。独自因子の割合＝独自性です。Uniqueness と言います。因子分析で得られる指標図 3 因子得点

因子得点は、各因子ごとの各個体（対象者）のスコアを表します。因子得点が高い人は、その因子に影響されている度合いが高いといえます。下記の表は、適性検査の成績を因子分析した結果の一部です。「計算能力」、「図形処理能力」、「言語能力」、「記憶能力」という 4 つの因子が抽出され、対象者ごとの因子得点を求めたものです。因子得点から、対象者を 3 つのグループに分けることができました。因子得点図 4

バリマックス回転　因子分析における直交回転法のひとつで、もっともよく利用されてきた。回転の目的は因子の解釈を容易にすることであり、バリマックス回転（varimax rotation）が解釈しやす

い結果を与えることが多かったために、研究者や実務家に非常に頻繁に利用されてきた。実は回転方法は無数にある。　素朴な疑問として「回転」とは何なのか、なぜ「回転」するのか、ということを理解したいが、それに先立って回転前（初期解）と回転後（回転解）の実例を示す。これはブランド戦略サーベイの企業イメージ 25 変数の因子分析である。どのような変化が回転前後で生じているであろうか。

初期解（回転前）の因子負荷行列

バリマックス回転後（回転解）の因子負荷行列

5.2　回転とは何か

回転は幾何学的概念である。一方、因子負荷量は解析的概念である。データ解析では、しばしば幾何学的表現と解析的表現が、同じ文脈で混在するので、慣れていないと混乱するであろう。因子負荷行列を図的に表現すれば下図のようになる。　因子は２個としてあるので、因子を縦軸と横軸にすれば平面を描くことができる（３因子による空間表示でもかまわない）。変数は６個にして色分けしてある。●は因子１と因子２の因子負荷量の値を座標値とした位置にあるが、見やすいように原点からのベクトルで表現してある。これが因子負荷行列の図的表現である。　回転とは、この平面つまり座標空間で因子（軸）を「回転させる」という幾何学的イメージに準拠している。６変数の相対的位置は変わっていない。下図の回転前後は因子（軸）ではなく、変数が回転しているように見えるが、因子の方を回転しているのである。そして、この回転は 360 度、どのように回転することもできる。無数の回転解が存在する。因子分析は座標空間だけを定めたのである。ちなみに、これをネガティブに「因子の不定性」とか、ポジティブに「回転の自由度」などという。

因子負荷行列の図的表現

＜なぜ回転するのか？＞　解釈しやすい解を得るためである。解釈しやすいとは、どういう状態であろうか。それは単純構造の時である。単純構造という概念は Thurstone が提案したのだが、これを解析的に実現したのが Kaiser で、1958 年に”The varimax criterion for analytic rotation in factor analysis”という論文として Psychometrika に発表した。この時、バリマックス回転が成立した。

因子負荷行列の図的表現をみると、回転前は 6 変数のすべてが因子 1 と関係している。回転後では最初の 3 変数は縦軸と、後の 3 変数は横軸と強い関係を持つように分離している。別の見方をすると、6 変数が 3 変数ごとにグループ化された。関係の強さは幾何学的には因子の軸と各変数ベクトルとの角度の小ささである。回転後は、3 変数はある因子と強く関係し、他の因子とは弱い関係になった。回転によって単純構造に接近したのである。　Kaiser は単純構造を得るには、因子負荷行列の要素の分散を（規準化したうえで）最大化すれば実現できる、というアイデアを得た。大きい負荷はより大きく、小さい負荷はより小さくなるような規準に向かって回転させるので、分散（variance）の最大化（max）、すなわち varimax という名前にしたのである。　最初に示した「ブランド戦略サーベイ」の初期解（回転前）は、因子 1 にほぼすべての変数は高い負荷を持つ。しかし、バリマックス回転後は単純構造に向かって、因子と変数とのコントラストが強化されていることが分かる。これで 4 因子についての解釈は容易な方向に改善されたのである。　＜因子負荷量の計算＞　因子分析の数理的な目標は、因子負荷量の推定である。しかし因子分析モデルは強い制約条件をもつ統計モデルである。そのため、まず計算しやすいような解を最初に求めている。それが「初期解」という名前の意

味である。初期解は第 1 因子の分散が最大になるように計算し、次に第 2 因子の分散、という順番に解を求めているので、因子 1 の負荷量がすべて大きかったのである。　因子分析モデルの制約条件が多い理由は、因子が観測されていない潜在変数であるためである。方程式の本数よりも未知数の方が多いと、一意に解を求めることができないので制約条件を設定して計算している。回転解とはそのような制約のあとに「有用な」解を求めていくことである。なお、回転しても因子分析モデルの共通性や独自性、モデルの適合度などは変化しない。

5.3　抽出された共通因子

Table. 5.1　女性の FA

	var_name_Jpn	F1	F2
1	受療率_外来_脳血管疾患_2017	0.28	0.64
2	人口・世帯_老年人口割合 2020	0.01	0.86
3	人口・世帯_生産年齢人口割合 2020	0.13	-0.92
4	自然環境_年平均気温	-0.25	-0.34
5	労働_完全失業率	0.58	-0.01
6	居住_都市公園数 (可住地面積 100km^2 当たり)	0.17	-0.85
7	高血圧疾患_外来 2014 年	0.36	-0.84
8	悪性新生物 (大腸)_年齢調整死亡率 2015	0.56	-0.07
9	ボランティア総行動率－総数	-0.80	0.08
10	受療率_外来_心疾患_2017	-0.02	0.47
11	居住_一戸建住宅比率	-0.31	0.82
12	75 歳未満調整死亡率_悪政新生物_2019	0.80	0.33
13	診療所数_2019	-0.19	0.08
14	バリアフリー_手すりがある 2018	0.33	-0.86
15	循環器専門医数_2020	0.27	-0.84
16	家計_スマートフォン所有数量（千世帯当たり）	-0.32	-0.73
17	ボランティア総行動率－高齢者を対象とした活動	-0.77	0.24

Table. 5.2　男性の FA

	var_name_Jpn	F1	F2
1	受療率_入院_心疾患_2017	0.02	-0.61
2	自然環境_年平均気温	0.50	0.16
3	健康・医療_保健師数（人口 10 万人当たり）	-0.36	-0.76
4	家計_貯蓄現在高	-0.43	0.65
5	人口・世帯_高齢単身者世帯の割合	0.20	-0.51
6	悪性新生物 (大腸)_年齢調整死亡率 2015	0.65	-0.11
7	自己啓発・訓練－パソコンなどの情報処理	0.08	0.90
8	一定のバリアフリー化率_2018	-0.93	0.07
9	自己啓発・訓練－芸術・文化	-0.11	0.87
10	自己啓発・訓練－英語以外の外国語	0.17	0.82

第 6 章　glm による分析

一般化線形モデルとは、目的変数 Y の期待値 μ と説明変数 F_1, F_2 の関係を

$$g(\mu) = \beta_0 + \beta_0 F_1 + \beta_2 F_2 \tag{6.1}$$

のように表し、データを用いて関数 β_0, β_1 を推定することをその目的とするモデルである。ここで、式の左辺が目的変数の観測値ではなく、平均 (期待値) でモデリングされることに注意する。リンク関数と呼ばれる左辺の関数 $g(x)$ は研究者が決めるものであり、本研究では $g(x)$ をに gamma 関数と　logit 関数を使うモデリングをおこう。

したがって本研究で用いられる寿命変数の平均 (期待値) がモデリングに組み込まれると考えればよい。期待値を仮定するためには必然的に目的変数の分布を仮定する必要がある。上記に回帰モデルには目的変数が正規分布を従うと仮定するモデルである。

以上の設定の基で男性の平均寿命に対する回帰分析モデリングは男性の平均寿命は正規分布を従うと仮定の上で、平均寿命の (条件付き) 期待値が 160 個の説明変数により

$$\text{男性の平均寿命の期待値} = \beta_0 + \beta_0 x_1 + \cdots + \beta_{160} x_{160} \tag{6.2}$$

のようなモデルとなる。

のを変数同士の相関構造を保ちんがら、されに

統計的関連性が認められる変数を選択し、

目的変数と

その基準を通過した変数を同時に考慮して２つの共通因子を抽出することにある。

般化線形モデルとベイズモデルにより分析を進める。

一般化因子

具合的には 160 変数を技術的に回帰分析が可能な数の複数のグループに分けて、目的変数と組み合わせて回帰分析を施した後、stepwise AIC を用いて変数の選択を行う（4 章参照)。

その次、AIC 基準で残った説明変数に対して、多変量分析手法の因子分析を使い２つ共通因子を抽出し、この共通因子を説明変数として扱う (図 3.2 参照)。

のを変数同士の相関構造を保ちんがら、されに

統計的関連性が認められる変数を選択し、

目的変数と

その基準を通過した変数を同時に考慮して２つの共通因子を抽出することにある。

般化線形モデルとベイズモデルにより分析を進める。

一般化因子

具合的には 160 変数を技術的に回帰分析が可能な数の複数のグループに分けて、目的変数と組み合わせて回帰分析を施した後、stepwise AIC を用いて変数の選択を行う（¥refchapter:VarSelection 章参照)。

その次、AIC 基準で残った説明変数に対して、多変量分析手法の因子分析を使い２つ共通因子を抽出し、この共通因子を説明変数として扱う (図¥refModelSuppression 参照)。

統計では、一般化線形モデル（GLM）は、通常の線形回帰を柔軟に一般化したものであり、応答変数に正規分布以外の誤差分布を持たせることができます。 GLM は、線形モデルをリンク関数を介して応答変数に関連付けることを許可し、各測定値の分散の大きさをその予測値の関数にすることにより、線形回帰を一般化します。

一般化線形モデルは、線形回帰、ロジスティック回帰、ポアソン回帰など、他のさまざまな統計モデルを統合する方法として、ジョンネルダーとロバートウェダーバーンによって策定されました。[1] 彼らは、モデルパラメータの最尤推定のために繰り返し再重み付けされた最小二乗法を提案しました。最尤推定は依然として一般的であり、多くの統計計算パッケージのデフォルトの方法です。ベイジアンアプローチや分散安定化応答への最小二乗適合など、他のアプローチが開発されています。

6.1 線形回帰分析

	term	estimate	std.error	statistic	p.value
1	(Intercept)	87.02	0.04	1988.52	0.00
2	d_f_.FA$OBS.rotate1	-0.26	0.04	-5.91	0.00
3	d_f_.FA$OBS.rotate2	-0.09	0.04	-2.13	0.04

Table. 6.1 女性の回帰 withFA(平均寿命)

	term	estimate	std.error	statistic	p.value
1	(Intercept)	80.65	0.06	1393.83	0.00
2	d_m_.FA$OBS.rotate1	-0.25	0.06	-4.29	0.00
3	d_m_.FA$OBS.rotate2	0.34	0.06	5.89	0.00

Table. 6.2 男性の回帰 withFA(平均寿命)

	term	estimate	std.error	statistic	p.value
1	(Intercept)	74.94	0.09	792.01	0.00
2	d_f_.FA$OBS.rotate1	-0.13	0.10	-1.35	0.18
3	d_f_.FA$OBS.rotate2	0.07	0.10	0.73	0.47

Table. 6.3 女性の回帰 withFA(健康寿命)

	term	estimate	std.error	statistic	p.value
1	(Intercept)	72.06	0.07	1000.10	0.00
2	d_m_.FA$OBS.rotate1	-0.10	0.07	-1.41	0.17
3	d_m_.FA$OBS.rotate2	0.15	0.07	2.11	0.04

Table. 6.4 男性の回帰 withFA(健康寿命)

6.2 一般化線形モデル: Gamma dist

	term	estimate	std.error	statistic	p.value
1	(Intercept)	4.466	0.001	8876.243	0.000
2	F1	-0.003	0.001	-5.913	0.000
3	F2	-0.001	0.001	-2.138	0.038

Table. 6.5 女性の一般化線形モデル withFA(平均寿命)

	term	estimate	std.error	statistic	p.value
1	(Intercept)	4.390	0.001	6101.247	0.000
2	F1	-0.003	0.001	-4.279	0.000
3	F2	0.004	0.001	5.881	0.000

Table. 6.6　男性の一般化線形モデル withFA(平均寿命)

	term	estimate	std.error	statistic	p.value
1	(Intercept)	4.317	0.001	3418.628	0.000
2	F1	-0.002	0.001	-1.354	0.183
3	F2	0.001	0.001	0.731	0.469

Table. 6.7　女性の一般化線形モデル withFA(健康寿命)

	term	estimate	std.error	statistic	p.value
1	(Intercept)	4.278	0.001	4278.460	0.000
2	F1	-0.001	0.001	-1.411	0.165
3	F2	0.002	0.001	2.110	0.041

Table. 6.8　男性の一般化線形モデル withFA(健康寿命)

6.3　一般化線形モデル: logit model

	term	estimate	std.error	statistic	p.value
1	(Intercept)	-0.185	0.350	-0.528	0.598
2	d_f_.FA$OBS.rotate1	-1.447	0.520	-2.784	0.005
3	d_f_.FA$OBS.rotate2	-0.575	0.371	-1.548	0.122

Table. 6.9　女性の一般化線形モデル (logit)withFA(平均寿命)

	term	estimate	std.error	statistic	p.value
1	(Intercept)	-0.335	0.455	-0.737	0.461
2	d_m_.FA$OBS.rotate1	-1.778	0.658	-2.703	0.007
3	d_m_.FA$OBS.rotate2	2.566	0.798	3.213	0.001

Table. 6.10　男性の一般化線形モデル (logit)withFA(平均寿命)

	term	estimate	std.error	statistic	p.value
1	(Intercept)	-0.098	0.322	-0.304	0.761
2	d_m_.FA$OBS.rotate1	-0.782	0.412	-1.896	0.058
3	d_m_.FA$OBS.rotate2	0.658	0.349	1.888	0.059

Table. 6.11　男性の一般化線形モデル (logit)withFA(健康寿命)

第 7 章　ベイズモデルによる推定

bayesian multilevel モデル

https://www.stata.com/features/overview/bayesian-multilevel-models/

何のことか？

multilevel モデルは、グループ固有の効果を組み込んだ回帰モデル。

グループは、病院、病院内にネストされた医師、病院内にネストされた医師内にネストされた患者など、グループ固有の効果は、いくつかの事前分布、(通常は正規分布に従って)、グループ間でランダムに変化すると想定。

さまざまなレベルの階層を表す場合があ。この仮定により、multilevel モデルはベイズ分析の自然な候補にな。bayes multilevel モデルはさらに、回帰係数や分散成分（グループ固有の効果の分散）などの他のモデルパラメーターもランダムであると想定。

bayesmultilevel モデルを使用する理由 bayes 分析の標準的な理由に加えて、bayesmultilevel モデリングは、グループの数が少ない場合、または多くの階層レベルが存在する場合によく使用されます。

逸脱度情報量基準（DIC）などのベイズ情報量基準も、multilevel モデルの比較によく使用されます。グループの比較が主な関心事である場合、bayesmultilevel モデリングは、グループ固有の効果の分布全体を提供できます。

multilevel コマンドの前にベイズを付けるだけ。

bayesmultilevel モデルを Stata に適合させることができ、これを簡単に行うことができます。。ベイズ：混合 y x1 x2 —— id：もちろん、「簡単に」と言うときは、モデルの定式化ではなく、モデルの仕様を指します。他のモデリングタスクと同様に、bayesmultilevel モデリングでは慎重に検討する必要があ。

連続、打ち切り、バイナリ、序数、カウント、GLM、および生存の結果がサポートされています。

サポートされている multilevel コマンドの完全なリストを参照してください。

複数レベルの階層、ネストおよびクロスされたランダム効果、ランダム切片と係数、ランダム効果共分散構造など、

すべての multilevel 機能を利用できます。

[BAYES] bayesmh コマンドによって提供されるすべてのベイズ機能は、

multilevel コマンドでベイズプレフィックスを使用する場合にサポートされます。

bayesmultilevel モデリングの新機能もご覧ください。

$$p(\theta|y) \propto p(\theta)p(y|\theta) \tag{7.1}$$

$$\text{寿命変数} = \beta_0 + \beta_0 F_1 + \beta_2 F_2 + \varepsilon \tag{7.2}$$

事前分布は

$$\beta_0 \sim N() \tag{7.3}$$

$$\beta_2 \sim N() \tag{7.4}$$

$$\beta_3 \sim N() \tag{7.5}$$

7.1 bayes モデルによる係数の事後分布推定

	mean	sd	5.5%	94.5%
beta0	87.022	0.042	86.954	87.090
beta1	-0.261	0.043	-0.330	-0.193
beta2	-0.094	0.043	-0.163	-0.026
sigma	0.290	0.030	0.242	0.338

Table. 7.1　女性の Bayes(平均寿命)

	mean	sd	5.5%	94.5%
beta0	80.652	0.056	80.563	80.742
beta1	-0.251	0.057	-0.342	-0.161
beta2	0.345	0.057	0.254	0.435
sigma	0.384	0.040	0.321	0.447

Table. 7.2　男性の Bayes(平均寿命)

	mean	sd	5.5%	94.5%
beta0	74.940	0.092	74.794	75.086
beta1	-0.129	0.093	-0.277	0.018
beta2	0.070	0.093	-0.078	0.218
sigma	0.628	0.065	0.524	0.731

Table. 7.3　女性の Bayes(健康寿命)

7.2 因子による寿命の生存曲線への影響

7.3 bayes モデルによる寿命の事後分布の推定

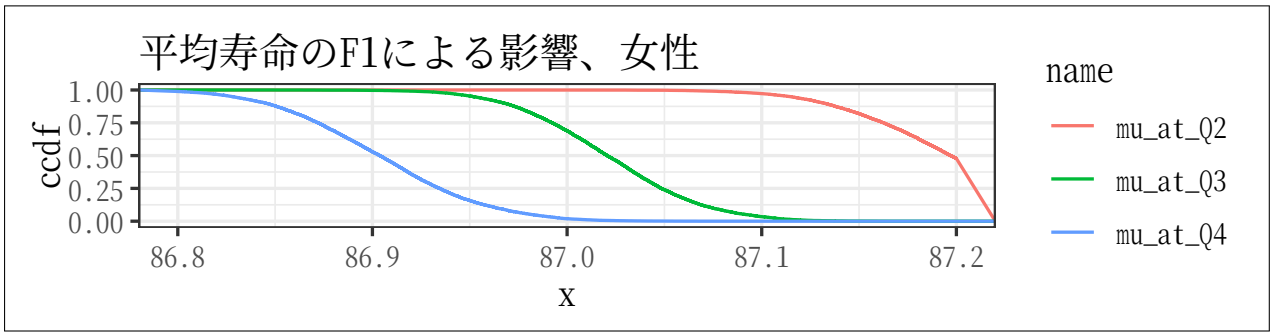

Fig. 7.1　Bayes_LE_f_ccdf_F1

	mean	sd	5.5%	94.5%
beta0	72.064	0.070	71.953	72.175
beta1	-0.103	0.070	-0.215	0.010
beta2	0.153	0.070	0.041	0.266
sigma	0.478	0.049	0.399	0.557

Table. 7.4　男性の Bayes(健康寿命)

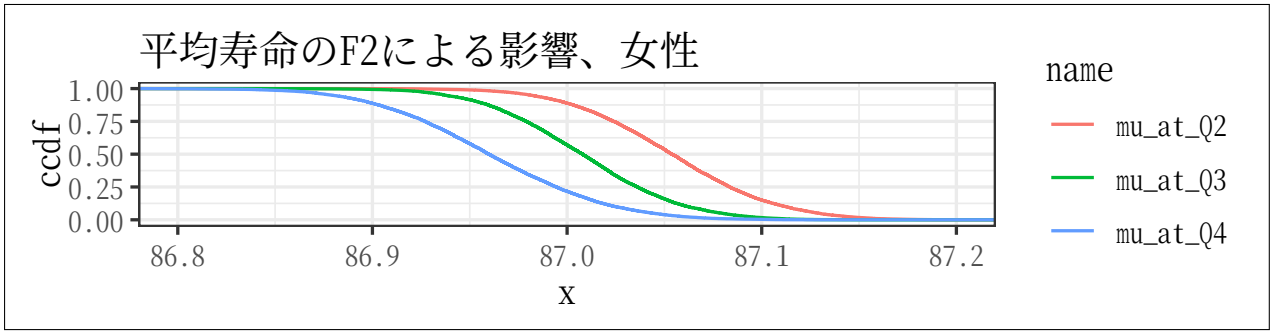

Fig. 7.2　Bayes_LE_f_ccdf_F2

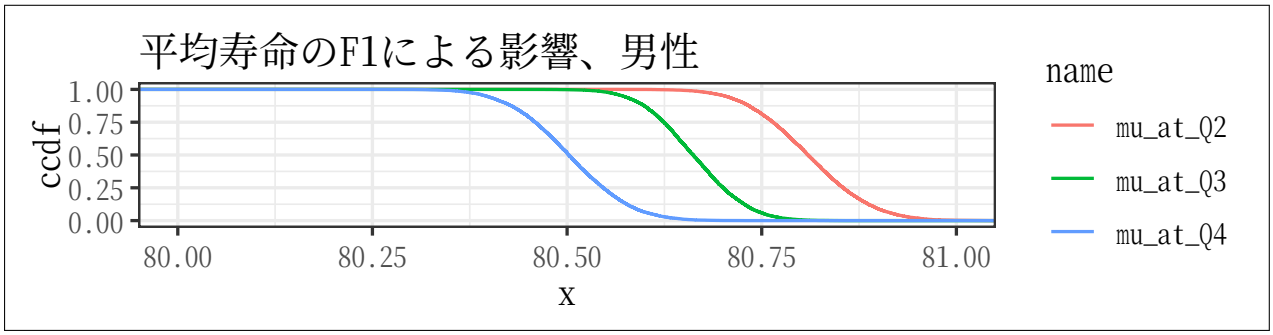

Fig. 7.3　Bayes_LE_m_ccdf_F1

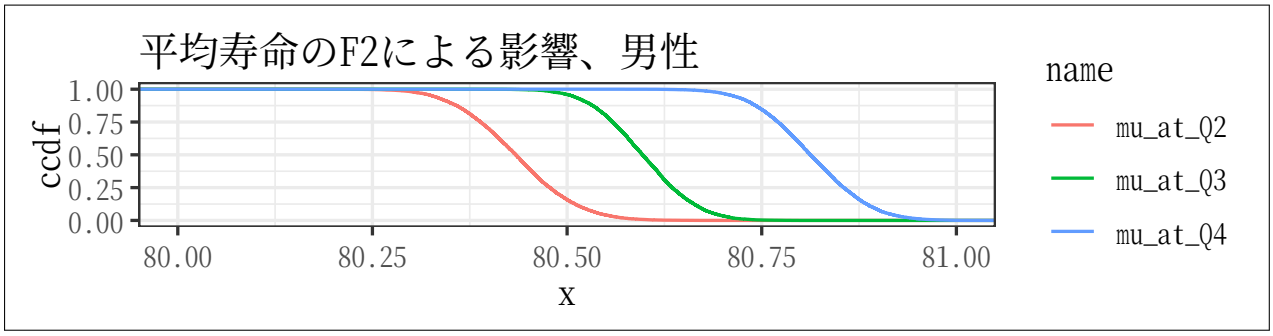

Fig. 7.4　Bayes_LE_m_ccdf_F2

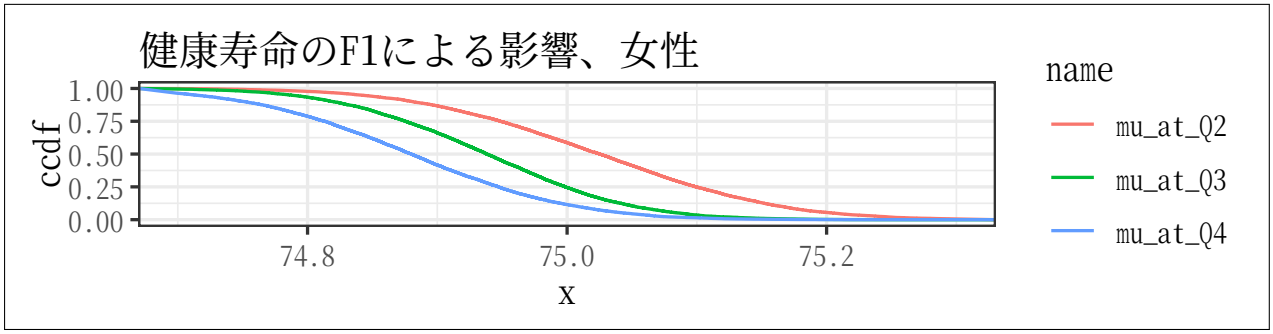

Fig. 7.5　Bayes_HLE_f_ccdf_F1

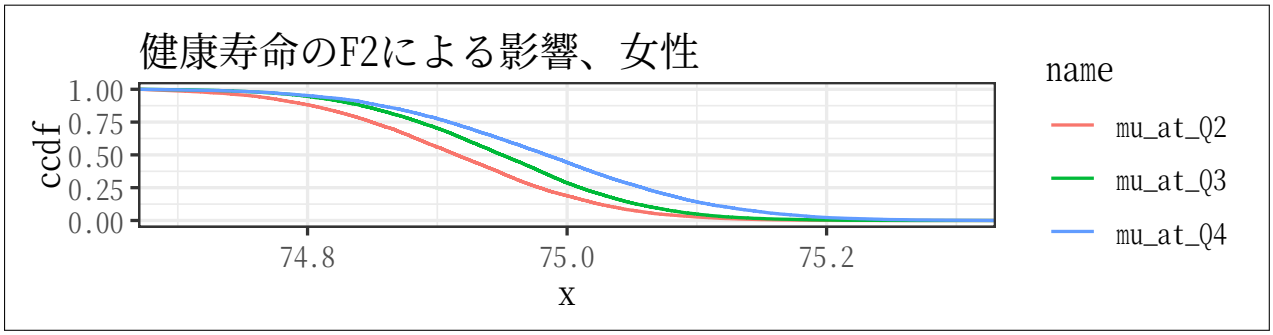

Fig. 7.6　Bayes_HLE_f_ccdf_F2

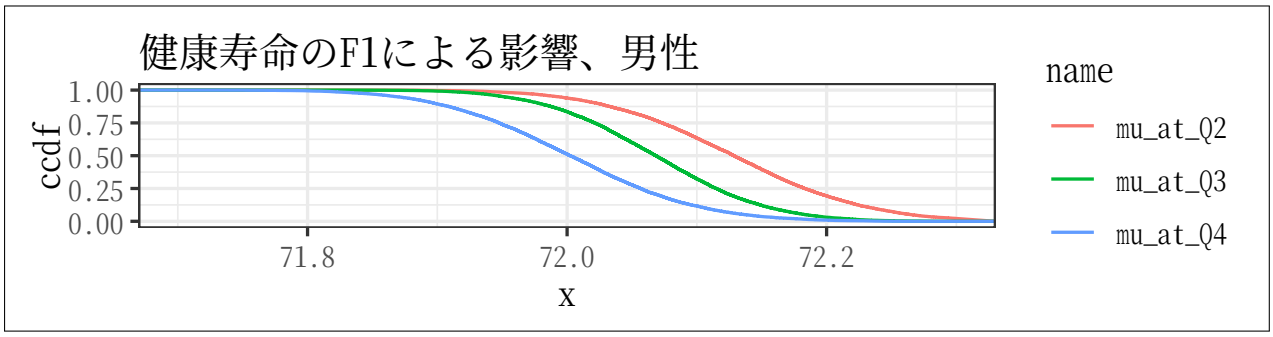

Fig. 7.7　Bayes_HLE_m_ccdf_F1

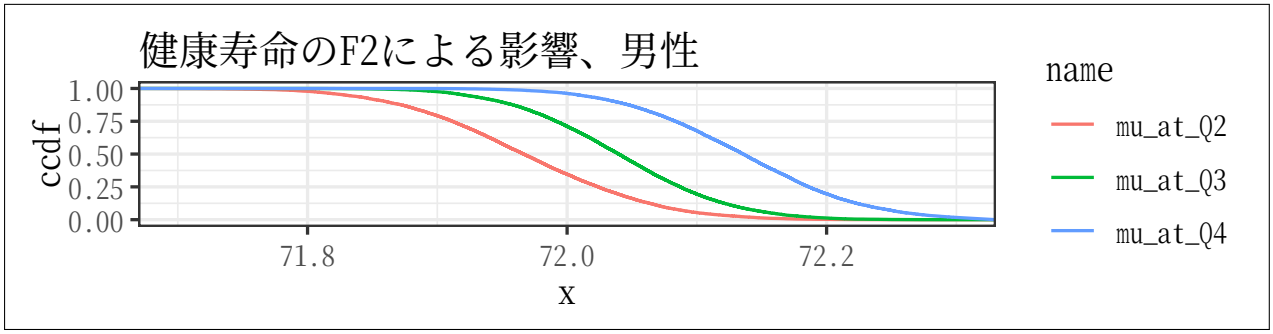

Fig. 7.8 Bayes_HLE_m_ccdf_F2

第 8 章　まとめ

結果、

今後の展望・分析の提言

研究の限界

5 参考資料（Reference）　引用文献、参考文献

6 添付（Appendix）