2022年度　統計データ分析コンペティション

審査員奨励賞［大学生・一般の部］

欠測値を含むパネルデータを用いた健康寿命の要因分析

高須 尚哉、近藤 雅哉、山中 美幸、丹羽 美歩
（南山大学総合政策学部）

欠測値を含むパネルデータを用いた健康寿命の要因分析

高須尚哉・近藤雅哉・山中美幸・丹羽美歩
（南山大学総合政策学部総合政策学科）

1. 研究のテーマと目的

近年、医療の発達とともに世界的に平均寿命が伸びている。内閣府の「令和 3 年版高齢社会白書」[1]では日本の高齢化率は 28.8%となっており、人数に直すと 3,619 万人いると示されている。日本の総人口の約 3 割を占める高齢者は世代で人口が異なり、前期高齢者年齢である 65~74 歳人口は 1747 万人、後期高齢者年齢である 75 歳以上人口は 1872 万人となっており、後期高齢者人口が前期高齢者人口を上回る。厚生労働省の e-ヘルスネット[2]によると、平均寿命と健康寿命の差は男性において約 9 年、女性においては約 12 年であり、「2001 年と比べても、平均寿命と健康寿命は男女ともに延伸しているが差は縮小していない」とも述べられている。これは平均寿命と健康寿命が年々伸長する中で、持病や加齢により寿命そのものは伸びているものの、過去と比較して健康的な生活を送る事が出来ない期間が長い人が増えていると考えられる。この状態で日本における最大の社会問題の一つである超少子高齢社会が続けば、年金問題や老後の資産問題も現在より更に深刻になるだろう。就業年数についても、一般的に 60 歳であった定年退職年齢が、2021 年の高齢者雇用安定法の改定により、65 歳までの雇用確保の義務化、70 歳までの高年齢者の就業機会の確保が努力義務として定められた。また、図 1 の平均寿命の推移については緩やかではあるものの、今後徐々に伸び続けていくと予想されている。そのため、それに伴い人々の就業年数が現在よりもさらに伸びる可能性が考えられる。高齢者がより長期間働くことや、自助可能な健康寿命の伸長は超少子高齢社会へ対応するためにも必要不可欠であり、実際これまでも健康寿命の延伸に向けた数々の先行研究がなされてきた。

健康寿命の要因分析についての先行研究としては、西條（2022）[3]では健康寿命と各自治体の取り組みにおいての関係についての研究が行われたが、特徴的な相関は見られなかった。また、Hosokawa at al.（2020）[4]では説明変数に病床数や医療従事者数等の医療資源に焦点を当てたクロスセクションデータを使用した分析研究行った結果、男女ともに在宅医療サービスと医療費が正に有意で健康寿命に影響与えるとして、健康寿命延伸のために医療資源の最適な配分を提言している。これらの先行研究にみられる様に健康寿命についての要因は多岐に渡り、気温や食文化などの地域的なものから運動習慣、食生活、睡眠時間などの習慣的なものまで様々である。そうした分野毎の健康寿命の要因分析を行っている先行研究は多くなされているが、複数分野の説明変数を使用したパネルデータを用いた先行研究は少ない。そのため本研究では一つの分野についてではなく複数の分野について都道府県別のパネルデータを用いて重回帰分析をおこない、健康寿命の規定要因の解明を試みた。また各先行研究でも述べられている様に男女間で平均寿命や健康寿命が異なってくるため本研究における分析も男女別で行うこととした。

また、本研究における健康寿命の定義は「日常生活に制限のない期間の平均」を指す。基礎資料として、健康情報は国民生活基礎調査、死亡情報は人口動態統計から算出されている。

加えて、今回、分析のために収集した統計データの中に数ヵ所ではあるが欠損が見られたことから、推定結果の偏りや、それに伴う分析精度の低下が生じる懸念から、本研究では欠損値をリストワイズ法と多重代入法で処理したデータセットの検証を併せて行う。

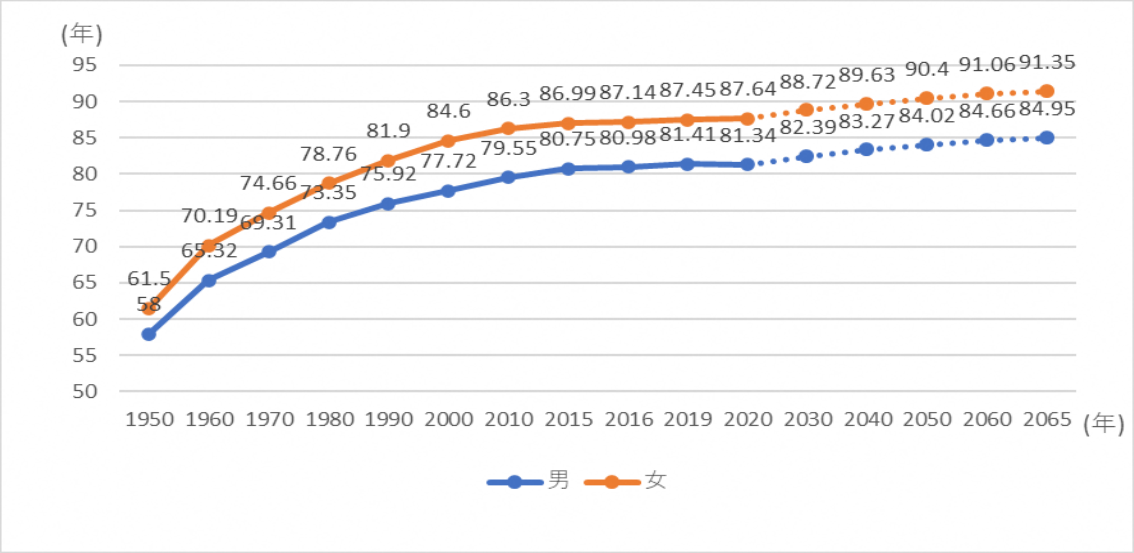

図 1　男女別平均寿命の推移.

1950 年は厚生労働省「簡易生命表」、1960 年から 2015 年までは厚生労働省「完全生命表」、2018 年は厚生労働省「簡易生命表」、2020 年以降は、国立社会保障・人口問題研究所「日本の将来設計人口（平成 29 年推計）」の出生中位・死亡中位仮定による推計結果。1970 年以前は沖縄県を除く値である。0 歳の平均寿命が「平均寿命」である。

2.研究の方法と手順

本研究では、各都道府県の 2010、2013、2016、2019 年の 4 時点における平均健康寿命を被説明変数として、説明変数は各年の 3 年前(2007、2010、2013、2016 年)のラグ変数を使用してパネルデータ分析を行い、健康寿命の規定要因と、各変数との因果関係を明らかにすることを目的としている。

今回収集した統計データの中に数ヶ所ではあるが欠測がみられたことから、欠測値を含む対象をリストワイズ法により欠測値を含むデータ項目を削除したデータでの分析を検討する。しかし、高橋・渡辺(2017)[5] によると、リストワイズ除去を行ったデータセットによる分析は分析結果の偏りや精度の低下の懸念があることから、本研究においても極少数とは言え欠測が生じている以上は分析結果への信頼性が揺らぐ可能性を考慮し、簡単にではあるが併せて多重代入法による分析を行い、リストワイズ除去と多重代入法の比較検討を行う。「3 データセットの加工」にて詳しく述べるが、本研究で扱うデータセットにおける欠測メカニズムは Missing At Random(以下 MAR)であると推定出来ると考えられるため、欠測値への対処として多重代入法を採択することへの妥当性については問題ないと考えられることから、本研究では連鎖方程式による多重代入法（multiple imputation by chained equation. 以下 MICE）[6] を用いた分析を行った。なお、本研究においては、多重代入法によって欠損値を補完した 50 個のデータセットを作成し、統合した結果を分析した。代入モデルは以下の通り（式 1）である。

$$\tilde{y}_{i,m} = \tilde{\beta}_{0,m} + \tilde{\beta}_{1,m}x_i + \cdots + \tilde{\beta}_{7,m}x_i + \tilde{\varepsilon}_{i,m} \qquad (1)$$

*mは 1～50、$\tilde{y}_{i,m}$は欠損している説明変数（喫煙率、酒類消費量）、βは回帰係数((1)、

(2)共通))、x_iは欠損していないその他の変数、ε_iは誤差項となっている((1)、(2)共通)、観測数は 188 となっている。

リストワイズ除去を行ったデータセットにおけるパネルデータ分析では以下の回帰モデル(式２)を想定した。

$$y_{it} = \beta_0 + \sum_{k=1}^{7} \beta_k x_{k,i(t-3)} + \varepsilon_i \qquad (2)$$

* yは健康寿命、iは各都道府県番号(i=1、2、…、46、欠測の関係で沖縄を除くため 46)、tは年度(t=2010、2013、2016、2019)、β_0は切片、$x_{k,i}$は説明変数となっている。また、先行研究の結果からも分かるように、健康寿命に影響する要因は男女間で異なるため本研究においても男女別の要因分析を行った。

なお、このモデルにおいて年度による影響を考慮するために時点効果は年度ダミーとして用いられている。

2.1 仮説

以下の表 1 は各説明変数の符号を予想したものである。被説明変数である健康寿命に正の相関が予想される説明変数には＋を、負の相関が予想される説明変数には－を記入した。

表 1：符号の予想

説明変数	符号の予想
年平均気温	－
医療費	－
体育館数（公共）	＋
多目的運動広場数（公共）	＋
喫煙率	－
酒類消費	－
降水日数(年間)	－

３．データセットの加工

本研究では、教育用標準データセットに加え、他の出典からもデータを取得した。各データは全て都道府県別に取得した。男女別データが入手できた健康寿命と喫煙率のみ、男女別データを用いた。ただし、男女ともに熊本県（2016 年）の健康寿命と喫煙率のデータが欠測している。これは 2016 年に発生した熊本地震の影響によるもので、国民生活基礎調査が熊本県を調査していないため、熊本県が含まれていない。また、沖縄県（全時点）の酒類消費のデータも欠測している。この欠測については統計データの出典である国税庁にも直接問い合わせたが、「統計データを取集した業者に何らかの意図があっての事だろうが、国税庁としては不明」とのことだった。調査年度については、被説明変数である健康寿命は 2010、2013、2016、2019 年とした。説明変数の調査年度は被説明変数のそれより以前のものを使用した。体育館数と多目的運動広場数は 2008、2011、2015、2018 年、その他説明変数に関しては 2007、2010、2013、2016 年とした。分析に使用したデータは以下表 2 の通りである。

医療費は「一人当たり実績医療費（千円）」の合計金額を指す。小数点以下第一位を四捨五入して算出したデータを用いた。喫煙率は国民生活基礎調査によるもので、成人が対象である。各時点で喫煙率の定義が異なる。2003~2010 年の喫煙率の定義は、『現在習慣的に喫煙している者（これまで合計 100 本以上又は 6 カ月以上たばこを吸っている（吸っていた）者のうち、

「この 1 か月間に毎日又はときどきたばこを吸っている」と回答した者）』の割合（%）である。対して、2011 年以降の喫煙率の定義は、『これまで習慣的にたばこを吸っていたことがある者のうち、「この 1 カ月間に毎日又は時々たばこを吸っている」と回答した者』の割合である。酒類消費は「成人一人当たりの酒類販売（消費）数量（ℓ）」を指す。ただし、「みりん」の項目を除いて算出したデータを用いた。

表 2：分析に使用したデータとその出典一覧

項目名	単位	出典	調査年度
a 健康寿命(hle)	年	厚生労働省 第16回健康日本21(第二次) 推進専門委員会資料3-1	2010～2019
b 年平均気温(tmp)	℃	教育用標準データセット SSDSE-B	2007～2016
c 実績医療費の合計(medicsp)	千円	厚生労働省 医療保険データベース	2007～2016
d 体育館数(公共)(gym)	施設/100万人	文部科学省 社会教育調査	2008～2018
e 多目的運動広場数(公共)(multi)	施設/100万人	文部科学省 社会教育調査	2008～2018
f 喫煙率(smoke)	%	国立がん研究センター	2007～2016
g 酒類消費(alc)	ℓ	国税庁 酒のしおり	2007～2016
h 年降水日数(rainday)	日	教育用標準データセット SSDSE-B	2007～2016

以下表 3・4 は分析に使用した変数の項目と記述統計を男女別に示したものである。VIF は男女ともに 10 を超える数値はないため、今回の分析では多重共線性の問題を考慮する必要はないと言える。

表 3：分析に使用した変数の項目と記述統計（女性）

変数	観測数	平均	標準偏差	最小値	最大値	VIF
健康寿命	187	74.66	0.99	72.37	77.58	-
平均気温	188	15.87	2.37	9.20	24.10	2.25044
医療費	188	357.73	59.91	249.00	518.00	3.278893
体育館数	188	82.57	44.55	16.5	233.90	4.925851
多目的運動広場数	188	84.30	40.39	8.70	194.60	3.22122
喫煙率	187	9.64	2.38	5.00	20.60	2.774409
酒類消費量	184	83.54	11.8	61.10	127.40	1.683816
降水日数	188	117.35	28.09	67.00	197.00	2.190629

表 4：分析に使用した変数の項目と記述統計（男性）

変数	観測数	平均	標準偏差	最小値	最大値	VIF
健康寿命	187	71.58	1.03	68.95	73.72	-
平均気温	188	15.87	2.37	9.20	24.10	1.8815
医療費	188	357.72	59.91	249.00	518.00	3.212694
体育館数	188	82.57	44.55	16.5	233.90	4.450748
多目的運動広場数	188	84.30	40.39	8.70	194.60	3.693569
喫煙率	187	34.68	3.88	27.00	45.30	4.156043
酒類消費量	184	83.54	11.8	61.10	127.40	1.470916
降水日数	188	117.35	28.09	67.00	197	2.207583

4．データ分析の結果

表 5：分析結果

	女性				男性			
	多重代入法	リストワイズ法			多重代入法	リストワイズ法		
	プールド	プールド	固定効果	変量効果	プールド	プールド	固定効果	変量効果
定数項	75.5156	74.776***	ー	74.971***	73.5210	72.987***	ー	72.008***
	(0.966)	(1.013)	(ー)	(1.379)	(1.085)	(1.094)	(ー)	(1.323)
年平均気温	0.0478*	0.072**	−0.302	0.038	0.0249	0.061**	−0.181	0.055
	(0.029)	(0.035)	(0.223)	(0.051)	(0.022)	(0.025)	(0.184)	(0.038)
医療費	−0.0056***	−0.006***	−0.001	−0.003*	−0.0052***	−0.006***	−0.001	−0.005***
	(0.001)	(0.001)	(0.003)	(0.002)	(0.001)	(0.001)	(0.002)	(0.001)
体育館数	−0.0013	−0.001	0.01	0.002	−0.0057***	−0.006***	0.003	−0.004*
	(0.002)	(0.002)	(0.008)	(0.003)	(0.002)	(0.002)	(0.007)	(0.003)
多目的運動広場数	0.0059***	0.006***	−0.012	0.002	0.0059***	0.006***	−0.005	0.004
	(0.002)	(0.002)	(0.007)	(0.003)	(0.002)	(0.002)	(0.006)	(0.003)
喫煙率	0.0099	0.017	−0.0004	0.012	−0.0198	−0.016	−0.002	−0.007
	(0.033)	(0.034)	(0.050)	(0.039)	(0.020)	(0.020)	(0.026)	(0.021)
酒類消費	−0.0145***	−0.014***	−0.023**	−0.017**	−0.0164***	−0.015***	−0.002	−0.012**
	(0.005)	(0.005)	(0.012)	(0.007)	(0.004)	(0.004)	(0.010)	(0.005)
年降水日数	0.0085***	0.010***	0.002	0.005*	0.0077***	0.009***	0.008**	0.008***
	(0.002)	(0.003)	(0.004)	(0.003)	(0.002)	(0.002)	(0.004)	(0.003)
2013年ダミー	−0.00004	−0.043	0.745**	0.356	0.0772	−0.089	0.486	0.182
	(ー)	(0.266)	(0.371)	(0.282)	(ー)	(0.244)	(0.333)	(0.257)
2016年ダミー	0.6527***	0.626***	1.015***	0.862***	1.0035***	0.939***	1.464***	1.168***
	(ー)	(0.207)	(0.281)	(0.210)	(ー)	(0.197)	(0.266)	(0.202)
2019年ダミー	1.3073***	1.296***	1.82***	1.514***	1.5628***	1.520***	2.137***	1.755***
	(ー)	(0.197)	(0.263)	(0.187)	(ー)	(0.217)	(0.285)	(0.217)
観測数	ー	182	182	182	ー	182	182	182
決定係数	0.578	0.582	0.787	0.565	0.757	0.770	0.894	0.765

注：()内の数値は、各モデルの標準誤差を示す。***,**,*はそれぞれ 1%、5%、10%水準で統計的に有意であることを示す。また、本研究では多重代入を行うにあたって、統計解析ソフト「R」内の mice パッケージを使用したが、van Buuren(2012)[7] によると、当パッケージの仕様上、パネルデータに対応しきれておらず、モデル診断が不可能なため単純なプールドモデルによる重回帰分析を行った。

表 5 より、多重代入法とリストワイズ法それぞれのプールドモデルの結果を比較すると、男女共に一般的に多重代入を行うと数値が大きくなりがちな標準誤差に大差はなく、分析結果もそれ程差は認められなかった。そのため、その他モデルの比較についてもプールドモデルと同様の結果であると予想出来ることから、本研究の場合ではリストワイズ法によって欠損値を処理したデータセットの分析を行っても問題はないと言える。したがって、以下ではパネルデータを各種検定にかけて採択されたモデルについての結果を述べる。

・女性

式（2）について F 検定を行ったところ、F 値は 6.1812 で 1%水準で個別効果が存在する。Hausman 検定の結果、カイ二乗値は 11.14 であり、p値は 0.3465 であるため変量効果モデルが選ばれた。

リストワイズ法の変量効果モデルでは、医療費の係数推定値は約－0.003 で、10%水準で有意であった。これは実績医療費の合計が 1000 円増えると健康寿命が約 0.003 年減少することを示している。酒類消費の係数推定値は約－0.017 で、5%水準で有意である。これは酒類消費

が 1 リットル増えると健康寿命が約 0.017 年減少することを示している。年降水日数の係数推定値は約 0.005 で、10%水準で有意であった。これは年降水日数が 1 日増えると健康寿命が約 0.005 年増加することを示している。

・男性

女性と同様に式（2）について F 検定を行ったところ、F 値は 5.1894 で 1%水準で個別効果が存在する。Hausman 検定の結果、カイ二乗値は 9.64 であり、p値は 0.4722 であるため、女性と同じく変量効果モデルが選ばれた。

リストワイズ法の変量効果モデルでは、医療費の係数推定値は約－0.005 で、1%水準で有意であった。これは実績医療費の合計が 1000 円増えると健康寿命が約 0.005 年減少することを示している。また、体育館数の係数推定値は約－0.004 で、10%水準で有意であった。これは体育館数が人口 100 万人当たり 1 棟増えると健康寿命が約 0.004 年減少することを示している。酒類消費の係数推定値は約－0.012 であり、5%水準で有意であった。これは酒類消費が 1 リットル増えると健康寿命が 0.012 年減少することを示している。年降水日数の係数推定値は約 0.008 で、1%水準で有意であった。これは年降水日数が 1 日増加すると健康寿命が 0.008 年増加することを示している。

5.結果の考察

まず、男女ともに符号が一致し、統計的に有意であった変数として医療費・多目的運動広場数・酒類消費・降水日数がある。医療費は仮説と同様に負で有意となった。医療技術の向上などにより平均寿命・健康寿命は年々延伸してきてはいるが、加齢に伴って慢性的な疾病に罹るリスクが高まることから、歳を重ねる毎に自然と医療費がかかる様になり、その結果として医療費では負の結果になったと考えられる。次に、多目的運動広場数について、これまで多くの研究において運動・スポーツが健康寿命、平均寿命に与える影響は大きいとされており、本研究においては多目的運動広場数という施設の数に注目して運動と健康寿命の関係を分析した結果、正に有意という結果が得られた。昨今、ボール遊び禁止の公園の増加等から近所の公園でボールを使用したスポーツをすることは難しい状況にある人々が多い中で、身近に手軽に利用することの出来る多目的運動広場が増えれば、運動を通じた家族・友人との交流や個人が習慣的な運動を行う事が容易となり、健康寿命の延伸にとって良い影響をもたらすのではないだろうか。逆に、一般的な通説として健康寿命に悪影響を与えるとされている酒に関する変数である酒類消費量については、本研究においても推定値に負の符号が現れたことから、健康寿命の延伸の為には酒類の摂取量を減らすことは重要であると言える。また、降水日数について、仮説における降水日数が多ければ運動する機会が減少し、健康寿命に負の影響を与えるという予想に反して正に有意の結果となった。降水日数の多い都道府県が北陸・東北の日本海側地域に集中して分布していることから、他の地域では人々が外を出歩かなくなる冬季に雪かきで日常的に体を動かしているからこの様な結果になったと考察出来るのではないだろうか。或いは、降水日数に関係する環境要因等、本研究で説明変数として用いなかったその他の要因を代替している可能性も考えられる。体育館数については、仮説の段階では多目的運動広場数と同様に運動習慣に関係する変数として正であると予想していたが、女性は統計的に有意ではなく、男性は 1%水準で負に有意であるという結果となった。

6. 貢献と限界

本研究では欠測値を含む都道府県パネルデータについてリストワイズ法と多重代入法による比較検証の後、データ分析を行った。その結果、リストワイズ法と多重代入法では特筆すべき差は見られなかった。このことから、今回使用したデータセットには欠損が数ヵ所見られたものの、多重代入を行う必要はなく、リストワイズ法による分析で対応可能であることが分かった。また、健康寿命に影響を与えるとされた要因として示された、医療費・多目的運動広場数・酒類消費量・年降水日数のうち、多目的運動広場数と酒類消費量については、それぞれ、公営の多目的運動広場を増やす、これまで通り、過度な飲酒は健康に悪影響を及ぼす事を周知させる。といった具体的な政策立案によって健康寿命の延伸に寄与する事が可能であると考えられる。

一方で、本研究の限界として、リストワイズ法と多重代入法の比較に際して、多重代入法を行うために使用した mice パッケージでは仕様上、パネルデータへの対応が難しかった事から、リストワイズ法のモデルとして採用された変量効果モデルとの厳密な比較が出来たとは言えず、今後の課題としては、Amelia パッケージなどのパネルデータに対応した手段を用い、分析を行う事がまず一つ挙げられる。次に、説明変数のうち、医療費、年降水日数には選択、及び処理に問題があったと考えられる。まず、医療費は厚生労働省の医療保険データベースからの出典であるが、本研究においては合計金額を統計データとして使用しており、歯科や内科などどういった分野の医療費が健康寿命に影響を与えているのかを明確にする事が出来なかった。また、ラグ変数を使用したものの、医療費が健康寿命に影響を与えているのか、その逆で健康寿命が医療費に影響を与えているかの因果関係が不明であるため、操作変数法を用いて因果関係を明確にする必要性があったと考えられる。

また、降水日数については、５節で述べた様に関連するその他の環境要因を代替している可能性が考えられるため、新たにいくつかの環境に関する変数を加える事で分析の精度を上げる事が可能だろう。

参考文献

1. 内閣府「令和 3 年版高齢社会白書（概要版）」
https://www8.cao.go.jp/kourei/whitepaper/w-2021/html/gaiyou/s1_1.html

2.厚生労働省,e-ヘルスネット「健康寿命延伸プラン」
https://www.e-healthnet.mhlw.go.jp/information/hale/h-01-004.html

3.西條ひろみ（2022）,「健康寿命延伸施策に関する自治体間の比較研究」大阪商業大学共同参画研究所紀要,第 3 号,pp49-81

4. Rikuya Hosokawa, Toshiyuki Ojima, Tomoya Myojin, Jun Aida, Katsunori Kondo, Naoki Kondo, “Associations between Healthcare Resources and Healthy Life Expectancy: A Descriptive Study across Secondary Medical Areas in Japan.” International Journal of Environmental Research and Public Health, 2020

5.高橋将宜・渡辺美智子,「欠測データ処理　R による単一代入法と多重代入法」共立出版（2017）

6.野間久史,「連鎖方程式による多重代入法」,『応用統計学』,vol.46,no2,(2017 年)pp67-86
7. Van Buuren, S. (2012) Flexible Imputation of Missing Data, Chapter 2. Multiple Imputation. Chapman and Hall/CRC Press, Boca Raton.