研究ノート

屈曲点を持つことが予想される場合のデータ補間アルゴリズム[1][2]

小川　浩

Simple interpolation algorithm for kinked data

Hiroshi Ogawa
Kanagawa University

【要約】　マイクロデータの分析においては、欠損値をどのように処理するかは重要な意思決定となる。特にパネルデータを扱っている場合は、元々のデータ件数が少ないことが多い上、調査期間が長期にわたるため欠損値が全くないデータはむしろ珍しくなる。欠損値が存在するサンプルをすべて捨てても十分なサンプル数が確保できるデータセットであれば特段の問題はないが、欠損値があるサンプルであっても利用せざるをえない状況は少なくない。

本稿では、家計経済研究所が実施している「消費生活に関するパネル調査」の個票データのなかで、調査対象となっている女性の両親の収入に関する変数を用いて、定年などの理由によって大きな収入低下（屈曲点）が存在するという経験的な知識を用いて推定するアルゴリズムを試した。このアルゴリズムおよびベンチマークとして単純な前後のデータ平均をとる方法を、欠損値がないサンプルセット（つまり、真の値がわかっているデータ）から乱数的に欠損値を作成したサンプルセットに適用し、推定値を真の値と比較した。その結果、本稿で提案するアルゴリズムの方が真の値との誤差の標準偏差が小さくなり、優位性が示せた。

【キーワード】　データ補間　マイクロデータ　アルゴリズム評価

目　次

1　本稿は、公益財団法人家計経済研究所が実施した「消費生活に関するパネル調査」の個票データを用いた。

2　本研究は2016年度神奈川大学経済貿易研究所研究支援補助金の助成を受けて実施した。

1．はじめに

マイクロデータの分析においては、データ収集段階で発生したデータの欠損をどのように回避、補間するかが重要となる。基本的な方針は大きく、

1．分析に使う変数に欠損があるサンプルは使わない（ドロップする）。

2．他のデータを用いて欠損値を補間する。

の2つとなる。

元々のデータセットに含まれるサンプル数が充分大きく、欠損値を含むサンプルが相対的に充分小さく、欠損がランダムである場合には1の方針で問題ない。たとえば調査票を調査員が確認した上で集計に回すような大規模調査であればこの条件を満たすことが多い。しかしながら、郵送調査で行っているパネル調査などでは欠損が1つでもあるサンプルを分析から外すとサンプル数が大幅に減少する可能性が高い。これは、(a) 長期にわたる調査の場合は調査項目の入れ替えが行われることがある、(b) 回答者の都合により回答しないことがある、などの理由による。本稿でデータ例として扱っている家計経済研究所のパネルデータでは、調査項目の入れ替えにより特定調査年では全データから特定変数が落ちているケースもあり、欠損値があるサンプルは使わないという選択肢はパネルデータとして扱う場合選択不能である。

2の補間を行う場合、単純な移動平均や回帰モデルを用いた推定などの方法がよく使われるが、いずれにせよ存在しないデータをサンプルに追加することになり、補間方法の妥当性については充分な検証が必要となる。

補間の方法によってどの程度の差が出るかをまず模式図で示しておこう。図1は欠損値の部分（観測期＝7）前後でなめらかにデータが変化しているため前後のデータの移動平均（図中の○を付したデータで計算）でもそれらしい補間が行えている。一方、図2のケースでは観測期＝5と6の間に大きな変動が発生しているため、図1と同じ方法で移動平均を取ると補間値が過大（なように見える）。

図1　データがなめらかに変動しているケース（模式図）

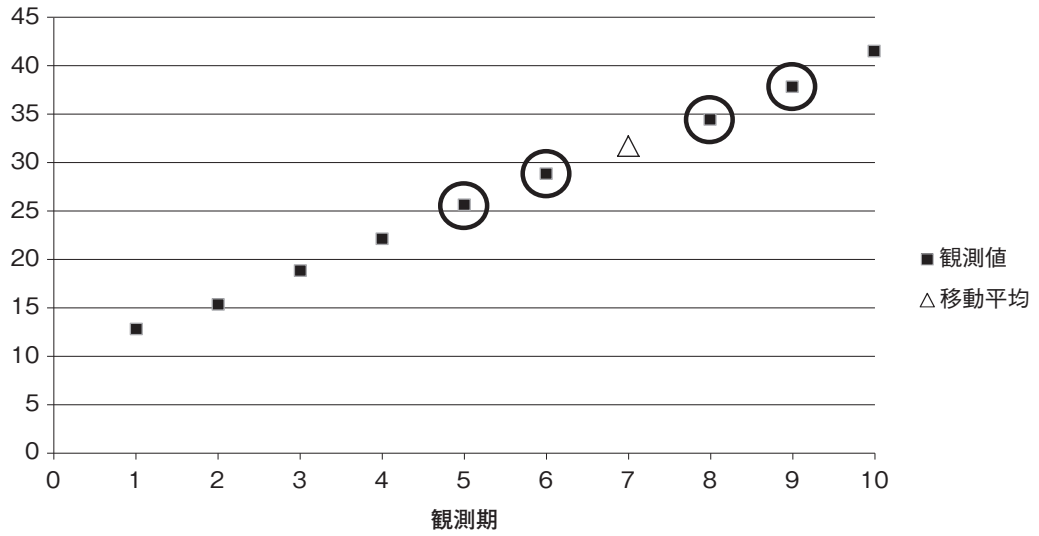


図2　データに急峻な変化が生じているケース（模式図）

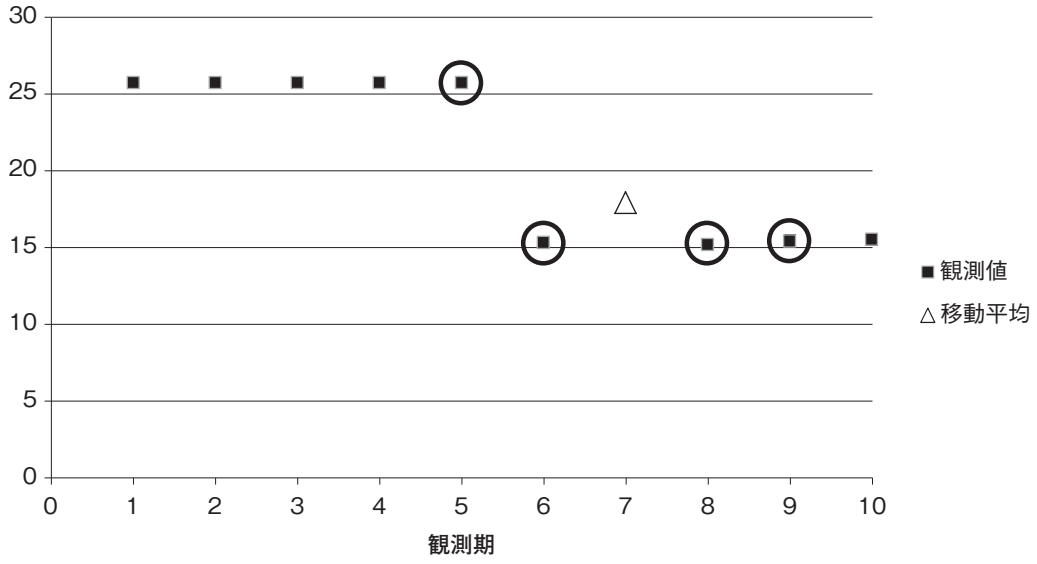
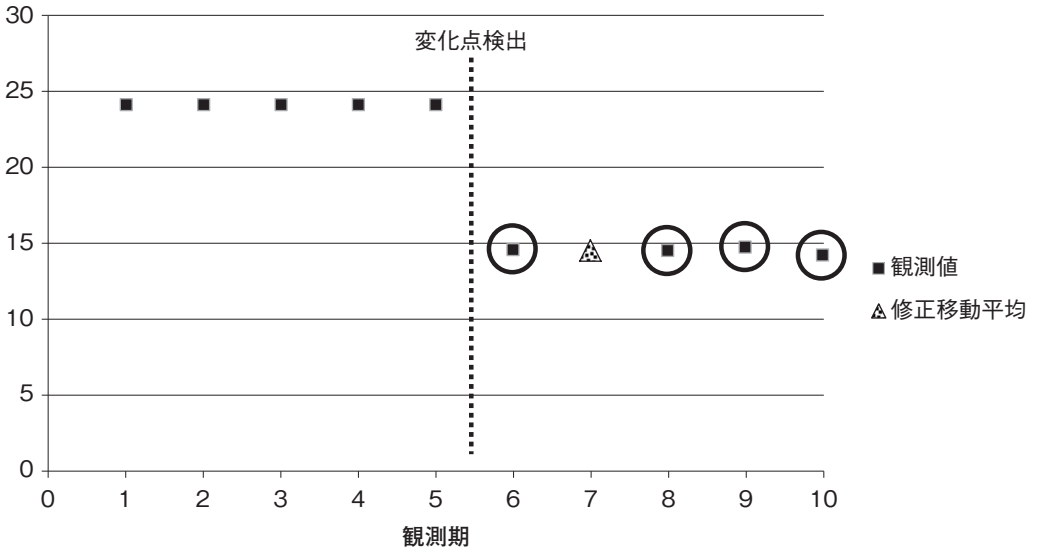


図3　変化後データで修正補間したケース（模式図）



本稿では、事前にデータの性質として急峻な変化点が存在することが予想される場合に、図3のように、まず変化点を検出し、変化点をまたがない形で移動平均をとることにより「もっともらしい」補間値を求めることができることを示す。なお、実際に欠損したデータを推定した場合には補間による誤差を計測不能であるため、誤差の評価に当たっては「実際には欠損が存在しないサンプル」を用い、乱数的にデータをマスクして「欠損値」として補間した結果と、正しいデータとの差がどの程度改善されるかで評価する。

2．データ

本稿で例として取り上げるデータ（「消費生活に関するパネル調査」家計経済研究所）は1993年から2016年まで毎年１回、24回分のデータを収集開始時期の違う５つのコホート（A～E）で収集している（表１）。ただし、本論文執筆時に外部提供されていた個票は2013年までの21回分である。

本稿で補間の対象として用いる親年収は、図４のような形式で調査されている。調査されている年については同じ形式で質問しているが、３回目、５回目（コホートＡのみ）、７回目の調査でこの設問自体が存在しない。そのため家計研データに含まれるA～Eの５つのコホートのうち、初期の２つのコホートA、Bにはそれぞれ３件、１件の欠損が必ず含まれている。また、この初期の２つのコホートは調査に含まれる全個人の７割[3]を含んでいるため、欠損値が存在するサンプルは利用しないという方針を採用することによるサンプル数減少は許容できる範囲ではない。

また、この設問は娘に対して親の収入を聞くという性質上、回答率がかなり悪い設問となっている。調査が行われている回であっても回答率は８～９割程度であり[4]、全調査期間について回答している人の割合はコホートCで70.7%、コホートDでは75.1%となる（表２）。欠損があるサンプルを使わない場合は更に３割近いサンプルを失うことになるため、適切な補間が行えるのであれば補間することが望ましい。

コホートCとDのうち、全ての調査でこの設問に回答しているサンプルを用いてデータの挙動をチェックすると、父親の年齢別平均年収は図５のように変動しており、60歳の定年付近で大

表１　コホートごとの収集開始時期とサンプル人数

コホート名	収集開始	含まれる人数	初期年齢
A	１回目	1500	24～34歳
B	５回目	500	24～27歳
C	11回目	836	24～29歳
D	16回目	636	24～28歳
E	21回目	625	24～28歳

図４　親年収の調査票記載例（2013年調査票より転載）

あなたの親の**昨年１年間（平成24年1月～平成24年12月）**の収入額（税込み額）はおおよそいくらぐらいですか。勤め先からの収入、事業収入、社会保障給付、財産収入などを合計し、あてはまる金額に○をして下さい。

1	2	3	4	5	6	7	8
249万円以下	250～499万円	500～749万円	750～999万円	1000～1249万円	1250～1499万円	1500万円以上	自分の両親は死亡

3　コホートEも含めると約５割であるが、コホートEは外部提供データには１回分しか含まれていないためパネルデータとして分析することは困難である。

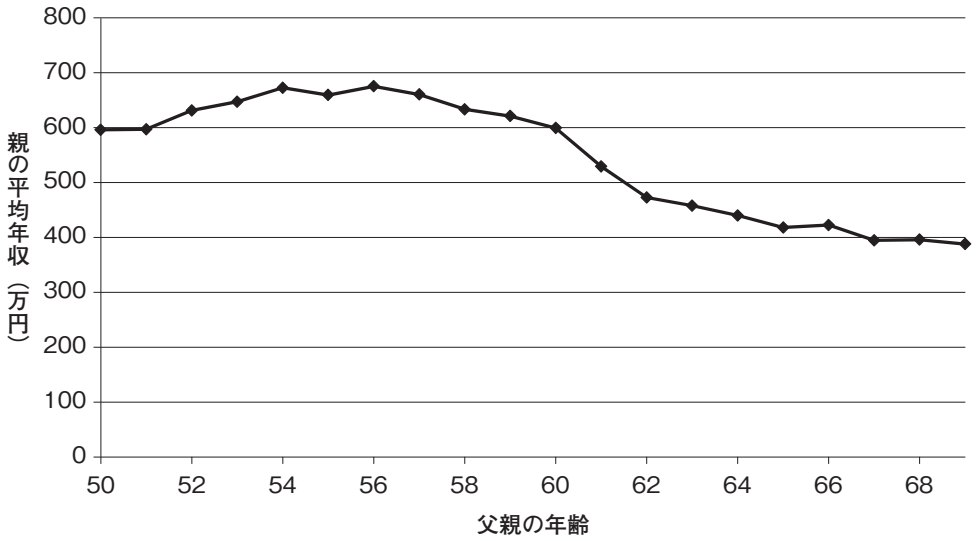
4　９回目の調査だけ、回答率が２割程度となっている。

表 2　調査対象年数と親年収継続回答割合

経過年数	コホート C	コホート D
1	81.3%	66.7%
2	88.2%	76.7%
3	83.0%	85.7%
4	76.5%	72.0%
5	72.4%	60.0%
6	73.9%	**75.1%**
7	94.4%	
8	64.3%	
9	45.5%	
10	82.4%	
11	**70.7%**	

資料：「消費生活に関するパネル調査」（家計経済研究所）個票より筆者集計

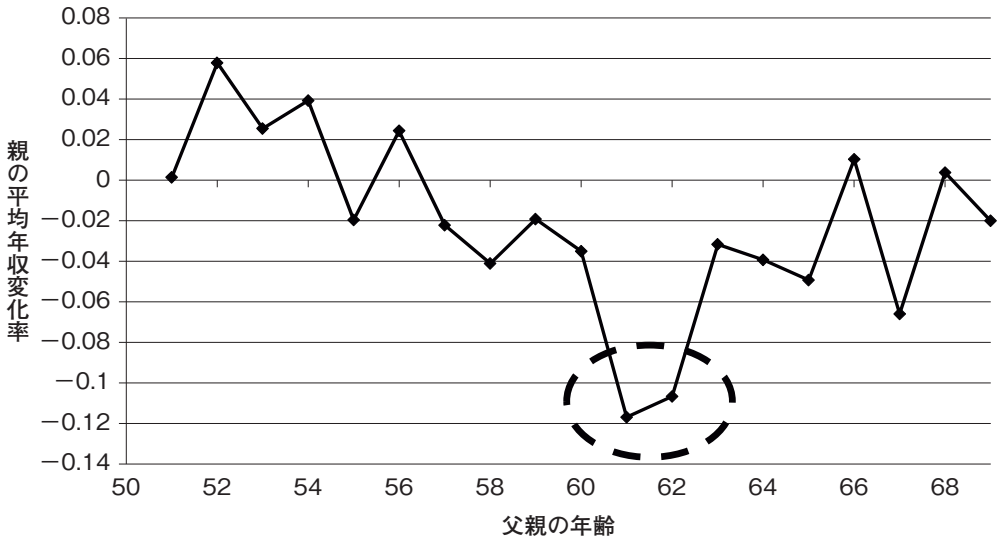
図 5　父親の年齢別親の平均収入（コホート C および D）



資料：「消費生活に関するパネル調査」（家計経済研究所）個票より筆者集計

きく減少していることがわかる。このように大きな変動がある場合には、上述のように単純な平均などによる補間では補間による誤差が大きくなる可能性が高い。また、図 6 に示した対前年比での平均年収変化率から分かるように、定年近辺での収入減少は特徴的であり、このタイミングを適切に検出して補間アルゴリズムに生かすことで補間結果が改善されると予想できる。

図６　父親の年齢別親の平均収入対前年変化率（ コホート C および D）



資料：「消費生活に関するパネル調査」（家計経済研究所）個票より筆者集計

３．アルゴリズム

父親の年齢を a 歳、平均を取る前後の年齢幅を s とした場合のフローチャートを図７に示す。図３の「変化点検出」がフローチャート中の①に相当する。また、②に含まれる、平均を取る幅である s については、s が長ければ補間が成功する確率は上がるが、広い範囲での平均となるため真の値との誤差も拡大することが予想される。そのため、s については後述のシミュレーションを行って決定することとする。

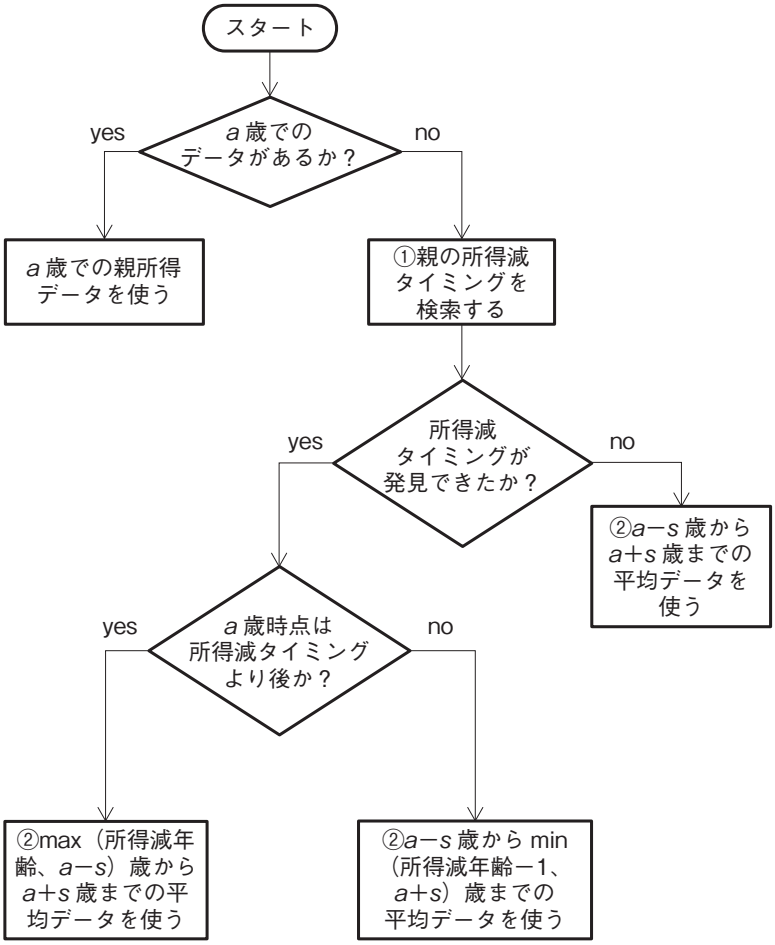
3.1. 親の所得減タイミングの検索方法

本論文では、経験則として親の収入は図５のように定年で大幅に低下することを仮定して補間計算を行う。そのため、

１．親が若い時点から順番に年収を読み取り、着目している年の収入＜前年の収入×最低減額率 d となっている年を全てリストアップする。
２．上で作成した収入減発生年リストから、収入低下金額が最大となっている年を「所得減が発生した年」とする。
３．収入低下金額が同じ年が２回以上ある場合は、先に低下した年を「所得減が発生した年」とする。

という手続きで、親の所得減タイミングを算出している。「最低減額率 d」は、収入変動がある程度大きい年以外はスキップするために導入するパラーメータである。この値についても適切な値が事前に分かっているわけではないため、後述のシミュレーション結果から決定する。

図７　親の所得減タイミングを考慮したデータ補間アルゴリズム[5]



3.2. 所得減少タイミングの計算例

上記のアルゴリズムで実際に所得減少タイミングを算出すると以下のケース１、２のようになる。

この場合、親の収入が減少している56歳（娘の年齢では27歳）が収入減少タイミングとなる。

ケース２では、ケース１に加えて64歳になるときにもう一度賃金低下が発生しており、変化額は55→56歳と同じく250万円である。変化額が同じケースが複数あるため、若い時に発生した変化を収入減少タイミングとして、56歳（娘の年齢では27歳）が収入減少タイミングとなる。

5　②の過程で平均に使うデータが存在しないケースは、このアルゴリズムでは補間不能となる。

ケース１　収入低下が１回しか発生しないケース

娘年齢	父年齢	親収入	変化率	変化額
26	55	624.5		
27	**56**	374.5	−0.40032	−250
28	57	374.5	0	0
29	58	374.5	0	0
30	59	374.5	0	0
31	60	374.5	0	0
32	61	374.5	0	0
33	62	374.5	0	0
34	63	374.5	0	0
35	64	374.5	0	0
36	65	374.5	0	0

ケース２　収入低下が２回以上発生しているケース

娘年齢	父年齢	親収入	変化率	変化額
26	55	624.5		
27	**56**	374.5	−0.40032	−250
28	57	374.5	0	0
29	58	374.5	0	0
30	59	374.5	0	0
31	60	374.5	0	0
32	61	374.5	0	0
33	62	374.5	0	0
34	63	374.5	0	0
35	64	124.5	−0.66756	−250
36	65	124.5	0	0

表３　親の年収階級値と変化率

収入階級	249万円以下	250～499万円	500～749万円	750～999万円	1000～1249万円	1250～1499万円	1500万円以上
階級値（万円）	124.5	374.5	624.5	874.5	1124.5	1374.5	1800
１階級低下時変化率	−67%	−40%	−29%	−22%	−18%	−24%	

４．補間アルゴリズムの性能検証

本論文で提案しているような補間アルゴリズムが、単純な平均よりも優れているかどうかは自明ではない。ここでは、

１．データセットに含まれている範囲では親の収入データに欠損がないサブデータセットをコホートＣおよびコホートＤより作成する（1095サンプル）。

2．上記の無欠損データセットに含まれる各ケースから、ランダムに1年選ぶ。
3．2で選んだ年のデータが欠損していると仮定して、図7のアルゴリズムに従い補間を行う。
4．真の値との差を計算する。

という方法で「正しいデータと比較する」ことにより、アルゴリズムの評価を行っている。ただし、上述の通り平均を取る期間 s と、どの程度収入が低下したら「下がった」と扱う基準である最低減額率 d については、それぞれ（1～4）、（0.5～0.8）の範囲で変えながら計算を行った。最低減額率の上限が1ではなく0.8となっている理由は、元々のデータが階級データであり、表3に示すように1つでも階級が下がると少なくとも18%は減額するため、その範囲より最低減額率が大きい部分は考慮する必要がないことによる。

全てのデータが観測されているサブデータセットに含まれるサンプル数は1095、平均期間 s と最低減額率 d の組み合わせごとに500回計算を繰り返し、乱数の偏りによる影響を減らしている。

4.1. 復元可能なサンプル割合

まず、補間によって復元可能だったサンプルの割合を表4に示す。予想されていたように平均期間 s が長い方が復元可能な割合は若干高くなっている。最低減額率 d が小さい方が復元可能割合が高くなっているのは、図7のアルゴリズムでは所得減少ポイントが発見できなかった場合、単純平均で計算するためと考えられる[6]。平均期間 s が1年の場合は、最低減額率 d が大きくなると若干復元可能割合が低下するが、2年以上の場合は8割以上の復元可能割合が維持できる。

4.2. 推定誤差

本論文は、定年などによって急激に収入が減少することを経験則として組み込んだ補間によって、単純な補間よりも改善された推定値が得られることを示すことを目的としている。表5は真の値と、補間によって得られた値の差の標準偏差を計算したものである。全体的な傾向としては、

1．平均期間 s が長くなると推定誤差の標準偏差は大きくなる。
2．最低減額率 d が大きくなると推定誤差の標準偏差は小さくなる。

表4　欠損が1年分だったときの復元可能割合

		最低減額率 d				単純平均
		0.5	0.6	0.7	0.8	
平均期間 s	1	80.7%	78.2%	77.9%	75.5%	87.5%
	2	84.6%	83.0%	82.8%	81.1%	87.7%
	3	84.6%	83.0%	82.8%	81.1%	87.8%
	4	84.6%	83.1%	82.8%	81.1%	87.8%

資料：「消費生活に関するパネル調査」（家計経済研究所）個票より筆者集計

6　たとえば上に例としてあげたケース 1 で父親55歳（娘26歳）のデータが欠損していた場合、単純平均であれば収入減のタイミングを無視して計算するため計算可能だが、図7のアルゴリズムだと収入減少前のデータが1つもないため計算できない。

表５　補間した値と真の値の差の標準偏差（万円）

		最低減額率 d				単純平均
		0.5	0.6	0.7	0.8	
平均期間 s	1	198.4	195.9	192.0	189.1	235.3
	2	200.4	198.5	195.4	193.9	226.1
	3	200.2	197.7	195.9	194.5	227.5
	4	201.0	198.9	195.9	194.9	230.3

資料：「消費生活に関するパネル調査」（家計経済研究所）個票より筆者集計

の２つの特徴が観察できる。１については、平均期間 s が長くなれば欠損しているデータから時間的に離れたデータまで平均に含まれることになるため、誤差が大きくなることは予想通りである。また２についても4.1で論じたように、最低減額率 d が小さくなると収入減の変化点検出にはより大きな収入減が必要となり、全体としては単純平均に近づくことが予想される。

表５の最右列に示した単純平均では標準偏差が225～235万円程度の範囲で推移しており、本論文で提案した経験的な知識を組み込んだ補間アルゴリズムを用いることによって、推定誤差を小さくできることが示せた。

５．まとめ

本論文では、家計経済研究所が実施している「消費生活に関するパネル調査」の個票データを用い、欠損値を含むデータの補間時に経験則として知られているデータの特性を考慮した補間アルゴリズムの検証を行った。その結果として、

１．定年などによって高齢者の収入は大幅に減少するタイミングを持つ、という経験則を組み込んだ補間アルゴリズムによって、単純な平均に比べて、真の値との推定誤差の標準偏差で評価して35～45万円程度の改善が得られた。

２．本論文で提案した補間アルゴリズムは単純な平均による補間よりも補間可能なケースが少ないが、おおむね８割程度のケースで補間可能である。そのため、補間精度を上げつつ広い範囲で適用可能と考えられる。

という知見を得た。