参議院常任委員会調査室・特別調査室

論題	データ分析の手法に関する一試論 ～分析方法の違いにより同じデータから異なる結果が生ずる例～
著者 / 所属	前田　泰伸 / 調査情報担当室
雑誌名 / ISSN	経済のプリズム / 1882-062X
編集・発行	参議院事務局 企画調整室（調査情報担当室）
通号	230 号
刊行日	2023-11-20
頁	15-24
URL	https://www.sangiin.go.jp/japanese/annai/chousa/keizai_prism/backnumber/r05pdf/202323002.pdf

データ分析の手法に関する一試論

～分析方法の違いにより同じデータから異なる結果が生ずる例～

調査情報担当室　前田　泰伸

《要旨》

本稿では、パネルデータの分析から、同じデータを用いても分析の方法が異なれば結果が異なる例を紹介する。使用するデータは、テレビ・ラジオ・新聞・雑誌の時間（被説明変数）、高齢化率、完全失業率（説明変数）の都道府県別データである。これらのデータをパネルデータとして、統計ソフトを使用し、固定効果モデル、時間固定効果モデルで分析すると、テレビ・ラジオ・新聞・雑誌の時間と高齢化率の関係については、固定効果モデルでは逆相関、時間固定モデルでは順相関という異なる（正反対の）結果となる。また、固定効果と時間固定効果の両方を同じ回帰式に加えて分析を行うと、テレビ・ラジオ・新聞・雑誌の時間に対して有意な影響を与えるのは高齢化率のみとなる。分析方法が異なれば、使用するデータが同じでも異なった結果が生じ得るが、そうした場合には、その原因・理由について更に掘り下げて検討する姿勢が重要かと思われる。なお、そもそも論として、統計ソフトを使用した複雑な分析を行うより、棒グラフなどにより誰が見ても分かりやすい説明を行う方が適切な場合もあるのではないかとも思われる。

１．はじめに

本稿では、データ分析の手法に関して、同じデータを用いても分析の方法が異なればその結果も異なる（正反対ともなる）例を紹介することとしたい。

政府の白書やシンクタンクのレポート等では、経済・社会指標を利用した多くの分析が行われている。具体的には、単純にデータをグラフの形で表示してその傾向を示すといったものから、統計ソフトを利用して複雑・高度な分析を行うなど、様々な手法がとられている。ただ、このような複雑・高度な分析の場合には、白書やレポート等の文中では分析の結果だけが示され、その分析がどういった考え方に基づき、どのような計算によって行われるのか等について

は、専門的な知識がなければブラックボックスのように理解が難しくなっている場合も多い。

本稿では、分析手法そのものに着目することとして、全く同じデータを用いても分析の手法を変えると異なる（正反対の）結果となる例について紹介し、そうした結果が生じる原因・理由について検討することとする。具体的には、都道府県別のデータを組み合わせてパネルデータを構築し、固定効果モデル等により分析する（詳細は後述）。また、これらの分析とは別に、シンプルなグラフにより誰が見ても分かりやすい説明を行う方が適切な場合もあるのではないかという点についても示すこととしたい。

２．使用するデータと分析の手法

（１）使用するデータ

本稿では、都道府県別のデータからパネルデータを構築するが、その際に使用するデータは、週全体平均での１日の生活時間のうちテレビ・ラジオ・新聞・雑誌の時間（総務省「社会生活基本調査」）、高齢化率[1]（総務省「人口推計」）、完全失業率[2]（総務省「労働力調査」）である。そして、それぞれの2001年、2006年、2011年、2016年、2021年のデータからパネルデータを構築する[3]。

これらのデータを選んだ理由としては、テレビ・ラジオ・新聞・雑誌の時間については、かねてから継続的に調査が行われており、調査の最新年（2021年）も比較的新しいこと、おそらく大多数の人たちにとっては日常生活の一部であり、経済や産業に関する専門的な指標を例とする場合と比べると分かりやすくなるのではないかと考えられることである。

また、高齢化率については、次頁の図表１に示すように、2021年の単年のデータ（都道府県別のクロスセクションデータ[4]）から回帰分析を行うと、高齢化率が高い都道府県ではテレビ・ラジオ・新聞・雑誌の時間が長くなる順相関の傾

---

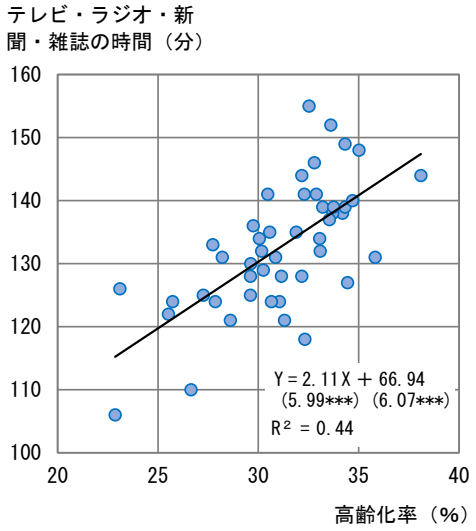
[1] 高齢化率とは、65歳以上の人口の総人口に占める割合のことである。なお、高齢者については、一般的にはＷＴＯの定義により、65歳以上の人とされることが多い。

[2] 完全失業率とは、労働力人口（15歳以上の働く意欲のある人）のうち、完全失業者（職がなく、求職活動をしている人）が占める割合のことである（労働力調査）。

[3] パネルデータとは、クロスセクションデータ（後掲注４参照）と時系列データ（後掲注５参照）を組み合わせたもののことである。一般的に、パネルデータから分析を行う場合には、情報量が多くなることにより推計の精度が向上する等の利点があると考えられている。

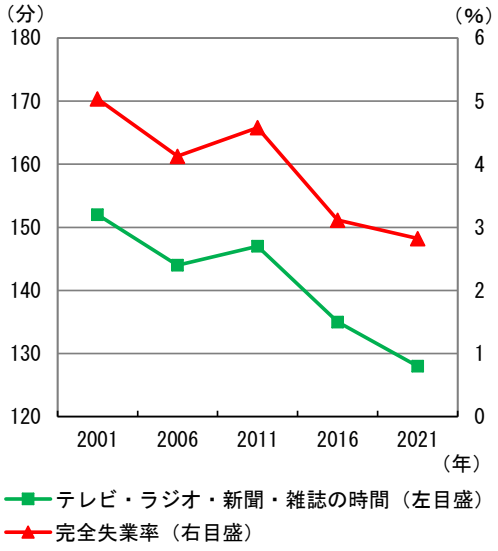
[4] クロスセクションデータとは、時間を１時点に固定し、その時点における観察個体（人、世帯、場所等のことであり、本稿では都道府県である）において発生しているデータを記録したもののことである。

図表１　テレビ・ラジオ・新聞・雑誌の時間と高齢化率の関係



（注）1.テレビ･ラジオ･新聞･雑誌の時間、高齢化率ともに、都道府県平均のデータである。
2.Xの係数及び定数項の下の()内の数値はｔ値であり、「***」はｔ値が１%の有意水準を満たすことを示す。また、R^2は決定係数である。
（出所）総務省「社会生活基本調査」、「人口推計」より作成

図表２　テレビ・ラジオ・新聞・雑誌の時間と完全失業率の関係



（注）テレビ･ラジオ･新聞･雑誌の時間、完全失業率ともに、全国平均のデータである。
（出所）総務省「社会生活基本調査」、「労働力調査」より作成

向が見られることである。完全失業率については、テレビ・ラジオ・新聞・雑誌の時間とともに（両者とも全国平均であり、2001年から2021年までの時系列データ[5]である）、同じ座標平面上に折れ線グラフで描くと、図表２のように、両者はほぼ並行した動きとなり、大まかな傾向としては、テレビ・ラジオ・新聞・雑誌の時間が短くなり、同様に完全失業率も低くなる順相関の関係にあることが言えよう。これらの関係がパネルデータの分析ではどのような結果となるか、実際に分析を行ってみようということである[6]。

[5] 時系列データとは、ある事象について時間の経過にしたがって観測されたデータを記録したもののことである。

[6] なお、詳細は本文において詳述するが、図表１と図表２の順相関の関係は、必ずしもパネルデータの分析においても同様に見られるとは限らない。図表１（クロスセクションデータ）では時間の経過による影響は図表に現れず、また、図表２（時系列データ）ではクロスセクション面での影響は図表に現れないが、パネルデータはクロスセクションデータと時系列データを組み合わせたものであるため（前掲注３参照）、そうした影響が（分析の手法にもよるが）パネルデータの分析の結果に現れる可能性がある。

（２）分析の手法～固定効果モデルと時間固定効果モデル

本稿では、パネルデータの分析においてよく用いられる固定効果モデルとともに、それと似た手法である時間固定効果モデルによって分析を行う。これらの概要について説明すると、次のようになる。

まず、固定効果モデルとは、観察個体ごとに異なるが時間を通じて一定である変数を仮にαとして（これを「固定効果」という）、このαを回帰式に加えて分析を行うものであり、回帰式は「$Y_{it} = \alpha_i + \beta X_{it} + u_{it}$」のような形で表される。回帰式のＹは被説明変数、Ｘは説明変数であり、それぞれの変数の右下の添字は、iが各都道府県、tが時点を示すものとなっている。固定効果αについては、αが時間を通じて一定であるため、αの添字はiだけであり、また、ｕは誤差項に当たる。固定効果モデルでは、観察個体における観察されない異質性を統御することができる[7]などの利点もある。なお、本稿では都道府県別のデータから分析を行うが、この場合の固定効果α_iについては、これを別の視点から見ると、各都道府県に対して個別に割り当てたダミー変数と考えることもできる。おそらくは、こう考える方が、直観的にも分かりやすい説明となるのではないかと思われる。

また、時間固定効果モデル[8]とは、前述の固定効果モデルとは対照的に、観察個体を通じて一定であるが時間とともに変化する変数を回帰式に加えて分析を行うものである。そうした変数を便宜的にγとすると、回帰式は「$Y_{it} = \gamma_t + \delta X_{it} + u_{it}$」のような形となり、固定効果モデルとの違いは、γの添字が（各都道府県を示すiではなく）時点を示すtとなっていることである。なお、時間固定効果モデルの直観的な理解としては、本稿のように複数年（年単位）のデータの場合には、年ごとに個別のダミー変数を割り当てたと考えるのが簡明かと思われる[9]。

---

[7] 若干敷衍して言うと、モデルでは変数として現れない（観察されない）が、観察個体（本稿の分析では都道府県）間での差や違い（異質性）が被説明変数に対して影響を及ぼしている場合、そうした影響を取り除くことができるということである。

[8] 名称については、時間効果固定モデルや時間効果モデルなどと呼ばれる場合もあるが、本稿では、便宜的に時間固定効果モデルと呼ぶこととしたい。

[9] 分析方法としては、本文に挙げたもの以外に、変量効果モデルも用いられている。変量効果モデルでも、観察個体ごとに異なるが時間を通じて一定の変数（α_i）を回帰式に加えて分析を行うが、α_iと説明変数が無相関という仮定が置かれる（固定効果モデルでは、両者は相関するとされる）。ただ、本稿では、議論が複雑になることもあり、変量効果モデルの詳細な説明等は割愛することとしたい（詳細については、西山慶彦ほか『計量経済学』有斐閣（2019）第６章（210～265頁）など、計量経済学の教科書を参照）。

３．分析と結果とその検討

（１）固定効果モデル、時間固定効果モデルによる分析の結果

以上のようなパネルデータについて、被説明変数をテレビ・ラジオ・新聞・雑誌の時間､説明変数を高齢化率と完全失業率(いずれも都道府県別の2001年、2006年、2011年、2016年、2021年のデータ）として、まず、固定効果モデルによる分析を行うと、結果は次の推計式１のようになる[10]。

推計式１　固定効果モデル

$$Y_{it} = \alpha_i - \underset{(-3.98^{***})}{0.70X_{1it}} + \underset{(6.30^{***})}{4.51X_{2it}} + u_{it} \qquad (R^2 = 0.83)$$

（注）1. Ｙはテレビ・ラジオ・新聞・雑誌の時間（分)、X_1は高齢化率（%)、X_2は完全失業率（%)、αは固定効果、ｕは誤差項であり、それぞれの変数の右下の添字ｉは都道府県、ｔは時間（年）を示す。
2. X_1、X_2の下の()内の数値はｔ値（クラスター構造に対して頑健な標準誤差による)。右肩の「***」は、ｔ値が１%の有意水準で有意であることを示す。R^2は決定係数。
3. データの出所は、総務省「社会生活基本調査」、「人口推計」、「労働力調査」である。

推計式１を見ると、X_1（高齢化率）の係数はマイナス、X_2（完全失業率）の係数はプラスとなっており、このことから、テレビ・ラジオ・新聞・雑誌の時間は、高齢化率が高くなると短くなり、完全失業率が高くなると長くなると考えることができる。

そして、時間固定効果モデルの結果を示したものが、次の推計式２である。

推計式２　時間固定効果モデル

$$Y_{it} = \gamma_t + \underset{(4.81^{***})}{1.76X_{1it}} + \underset{(3.28^{***})}{3.68X_{2it}} + u_{it} \qquad (R^2 = 0.60)$$

（注）1. Ｙはテレビ・ラジオ・新聞・雑誌の時間（分)、X_1は高齢化率（%)、X_2は完全失業率（%)、γは時間固定効果、ｕは誤差項であり、それぞれの変数の右下の添字ｉは都道府県、ｔは時間（年）を示す。
2. X_1、X_2の下の()内の数値はｔ値（クラスター構造に対して頑健な標準誤差による)。右肩の「***」は、ｔ値が１%の有意水準で有意であることを示す。R^2は決定係数。
3. データの出所は、総務省「社会生活基本調査」、「人口推計」、「労働力調査」である。

推計式２を見ると、X_1、X_2ともに係数はプラスであり、高齢化率、完全失業率のいずれについても、これらの数値が高くなるとテレビ・ラジオ・新聞・雑誌の時間が長くなるという結果となっている。なお、X_1、X_2に係るｔ値を見ると、いずれも１%の水準で有意となっており、このことは、推計式１の場

[10] 分析においては、統計ソフトEViews11を使用した。

合においても同様である。

ここで、推計式１（固定効果モデル）と推計式２（時間固定効果モデル）を比べてみると、X_1（高齢化率）については、推計式１と推計式２では符号が異なっている（プラスとマイナスで正反対である）ことが分かる。つまり、推計式１では、高齢化率が高くなるとテレビ・ラジオ・新聞・雑誌の時間が短くなる逆相関の関係であるのに対し、推計式２は、高齢化率が高くなるとテレビ・ラジオ・新聞・雑誌の時間も長くなる順相関の関係であるが、これについては、どう理解すべきであろうか。

（２）推計式１と推計式２でX_1の符号が異なるのはなぜか

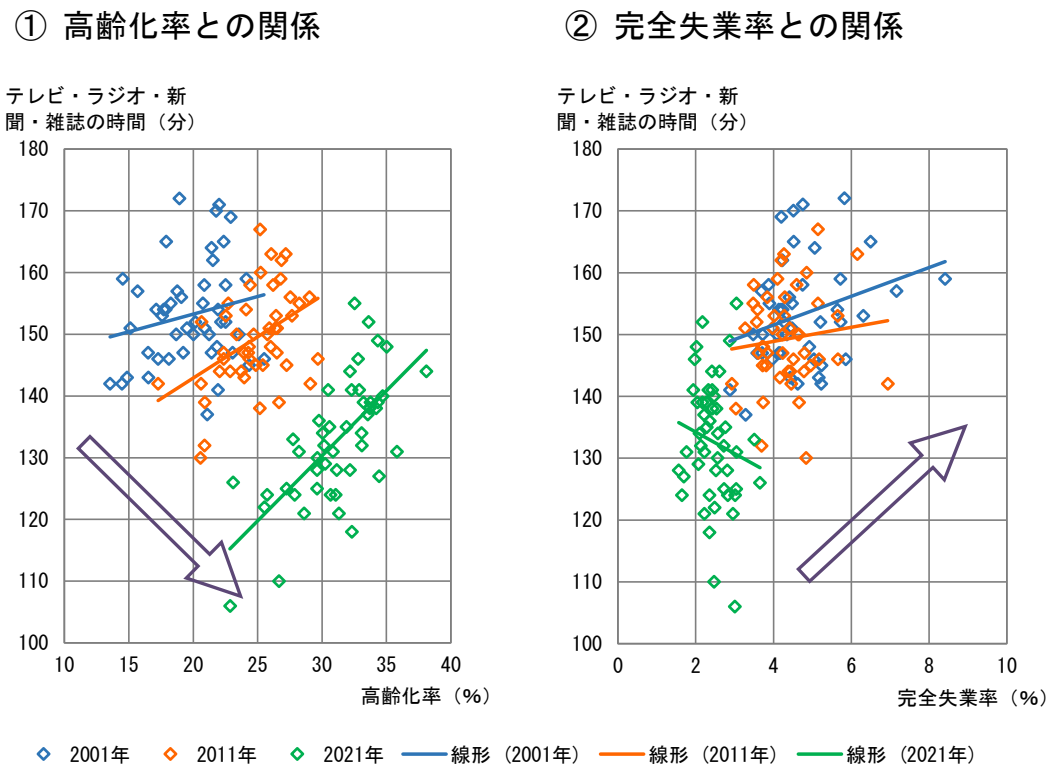
推計式１（固定効果モデル）と推計式２（時間固定効果モデル）でX_1（高齢化率）の符号が異なる理由は分析方法の違いによるものであるが、その違いについて大まかに言うと、次のようにまとめることができよう。

推計式１は固定効果モデルであり、X_1の係数は、テレビ・ラジオ・新聞・雑誌の時間（被説明変数）、高齢化率（説明変数）それぞれの時間の経過によって変化した部分の関係を表すと考えることができる（変化しない（時間を通じて一定の）部分は、固定効果（α_i）に含まれるため）。そして、テレビ・ラジオ・新聞・雑誌の時間と高齢化率の時間の経過に伴う変化としては、前者は時間の経過につれて（調査年が新しくなるほど）短くなり（図表２の緑の折れ線（テレビ・ラジオ・新聞・雑誌の時間）を参照）、後者は少子高齢化の進行により全国的に年々その割合が高まっている[11]ことが考えられる。そのため、（X_1の係数に関係する）時間の経過によって変化した部分で見ると、高齢化率（説明変数）が高くなればテレビ・ラジオ・新聞・雑誌の時間（被説明変数）が短くなる関係となり、推計式１ではX_1の係数がマイナスとなったということが言えよう。

次頁の図表３①は、こうした関係を視覚的・模式的に示すため、2001年、2011年、2021年の（2006年と2016年については、図表を見やすくするため、割愛した）テレビ・ラジオ・新聞・雑誌の時間と高齢化率（のみ）の関係について散布図を描き、最小二乗法による回帰直線を引いたものである。散布図の全体的な傾向としては2001年から2021年にかけて左上から右下にシフトしており、時間の経過とともにテレビ・ラジオ・新聞・雑誌の時間が短くなり、同時に高齢

[11] 日本全国で見ると、高齢化率は、2001年の18.0％から2021年には28.9％と上昇している（総務省「人口推計」による）。図表等については、紙幅の関係もあり、割愛することとしたい。

図表３　テレビ・ラジオ・新聞・雑誌の時間、高齢化率、完全失業率の関係



（注）「線形」とは、最小二乗法による回帰直線を示す。回帰式等については、図表が煩雑になるため、割愛した。
（出所）総務省「社会生活基本調査」、「人口推計」、「労働力調査」より作成

化率が上昇している逆相関の関係がうかがえよう（図表３①の右下がりの矢印を参照）。他方で、図表３①の散布図について、2001年、2011年、2021年で個別に回帰直線を引くと、いずれの回帰直線も右上がりとなっている。これらの個別の回帰直線には、それぞれの年のクロスセクション面での関係性が現れていると考えることもできるため、時間の経過による影響を時間固定効果とする時間固定効果モデルでは、こうしたクロスセクション面での（図形的には右上がりの）関係性がX_1（高齢化率）の係数に引き継がれるような形となり、推計式２のX_1の係数がプラスとなったと整理することができよう。

また、図表３②は、テレビ・ラジオ・新聞・雑誌の時間と完全失業率（のみ）の関係についても同様に散布図を描き、最小二乗法による回帰直線を引いたものである。こちらの場合、散布図は、図形的には左下から右上へと（時系列的に厳密に言えば右上から左下に）描くことができる。図表３②では、時間の経過につれてテレビ・ラジオ・新聞・雑誌の時間が短くなると完全失業率も低下する順相関（図形的には、図表３②の右上がりの矢印を参照）の傾向となって

おり、こう考えると、推計式１（固定効果モデル）ではX_2の係数がプラスとなったと見ることができよう。なお、推計式２（時間固定効果モデル）との関係では、それぞれの年の個別の回帰直線を見ると、2021年は例外であるが、それ以外の年では右上がりであり（なお、図表３②に描かれていない2006年、2016年も右上がりである）、やや強引な見方かもしれないが、基本的には右上がりと見ることもできるのではないかと思われる。こう考えた場合には、このことは、推計式２のX_2の係数がプラスとなっていることの説明ともなろう。

（３）固定効果と時間固定効果の両方を１つの回帰式に加えたモデル

ここまでは、固定効果モデルと時間固定効果モデルそれぞれで分析を行ってきたが、固定効果モデルの固定効果と時間固定モデルの時間固定効果は相互に排他的なものではなく、両方を同じ回帰式に加えた分析も可能である[12]。

そこで、本稿での分析の最後に、こうした両方の効果を加えたモデルによる分析を行うこととしたい。なお、使用するデータや被説明変数、説明変数は推計式１、推計式２と同じである。そして、その分析の結果を示したものが、次の推計式３である。

推計式３　固定効果と時間固定効果を加えたモデル

$$Y_{it} = \alpha_i + \gamma_t + \underset{(2.47^{**})}{1.73X_{1it}} + \underset{(1.09)}{1.03X_{2it}} + u_{it} \qquad (R^2 = 0.88)$$

（注）1. Ｙはテレビ・ラジオ・新聞・雑誌の時間（分）、X_1は高齢化率（％）、X_2は完全失業率（％）、αは固定効果、γは時間固定効果、ｕは誤差項であり、それぞれの変数の右下の添字ｉは都道府県、ｔは時間（年）を示す。
2. X_1、X_2の下の()内の数値はｔ値（クラスター構造に対して頑健な標準誤差による）。右肩の「**」は、ｔ値が５％の有意水準で有意であることを示す。R^2は決定係数。
3. データの出所は、総務省「社会生活基本調査」、「人口推計」、「労働力調査」である。

推計式３を見ると、X_1（高齢化率）、X_2（完全失業率）のいずれも係数の符号はプラスであるが、それぞれのｔ値については、X_1では５％の水準で有意であるものの、X_2では水準を仮に10％としても有意ではないということが分かる。つまり、高齢化率の場合には、固定効果と時間固定効果の両方を加えた場合にも順相関の（高齢化率が高くなれば、テレビ・ラジオ・新聞・雑誌の時間が長くなる）関係となるのに対し、完全失業率の場合はそうではなく、本

[12] こうした方法も、さほど珍しいものではなく、実証分析において広く使用されている（回帰式の具体的な導出方法については、『計量経済学』（前掲注９）237～239頁など、計量経済学の教科書を参照）。

当のところは、完全失業率とテレビ・ラジオ・新聞・雑誌の時間の間には特段の関係はないのではないか、例えば、推計式１（固定効果モデル）では、回帰式に現れない時間固定効果がX_2に影響を与え、また、推計式２（時間固定効果モデル）では、回帰式に現れない固定効果がX_2に影響を与えたことにより、それぞれの場合のX_2のｔ値が有意となったのではないか、という可能性も考えられるかもしれない。

４．おわりに

（１）これまでのまとめ

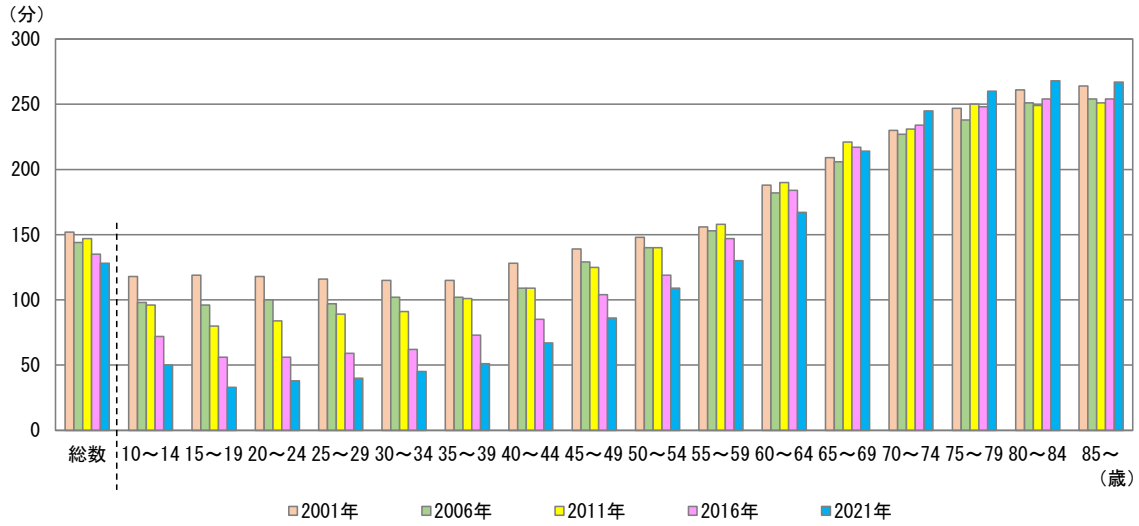
本稿では、都道府県平均でのテレビ・ラジオ・新聞・雑誌の時間、高齢化率、完全失業率からパネルデータを構築し、固定効果モデル（推計式１）や時間固定効果モデル（推計式２）、更にはそれら両方の効果を同じ回帰式に加えたモデル（推計式３）により分析を行った。固定効果モデルと時間固定効果モデルでは、一見すると正反対の結果が生じたようにも思われるが、この点については、分析方法が異なれば、使用するデータが同じであっても異なった結果が生じ得ることを意識しつつ、そうした結果が生じた場合には、その原因・理由について更に掘り下げて検討するという姿勢が重要であろう。現在では、データを収集して統計ソフトを使用すれば、複雑な（手計算ではほぼ不可能な）分析も苦労なく行うことができる。しかし、そうした統計ソフトを使用して適切な結論を導き出すためには、統計ソフトが機械的に計算等の処理をする分析手法（本稿の例では、固定効果モデルなど）そのものに対する深い理解も必要不可欠と言えよう。

（２）棒グラフによるシンプルな説明

また、そもそも論であるが、統計ソフトを使用して分析を行うべき場合かそうでないかの判断も重要かと思われる。例えば、推計式３からは、高齢化率が高くなればテレビ・ラジオ・新聞・雑誌の時間が長くなる傾向がうかがえるが、このことのみを示したいということであれば、次頁に示す図表４からも、ほぼ同様のことが言えるのではなかろうか。

図表４は、2001年から2021年までのテレビ・ラジオ・新聞・雑誌の時間（全国平均における週全体平均での時間）の推移を示したものである。これを見ると、いずれの年においても、年齢階級が高くなると（65歳以上の高齢者については特に）テレビ・ラジオ・新聞・雑誌の時間が長くなっていることが分かる。また、それぞれの年齢階級について、2001年から2021年までの（時間の経過に

図表４　年齢階級別に見たテレビ・ラジオ・新聞・雑誌の時間の推移



（注）1. 全国平均、週全体平均での時間（単位は分）である。
　　2. 総数とは、10 歳以上の総数での全国平均の時間を示す。
（出所）総務省「社会生活基本調査」より作成

伴う）変化を見ると、年齢階級の低い人たちを中心としてテレビ・ラジオ・新聞・雑誌の時間が短くなる傾向があり、このことは、総数で見た場合でも同様である[13]。なお、こうした傾向は、年齢階級が高くなると当てはまらなくなってくるが、高齢者の場合には、仕事からリタイヤすると自由に使える時間が多くなり、そうした時間を昔から馴染みの深いテレビや新聞等に充てているといったことが考えられよう。

統計ソフトは、統計学等の知識のある人がこれを適切に利用すれば、非常に有益なツールとなるであろう。ただ、場合によっては、そうした知識のない人にとってはブラックボックスのような統計ソフトを使用するより、誰が見てもシンプルで分かりやすい棒グラフ等を使って表現する方が適切な説明となることも多いのではなかろうか。データ分析においては、必要に応じて適切な手法を選択することも重要であろう。

（内線 75044）

[13] 社会生活基本調査（2021）のうち「生活時間及び生活行動に関する結果」の「結果の概要」14頁では、スマートフォンやパソコンなどの使用時間が長くなるほど、テレビ・ラジオ・新聞・雑誌の時間が短くなる傾向となることが示されている。