参議院常任委員会調査室・特別調査室

論題	回帰分析によるミスリードについて ～新型コロナウイルス感染症の死亡リスクを題材として～
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	203 号
刊行日	2021-8-11
頁	39-48
URL	https://www.sangiin.go.jp/japanese/annai/chousa/keizai_prism/backnumber/r03pdf/202120303.pdf

回帰分析によるミスリードについて
～新型コロナウイルス感染症の死亡リスクを題材として～

調査情報担当室　前田　泰伸

《要旨》

本稿では、新型コロナウイルス感染症の死亡リスクを題材として、回帰分析によるミスリードやそうしたことが起こる理由について示すこととする。被説明変数を都道府県別の新型コロナの死亡率、説明変数を都道府県別の高齢者の割合、高齢者10万人当たりの特別養護老人ホームの定員数、三世代世帯の割合として、ぞれぞれで回帰分析を行うと、いずれも係数のｔ値は統計的に有意な値となる。しかし、説明変数に都道府県別の人口を加え、合計４つの説明変数から重回帰分析を行うと、統計的に有意となるのは人口のみとなる。こうしたことが起こる理由としては、都道府県別の高齢者の割合等の３つがそれぞれ人口と関係し、人口が新型コロナの死亡率とも関係していることが考えられ、そのため、高齢者の割合等のそれぞれの回帰分析だけでは、背後にある人口という要因を見落とす可能性がある。結論としては、１回の回帰分析によって統計的に有意な結果が得られても、それだけで満足するのではなく、その有意な結果が生ずる原因についても考えを巡らせ、可能であれば別の方法も試してみるなどのことが重要であるといえよう。

１．はじめに[1]

本稿では、新型コロナウイルス感染症（以下、「新型コロナ」という）の死亡リスクを題材として、回帰分析を行う場合のミスリードやそうしたことが起こる理由について示すこととする。なお、筆者は医療や感染症の専門家ではない。本稿の基本的な目的は、新型コロナの死亡リスクについて科学的知見を示そうというものではなく、疑問や疑念が生ずるような回帰分析をあえて例として取り上げることにより、データを統計的に解析して推論などを行う際に留意すべ

[1] 本稿は、2021 年 7 月 28 日までの公開情報に基づいて執筆している。

き事項、つまり、統計的な見方や思考プロセスの在り方について述べていこうとするものである。

なお、新型コロナの“リスク”に関しては、いわゆる後遺症についても多くの報道等がなされている。こうした後遺症についても科学的な検証が必要なことはいうまでもないが[2]、本稿では前述の目的との関係もあり、リスクを死亡リスクに限定し、後遺症についてはひとまず措くこととしたい。

２．年齢層別に見た新型コロナの死亡リスク

まずは、年齢層別に見た新型コロナの死亡リスクについて示すこととする。図表１①は、2021年６月末までの新型コロナの死亡者数の累計を年齢層別・男女別に示したものである[3]。これを見ると、男性・女性のいずれも、新型コロナの死亡者の大部分は高齢者が占めており、20歳代や30歳代などの若者が新型コロナにより死亡する可能性は、完全にゼロではないが非常に小さいということが分かる[4]。

また、図表１①では、90歳以上の死亡者数が80歳代に比べて少なくなっているが、この辺りの年齢となると、いかに現在では平均寿命が延伸しているとはいえ[5]、こうした年齢層に属する人口が寿命を迎えるなどにより減少していることが考えられる。そこで、死亡者数と各年齢層の人口から年齢層別に、人口10万人当たりの死亡者数（以下、「死亡率」という）を計算して示したものが、図表１②である。これを見ると、男性では50歳代から、女性では60歳代から死亡率が上昇しており、70歳代以降になると、男性、女性とも加速度的に死亡率が高くなっていくことが分かる[6]。

---

[2] 新型コロナの後遺症については、厚生労働省で実態調査が行われており、第39回新型コロナウイルス感染症対策アドバイザリーボード（2021年6月16日）において、その中間報告が公表されている（https://www.mhlw.go.jp/content/10900000/000798853.pdf）。

[3] データの期間については、新型コロナのいわゆる第４波が落ち着いたと見られる６月末を区切りとした。なお、厚生労働省の集計は１週間単位で行われ、集計がなされる曜日の関係で６月 29 日までとなっている。また、原稿脱稿時の７月末現在、今度は第５波の動向が懸念されているところである。

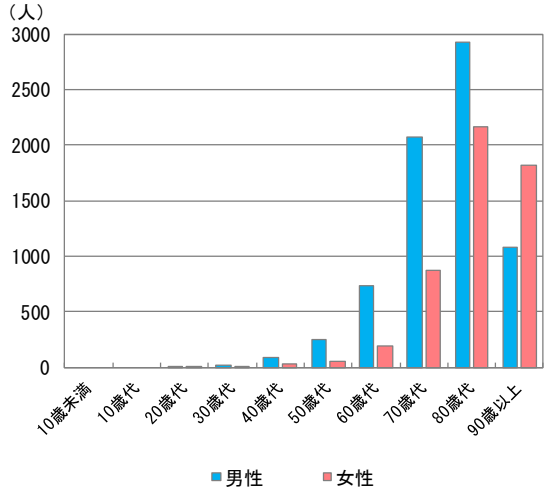
[4] 新型コロナの比較的若い世代での死亡者は、2021年６月29日までの累計で、10歳以下と10歳代では男性、女性ともゼロ、20歳代では男性７人、女性１人、30歳代では男性17人、女性９人となっている。

[5] 平均寿命は、男性81.41歳、女性87.45歳となっている（厚生労働省「令和元（2019）年 簡易生命表」）。

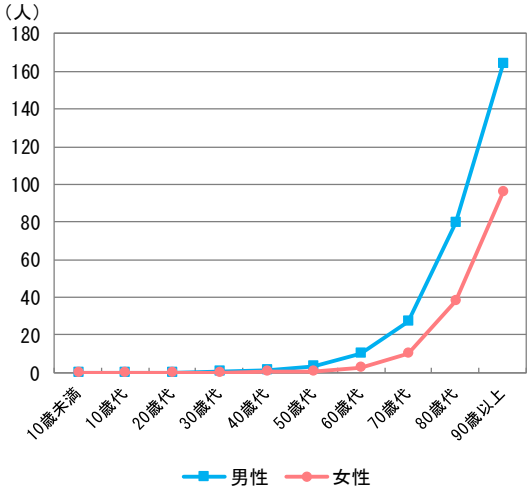
[6] なお、新型コロナによる死亡については、別府志海「新型コロナウイルス感染拡大期における死亡・死因の状況」（国立社会保障・人口問題研究所『新型コロナウイルス感染拡大と人口動態』（2021.7）12頁）において詳細な分析が行われている。

図表１　年齢層別に見た新型コロナの死亡リスク

① 死亡者数



② 死亡率（10 万人当たり）



（注）死亡者数は 2021 年６月 29 日までの累計である（１週間単位で集計が行われる）。死亡者数に性別・年代不明・非公表等は含まれない。
（出所）総務省「人口推計」、厚生労働省「データからわかる－新型コロナウイルス感染症情報－」（https://covid19.mhlw.go.jp/）より作成

以上のように、新型コロナの死亡リスクは高齢者（とりわけ70歳代以上）で高くなっている一方で、20歳代や30歳代以下の若者が新型コロナで死亡するリスクは統計上の数値としてはゼロに近い。そのため、死亡リスクに限ってみると、新型コロナに対する警戒が特に必要なのは高齢者ということになる。このこと自体はかなり以前からいわれていることであるが、これからの論述の前提でもあり、念のため確認しておくこととしたい。

３．都道府県別の新型コロナの死亡率についての回帰分析

ここからは、新型コロナの死亡者の大部分が高齢者[7]であることを念頭に置きつつ、被説明変数を都道府県別の新型コロナによる死亡率[8]、説明変数を都道府県別の（１）高齢者（70歳以上）の割合、（２）高齢者（70歳以上）10万人当た

[7] 高齢者については、一般的には国連の世界保健機関（WHO）の定義による 65 歳以上とされることが多いと思われる。しかし、本稿では、新型コロナによる死亡者数が大きく増加する 70 歳代以上をもって区切りとすることとする（図表１参照）。
[8] 都道府県別の死亡率の計算に用いた死亡者数（NHK「特設サイト 新型コロナウイルス」（https://www3.nhk.or.jp/news/special/coronavirus/）による）は、2021 年６月 30 日までの累計である。図表１と同様、新型コロナのいわゆる第４波が落ち着いたと見られる６月末を区切りとした。ただし、都道府県別人口は、総務省「人口推計」による 2019 年 10 月１日現在の人口である。

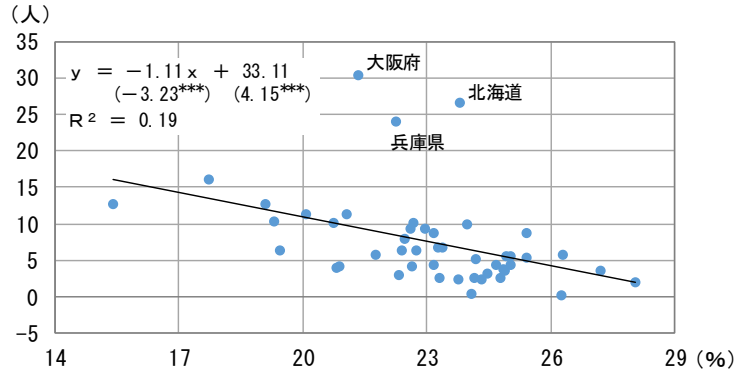
りの特別養護老人ホームの定員数、（３）三世代世帯の割合として、それぞれで回帰分析を行っていくこととする。なお、結果としては、いずれの回帰分析でも統計的に有意な関係が見られることとなるが、後に詳述するように、本当のところはどうかという点については疑問もあり得る。

（１）新型コロナの死亡率と高齢者の割合の関係

最初に、都道府県別の新型コロナの死亡率と70歳以上の高齢者の割合について回帰分析を行うこととする。事前の予想としては、若者や現役世代に比べて高齢者が新型コロナにより死亡するリスクが高いとすれば、人口のうち高齢者が占める割合が高い都道府県ほど新型コロナの死亡率が高くなっていることが考えられる。

ところが、実際に回帰分析を行ってみると、予想に反し、高齢者の割合が高い都道府県ほど新型コロナの死亡率が低くなるという統計的に有意な（説明変数の係数のｔ値が１％の有意水準を満たす）関係が見られる（図表２）。リスクが高い高齢者の割合が高い都道府県ほど新型コロナの死亡率が低くなるのは不可解にも思えるが、その理由は後に詳述することとして、ここではとりあえず話を先に進めることとしたい。

図表２　新型コロナの死亡率と高齢者の割合の関係



（注）1. 縦軸は新型コロナの死亡率（人）、横軸は高齢者の割合（%）である。
2. 定数項及びｘの係数の下の数値はｔ値。***はｔ値が１％の水準で有意であることを示す。また、R^2は決定係数。
（出所）総務省「人口推計」、ＮＨＫ「特設サイト 新型コロナウイルス」より作成

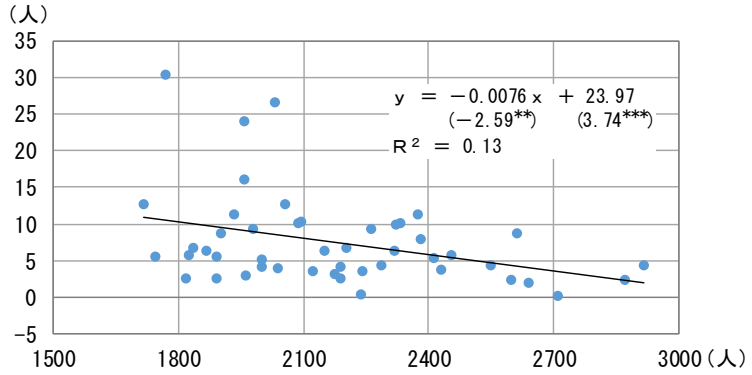
なお、図表２では、散布図に表示された点のうち３つ（北海道、大阪府、兵庫県）が離れ小島のように他の点から離れた位置に来ている（図表３、図表４も同じ）。これらの道府県の新型コロナの死亡率は、北海道で26.6人、大阪府で

30.3人、兵庫県で23.9人と突出して高くなっているが、この点についても後に少々触れることとする[9]。

（２）新型コロナの死亡率と特別養護老人ホームの定員数の関係

次に、都道府県別の新型コロナの死亡率と高齢者10万人当たりの特別養護老人ホームの定員数について回帰分析を行う。高齢期の住まいには様々なものがあるが、ここでは、日常的に介護が必要な高齢者を介護のプロフェッショナルが世話することで、新型コロナの関係でも感染リスク（更には死亡リスク）が低くなることが想定（期待）される特別養護老人ホームを代表として取り上げることとしたい。特別養護老人ホーム（介護福祉老人施設）とは、要介護３以上の認定を受けた人（65歳以上）を対象として、入浴、排泄、食事等の介護その他日常生活の世話、機能訓練、健康管理及び療養上の世話を行う施設のことである[10]。事前の予想としては、高齢者10万人当たりの特別養護老人ホーム定員数が多い都道府県ほど、行き届いた介護を受けることができる高齢者が多くなり、結果的に新型コロナの死亡率が低くなることが考えられる。

図表３　新型コロナの死亡率と特別養護老人ホームの定員の関係



（注）1. 縦軸は新型コロナの死亡率（人）、横軸は高齢者（70 歳以上）10 万人当たりの特別養護老人ホーム定員数（人）である。
2. 定数項及びｘの係数の下の数値はｔ値。**はｔ値が５％の水準で、***はｔ値が１％の水準で有意であることを示す。また、Ｒ²は決定係数。
（出所）総務省「人口推計」、厚生労働省「令和元（2019）年 介護サービス施設・事業所調査」、ＮＨＫ「特設サイト 新型コロナウイルス」より作成

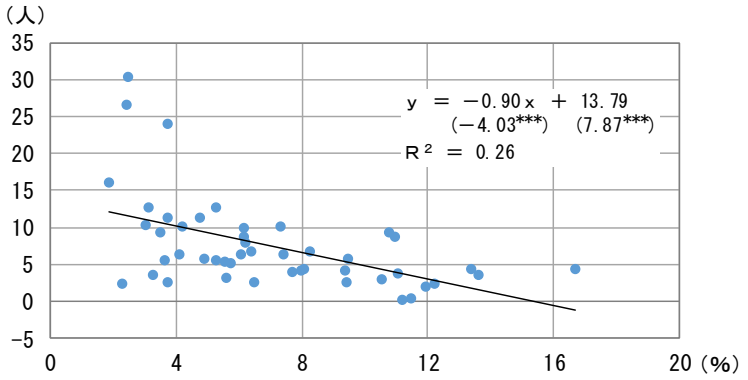
[9] これらは外れ値あるいは異常値などと呼ばれ、厳密な分析のためには外れ値の処理等も必要かと思われる。しかし、本稿の目的には直接の関係はなく、本稿ではあえて特段の操作は行わずに回帰分析を行うこととした。
[10] 高齢期の住まいについては、矢田尚子「高齢期の住まいの種類と特徴について」（国民生活センター『国民生活』（2021.3）１頁）参照。

図表３は、この回帰分析の結果を示したものであるが、この場合は予想の通り、高齢者10万人あたりの特別養護老人ホーム定員数が多い都道府県ほど、新型コロナの死亡率が低くなるという統計的に有意な（説明変数の係数のｔ値が５％の有意水準を満たす）関係が見られる。

（３）新型コロナの死亡率と三世代世帯の割合の関係

さらに、都道府県別の新型コロナの死亡率と三世代世帯[11]の割合について回帰分析を行うこととする。少々唐突感があるかもしれないが、こうした回帰分析を行うのは、次の理由による。世間一般では、行動範囲の広い若者が自分で気が付かないうちに新型コロナに感染し、ウイルスを家庭内に持ち帰って感染を広げているのではないかという疑いが持たれているようである。本当に若者が感染を広げているのであれば、これは一つの仮説、可能性であるが、若者が高齢者と同居する三世代世帯では、高齢の祖父母が子や孫の持ち帰ったウイルスによって感染（さらに、場合によっては死亡）するリスクが相対的に高くなり、そのため、三世代世帯の割合が高い都道府県では新型コロナの死亡率も高くなっていることが考えられる。

図表４　新型コロナの死亡率と三世代世帯の割合の関係



（注）1. 縦軸は新型コロナの死亡率（人）、横軸は三世代世帯の割合（%）である。
2. 定数項及びｘの係数の下の数値はｔ値。***はｔ値が１%の水準で有意であることを示す。また、R^2は決定係数。
（出所）総務省「人口推計」、厚生労働省「令和元（2019）年 国民生活基礎調査」、ＮＨＫ「特設サイト 新型コロナウイルス」より作成

そこで、新型コロナの死亡率と三世代世帯の割合についての回帰分析の結果を示したものが、図表４である。これを見ると、三世代世帯の割合が高い都道

[11] 三世代世帯とは、世帯主を中心とした直系三世代以上の世帯をいう。

府県では、予想とは反対に新型コロナの死亡率が低くなるという統計的に有意な（説明変数の係数のｔ値が１％の有意水準を満たす）関係となっている。この結果に即して考えるとすれば、若者が家庭内で祖父母等（つまり高齢者）に対する新型コロナの感染を広めているという言説には、少々疑問の余地があり得るということになろう[12]。

以上のように、都道府県別の新型コロナの死亡率に対し、高齢者の割合、高齢者10万人当たりの特別養護老人ホームの定員数、三世代世帯の割合という３つの間で回帰分析を行うと、説明変数の係数の符号が事前の予想とは異なるものもあるが、その係数のｔ値は、いずれも統計的に有意な値となっている。ところが、次に詳述するように、人口も説明変数に加えて４つの説明変数から重回帰分析を行った場合には、図表２～４の回帰分析で見られた統計的に有意な関係は、一転して失われることとなる。そうであれば、これまでの回帰分析で見られた関係については、本当のところは存在しないのではないかという可能性もあり得ることとなる。

４．重回帰分析による新型コロナの死亡率の分析

（１）都道府県別の人口を加えた重回帰分析

ここでは、今までに挙げた３つの説明変数に更に都道府県別の人口を加え、合計４つの説明変数を用いて重回帰分析を行うこととする。重回帰分析とは、同時に複数の説明変数を用いて分析を行うことであり[13]、重回帰分析を行う主な目的は、欠落変数の問題を避けるためである。欠落変数の問題とは、説明変数と相関があり、更に被説明変数に影響を与えるような要素を誤差項が含んでいる場合には、説明変数と誤差項が相関を持つこととなり、説明変数の効果を正しく推定できなくなるということである[14]。

被説明変数を都道府県別の新型コロナの死亡率、説明変数を高齢者の割合、高齢者10万人当たりの特別養護老人ホームの定員数、三世代世帯の割合及び人口（単位は万人）として重回帰分析を行うと、結果は次のようになる。

---

[12] なお、一般論としては、新型コロナに限った話ではないが、三世代世帯であれば、高齢者の容体の急変や突然の意識喪失といった場合でも、家族が救急車を呼ぶ、場合によっては息子・娘が運転する自動車で病院に向かうなどの対応ができ、高齢者の単独世帯などと比べると相対的に死亡リスクが低くなることも考えられる。

[13] これまで行ってきた説明変数が１つの回帰分析は、重回帰分析と区別する場合には、単回帰分析と呼ばれる。

[14] こうした欠落変数の問題を起こすような変数は、交絡因子と呼ばれる。詳細については、西山慶彦ほか『計量経済学』有斐閣（2019.7）141 頁など、計量経済学の教科書を参照。

推計式１　都道府県別の新型コロナ死亡率の重回帰分析

$$y = 11.11 - 0.016x_1 - 0.002x_2 - 0.357x_3 + 0.012x_4 + u \quad (R^2 = 0.50)$$
(1.154)　(-0.045)　(-0.666)　(-1.429)　(3.679***)

（注）1. 被説明変数ｙは新型コロナによる人口 10 万人当たりの死亡率、説明変数x_1は 70 歳以上の高齢者の割合、x_2は人口 10 万人当たりの特別養護老人ホーム定員、x_3は三世代世帯の割合、x_4は人口（単位は万人）、ｕは誤差項である。
2. 定数項とx_1～x_4の各係数の下の()内の数値はｔ値。***は、ｔ値が１％の有意水準で有意であることを示す。R^2は決定係数である。
3. データの出所は総務省「人口推計」、厚生労働省「令和元（2019）年 介護サービス施設・事業所調査」「令和元（2019）年 国民生活基礎調査」、ＮＨＫ「特設サイト 新型コロナウイルス」。

重回帰分析の結果を見ると、これまでの（単）回帰分析では統計的に有意な関係にあった高齢者の割合、高齢者10万人当たりの特別養護老人ホームの定員数、三世代世帯の割合は有意ではなくなり、都道府県の人口のみが統計的に有意となっている。このことは、換言すれば、都道府県別に見た新型コロナの死亡率に関係があるのは都道府県別の人口のみであり、高齢者の割合などの３つの説明変数は本当のところは無関係かもしれないということになる。

（２）それぞれの回帰分析での有意性が重回帰分析では失われる理由

では、なぜそれぞれの回帰分析（図表２～４）で見られた有意性が重回帰分析では失われているのであろうか。理由としては、それぞれの回帰分析の誤差項の中に実は都道府県別の人口の要素が含まれており、その都道府県別の人口が、高齢者の割合、高齢者10万人当たりの特別養護老人ホームの定員数、三世代世帯の割合という３つの説明変数とそれぞれに関係し、更に新型コロナの死亡率とも関係していたこと（欠落変数の問題）が考えられる。

そこで、都道府県別の人口（単位は万人）を説明変数（ｘ）とし、図表２～４の説明変数を今度は被説明変数（ｙ）として回帰分析を行うと、その結果は推計式２のようになる。これを見ると、人口は図表２～４の説明変数に対してはマイナスの影響を与えており、人口が多い都道府県ほど、高齢者の割合は小さく、高齢者10万人当たりの特別養護老人ホームの定員数は少なく、三世代世帯の割合も小さくなることが分かる。

これを多少敷衍していうと、まず、高齢者の割合については、進学や就職等に際して若者が人口の多い大都市に集中する人の流れが長年にわたり続いており、その結果、人口の多い都道府県では高齢者の割合が小さくなり、人口が少ない都道府県では高齢者の割合が大きくなる。次に、特別養護老人ホームについては、人口の多い都道府県では高齢者の人数が多く、近年はいわゆる介護難

推計式２　都道府県別の高齢者の割合等と人口との関係

○高齢者の割合との関係

$y = 24.46 - 0.0053x + u$　　$(R^2 = 0.37)$

(61.43***)　(-5.12***)

○高齢者 10 万人当たりの特別養護老人ホームの定員数との関係

$y = 2261.61 - 0.31x + u$　　$(R^2 = 0.09)$

(38.65***)　(-2.06**)

○三世代世帯との関係

$y = 8.64 - 0.0062x + u$　　$(R^2 = 0.23)$

(13.45***)　(-3.69***)

○新型コロナ死亡率との関係

$y = 3.47 + 0.015x + u$　　$(R^2 = 0.45)$

(3.63***)　(6.03***)

(注) 1. 説明変数 x は都道府県別の人口（単位は万人)、被説明変数 y はそれぞれの小見出しに掲げられた事項、u は誤差項である。
2. 定数項と x の各係数の下の()内の数値は t 値。**は t 値が５％の水準で、***は t 値が１％の有意水準で有意であることを示す。R^2は決定係数である。
3. データの出所は推計式１と同じ。

民[15]の問題が持ち上がるなど、高齢者10万人当たりの特別養護老人ホームの定員数が少なくなるが、人口が少ない都道府県では高齢者の人数も少なく、特別養護老人ホームの定員にはまだ余裕がある。さらに、三世代世帯の割合については、地方では大都市と比べると先祖伝来の土地・家屋に祖父母・孫が同居することも比較的多く、そのため、人口の多い都道府県ほど三世代世帯の割合が小さくなる。都道府県の個別事情は様々であるが、総じて見た場合の傾向としては、こうしたことがいえるのではないかと思われる。

他方で新型コロナの死亡率と人口との関係については、人から人に感染する新型コロナの性質上、人口が多い都市部ほど感染リスクが高くなり、ひいては（特に高齢者にとって）死亡リスクも高くなることが考えられる。そのため、基本的には、人口が多い都道府県では新型コロナの死亡リスクも高くなるということがいえよう。

５．おわりに

以上述べてきたように、都道府県別の新型コロナの死亡率については、都道府県別の高齢者の割合、高齢者10万人当たりの特別養護老人ホームの定員数、

[15] 介護難民とは、介護が必要であるにもかかわらず、自宅でも病院でも介護施設でも介護を受けることができない人のことである。

三世代世帯の割合に対してそれぞれ回帰分析を行った場合には統計的に有意な関係が見られるが、それらに都道府県の人口を加えて４つの説明変数で重回帰分析を行うと、新型コロナの死亡率との間で有意な関係があるのは人口だけという結果になってしまう。すなわち、高齢者の割合などのそれぞれに回帰分析を行うだけでは、背後にある人口という要因を見落とす可能性があるということである。したがって、こうしたミスリードを避けるためにも、結論としては、１回の回帰分析によって統計的に有意な結果が得られても、それだけで満足するのではなく、その有意な結果が生ずる原因についても考えを巡らせ、可能であれば別の方法も試してみるなどのことが重要であるといえよう。

なお、本稿での重回帰分析の結果は、人口以外の要因が新型コロナの死亡率に影響を与えないことまで意味するものではない。実際には、人口以外のもので、かつ、人口とは無関係の幾つもの要因が新型コロナの死亡率に影響を及ぼしていることが考えられる。前述のように、北海道、大阪府、兵庫県では新型コロナの死亡率が突出して高くなっているが、これらの道府県については、新型コロナの感染拡大に伴い、医療提供体制の逼迫やいわゆる“医療崩壊”などが報じられてきたところである[16]。その意味では、新型コロナに係る医療提供体制の逼迫の度合いが新型コロナの死亡率に関係している可能性を考えることができよう[17]。また、これ以外にも、都道府県別の人々の外出率、マスクの着用時間、居酒屋の利用頻度などを正確に測定して厳密に指標化・数値化することができれば、これらの要因が都道府県別の感染者数に影響を及ぼし、更に死亡率にも関係するという結果もあり得るかもしれない。ただ、ここで挙げたような新型コロナに影響を及ぼす要因そのものに関しては、医療や感染症などの専門的な見地に基づく慎重な検討が必要であり、統計的な手法についても更に緻密なものが求められるかと思われるため、本稿ではこの点についての結論は留保し、今後の課題とすることとしたい。

（内線75044）

---

[16] 医療崩壊とは、本来あるべき医療を受けることができないことである。なお、「新型コロナ 大阪第４波／上 変異株対応、負の連鎖」『毎日新聞』（2021.5.23）などを参照。

[17] 死亡率ではなく、人口 10 万人当たりの感染者数と人口の関係について、都道府県別に回帰分析を行うと、この場合も両者の間には統計的に有意な相関関係（人口が増えれば人口 10 万人当たりの感染者数も増える）が見られる（図表等は、スペースの関係もあり割愛する）。なお、北海道、大阪府、兵庫県の人口 10 万人当たりの感染者数について図形的に見ると、回帰直線からの乖離は、死亡率の場合（例えば図表２を参照）にように一見して外れ値と思えるほどに大きくはなく、ほぼ人口に見合った程度に収まっているように思われる。