回帰分析におけるｔ値とｐ値の意味について

調査情報担当室　前田　泰伸

《要旨》

回帰分析とは、統計的手法によって説明変数と被説明変数の関係を推計する分析方法である。ｔ値及びｐ値は、係数の有意性を示す重要な指標であるが、有意性ばかりに気を取られていると、思わぬ陥穽に陥る可能性もある。すなわち、回帰分析の際、データの個数が極端に多くなった場合には、係数や決定係数が小さく、説明変数と被説明変数の間に関係がありそうに見えなくても、係数のｔ検定の結果、ｔ値やｐ値が有意となることがあり得る。これは、データの個数が増えたため推計の精度が上がり、係数がゼロに近い値であっても、それがゼロであるとはいえないという意味でｔ値（及びｐ値）が有意となるからである。

今はビッグデータの時代であり、億単位のデータの収集も困難ではないが、そうして得られたデータから手当たり次第に回帰分析を行っていった場合には、こうしたことが現実に起こる可能性もあり得よう。回帰分析では、ｔ値（及びｐ値）が有意であれば万事解決というように単純に考えるのではなく、分析結果について経済的・社会的な意味など様々な角度から吟味し、妥当な結論を導くということが重要である。

１．はじめに～有意性の指標としてのｔ値及びｐ値

回帰分析とは、統計的手法によって説明変数と被説明変数の関係を推計する分析方法であり、とりわけ説明変数が１つの単回帰分析は、散布図や回帰直線を図表として示すことにより両変数の関係を視覚的に分かりやすく伝えることができるため、政府の白書や経済レポートでも比較的よく用いられている。ただ、こうした回帰分析は、使い方によっては誤解やミスリードを招く可能性があり、拙稿「回帰分析の不適切使用に注意」[1]では、完全失業率と自殺死亡率について、時系列データでは強い関係がありそうに見えても都道府県別クロスセ

[1] 参議院事務局企画調整室『経済のプリズム』第 187 号（2020.5）45 頁参照。

クションデータ[2]では関係がなさそうに見えることに関する考察を行った。

本稿では、回帰分析における重要な指標であるｔ値及びｐ値について取り上げることとする。ｔ値及びｐ値とは、係数の有意性を示す指標であるが、ただ単に有意であるかそうでないかばかりに気を取られていると、思わぬところで陥穽に陥ることもあり得る。本稿では、パソコンでランダムに発生させたデータから回帰分析を行い、そのような本来的に無関係なデータでもｔ値やｐ値が有意な値となり得る例を示すことで、逆説的にこうした陥穽について説明することとしたい。

２．ランダムなデータの回帰分析でもｔ値やｐ値が有意となり得る

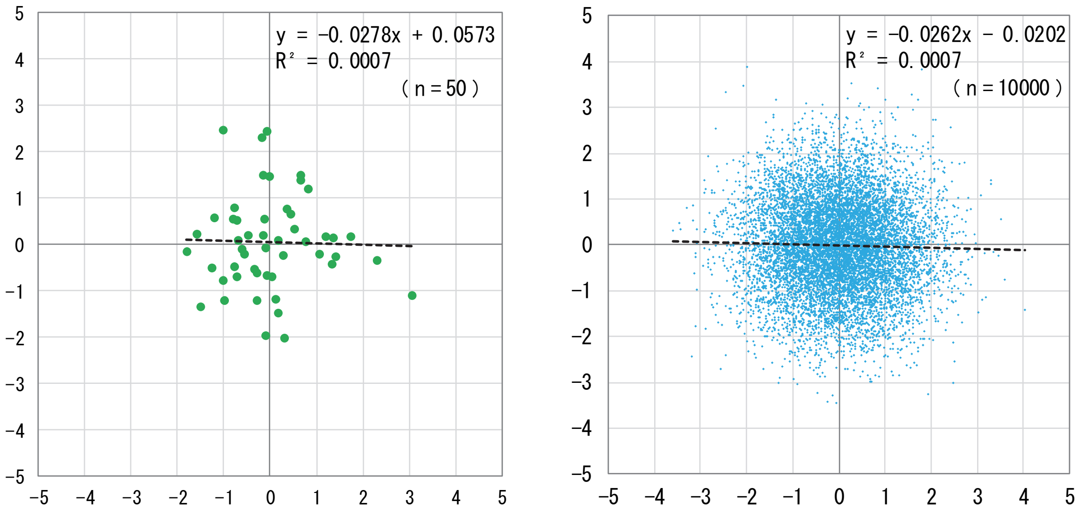
回帰分析による推計結果を示す際には、例えば“$Y＝AX＋B$”といった回帰式[3]のほか、決定係数（R^2）やｔ値やｐ値などの統計的指標も示されるのが通例である。決定係数とは、推定された回帰式の当てはまりの良さの度合いを示す指標であり、０から１までの値を取り、１に近いほど回帰式が観測されたデータに良く当てはまることを示している。また、ｔ値及びｐ値とは、前述のように説明変数の係数が統計的に意味を持つかどうかの有意性を示す指標である。ｔ値及びｐ値は、回帰分析において非常に重視される指標であり、具体的には、「ｔ値が２より大きい（又は、ｐ値が0.05より小さい）ため、ｔ値（又はｐ値）が５％の有意水準を満たしている。したがって、説明変数の係数が０であるとはいえない、つまり、説明変数と被説明変数の関係が統計的に推認できるということだ」といった使い方がなされる[4]。

これらのことを念頭に置きながら、次に、２つの回帰分析を示すこととする。図表（①、②）は、正規分布（平均０、標準偏差１）に従うＸとＹをランダムに発生させ、Ｘを説明変数、Ｙを被説明変数として回帰分析を行ったものである。なお、図表（①、②）は、データを繰り返し何度も（何十回、何百回と）発生させ、その中からＸの係数、切片、決定係数にあまり大きな差がなく、目立った違いはデータの個数だけである（図表上の点の数でもある）というもの

[2] クロスセクションデータとは、時間を一時点に固定し、その時点における場所、人、世帯等において発生したデータを収集・記録したものである。
[3] 回帰式が“$Y＝AX＋B$”となるのは単回帰分析である。本稿で「回帰分析」という場合には基本的に単回帰分析を指しているが、ｔ値及びｐ値に関して本稿で説明する内容は、説明変数が２つ以上の重回帰分析にも当てはまる。
[4] 本文に挙げた以外にはＦ値という指標もよく用いられる。これは、主に重回帰分析において、説明変数の係数の少なくとも１つはゼロではないことをテストするためのＦ検定で用いられる指標である。

図表　ランダムなデータによる回帰分析

① データ 50 個　　　　② データ１万個



①

	係数	標準誤差	t 値	p 値
X	-0.0278	0.1504	-0.1850	0.8540
切片	0.0573	0.1466	0.3912	0.6974

②

	係数	標準誤差	t 値	p 値
X	-0.0262	0.0100	-2.6136	0.0090
切片	-0.0202	0.0100	-2.0236	0.0430

（出所）Microsoft Excel によりランダムなデータを発生させて筆者作成

を選んでいる。データの個数は、①では 50 個（n=50）、②では１万個（n=10000）である。

　ここで、まず、図表（①、②）のデータや回帰分析には取り立てて意味がないことは、説明を要しないであろう。データは現実の経済統計などから得られたものではなく、パソコンでランダムに生成された、いわば偶然の産物である。また、それぞれの散布図の形状を見ても、いかにも偶然らしく、点が団子状態であり（特に、図表②）、回帰式のＸの係数はゼロに近く、回帰直線（図表（①、②）における点線）はほぼ水平となっている。回帰式の当てはまりの良さを示す決定係数もほぼゼロとなっており、これは、回帰式がＸとＹの関係についてほとんど何も説明できていないということを示すものである。このように、図表（①、②）のＸとＹは、ランダムな生成という来歴に加え、純粋に数値と数値の関係として見た場合でもやはり関係がなさそうに見える。さらにいえば、こうした回帰分析には、意味があるのかどうかという疑問を呈せざるを得ないということもいえよう。

　ところが、回帰式の係数（あるいは切片）についてｔ値（及びｐ値）を見る

と、あろうことか、図表②では、ｔ値（及びｐ値）が有意となっており、特にＸ（説明変数）は、１％の水準でもｔ値（及びｐ値）が有意となっている。であれば、図表①はさておき、図表②に限っていえば、ＸとＹの関係には統計的に意味があると結論付けてよいのだろうか[5]。

３．そもそもｔ値やｐ値とは何か

そこで、次には、どうして図表②においてｔ値やｐ値が有意となったのかを考えることになるが、ただ、その前に前提事項として、そもそもｔ値やｐ値がどういうものかについて確認しておくこととしたい。回帰分析を行う際には、説明変数の係数についてｔ検定を行い、係数がゼロではないということのテストをするのが、いわば必須の手続となっている。これは、係数がゼロであるということになれば、例えば“$Y=AX+B$”という回帰式は、“$A=0$”であることから“$Y=B$”と同じになり、被説明変数（Ｙ）が説明変数（Ｘ）に関係しないことが算術的にも明らかになるためである。

回帰分析におけるｔ検定の具体的な方法は、①説明変数の係数が本当はゼロであるという仮説（これは「帰無仮説」と呼ばれる）を立てる、②検定統計量“ｔ値”を計算する（ｔ値の計算方法については後述）、③有意水準を決め[6]、その有意水準に対応するｔ値の臨界値[7]と②で求めた検定統計量ｔ値を比較する[8]、④検定統計量ｔ値の絶対値が臨界値の絶対値より小さい場合には、帰無仮説を正しいとして採択するが、そうでなくｔ値の絶対値が臨界値の絶対値より大きくなった場合には、帰無仮説を誤りとして捨て去り（棄却）、それとは反対の「係数がゼロとはいえない」という仮説（対立仮説）を採択する、といった流れとなる。

以上がｔ検定の方法・手順であるが、そもそもｔ検定は何をテストしているのであろうか。その点をかいつまんで説明すると、次のようになる。回帰分析では基本的に、説明変数、被説明変数及びそれらから計算される回帰式の係数は確率変数であり、母集団からランダムに選ばれた数値であるという考え方を

---

[5] なお、図表（①、②）の①と②を比べてみると、標準誤差にも比較的大きな違いがあるが、この点については、本文において後に詳述する（後掲注９参照）。
[6] 有意水準については、特にこの数値でなければならないという決まりなどはないが、慣例的に５％あるいは１％が用いられることが多い。
[7] 臨界値は、データの個数による“自由度”に応じて少々数値が変わるが、本稿では割愛する。詳細は、森棟公夫ほか『統計学[改訂版]』有斐閣（2015）など統計学の教科書を参照。
[8] 現在では、検定統計量ｔ値のほか、有意水準及び自由度に対応したｔ値の臨界値等については、統計ソフトによって自動で計算がなされる。

採る。有意水準とは、単純にいうと、帰無仮説が正しい（本当のところは係数がゼロである）との前提の下においてそうした数値がランダム（偶然）に選ばれる確率のことである。例えば、有意水準を５％とした場合において、検定統計量ｔ値がその有意水準に対応する臨界値以上となった（つまり、５％の有意水準を満たした）とすると、それは、帰無仮説が正しいという仮定の下では５％以下の確率でしか起こり得ないレアな数値が選ばれてしまったということである。また、１％の有意水準を満たした場合には、１％以下の確率でしか起こりえない珍事が発生したということになる。

ただ、ここから先は、いわば常識的な話であるが、そうした発生確率が極めて低いレアケースが本当に発生したと考えるのが合理的かどうかである。つまり、回帰式の係数は本当のところはゼロであるにもかかわらず、いわば運命の女神の悪戯により発生確率が極めて低い数値が偶然に選ばれることになったと解釈するのか、そうではなく、そもそも前提である帰無仮説が誤りであり、回帰正規の係数は実はゼロではないと解釈するのか、いずれが合理的なのかということである。この場合、ｔ検定の考え方としては、常識的にはおそらく前者は考えにくいだろうということで、５％や１％などの有意水準が満たされた場合には帰無仮説を棄却し、それとは反対の「係数がゼロではないであろう」という対立仮説を採択するという流れとなる。なお、ｐ値については、ｔ値の絶対値をパーセント表示（p 値の１＝100%）に直して示した数値であると考えると分かりやすい。

４．図表②でｔ値が有意水準を満たしてしまった理由

前置きが少々長くなってしまったが、図表②において、回帰式の傾き、すなわち説明変数の係数が非常に小さく、説明変数と被説明変数に関係がありそうに見えないにも関わらず、図表②のｔ値（及びｐ値）が有意となっている理由は、結論的にいえば、②ではデータの個数が非常に多かったためである。ｔ値は、計算上、係数を標準誤差[9]で割ることで求められるが、式を変形していくと次のようになる。

$$t\text{ 値} = \frac{\text{係数}}{\text{標準誤差}} = \frac{\text{係数}}{\text{標準偏差} \div \sqrt{\text{データの個数}}} = \frac{\text{係数} \times \sqrt{\text{データの個数}}}{\text{標準偏差}}$$

[9] 図表②の標準誤差は、図表①の標準誤差に比べてかなり小さくなっている（前掲注５参照）。

なお、標準偏差とはデータのちらばり具合を表す指標であるが、図表（①、②）は、いずれも平均０、標準偏差１の正規分布に従ってランダムに発生させたデータであり、また、両者の係数には大きな違いがなく（図表（①、②）参照)、数値としてはどちらも非常に小さくなっている。図表①と②の違いは、結局、データの個数ということになり、②でｔ値（及びｐ値）が有意になっている理由は、データの個数が非常に多いことによりｔ値（及びｐ値）の数値が引き上げられ、推計の精度が非常に高まったためということがいえる。

ただ、この場合に注意すべきことは、データの個数が増えていくとｔ値（及びｐ値）も大きくなる可能性があるが、だからといって、説明変数と被説明変数の間に強い関係性が推認されるとは限らないということである。繰り返しになるが、回帰分析のｔ検定においてｔ値（及びｐ値）が有意水準を満たすということは、係数がゼロであるという前提で見た場合には発生確率が極めて低いレアケースが起こったということであり、それ以上のものではない。データの個数を極端に大きく増やしていった場合には推計の精度が非常に大きくなるため、図表②のように、回帰式の係数がほんの少しゼロから離れただけでも、ｔ値（及びｐ値）が有意となることはあり得る。ただ、こうしてｔ値（及びｐ値）が有意となったとしても、それは、係数がゼロであるという帰無仮説を棄却するという意味において説明変数と被説明変数が無関係ではない（いわば、関係が「有る」か「無い」の二者択一では「有る」）ことを示すものに過ぎず、説明変数が被説明変数にどの程度の影響を及ぼしているか（つまり、関係が「有る」としても、それはどの程度か）は、ｔ値（及びｐ値）とは別の話となる[10]。

５．むすび

回帰分析に用いるデータの個数が多くなることは、分析の精度を向上させることにもつながり、基本的には好ましいことである。しかし、本稿で示したように、データの個数を１万や10万など極端に大きく増やしていった場合には、ｔ値（及びｐ値）が有意であっても、係数がかなりゼロに近く、グラフ上の傾きもほとんど水平となり、決定係数も非常に低くなるということがあり得る。その場合には、ｔ値（及びｐ値）は有意であり説明変数の係数がゼロではない

[10] 廣野元久『目からウロコの多変量解析』日科技連（2019）では、有意差検定でｐ値が１％を下回るなど高度に有意であるということは、帰無仮説を棄却する場合に誤る可能性が極めて小さくなったという意味であり、相関（つまり、関係や影響）が強いという意味ではないとしている。

という点では一応の意味があるとしても、係数の値がほとんどゼロと同視できる程度であり説明変数が被説明変数にほとんど影響を及ぼさないという点では、こうした回帰分析に意味があるのかという疑問が湧いてくることもあるかもしれない。あるいは、本来的には何の関係もない２つの変数が、５％以下あるいは１％以下の偶然によりたまたまｔ値やｐ値が有意になったに過ぎないということも、可能性としては考えられなくはない[11]。特に今はビッグデータの時代であり、収集するデータによっては、その個数が数千万あるいは億単位に上ることもあろうかと思われる[12]。そうして得られたデータから手当たり次第に回帰分析を行っていくとすると、理論的には何の関係もなさそうに見えるのにｔ値（及びｐ値）が有意になっているということが現実に起こるかもしれない[13]。

結局のところ、重要なことは、ｔ値やｐ値とは有意性を示す重要な指標であるが、それらが有意であれば万事解決と単純に考えるのではなく、回帰分析の結果について、経済的・社会的な意味など様々な角度から吟味し、妥当な結論を導くということであろう[14]。また、その際には、世間一般の意味での常識を働かせるということも必要となるのではないかと思われる。

（内線75044）

[11] 偶然にｔ値（及びｐ値）が有意になったとすると、棄却されてはいけない帰無仮説が棄却されるという意味では“第１種の過誤”が起こっていると考えることもできる。なお、図表②は、Ｘ、Ｙともランダムに発生させたデータ（それゆえ本来的に無関係）であり、偶然にｔ値（及びｐ値）が有意になった例でもある。

[12] 身近なビッグデータの例としては、店のレジで販売がなされた時に記録されるＰＯＳデータなどが挙げられよう。なお、総務省「情報通信白書」（2017）には、ビッグデータの定義や範囲についての言及がある。

[13] ビッグデータを活用し、同じ母集団からランダムに数万個のデータを選んで回帰分析を行うということを何度も繰り返した場合において、常に係数が非常にゼロに近い同じような値を取り、しかも常にｔ値やｐ値が１％の有意水準を満たしたということがあったとすると、説明変数と被説明変数は非常に弱いながらも関係があるという結論にならざるを得ない。ただ、そのことが経済的・社会的に意味を持つかどうかについては、一概にいえないように思われる。

[14] 豊田秀樹『瀕死の統計学を救え！』朝倉書店（2020）では、二項検定の例を示しつつ、多くの有意性検定での共通の課題として、帰無仮説が否定されたとしても、それは科学的な発見を意味するものではなく、科学的発見のための必要条件の確認に過ぎないとしている。