研究員の眼

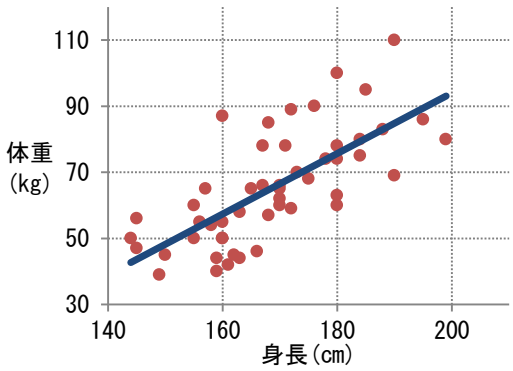
回帰分析の落とし穴

分析結果は、推論の正しさを裏付けているか?

保険研究部 主任研究員　篠原　拓也
(03)3512-1823　tshino@nli-research.co.jp

実験や観測、アンケートなどから得られるデータをもとに、○○が原因で、□□という結果になる、という推論をしたとしよう。例えば、よく使われる例で、身長と体重の関係がある。ある成人男性の集団をもとに、「身長が高い人は、体重が重い」という推論をする。横軸に身長を、縦軸に体重をとって分布図として、各データを表してみると大体の傾向がわかる。体格は人それぞれで、中には、身長は高いが体重は軽いという人や、身長は低いが体重が重いという人もいるが、一般的には大きな体の人は小さな体の人に比べて身長が高く体重は重い、という傾向にあり、「身長が高い人は、体重が重い」という推論は、概ね間違っていないと考えられる。

身長と体重の分布図（イメージ）



それを図示するのに用いられるのが、回帰分析である。統計的な手法を用いて、分布図に、データの分布傾向を示す直線を引く。この線が右上がりの場合、身長が高いと体重が重い、という関係が見えてくる。

この直線と各データの間のずれが小さいほど、当てはまりのよい直線が引けることになる。横軸と縦軸の相関の程度は、相関係数という、1 から-1 までの間の数値で表される。正の値の場合、一方の数量が増えると、もう一方の数量も増えるという正の相関となる。身長と体重の関係は、正の相関である。逆に、一方の数量が増えると、もう一方の数量が減る場合は、負の相関となる。相関係数の値が 1 や-1 に近いときは相関が強い、0 に近いときは相関が弱いと言われる。

現在、回帰分析は、表計算ソフトや各種統計ツールで簡単に行うことができ、様々な統計分析で活用されている。しかし、回帰分析には、気をつけておくべき落とし穴がいくつか潜んでいる。

まず、1 つ目は、データを区分すればするほど相関は強くなるが、結果が複雑になる点である。例えば、身長と体重の例で、集団を 20～39 歳、40～59 歳、60 歳～、などと年齢ごとの群団に分けて、それぞれ回帰分析をすると、群団に分ける前よりも、相関を強めることができる。これは一見、良い

ことのように見える。しかし、分析結果が複数に分かれて、複雑になることに注意する必要がある。ここで更に、各年齢群団を、肥満にならないよう食事制限や運動に努めている人と、そうではない人に分けて、それぞれのグループで回帰分析をすれば、もっと強い相関が得られるかもしれない。しかし、このようにして、区分を細かくして得られた分析結果は、複雑で理解しにくい。

2 つ目は、原因と結果を逆にすると、奇妙な推論になってしまう点である。回帰分析は、両者の関係を直線で表示するが、因果関係については何も示さない。例えば、様々な都市で、警察官の数と犯罪率の関係を見てみると、両者には負の相関がある。これを、「警察官が多いと、犯罪率は低下する」と推論するのは妥当であろう。しかし、「犯罪率が低いと、警察官は多くなる」と推論するのは奇妙である。

3 つ目は、無理やりに直線を当てはめても意味がない点である。野球の試合での控え投手を例に、ブルペンでの投球数と、試合での投球結果の関係を考えてみよう。控え投手は、ある程度ブルペンで投球をしないと試合でいい結果が出せないが、ブルペンで投げ込み過ぎると疲労してしまい悪影響となる。つまり、ブルペンでの投球数と、試合での投球結果の関係を、単純な直線で表すことはできない。このような場合には、直線にこだわらずに、曲線で近似することを模索すべきであろう。

4 つ目は、回帰分析は有効なツールだが、これだけで無理に推論を進めるべきではないという点である。例えば、2000 年代に、日本の 65 歳以上人口と、アメリカの携帯電話契約数はいずれも増加した。回帰分析をすると、両者には、強い正の相関が見られることとなる。しかし、だからと言って、「2000 年代は、日本の 65 歳以降の人口が上昇したから、アメリカの携帯電話の契約数が伸びた」などと推論することは、ナンセンスであろう。

最後に、5 つ目の点は、応用編で、複数の原因を想定して分析をする場合に生じ得る「多重共線性」といわれる問題である。例として、先ほどの警察官の数と犯罪率に、パトカーの数も入れて、「警察官やパトカーの数が多いと、犯罪率は低下する」と推論してみよう。これは重回帰分析といわれ、犯罪率を、警察官とパトカーの数を用いた算式で、より精緻に表現しようとするものである。

ここで、「パトカーの数が多いと、犯罪率は上昇する」という分析結果が得られることがある。これは、原因として想定した、警察官の数と、パトカーの数の間に強い相関がある場合に発生する。算式上、「警察官の数が多いと、犯罪率は低下する」という関係が強く出過ぎてしまい、パトカーの数と犯罪率の関係が、これを打ち消すように、本来とは逆の関係として表現されてしまうのである。この場合、例えば警察官の数を除いて、パトカーの数と犯罪率の関係を、再度分析することが必要となろう。

以上のとおり、回帰分析には落とし穴がある。常に、データの分布図を参照して、分析結果の妥当性を確認する必要がある。また、回帰分析は推論を裏付ける証拠の 1 つにはなり得るが、回帰分析だけで推論の正しさが証明できる訳ではない。回帰分析を用いた分析結果を把握する際には、これらのことに注意する必要があると思われるが、いかがだろうか。

