回帰分析へのまえがきより

村上雅人

わたしが過ごした中学高や高校の理系の先生にはユニークな先生が多かった。中高ともに進学校ではあったが、受験一辺倒というより、受験そっちのけでよく一風変わった実験をやらされたことを覚えている。

確か生物の授業だったと思う。ある時、クラス全員が学校の裏庭の落ち葉の収集を命じられた。裏庭の掃除かと思っていたら、収集した葉っぱの重さと長さの測定をして表にまとめてみろとの指示である。その後、すべてのデータで頻度分布を棒グラフにしてみた。すると、平均値にピークがあり、その値からはずれるほど、標本の数が減っていくという分布になることが分かった。そして、世の中にあるなんらかのデータを集めると、標本数を増やせば正規分布と呼ばれる分布に近づくことを習った。

さらに、落ち葉の重さと長さに相関があるかどうかの検証も演習として行った。これには、x 軸に落ち葉の重さを、y 軸に長さをプロットして、どのような傾向を示すかを見ればよい。すると、多少のでこぼこはあるものの、これら変数はほぼ比例するということが分かった。

ところで、この比例関係をより定量化するには、どうしたらよいだろうか。もっとも簡単なのは、目分量で直線を引いてしまうことである。これが手っ取り早いし、意外とまともな結果が得られることが多い。しかし、より科学的なアプローチとするには、まず比例関係を数式で表して、それがどの程度の正当性があるかを数値で示すことである。

生物の先生は、落ち葉の解析のとき、最小 2 乗法という手法があって、それを使えば、所与のデータに対して、尤も相応しい比例関係を 1 次式で求めることができるということも教えてくれた。当時は、最小 2 乗法についての説明はそれほど詳細なものではなかったが、なるほどと関心したことを覚えている。

その後、大学に進んで誤差関数という、ちょっと風変わりな名前の関数を習った。当時は、誤差があること自体が間違いと思っていたので、なぜ不必要な（あるいは無い方がよい）誤差をわざわざ関数にあてはめる必要があるのかと疑問に思ったことを覚えている。

今にして思えば、これは、受験勉強の弊害のひとつであったかもしれない。世の中には、すべて正答があって、それに合わないものは間違いであるという先入観にはまっていたのである。よって、「誤差などは本来あるべきものではない、つまり、頑張れば、誤差などゼロにできる」と信じていたのである。

これは、古典力学と量子力学のせめぎあいに似ている。古典力学においては、量子力学の不確定性など許しがたいものである。量子力学では、物質の位置や速度は正確に決定できないと教えているのである。これは、常識では受け入れ

がたいことである。しかし、量子力学が古典力学にとってかわったように、自分の中でも、測定誤差があることは、むしろ本質的なことだと認識するようになった。誤差というよりもゆらぎと言った方がより正確かもしれない。

ところで、誤差の分布は正規分布にしたがうことが知られている。これを最初に提唱したのがガウスであり、この分布をガウス分布と呼ぶこともある。誤差関数は、このガウス分布関数を積分したものである。つまり、誤差といっても、でたらめなものではなく、きちんとした数学的解析ができるものなのである。

その後、卒業研究や大学院での研究で、そのデータ解析において最小 2 乗法をひんぱんに使うようになった。この手法は、その名の通り「誤差の 2 乗和を最小にする方法」である。例えば、落ち葉の長さと重さの関係に $y=ax+b$ という関係があると仮定すると、実際の測定値との間の誤差の 2 乗和が最小になるように、係数 a と b を決める手法である。これが本書の主題である回帰分析である。

回帰分析は、理工系の研究室では、ごく一般に使われている手法である。かつては研究者が苦労して計算で求めていたが、いまでは市販ソフトでいとも簡単に結果が得られるようになった。世の中便利になると、必ず弊害があらわれる。回帰分析の過程がブラックボックスになったために、得られた結果が本当に信憑性のあるものかどうかの検証ができにくくなったのである。場合によっては、回帰分析の手法そのものを知らない研究者も多くなってしまった。

喩えはよくないかもしれないが、携帯電話を使いこなせても、それがどのような原理で通信を可能にしているのかを理解していないようなものである。もちろん、多くのひとにとっては、それで構わないのであろうが、回帰分析が自分の研究の結論を左右するような場合には、原理を知らないですますことはできないであろう。

最近の学会で、こんな会話を聞いた。あるデータのフィッティング結果を発表した研究者に、会場から「このフィッティングにどの程度の信頼性が置けるのか」という質問があった。発表者は、コンピュータの出力結果の決定係数を示し、0.9 と 1 に近いので、このフィッティングはかなり信用が置けると答えていた。これで、質問者も納得して終わってしまった。

しかし、回帰分析を学び、さらに、その背後にある統計学について学ぶと、コンピュータにデータを入力して得られた結果をそのまま鵜呑みにしてしまうことがいかに危険かということを思い知らされる。

回帰分析は、誤差の 2 乗和を最小にする手法であるが、それで終わりではない。誤差が従う分布である正規分布を統計学的に分析すれば、いろいろな検証ができるのである。もちろん、そのためには、統計学をきちんと勉強する必要

がある。統計学については、すでに「なるほど統計学」で、その重要性も含めて基本を説明している。よって、くわしい統計手法については、そちらを参照していただきたい。ただし、回帰分析にとって必要な基本事項については、あえて重複するとは思いながらも、本書でも解説してある。

統計学の知識を利用すれば、回帰分析で得られた結果が、どの程度信頼の置けるものであるかを定量的に検証できるのである。特に、理系の場合は、実験データの数を多くするのが不可能な場合が多いので、統計的検証は重要になる。回帰分析でも、データ数が 4 個でフィッティングしていたものを、もう 1 点データをとったら、様相ががらっと変わることがよくある。転ばぬ先の杖ではないが、統計的な検証をしっかりしておけば、このような場合でもまごつかない。

残念ながら、統計の知識を前面に出すと、煙たがわれることも多い。例えば、データ数の少ない実験では、実験そのものに意味がないという検証結果が出てしまう場合がある。先日、ある会合で、10 個程度のデータ解析に正規分布を仮定して分析していたので、「数が少ない場合には、正規分布に従う変数でも *t* 分布という少し異なる分布に従うから、*t* 検定が必要となる」と指摘したら、「そんな細かいところまで気にする必要はない」と簡単に却下された。

実は、誤差が正規分布に従うと何度も強調してきたが、回帰分析においても、データ数が 30 個に満たない場合には、*t* 分布にしたがうので、*t* 検定が必要となる。理系の実験データでは、分析装置が自動的にデータを取得してくれるケースは別にして、データ数が 30 個を超えることはそれほど多くない。よって、*t* 分布を考える必要がある。うるさいといわれそうだが、貴重なデータを正しく処理するためにも、統計の知識は必要である。そうしないと、思わぬ過ちをおかす場合も出てくる。

ところで、回帰分析には、独立変数が 1 個の単回帰分析に対して、独立変数が 2 個以上の重回帰分析がある。経済学や人文系のデータ解析では、数多くの変数を扱う場合も多く、重回帰分析の方が主流である。この解析手法は多変量解析として、解説書が数多く出版されているので、詳細はそちらに譲るが、本書では、重回帰分析の基本的な考えを単回帰分析の手法をもとに説明している。その基本的考えについては、十分理解できるであろう。

ただし、ひとこと断っておくと、理系における実験データの解析では、変数が 2 個あれば、ほとんどのフィッティングが可能であるといわれている。このため、フィッティングパラメータとして変数を 2 個以上使うのを嫌う傾向がある。（あるいは、そのような解析にはクエスチョンマークがつけられる。）よって、できるだけ、他の条件は変えないようにして、一つの変数だけに着目し、その影響をみるというのが常套手段となっている。

社会現象を研究対象とすると、このような手法は使えないが、それでも、や

たらと変数を増やしたのでは、解析が複雑になるだけで、有用な結果は得られない。本書では紹介できなかったが、変数が多い場合には、その変数を使う意味があるかどうかの有意性を検証する方法も開発されている。この手法も、市販のプログラムに組み込まれていて、利用者は機械的に出力結果を参照すれば済むようになっている。しかし、繰り返すがブラックボックス的な解析では本質を見誤るおそれがある。やはり、基本を忘れてはならない。