単回帰分析はむずかしい

川瀬　雅也[1]*・松田　史生[2]

Aさん：私，撃沈しちゃったんでしょうか？さっきのセミナー？

B君：うーん．発表は練習通りでちゃんとしてたけどね．とにかく，X教授の意見を聞いてみようよ．

さて，Aさんが先ほどのセミナーで発表した研究報告について説明しておこう．Aさんの研究テーマは，環境汚染物質を効率よく分解する菌のスクリーニングと特性評価である．これまでに，菌を数種類分離し，その分解特性を調べているところだった．ある菌αの環境汚染物質の分解速度と温度の関係を調べていたところ図1のようなデータが得られた．

Aさんは，このデータをもとに「分解速度が温度に比例している」と指摘した．さらに，各温度でのデータは5回ずつとられていること，25℃でのデータは（23，24，24，34，24）であったこと，データの一つが34 mg/hと非常に大きいことから，25℃で分解速度の上昇はこの外れ値に起因するものであると考察した．B先輩とのディスカッション通り，うまく発表した若干ドヤ顔のAさんに対して，助教の先生は，データを冷静に見て，他の見方や考え方はないのかと指摘したわけである．

研究を行っていると，必ず一度は，このようなデータに出会うと思うが，皆さんは，どう扱うべきだと思われるだろうか．

回帰分析の落とし穴

X教授：Aさん，納得いかないような顔をしているが，何かおかしなことでもあったのかね？

Aさん：おかしなというか，実は，こういう次第なんです．

B君：データを見ても，Aさんの説明に問題があるとは思えないですよね．それで相談に来たわけです．

X教授：なるほど，助教の先生はAさんに研究者としてすごく期待しているみたいだね．

確かに，このデータを見た多くの人が「分解速度が温度に比例している」と感じると思うね．無意識のうちに回帰分析ができて，温度と分解速度の間に直線関係（線形の関係という）が成り立つと考えたからだね．回帰分析は知っているね．

Aさん，B君：はい．

X教授：君たちがよく使うのは単回帰分析だね．図1のように2変数（分解速度と温度）の間の関係を分析して，

分解速度＝f（温度）

という関係に表す分析だ．このような関係式を回帰式というんだ．これが「分解速度＝係数×温度＋定数」という1次関数になると線形単回帰ということになり，今，君たちが想定していることだね．試しに，Aさんのデータを単回帰分析にかけてみようか（図2）[1]．

Aさん：決定係数も大きいし，やっぱり，25℃のデータの一つがおかしかったんじゃないですか．

X教授：では，試しに，皆がおかしいというデータを外してみよう．図3のようになるね．

Aさん：あれ，なんか自信がなくなってきました．

X教授：では，もう一つこの図を見てもらおうか（図4）．さらに，外れ値と言われたデータを除いてみると（図

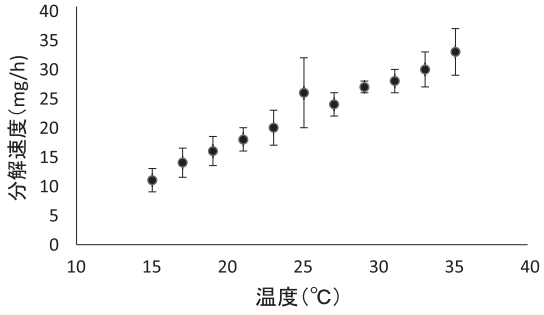

図1．分解速度と温度の関係

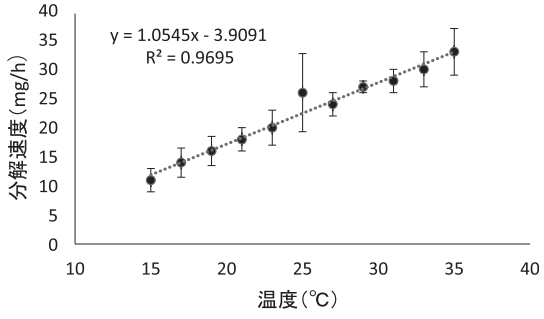

図2．単回帰の結果

著者紹介　[1]長浜バイオ大学（教授）　E-mail: m_kawase@nagahama-i-bio.ac.jp
[2]大阪大学大学院情報科学研究科（准教授）

5)，どうかな．

Aさん：ピッタリです．あれれ，どういうことですか．

B君：そもそもどうして3本の曲線なんですか？

X教授：まあ，図4，5だが，科学的な根拠はまったくなく，3曲線の近似でもフィットさせることができるというだけなんだ．もし，無理やり説明するなら，この菌には分解経路が三つあり，温度によって使われる経路が違うということになるのかな．

Aさん：助教の先生は，そのことを指摘したかったんでしょうか？データの読みが甘いって．

B君：ちょっと待ったー！今，まったく科学的根拠はないって言われたばかりだよ．よく考えよう．

X教授：そうだよ．簡単に，3曲線の方がフィットするから，三つの経路がなんて言うと，それこそ，助手の先生から総攻撃を受けたかもしれないね．

Aさん：？？？？でもフィットはしているんですよね．

X教授：まず，言いたいことは，Aさんのデータを見ると，意識していないが，人は直線関係を思い描いて，他の可能性を排除してしまうんだ．ところが，他の可能性はと考えると，ちょうど，図4や図5のようによりフィットした可能性を考えることができるんだ．しかし，これは，科学的ではないね．

B君：そう言われると，確かにそうですね．

X教授：もう一つ言うと，反応速度論は知っているね．

Aさん:実は，生物系なので，物理化学は弱いんです．

X教授：詳しいことは，自分で勉強してもらうことにして，反応速度は簡単に言えば$v = k[S]^n$と表すことができる．Sはこの場合，汚染物質の濃度で，nは反応次数，kは速度定数とよばれるものだ．$k = A\exp\left(-\frac{\Delta E}{RT}\right)$となる．これはアレニウスの式とよばれΔEは活性化エネルギー，Aは頻度因子だね．Aを理論的に求めようとする研究が今も行われている．ここで注目してほしいのは温度の入っている項だ．指数関数になっているね．

Aさん：はい．

X教授：もし，頻度因子や活性化エネルギーが実験した温度範囲で一定値なら，温度に対して反応速度が直線的に大きくなると思い込むのはおかしくないかね．

B君：本当ですね．まったく，気が付きませんでした．簡単に，データを見て直線関係があると思い込んでいました．

X教授：ただ，生きた生物を使う反応だから，何段階かの反応だろうし，反応の分岐もあるかもしれないし，簡単にアレニウスの式を持ってくることはできないと思うが，このデータから直線関係「だけ」を見つけるのはよくないね．あと，3本の曲線の例のようにフィットするからいいというのも良くない．グラフが示唆しているものは，その背後にある物理学的，生物学的な関係性だから，その点を議論してほしかったんだろうな，助手の先生は…．でも本当のところは，いろいろな角度から調べてみないと分からないけど，助手の先生はAさんならそれができると思ったのだと思うよ．

Aさん：少し，元気が出てきました．もっと，よく調べてみることにします．

X教授：回帰分析を使うときには，科学的に考えてみることが，まず必要だ．よく忘れてしまうことだけど，頭のどこかにないと科学をやっているとは言えないか

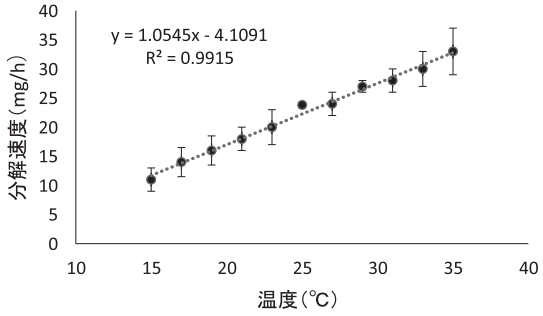

図3．25℃のデータの一つを外した結果

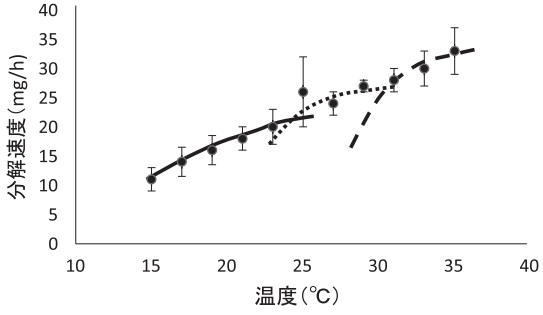

図4．データを別の見方で回帰したケース

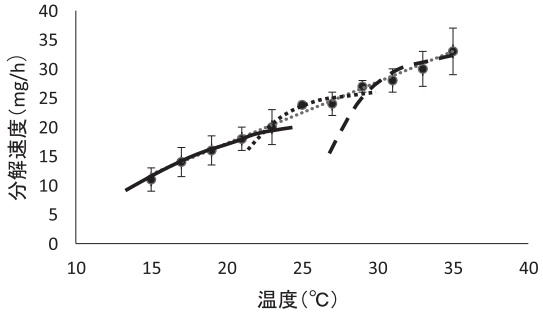

図5．25℃のデータの一つを外したケース

らね．それと，以前も言ったと思うが，簡単に，あるデータを外れ値だから省いてしまえということは，決して科学者が言ってはいけない言葉だ[2]．外れ値であることをきちんと説明できなければ，そのデータは外してはいけないんだ．図3を見てほしいんだが，何か気が付かないかね．

B君：そう言われれば，25°Cのデータがやっぱり直線にのっていませんね．かなり近くにはあるのですが．エラーバーも直線にかかっていないし．

X教授：そうだね．どう見るかによるが，やはり，少し気になるね．ここで，何かあると見るか，見ないかで，その後の展開が大きく違ってくる．もし，君たちが先生から言われたように，外れ値として扱ってしまったら，ものすごい発見の機会を失い，誰かに先を越されるかもしれない．

Aさん：誰かに先を越されたら，悔やんでも悔やみきれません．

X教授：そうだね．データをグラフ化すると，人間は単純だから，反射的に回帰分析的な考えをとってしまうんだ．でも，それでいいのか，立ち止まって考えることも必要だね．Aさんの場合，たまたま異常値があり，ここで考える機会ができたことは，まさに天の助けだね．でも，本当にAさんの実験の腕が悪く，単回帰でいい可能性も残っているので，まずは，いろいろと調べてから報告する方がいいよ．

B君：本当に，反応経路がいくつかあったりしたらすごい発見だよね．

Aさん：私も，これからいろんなデータを扱うと思うので，肝に銘じます．

X教授：君たちの先生が，こういう意味で「データを冷静に見ろ」と言ったのかどうかは分からないが，いい教訓が得られたと思っていいんじゃないかな．回帰分析をしていいのかどうかは，いろいろな可能性を考えて決断すべきだね．

B君：分かったような気がします．

Aさん：一つ疑問があるんですが．図2にある決定係数ですが，0.9695と，非常に大きいので，この値だけ見ると，単回帰でいいように思えるんですが，どう考えればいいんですか．

X教授：では，決定係数について勉強しようか．

決定係数

X教授：前回[3]，相関係数を勉強した時に，決定係数は単に相関係数の2乗と考えてはいけないという話をしたね．

Aさん：覚えています．

B君：でも，よく相関係数の2乗と説明されてますよ．

X教授：そうだね．決定係数の計算法から見ていこうか．これも前回言ったんだが，決定係数にはハッキリと合意された定義が定まっていないんだ[4]．一般的によく使われる定義[5]は

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}$$

だ．ここで，y_iはi番目のデータ（目的変数），f_iはi番目の回帰式による推定値，$\bar{y}$はy_iの平均値だ．

単に相関係数の2乗じゃないだろう．

Aさん，B君：確かにそうですね．

X教授：決定係数は，回帰式の当てはまりの良さを示す数値なんだ．本当は回帰式が間違っていても，たまたま，決定係数が大きくなるケースもあるということだ．この例を見てもらえれば，よくわかると思う（図6）．どうかね．

B君：こんなデータもよく見ます．直線で表すように言われますけど．

X教授：では，線形回帰してみよう（図7）．

どうだい，決定係数を見ると結構，合っているね．

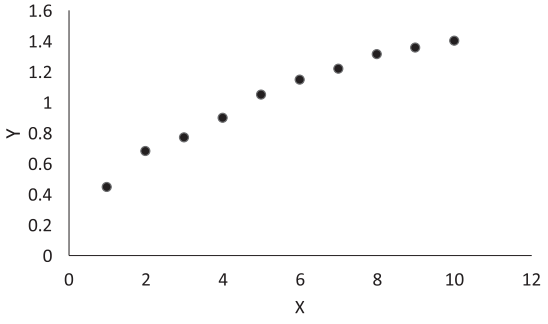

図6．線形回帰できない例

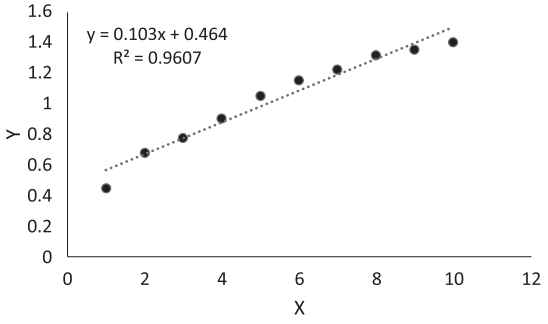

図7．線形回帰できない例（線形回帰）

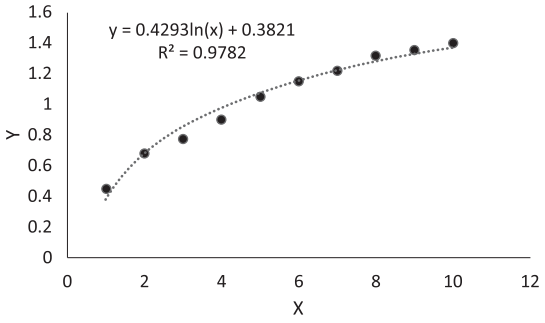

図8．図7を対数回帰したケース

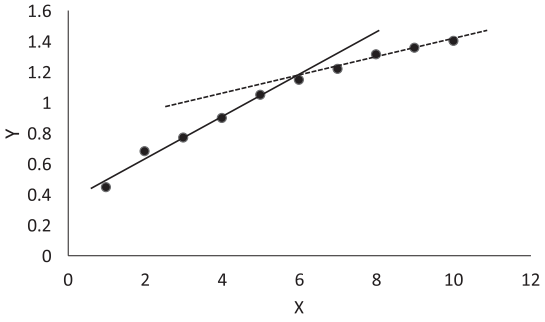

図9．図7を2直線で線形回帰したケース

ところが，データの点を見ると，両端は線の下にあるし，真ん中は線の上だ．こんなにある一定の傾向があるのはおかしいと思わないかね．

Aさん：そう言えばそうですね．

X教授：これも気を付けないといけない例なんだ．試しに線形回帰以外の回帰をしてみると（図8）．

Aさん：決定係数はあまり変わりませんが，こっちの方がよく合っている感じです．

X教授：この図も見てもらおうかな（図9）．

B君：これも，よく合っています．

X教授：図7の例は，少しずれを強調して分かりやすくしてあるんだが，単回帰分析＝線形回帰という考えを持たないということが大事だ．さっきも言ったようにもっと大事なことは，どのような回帰をするかということは，データのプロットの傾向ではなく，科学的な根拠に基づいてということが大事なんだ．もし，何かよりどころになる理論がないときは，いろいろな可能性を考えて，最初から一つに限定しないことだね．

B君：よくわかりました．

信頼区間

Aさん：他に気を付けないといけないことはありませんか？

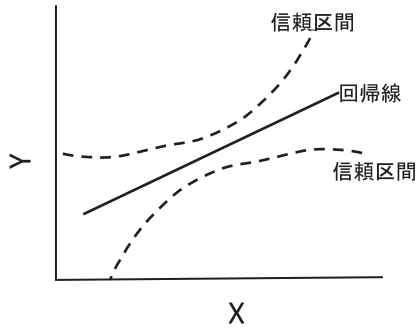

図10．線形単回帰分析の信頼区間

X教授：では，この図を見てもらおうか（図10）．

これは，分かりやすく極端に強調した模式図だが，線形単回帰分析の信頼区間を表しているんだ．回帰線の中央部は信頼区間の幅が狭く，回帰線の信頼度が高いが，端になるほど信頼区間の幅が広くなり信頼度も落ちるんだ．この点も，理解しておく必要があるね．

Aさん：そういえば，助手の先生から検量線を作成する時に似たような指摘を受けたことがあります．検量線は端のほうの精度が低いので，なるだけ濃度幅を広くとって実際の測定値が真ん中のほうに来るのがいい，と教わりました．

X教授：そのとおり，参考書を読むともっと詳しくなるよ[6,7]．それと，生物系だからといって，物理や化学もおろそかにしないようにね．

Aさん：はい．

B君：回帰分析なんて簡単だと思っていましたが，侮れませんね．

文　献

1) 回帰分析の原理については統計学のテキストを参照していただきたい．単回帰分析に関しては，RよりもExcelの近似線の挿入を使う方が簡単である．
2) 川瀬雅也，松田史生：生物工学，**95**, 96 (2017).
3) 川瀬雅也，松田史生：生物工学，**95**, 494 (2017).
4) Kvalseth, T. O.: *Am. Stat.*, **39**, 279 (1985).
5) Wikipedia「決定係数」: https://ja.wikipedia.org/wiki/%E6%B1%BA%E5%AE%9A%E4%BF%82%E6%95%B0 (2016/8/26)
6) Miller, J. N., Miller, J. C. 著, 宗森　信, 佐藤寿邦訳: データのとり方とまとめ方―分析化学のための統計学とケモメトリックス，共立出版 (2004).
7) 佐和隆光：回帰分析，朝倉書店 (1979).

（【第11回】は95巻12号に掲載予定です）