間違いから学ぶ

実践統計解析【最終回】

直交表と重回帰分析

川瀬　雅也[1]*・松田　史生[2]

直交表

Aさん：あの～，一つ いいですか.

X教授：何かな.

Aさん：統計処理に使うデータのとり方について困っているんです．私が見つけた環境汚染物質の分解菌を使って，分解効率の上がる培養条件を探し，重要な培養パラメータを見つけたいんです.

X教授：確かに，統計処理で目的は果たせそうだね.

Aさん：それでC先生に相談したら，温度，pH，炭素源濃度，窒素源濃度を今の条件（M）からふった実験で調べろって言うんです．それも生物工学会の要旨締め切りまでに．少し高め（H）と少し低め（L）っていう大雑把な感じにしても，全部で3^4 = 81通りの組合せがあって，とても間に合いそうにないんです.

B君：それはこないだ言ったみたいに，まず，温度だけふって決めて，温度はこれが最適として，他の条件も順に同じようにして決めていけばいいんじゃない.

Aさん：C先生は，それだと本当の最適条件を出しているかどうかわからないから，組合せで考えようって言うんです．個別に最適化すると（温度，pH，炭素源濃度，窒素源濃度）=（M，L，L，H）となったとしても，組合せを工夫して実験すると，本当の最適条件は（L，L，L，H）になるかもしれないということなんです.

X教授：なるほど，確かにC先生はいいところをついている．1回の実験にどの程度の時間が掛かるか，まったく考慮していないのも彼らしいね.

Aさん：笑い事じゃないですよ．生物工学会で発表は絶対したいですけど，でも，これだと要旨提出までお休みにデートもできないです.

X教授：…….「直交表」っていうのがあるんだけど，まずはそのお相手に教えてもらったら？

Aさん：もう相談したんですけど頼りにならないんです.

X教授：ふーん．えっと，今の工業製品は非常に複雑化しているが，開発期間は短くなっていると思わないかね．スマホがいい例だと思うが.

Aさん：確かにそうですね．でも急いで開発した結果，かえって危険性が増した例もありますね.

X教授：その通りだ．複雑化するほど，検討すべきパラメータは増えてくる．ところが，C先生から言われたように，多くのパラメータをすべての組合せで検討する（多次元配置実験という）のは時間的にも，コスト的にも現実的ではなくなってきているんだ.

Aさん：確かに.

X教授：そこで，各パラメータが独立していると考えていい場合（相互の関連性があっても，非常に小さい）は，この関連性（統計的には交互作用という）の評価を外して最小限の組合せの評価を行えばいいことになる．この組合せを見つけるときに使うのが直交表なんだ.

Aさん：すごい方法があるんですね．直交表が載っている参考書はありますか.

X教授：実験計画法の教科書には載っているには載っているが，自分の条件を当てはめて使うには，慣れないと結構難しいよ.

Aさん：へぇ～.

X教授：そう，がっかりしなくてもいいよ．自分の条件に合わせた直交表を作ればいいから.

Aさん：どうやって作るんですか．教えてください．時間がないんです.

X教授：分かったから，落ち着いて．これまで使ってきたRで作ることができるんだ.

Aさん：よかった.

X教授：Rで直交表を作る方法は二つある．DoE.baseというパッケージで直交表を作ることができる．まず，パッケージのインストールだね.

```
> install.packages("DoE.base")
```

ダウンロード先を聞かれると思うから，自分のいる場所の近くのミラーサイトを選べばいい．日本国内ならTokyoだろうね．次に，パッケージの読み込みだ.

```
> library(DoE.base)
```

いくつかエラーが出てくるが気にしないで進めるとい

*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）

い．データを入力しよう．

```
> Temp=c(“L”,“M”,“H”)
> pH=c(“L”,“M”,“H”)
> Csource=c(“L”,“M”,“H”)
> Nsource=c(“L”,“M”,“H”)
```

では，直交表を作ろうか．

```
>oaTable<-oa.design(factor.names=list(T=Temp,pH=pH,C=Csource,N=Nsource),seed=1)
```

どんな直交表ができたか確認してみよう

```
> print(oaTable)
```

表1．3水準の直交表

	T	pH	C	N
1	L	H	M	H
2	H	H	H	L
3	M	M	M	L
4	M	H	L	M
5	L	M	H	M
6	M	L	H	H
7	H	M	L	H
8	H	L	M	M
9	L	L	L	L

全部で9通りの組合せいいみたいだね．

Aさん：9通りなら，要旨の締め切りまでに実験ができます．

B君：でも，本当にこれだけでいいんですか．

X教授：いいところに気が付いたね．さっきも言ったように，あくまでも直交表の組合せは，各パラメータが独立か，交互作用があっても小さいと考えていい場合の最小限の組合せだ．この結果を見てもわかるように，最小限の組合せの評価だから，当然，これだけで正解が出るわけではないんだ．

Aさん：ちょっと，がっかりです．

X教授：よく考えてほしいんだが，81通りの実験をすべて行うのではなく，まず，9通りの実験を行い，最適条件の可能性のある条件を見つけ，その周辺を詳しく調べるという計画で進めることができると思うんだ．

それに，回帰分析の勉強をせっかくしたんだから，9通りの実験結果から回帰式を見つけて最大値になるパラメータの組合せをシミュレーションすることもできると思うよ．ただし，この方法は結構難しいけどね．

B君：確かにそうですね．どんな回帰式に当てはめるのがいいか考えないといけませんね．

X教授：そうだね．まあ，最初に試すのは重回帰分析だろうね．

重回帰分析

X教授：これまでに，重回帰分析を使っている論文を見たことはあるかな．

Aさん：はい，線形重回帰分析を使っているものです．

X教授：では線形重回帰分析に絞って，その基礎から話そうか．

線形重回帰分析は，一般式として

$$y = b + \sum_i a_i x_i$$

と表させるような関係があると考えていい場合に使われ，定数bと各変数の係数a_iを求める分析ということは分かっているね．当てはまりの良さは，決定係数で表されることもいいね．

Aさん：もちろんです．

X教授：yを目的変数（目的変量），x_iを説明変数（説明変量）といい，各変数の係数a_iを偏回帰係数というんだ．

具体的に次のようなデモデータを使ってみよう（表2）．データ数は9個で，説明変数が4個だ（mlr.csvというファイル名，学会HPからダウンロードできる）．

このデータで重回帰分析を行ってみよう．まずは，“data”というデータフレームにこのファイルを読み込んでみようか．

```
> data <- read.csv(“mlr.csv”,header=T,row.names=1)
```

表2．重回帰分析用のデモデータ

ID	x1	x2	x3	x4	y
1.0	1.2	14.5	2.2	21.1	4.7
2.0	1.3	15.1	2.4	17.9	4.9
3.0	1.3	16.0	2.3	18.5	5.8
4.0	1.2	15.2	2.1	16.9	5.6
5.0	1.1	13.9	1.9	19.1	5.4
6.0	1.4	14.6	2.5	20.7	4.4
7.0	1.3	14.9	2.3	18.1	4.6
8.0	1.2	14.5	2.1	22.5	4.8
9.0	1.3	15.3	2.4	19.2	5.1

次に，lmという関数で重回帰分析を行う．

```
> re1 <- lm(y~x1+x2+x3+x4,data=data)
```

lmの括弧の最初の項は，回帰式の形を示しているんだ．ただし，定数は省略している．y～としてもいい．2項目は使用するデータ名になる．結果がre1に入っているので，その要約を見ると

```
> summary(re1)
Coefficients:
             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  1.41487    3.64139   0.389    0.7174
x1          -0.40375    4.17441  -0.097    0.9276
x2           0.63573    0.22088   2.878    0.0451 *
x3          -1.97019    1.95875  -1.006    0.3714
x4          -0.04748    0.06462  -0.735    0.5032
Multiple R-squared:  0.8343, 見かけのR2値
Adjusted R-squared:  0.6687  調整済みのR2値
```

という結果が得られた．Estimateというのが係数なので

$$y = -0.403\,x1 + 0.636\,x2 - 1.970\,x3 - 0.047\,x4 + 1.414$$

という回帰式が得られたことになる．

B君：係数の大きさから重要な変数を見つけることができるんですよね．この場合x3ですね．

X教授：単に偏回帰係数の大小から重要変数を導くのは間違いのもとだ．各変数で単位もばらつきも違うからね．こういう場合は偏回帰係数を標準化して比べないといけないんだ．標準化の方法は[1]，各データの平均と分散がそれぞれ0と1になるようにするんだ．yの標準偏差をs_y，x_iの標準偏差をs_iとすると，標準偏回帰係数$a_i' = \frac{s_i}{s_y} a_i$となる．一番手っ取り早いのはデータそのものをscale関数で標準化してしまうといい．

```
> data_scale <-data.frame(scale(data))
```

として，標準化済みのデータをdata_scaleとして，重回帰分析を行う．

```
> re1_scale <- lm(y~x1+x2+x3+x4,data_scale)
> summary(re1_scale)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.741e-16  1.919e-01   0.000   1.0000
x1         -7.465e-02  7.718e-01  -0.097   0.9276
x2          8.037e-01  2.792e-01   2.878   0.0451 *
x3         -7.758e-01  7.713e-01  -1.006   0.3714
x4         -1.770e-01  2.408e-01  -0.735   0.5032
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Multiple R-squared:  0.8343, 見かけのR2値
Adjusted R-squared:  0.6687  調整済みのR2値
```

$$y = -0.074\,x1 + 0.803\,x2 - 0.776\,x3 - 0.177\,x4$$

Aさん：x2が一番大事な変数みたいですね．B先輩はやっぱり大事なものを見落とすのが得意ですよね．

X教授：それから，重回帰分析の結果の4つ目Pr(>|t|)が，「係数が0である」という帰無仮説に対するP値なんだ．x2は*がついているけど，これは優位水準αを0.05にしたとき帰無仮説が棄却できる．つまり，「係数が0ではない」という対立仮説を採用できます．という意味だね．

Aさん：となると，x2は説明変数として重要である．といえるわけですね．

X教授：さらに，重回帰分析を行うときに注意しないといけない事項があるから説明しておこう．

Aさん：説明変数の数でしたっけ？説明変数の数が多くなればなるほど，見かけ上当てはまりのいいモデルになるオーバーフィッティングが起きるんですよね．

B君：説明変数の数＜データの数でしたっけ？どういう風に考えたらいいんでしょうか？

X教授：ハッキリとした基準はないが，まずは，説明変数は少なければ少ないほどいい．そこで，説明変数間の相関を見て，相関が高いペアがあったらどちらかを除いてみよう．というもの重回帰分析では多重共線性に注意する必要がある．これは，説明変数の中に非常に相関関係の高い変数の組合せがある場合に起こる．たとえば，

```
> cor(data)
```

とすると，変数x1とx3の相関係数が0.96424405と非常に大きいので，根拠はないけどx1を削除して，

```
> re2 <- lm(y~x2+x3+x4,data= data_scale)
> summary(re2)
Multiple R-squared:  0.8339,
```

Adjusted R-squared: 0.7343

とすると調整済みのR^2はむしろ増加する．
さらに，根拠はないけど係数の小さなx4も削除すると

```
> re3 <- lm(y~x2+x3, data_scale)
> summary(re3)
Multiple R-squared:  0.8115,
Adjusted R-squared:  0.7487
```

とまたまた調整済みのR^2は増加する．

Aさん：変数を選ぶ客観的な基準がほしいですね．

X教授：そこで，一般には赤池情報量基準（AIC）を最低にするようなモデルがいいとされている．

B君：AICって，初めて聞きました．

Aさん：私も．どういう量なんですか．

X教授：AICはモデルの複雑さと，データとモデル（ここでは回帰式）の適合度のバランスを取る量だと言われているんだ．たとえば，今回のように，説明変数が多くなれば，決定係数は大きくなり，適合度がよくなってくるが，偶然に生じた誤差までもうまく取り込んでしまって，本当に減少を説明できているのかが怪しくなっているんだ．そこで，モデルの本当の適合度を評価するためAICが利用されるんだね．AICは次の式で計算できる．

$$AIC = n\left[\ln\left(2\pi\frac{Q^2}{n}\right)+1\right]+2(p+2)$$

nはデータ数，pは説明変数の数，Q^2は残差平方和だ．他の定義式もあるが，これがよく使われていると思う．

Aさん：単純に決定係数だけを頼りにすると，大変な間違いを起こすんですね．

X教授：Rでは，適切な変数を選ぶ方法があるので，その方法を教えようか．今回の例では，どの説明変数を使うのがいいのかを調べようとすると全部で17通りの組合せを調べないといけない（各自で確認していただきたい）．そこで，

```
> re1_scale <- lm(y~x1+x2+x3+x4,data_scale)
> re_step <− step(re1_scale)
```

AICが最小になるように，適切な説明変数を選択してくれるんだ．
実際にやってみるとわかるけど，スタートのy ~ x1 + x2 + x3 + x4のAIC=-7.24から，順番にAICが低下するように変数を減らし，y ~ x2 + x3のときがAIC=-10.08で最低になるという結果が出ている．
この時の回帰式は

```
> summary(re_step)
Call:
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.720e-16  1.671e-01   0.000  1.00000
x2           9.104e-01  2.049e-01   4.444  0.00436 **
x3          -8.938e-01  2.049e-01  -4.363  0.00475 **
```

$$y = 0.910\ x2 - 0.894\ x3$$

となる．

Aさん：AICからこの変数の組合せがベストである，といえる．さらに，x2, x3の標準偏回帰係数は両方とも統計的に有意であり，標準偏回帰係数の大きさがほぼ同じことから，x2, x3の寄与はほぼ等しい．つまりどちらも同じくらい重要だ，といえるわけですね．よくわかりました．要旨の締め切りに間に合いそうです．

B君：それはよかったね．こっちはぜんぜん間に合いそうにないよ．

Aさん：なにを言ってるんですか．先輩も学会にむけて頑張るって言ってたじゃないですか．

X教授：そうだ，忘れないうちに行っておくけど，来月から1年間，海外に調査に出るんだ．この続きは，戻ってからにしよう．

Aさん：いらっしゃらない間に，どうしても聞きたいことができたらどうすればいいんですか．

X教授：十分基礎はできているから，人に頼らず，自分で考えてごらん．オーバーフィッティングしすぎたときとか，どうしても困ったときのためにメールの連絡先は教えておくから．

参考文献

1) 向井文雄：生物統計学，p. 150，化学同人 (2011).

今回で連載は終了します．皆様のご意見を，是非，お寄せください．メールアドレスは本稿1ページ目のfootnoteにあります．