回帰モデリングの試み，R 言語によるデータ分析

――アクティブラーニングのケーススタディ――

ザン　ピン*・高田　正之**

要　約

データ分析は，文系・理系に関係なく，学生にとって必要な基礎力の一つである。しかし，様々な理由で，文系の学生に対応する典型的なデータ分析の習得について，十分な機会を与えられていないのが実情であり，珍しいことではない。

この教育実践は，２コマ前後という短い学習時間の後，受講学生全員に，「テーマは自由，実際のデータから，ある程度の相関係数値を持った面白いストーリーを作る」という課題に挑戦させたものである。

やや無謀に見えるが，面白い分析結果を得たと同時に，予測もしなかった現象が新たに起こった(1)。

キーワード：R 言語，回帰，数学モデリング，論理的な思考と解釈

1. はじめに

データ分析は，文系・理系に関係なく，学生にとって必要な基礎力の一つである。しかし，そもそも学生が基礎力に欠けていて長時間の理系的な学習に向いていないうえ，教える側の様々な事情もあり，典型的な統計(2)(3)分析法による習得については，文系学生に対して十分な機会を与えられていないのが実情であり，それは決して珍しいことではない。

ここで実施された教育実践の特徴は，長い時間をかけて，所謂基礎知識とテクニックに特化した学習を避け，短時間の学習の後，実践的なデータ分析に入り，主に以下のことを体得する。

- ➢ 人間の思考の曖昧さ・柔軟さと数学モデリングの間のギャップを認識する。
- ➢ 実際のデータを収集し，データの不完全さを認識する。
- ➢ 明らかではない，高い相関係数を得る難しさと面白さを体験する。
- ➢ テーマの選定・データ・分析・結果に基づいた理由づけと結論のプロセス及びアプローチ
- ➢ それに加えて，データ分析に対する姿勢として，正しく誠実であるべき点を強調する。
- ➢ 入手のしやすいデータだけに頼って安易な判断をくださない
- ➢ 統計結果を過剰に解説しない

テーマを自由とした目的は，答えのない課題は，実に面白いことである。つまり，テーマとデータ収集の模索過程の中，壁にぶつかり，その都度，新たな認識を得て，その繰り返しの中，理解と思考が深められる。『ここで，文系の適応主義や機能主義，つまり時間，環境の設定等のスパンを操作すれば，何でも説明「できちゃう」』(4)が通用しないところを体験する。

2016 年 11 月 30 日受付

* 江戸川大学 情報文化学科教授　数理計画
** 江戸川大学 情報文化学科教授　情報工学，ソフトウェア

2. 授業概要

2.1 授業概要

担当する授業は，2 年次の情報文化演習・実習という必修科目である。通年科目だが，学生の希望によるコースの選択や，複数の教員による割り当て，学科の方針など，年度により，実際に担当するコマ数は変動する。平成 27 年度は 7 コマ(データ分析以外の内容も少し取り入れた)，1 クラス，十数名の受講生だったが，平成 28 年度は 4 コマ，3 クラス（前期 1 クラス，後期 2 クラス)，合計 80 名弱になる。

2.2 R 言語とその学習

フリーの統計ソフトはさまざまな道のりを経て，ようやく R という形にたどり着いた。今，その知名度は統計有償ソフト SPSS，SAS に匹敵し，信頼性も高まり，教育，研究，企業で幅広く応用されている[(5)～(9)]。

R のスクリプト言語の利便的な操作性から，一般のプログラミング言語学習と違って，分析方法を限定すれば，短時間の入門練習に応用できる。

今回利用する R に関する実習内容は以下のように簡潔にまとめてある。

（ｉ）画面を説明した後，数個の変数，計算を練習する。

```
x <-（150，157，160，168，172）
summary（x）
```

（ⅱ）続いて，作業ディレクトリの変更，Excel のデータファイル（sample.csv）の用意・読み込み及び確認。

```
x <- read.csv（“sample.csv”）
head（x）
```

（ⅲ）単回帰（身長 ht と体重 wt）のひな型：散布図を描く，回帰（線形モデル）のパラメータを求め，回帰直線を散布図に追加，相関係数を示す。

```
plot（x$ht，x$wt）
reg <- lm（x$wt~x$ht）
abline（reg）
cor（x$ht，x$wt）
```

上に示した通り，実際に打つコマンドが限られている。もちろん，以上の内容に限定すれば，Excel でも可能だが，言語による分析は，文系の学生にとって，一つの刺激でもある。

初めて回帰分析を学習する場合，これだけでは理解には不十分である。Excel で簡単なデータを入力し，一部のデータを変更させ，それに連動する散布図・回帰直線・相関係数の様子も練習に付け加えた。また，各自で自分の身長を入力し，回帰パラメータによって計算される予測体重と実際の体重を比較することで，回帰の意味を理解させる。

時間の余裕があれば，不確実さ・ばらつき・分布など，いくつかの簡単な例を図で説明･質問し，ディスカッションしてもらった。さらに，Wikipedia で示された正規分布図を見せて，分布図の形状と標準偏差（σ）の関係も説明した。この授業の最後に書かせた授業の感想・印象および要望において，多くの学生は，「はじめて）正規分布の見方が分かった。ありがとうございます。これからも頑張るので，よろしくお願い致します。」と書いてくれた，中には「密度のある授業で，これからは一秒も気を逸らさない，頑張ります」と書いた学生もいた。

（それらを受けて，一部のクラスには，正規分布と回帰分析の関係，最小二乗法の基本的な考え方も，その次の授業で少し説明した。）

2.3 モデリングの練習

文系の学生は，数式を使って，予測するという学習経験が少ない。そのため，授業にはできるだけ身近なテーマを取り入れた。

まず，

明日の気温を予測する

という例を挙げ，受講者全員にディスカッションしてもらう。その際，「課題の意味が理解でき

ない，我々は何をすればよいのか」と，ある学生から逆に問いかけられた。そこで，「明日の気温について，前例に出した体重，身長の位置にベターと思われる要因をあてはめて考えてみる」と説明した。

学生たちは，湿度，気圧，晴れ，日差し，経度，今月の気温，去年の同じ日の気温等，次々と例を挙げていたが，やはり，未経験のため，ローカルな要因，あるいは，通常の天気予報等の情報に影響され，シンプルかつ有効な数式モデルに不可欠なベター要因とその表現に直ぐにたどり着けないケースがある。そのような議論後，学生たちに対し，季節を表現する平均気温や現在の気温情報を補足した。

また，モデリングについて，

明日の 気温　<=　（例年の）平均気温…

とホワイトボードに書き出し，印が付けられている気温について,「もっと限定したほうがよい」と付け加えた。これについて，直ぐに正しく答えられる学生が少ない。課題文の「気温」は表現として曖昧であり，実際には最高気温，最低気温，平均気温しかデータがない。

次に，気象庁が公開している過去データ，「地点」，関連の「項目」と「期間」を選択し，CSV ファイルをダウンロードして見せた。ダウンロードされたデータファイルにも,「気温」については，もちろん,「最高気温」,「最低気温」,「平均気温」という項目しかない。

具体的に計算する際，データには様々な欠損，不備があり, 書式を含む加工も必要な場合がある。気象データを R に読み込む前に, 不要な（先頭の）行・列の削除，フィールド（項目）名を処理しやすいように変更するなど，いろいろな処理が必要である。

上の例の平均気温を算出するよりも，さらに簡潔な回帰モデルを全員に練習させた。例えば,［直近 1 ヵ月］（2016/6/3-2016/7/3）という期間を選択し，気温差（最高気温 - 最低気温，diff）という新たな列を加えて，日照時間（sun）との散布図を描くと，比較的強い相関（この例の場合，0.910。6 月以外には，一般的に 0.9 を超えない）が得られる（図 1）。

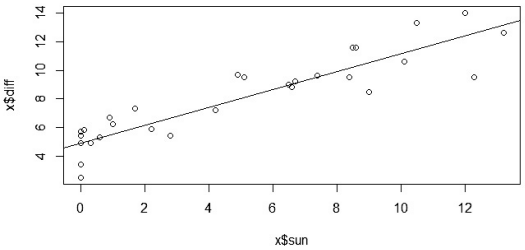

図 1　日照時間と気温差

2.4　課題

約 2 コマの学習・練習後，課題の時間になる。与えられた課題は :

「テーマは自由，実際のデータから，当たり前ではない，ある程度の相関係数値を持つ面白いストーリーを作る」というものである。

課題の補足説明として，データの出典を明確に記入すると同時に，誠実な姿勢で臨むことを繰り返し強調した。データの収集や分析の他に，統計結果に対する正しい解釈も次のようにつけ加えた。

『In many areas of scientific research, it is important to choose among various explanations of data observed.

This type of argument is a special case of Occam's Razor, a general principle governing scientific research, weighting possible explanations by their complexity…. William of Occam said Explanations should not be multiplied beyond necessity. In the end, we choose the simplest explanation that is consistent with the data observed』[10], pp488-490。

2.5　課題へのフォロー

課題を出した後，学生たちは放置されていると感じ，テーマを模索し始める。ここでは，たっぷり一コマを使って，自由に検討，相談，ディスカッションしてもらうが, ここでの学生の質問から，いろいろ気づかされたことがある。

(1) 自分が興味のあるものから探し，(少なくとも) 二組のデータが不可欠という認識が薄れている。
血液型や同性愛の許容度等である。しかしその後，国別の許容度のデータは得られたが，同性愛カップルの数が得られないため，このテーマの分析を断念せざるを得なかった。
(2) 二組データの関係について，何なのかがまだ掴めていない。
同じ「対象」の二つの属性ではなく，上位ランキングのデータを比較して回帰する。この間違いに最初に気づかされたのは，喫煙率と長寿の関係である。平均の喫煙率と寿命になぜか強い正の関係が得られていると不思議に思い，その後，複数の学生が似たような間違いをしたことが分かった。これはまったく想像していなかったケースである。

『まさに，多くの学的知識とは，眼鏡（理論）が先にあってものが見えるという「科学知」であろう。むしろ大学生にとって，突き動かされるアクチュアルな体験が先にある』[11]。想像が溢れると同時に，ロジック的にズレも生じる。

学習の初期段階に，「実在する何かが測定されたのか，それとも測定によってその何かが存在するということになったのか」[4]。授業が終わりに近づき，まだテーマを決めてない学生も複数おり，助けを求めてくるが，安易にアドバイスをせず，「新しいものを生み出す苦しみを味わうのが目的」と説明した。

個別の学生が授業中に，レポートの作業をしない。最初は学習の態度に問題あると思ったが，その後，自力でできないのが主な要因と分かった，個別的に補講を複数実行した。

3. 学生のリポートの分析

学生たちは様々なテーマを取り上げた。学生がテーマを選択するには，何かしらの理由があり，例えば身長と体重のデータに取り組むのは，前例の練習問題と似ているだけではなく，自分の身長が低いためであったりする。その他，人口と火事，落雷と死者（ただ，一部のレポートは相関係数を記録してなかった），男女別の浮気・未婚，カラオケ人数とルーム数など，ユニークな例も複数あった（表 1）。

表 1　多様な事例

要素	相関係数
睡眠時間と成績	0.100
標高と寿命	0.254
大学進学率と短大進学率 ***	-0864
一人当たりの収入とゴミの量	0.426
アルコール消費量と寿命	-0.229

*** 高校の卒業生総人数が一定のため，この二つのデータは従属の関係がある。ただ，就職人数もあり，また，学生にはこのように関係のある二組のデータを取り入れるのが望ましくないのは説明した。

スポーツに関連するデータに取り組む男子学生が数名いた（表 2）。これらの複数のレポートのデータを合わせてみると，やはり，直観と一致し，サッカーの得点 vs 勝ち数に関しては野球よりもっと強い相関があることが数量で評価できた。

表 2　スポーツ関係の例

要素	相関係数
サッカー J1 得点，勝ち数	0.871
サッカー J1 失点，勝ち数	-0.834
野球の得点，勝ち数	0.436

一部は練習問題の気象データからの影響で，日本特有の地理と気象データを取り上げた学生も複数いる（表 3）。ほぼ毎回，一番直截的な最高気温と最低気温の相関を求める学生がいる（最初に「練習問題とは別のテーマを探す」ように勧めている」）。

表 3　地理・気象の例

要素	相関係数
河口湖の水位，降水量	-0.291
海水温，台風数	-0.131
桜島の噴火数，地震の数	0.981
地震の回数，大きさ	0.930
最高気温，最低気温	0.925

短時間で，良いテーマを選択し，かつ説明できるデータに辿り着くことは容易ではないが，学生の中にも，それができた例も複数見られた（表 4）。

表 4　高い相関係数の社会問題の例

要素	相関係数
男女の年収の差，年齢（勤続年数）	0.873
寿命，喫煙率	-0.939
オンラインゲームジャンル男女利用率	-0.992
男女別の未婚率	0.98
年収と偏差値	0.691

表 4 の「男女の年収の差，年齢（勤続年数）」の例の場合，定年して 60 歳以上のデータを除けば，相関係数が 0.976 となり，データ数が少ないものの，相関が高い，非常に社会性の高いテーマである（図 2）。(https://doda.jp/guide/heikin/2014/age/)。

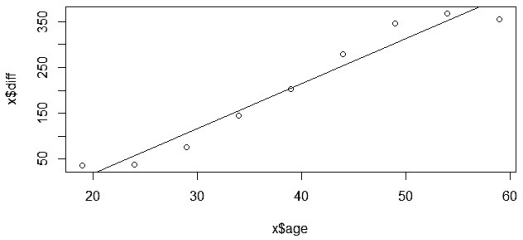

図 2　年齢（勤続年数）と年収の差

4. 考察と今後の課題

2 節，3 節で述べたように，短時間の学習で，文系の学生にもデータ分析について一定の成果を得られることが実証された。また，前期の最後のグループ発表会で，複数の学生の回帰分析についての発表は簡潔・適切であり，回帰分析とその目的を良く理解していることが分かった。

この教育実践は，文系学生にデータ分析がどういうものかという，学生の興味を引き付ける試みに過ぎない。効果を得るには，体系的に継続的なシステムが必要である。また，学生の感想文からは学問について憧れも感じた。ただ，学生たちが自分で書いたように予習・復習を実行した者は少ない，それが現状のようである。

ソフトウェアが発展した今，データ分析，基礎力やテクニックよりも，自らの感性・感覚，ロジック，考える力及びデータ分析への誠実な姿勢が重要である。

所謂ビッグデータの時代，これらの課題は教育側に与えられた責任と使命だと思っている。

謝辞

この教育実践で課題を真剣に取り組んできた情報文化学科の学生たち，参考文献(1)の原稿をチェックして頂いた学術情報課高橋恵美さんに感謝致します。

参考文献および関連 URL

(1) ザンピン，「文系学生の回帰モデリングの試み，R 言語によるデータ分析」，平成 28 年度 ICT 利用による教育改善研究発表会，公益社団法人　私立大学情報教育協会，pp2-5，2016
(2) 南風原朝和，心理統計学の基礎，統合的な理解のために，有斐閣，2002
(3) 南風原朝和，続・心理統計学の基礎，統合的な理解を広げ深める，有斐閣，2014
(4) 北田暁大，岸政彦，社会学はどこからきて，どこへ行くのか，書斎の窓 **No.644**，pp64-69，2016
(5) 金明哲，R によるデータサイエンス，森北出版，2007
(6) W.Richert，L.P.Coelho，斉藤康毅（訳），実践機械学習システム，オライリー・ジャパン，2014
(7) 片岡厳，データサイエンス養成読本，技術評論社，2013
(8) 片岡厳，データサイエンス養成読本，R 活用編，技術評論社，2015
(9) 統計ソフト R の使い方，https://sites.google.com/site/webtextofr/home
(10) Thomas M. Cover（著），Joy A .Thomas，Elements of Information Theory，2nd，Wiley Interscience 2006，P488-490
(11) 中村雄二郎，臨床の知とは何か，岩波，1992