統計的データ解析の教育における S-PLUS の利用

文部科学省統計数理研究所
統計計算開発センター
田村　義保

統計数理研究所という講義の義務がない（公開講座をやる義務と総合研究大学院大学の博士課程の学生を指導する義務はある）研究所にいるにも関わらず、講義が非常に好きなので、何カ所かで非常勤講師として教えている。統計の基礎、データ解析、時系列解析を教えることが多い。下の表は、データ解析で教えようと思っていることのテーマと内容である。

回数	テーマ	内容
1	記述統計学1	グラフや表を作ること。ヒストグラム、累積相対度数、箱ひげ図
2	記述統計学2	数値による記述。標本の問題。全データを用いた方が正確。モードの詳しい求め方。
3	分散分析と計算の精度	平均との差の検定。分散分析と精度について。
4	回帰分析	単回帰分析と予測　確率論の必要性をいう
5	簡単な検定	確率論の必要性と情報量の必要性をいう
6	KL情報量とAIC 1	Fisher情報量、最尤法
7	KL情報量とAIC 2	KL情報量、AICについて
8	重回帰分析1	入門　モデル選択　分散分析表などの見方
9	重回帰分析2	残差分析　非線形解析　ダミー変数
10	数量化1類とロジット回帰	数量化1類の基本的な考え方とダミー変数の重回帰分析
11	判別分析と数量化2類	判別分析を中心とするが少しはニューラルネットにもふれる。数量化2類が質的データについての判別分析であることにふれ、質的データを解析することの重要性にもふれる
12	主成分分析1	入門的な話
13	主成分分析2・数量化3類	バイプロットの見方などについて。質的データについての主成分にあたる数量化3類も説明する。
14	クラスター分析	分類法の基本である階層クラスター分析にふれる。Kmeans法にもふれる。
15	生存時間分析	生存時間分析
16	時系列分析1	基本的なモデルと予測法
17	時系列分析2	季節性を持つデータと予測法
18	シミュレーション法	乱数の発生法、モンテカルロシミュレーション
18	調査データの解析	調査法や調査データ解析法について
20	クロス表の分析	質的データの解析でもっとも重要なクロス表の解析について

易しすぎると思われる方もいるかもしれないが、データ解析の入門としては、この程度あればよいと考えている。

S-PLUS をどのように活用しているかについて説明する。計算に使うのはもちろんのことであるが、図を書くために多用している。そのようなことは誰でもやっていると言われるであろうが、私は次のように工夫している。TeX を使って講義資料の用意をしているが、例えば、平均値の区間推定を教える場合は、原稿に次のような S-PLUS の関数を書き込んでいる。

```
% fig0306a.sgr
%> plot(seq(-3,3,0.01),dnorm(seq(-3,3,0.01),0,1),xlab="x",ylab="",
%+ main="標準正規分布",type="l")
%> arrows(qnorm(0.05,0,1),dnorm(qnorm(0.05,0,1),0,1),qnorm(0.05,0,1),0)
%> arrows(qnorm(0.95,0,1),dnorm(qnorm(0.95,0,1),0,1),qnorm(0.95,0,1),0)
%> abline(h=0)
```

このようにしておくと、どのような考えで図を書いたかが原稿を見直せば分かるし、マニュアルを見直す必要がかなり減ってくる。また、マニュアルを読んでもよく分からないことを試行錯誤で習得しようとするときはだいたい work01.txt のようなファイルを作って、やったことを全て残すようにしている。下記は、ロジットモデルの説明のために glm を使おうとしたがフレームデータの作り方が良く分からなかったので下記のような作業をした。

```
write.table(Orange,"d:/tmp/lec/lec03/taisho/orange.dat")
```

でフレームデータをファイルに落とす。ただし、１行目に変数名以外に最初に行名が入るために正しく使えない。これを消して、かつ、区切り記号「,」を空白にする。

```
 Tree age circumference
1 1   118   30
2 1   484   58
3 1   664   87

orange.dat <- read.table("d:/tmp/lec/lec03/taisho/orange.dat")
```

でフレームデータはできるが第一変数も数値である。「""」でくくっても数値のままであった。

```
orange.dat[,1] <- as.factor(orange.dat[,1])
```

を実行すると Splus の解析が使える。

他に、もっとスマートなやり方、正しいやり方があるとは思うが、マニュアルや解説書を読んでも分からなかったので力業に頼った訳である。

講義は好きであるが、統計学の講義を聴く気は全くなかったので、過去に一度も統計学を体系的に習ったことはない。物理学の中の統計物理が専門であり、非線形確率微分方程式を使った研究を大学院の時は行っていた。このような経歴を持っているために、マハラノビスの距離を使った判別分析を教えようとした時に、結果が参考にしている本と何故、違うのかに気づくのに 1 時間くらいかかった。この時、いろいろとデータをいじくるために S-PLUS を用いた。本来の意味とはずれるが、探索的データ解析が訳に立つことが分かった。どのようなことであったかと他の解析法の講義（講義の準備）に S-PLUS をどのように使っているかは講演時に説明する。