S-PLUSの有効利用

上田　太一郎

1．はじめに

S-PLUS（およびS言語）はオブジェクト指向データ解析ソフトウェアである．92年の夏，偶然S-PLUSを知った．データ解析の可能性が無限に広がる思いであった．

15年前コンピュータは汎用機が全盛の時，ソフトはまだハードの付録のようなものであった．ある調査会社から計算機システムの問い合わせがあった．計算機システムのリプレースを考えているが，条件として①SPSSをサポートできること②数量化理論をサポートできること③統計がわかるSEがサポートしてくれることであった．（当時としては珍しくソフトサービスが条件であった．）幸い条件をすべてクリアでき，（当時の計算機の商談は半年～1年かかったが，）わずか2ヶ月でほとんど他社にきまっていた1億円の計算機システムを逆転受注した．SPSSのおかげである．その後リプレース時期がきたときはWS（ワークステーション）とSASを提案した．昨年の秋，日本に上陸したVisualstatは安価でかつ使いやすいソフトと思う．今後，注目すべきソフトと考える．

さて，S-PLUSであるが，筆者の担当業務では，人事評価，データ（ソフトウェア試験に使用）作成，ソフトウェア信頼度成長モデルなどを求めるのに役に立った．統計解析手法（アルゴリズム）もいくつか考案できた．現在，ソフトウェアの生産性・品質向上あるいはクライアント／サーバシステムの受注前活動のためにデータ解析を行なっている（[1]～[5]，[8]，[9]）．

S-PLUSでは人間の思考をそのままS言語で表現すればよく，データを自在に生成・解析できる．S-PLUSにおいてはデータも関数もすべてオブジェクトとして扱う．何（オブジェクト）をどうするかを関数で示す．関数で処理した結果はオブジェクトであり，この結果をどうするかも関数で指示すればよい．

うえだ　たいちろう　三菱電機東部コンピュータシステム㈱
生産管理部　〒244 横浜市戸塚区川上町87-1

統計解析でよく使用される回帰分析を例にとれば，回帰式の良さを確認するための残差のヒストグラムを求めるには以下のようになる．データx（説明変数），y（目的変数）に回帰式をあてはめるにはlsfit(x, y)とする．残差を求めるにはlsfit(x, y)$resとする．残差のヒストグラムを求めるにはhist（lsfit(x, y)$res）とすればよい．この過程は図のようになる．

S言語はAT&Tベル研究所の研究者J. M. チェンバースほかによって開発された問題解決型の不定形なデータ解析およびグラフィックスに大変有用な諸機能を備え，さらにオブジェクト指向を積極的にとり入れたデータ解析用言語および環境である．S-PLUSはさらに統計機能・グラフィック機能を充実している[7]．

S-PLUSの特徴をもとめると以下のようになる．

①WS，パソコンをターゲットにした，数値演算，統計解析，データの視覚化，探索的データ解析の1000以上の関数をサポートするオブジェクト指向対話データ解析ソフトウェアである．

②Sは対話型インタープリタ言語であり文法はC言語に似ている．スカラ，ベクトル，マトリクスの違いを意識しないで使用できる．

③UNIX，MS／DOS，WINDOWSとのインタフェースがとられ使いよくなっている．S-PLUSからC，FORTRANのプログラムを呼ぶことができる．

④テキスト（ASCII）ファイルのデータの読み込み，書き込みが可能である．

⑤通常の統計解析，多変量統計解析の他，以下のような高度なデータ解析機能がサポートされている．

線形モデル
実験計画
一般化線形モデル
一般化加法モデル
局所回帰モデル
樹形モデル
非線形モデル

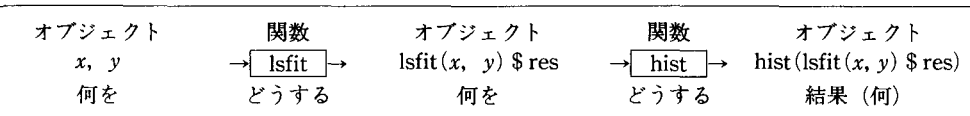

図1　オブジェクトと関数

2．利用例

2.1 “入社時に学力優秀デキルと限らず？”

－システムエンジニアリング企業における入社試験結果と数年後の能力評価との相関について－［3］

プログラム製作の生産性は個人差が大きく，またシステムの構築にあたっては，システムエンジニアリングの良否がシステムの出来不出来に関係すると言われている．そのため，有能な人材を採用時にどのようにして見分けるかということが課題としてある．すなわち入社試験の内容が将来の人材を見抜くものになっているかということである．そこで，入社試験の成績と入社後，何年か経ってからの能力評価との関係を調査・分析した．

能力評価制度では，「分野別専門技術」と「共通専門技術」および「業務遂行能力」の3つについて個々に評価している．「分野別専門技術」は設計・製作・試験技術など習得と訓練により培うもの，「共通専門技術」はソフトウェア開発にあたって必須なもの（プロジェクト管理，エンジニアリング，コンピュータの知識など），「業務遂行能力」は技術力ではなく業務を効果的に仕上げるための能力（責任感，コミュニケーション，リーダシップ）である．求める人材とは，これらを高い水準に維持していける者ということがいえる．

入社試験の内容は将来の人材を見分けることのできるものが望ましいが，これはなかなか難しい．そこで新卒採用にあたっては，いわゆる「良い学生」（成績が良く，性格が良い）をターゲットとしてきた．試験科目はプログラマ適性試験（適性，スピード，正確度）と「良い学生」を選別するための基礎学力（英語，数学，国語）である．これらにそれぞれ基準点を設け，これをクリアーした者を合格としている．この外に面接を行なっているが，面接評価は数値化していないので取り上げなかった．会社の発足時から能力評価の検討を行ない，昭和63年度から能力評価を行なってきた．

求める人材を能力評価が高い者ととらえ，この評価データと入社試験の成績データとの相関分析を行なった．S-PLUSの関数としては相関行列（cor），回帰分析（lsfit），正準相関分析（cancor）などを使用した．

結論として，日経新聞（93年12月28日夕刊）に載った見出し“入社時に学力優秀デキルと限らず？”，“発掘は面接で”，“適性試験は有用性アリ”のようになった．新聞の反響は大きかった．20社（あるいは人）弱から電話，手紙などによる問い合わせがあった．今後の課題として面接の評価に対しても同様に統計分析をしていきたいと考えている．また入社後の集合研修時の評価との相関も分析していきたい．

2.2 ソフトウェア（組合せ試験）のためのデータ作成［4］

某プロジェクトの支援をすることになった．膨大なデータを収集・加工・編集・出力するシステムの組合せ試験のためのデータを短期間（1ケ月）で作成する必要があった．統計1.5 MB（(300本)×(500コの数値／本)×(10バイト／数値)）のデータが簡単に作成でき，当初の計画に比べて5～10倍の効率向上が図れた．

以下はデータ作成のほんの数例である．わかりやすくするために簡単な事例とした．

①500から1までの数値を作成

data＜－500：1でdataというオブジェクトに500から1までの数値（整数）が作られる．（1：500とすれば1から500までの数値）

②30から120までの一様乱数を1000コ発生させ降順に並べる．

data 10＜－rev（sort（runif（1000，30，120）））

③（気温・雨量・SO_2のような）時系列データを5000コ発生させる．

timeseries＜－sar 4（5000）＊100

sar 4はユーザ関数で以下のようにして作った．

sar 4＜－function（x）{arima.sim（x, model＝list（ar＝c（1.2,－.8,.4,.15）））}データ（オブジェクト）を3.5インチフロッピーに書き込むにはdata 10ならwrite（data 10，“data 10”）

! copy data 10　a：

とすればよい．（注）! copyでパソコンのOS（MS-DOS）のコマンドを呼んでいる．

読者諸氏は上の課題に対してどう対処しますか？手作業でやるにはつらく根気がいる．CあるはFOR-

TRANでプログラムを作りますか？　上のようにさまざまな要求に対応するために，そのつどプログラムを修正・コンパイル・リンク・実行しますか？

2.3　ソフトウェア信頼度成長モデル

われわれソフトウェア生産に従事するものにとって大規模なソフトウェアプログラムのバグ収束時期およびその時の累積バグ数を予測することは切実な問題である．

ソフトウェアの累積バグ数をロジスティック，ゴンペルツおよび遅れS字に当てはめた実プロジェクトでの例である．$x=1, 2, \cdots\cdots, 11$ に対して $y=7, 23, 40, 61, 83, 99, 103, 125, 127, 146, 169$ であった．ロジスティックでは $y=175.86/(1+9.665\cdot\exp(-0.391x))$，ゴンペルツでは $y=194.4\exp(-3.06(0.788)^x)$，遅れS字では $y=181.62(1-(1+0.298x)\exp(-0.298x))$ となった．S-PLUSによる関数（遅れS字型曲線の係数を求める）は次のようになる．

```
function (x, y, p1, p2) {d<data. frame
(x=x, y=y)
parameters (d)<-list (p1=p1, p2=p2)
ms (~(p1 * (1-(1-p2 * x) * exp
(p2 * x))-y) ^2, d)}
```

ここで，x：1～n（n：カレンダータイムのポイント数），y：x に対応した累積バグ数，$p1$, $p2$：初期値

S-PLUSのユーザはこのコードを参考にして任意のソフトウェア信頼度成長モデルさらにより一般化した非線形回帰関数を作ることができる［4］．

2.4　重回帰分析変数選択基準

重回帰分析（数量化理論I類を含む）において最適な変数（数量化理論I類ではアイテム）を選択する問題は重要なテーマである．変数選択基準としてMallowsのCp（leaps），赤池のAIC，竹内のTIC，佐和の予測用修正重相関係数，芳賀ほかの自由度二重調整済重相関係数などが提案され実用に供されている．上の基準を含めた提案されている各種規準をS-PLUSで関数を作り数10ケースについて確認した．多少の相違はあるものの，ほとんど同じ変数の組を選択した．

変数選択規準には次のものがある．（竹内啓編「統計学辞典」を参照．規準値が大きくあるいは小さくなる変数の組合せを選択すればよい．＿＿は小さくなるほうがよいことを示す．）

AIC（赤池）：$n\log\sigma+2p(n\log(\Sigma e_i^2)+2p)$

TIC（竹内）：AIC$+p(p-2)/n$

MDL（Rissanen）：$(n/2)\log\sigma+(p/2)\log_2(n)$

佐和の予測用修正重相関係数：$1-(1-R^2)(n-2)(n-1)/\{(n-p-2)(n-p-1)\}$

竹内：$\sigma^2(n^2-n-p-2)/\{n(n-p-2)(n-p-1)\}$

芳賀ほかの自由度2重調整重相関係数：$1-(1-R^2)(n-1)(n+p+1)/\{(n-p-1)(n+1)\}$

Schwarz：$p\log n+n\log\sigma$

ここで，R：重相関係数，n：サンプル数，p：変数の個数，e_i：残差，σ：残差の標準偏差

2.5　分散分析表を求める万能関数と適用

筆者は統計解析手法のなかで重回帰分析，判別分析，そして実験計画法（および分散分析）が有用であると思い業務に生かしている．特に，ORマンは実験計画法をもっと活用することを強く希望する．実験計画法では人（ORマン）の知恵が発揮される．実験計画法は化学や医学などだけでなく企業の経営・営業支援などでもっと使用されてもよいと思う．

さて，実験計画法データを解析するには分散分析を行なう．分散分析は筆算あるいは電卓で可能であるが，電卓などによると計算の方ばかり目が向き，実験計画の理解がおろそかになるおそれがある．実験の計画（ORマンの知恵）に重点を置くべきであって，計算はコンピュータソフトウェアにやらせるのがベターであると考え，データと計画行列などを指定し分散分析表を求める万能関数を作成した．多元配置，直交表，擬因子法，分割法，直積法などによるデータはすべてこの関数が処理してくれる．［7］を読みながら延べ2週間でこの関数を作った．関数anovaallは以下のようになっている．わずか3ステップからなるこの関数は大変便利であり，愛用している．

```
anovaall<-function (x, y, ex)
{fac<-design (y)
data.fac<-data.frame (fac, x)
summary (aov (paste ("x~", ex), date.fac))}
```

ここで　x：データ，y：計画行列，ex：式　例　"A＋B"，"．"

適用例として，生産性向上のためにWSの導入効果データの統計的検定を紹介する．ソフトウェア製造のためのWSが旧式になってきたので新しいWSを導

入することを検討した．ある生産性向上を示す指標データ（大きい値ほどよい）をとった．

旧式 WS	20 33 31 29 27 30 26	平均 28.0
新 WS	41 29 38 28 40 39 45	平均 37.1

平均値を見ると向上しているように思われる．新WSが生産性向上に寄与しているかを統計的に検定した．2つの標本の平均値の差の有無は t 検定によってわかる．あるいはすこしおおげさかも知れないが1元配置データの分散分析によってもよい．

t 検定（t. test）は $ws1<-c$（20, 33, 31, 29, 27, 30, 26），$ws2<-c$(41, 29, 38, 28, 40, 39, 45) として t.test（$ws1$, $ws2$）とすればよい．危険率1％（p−value＝0.0079）で両者に差があることがわかる．分散分析（anovaall）では

anovaall（c($ws1$, $ws2$)，c(rep(1, 7)，rep(2, 7))，"."）と指定すると，次のような分散分析表が得られる．

分散分析表

	Df	Sum of Sq	Mean Sq	F Value	Pr (F)
A	1	292.571	292.571	10.122	0.0079
Residu	12	346.857	28.905		

分散分析表からも危険率1％(Pr(F)＝0.0079)で両者に差があることがわかる．検定結果を参考にし，その他導入効果などを検討し導入にふみ切った．

もう1つの適用例を紹介する．あるスーパーが行なった販促効果を実験計画法を用いて定量的に捉え売上の増加に成功した例である．そのスーパーでは牛乳，インスタントラーメン，清涼飲料の売上増加対策としてエンド陳列（有，無），値下げ（無，有），チラシなどによる広告（無，有）をとりあげた．

（（ ）内は水準）これらはスーパーでコントロールできる要因なので制御要因である．その他の要因として曜日（土，日，土日以外），時間帯（AM，PM，夜），天候（晴（曇を含む），雨）をとりあげた．割り付け表（要因の種類と，実験条件）は以下のようになった．

割り付け表

記号	A	B	C	D	E	G	データ(売上個数)(実測値を多少変更している)		
名称	エンド陳列	値下げ	広告	曜日	時間帯	天候			
NO							牛乳	インスタントラーメン	清涼飲料
1	有	無	無	土	AM	晴	33	19	40
2	有	有	無	日	AM	晴	26	19	41
3	有	無	有	土日以外	AM	晴	31	16	48
4	有	有	有	土	PM	晴	39	19	50
5	無	無	無	日	PM	雨	30	17	37
6	無	有	無	土日以外	PM	晴	37	21	41
7	無	無	有	土	夜	雨	30	12	12
8	無	有	有	日	夜	晴	44	10	42
9	有	無	無	土日以外	夜	雨	19	14	13
10	有	有	無	土	AM	雨	32	20	25
11	有	無	有	日	AM	雨	30	10	19
12	有	有	有	土日以外	AM	雨	38	22	33
13	無	無	無	土	PM	晴	29	16	20
14	無	有	無	日	PM	晴	38	12	30
15	無	無	有	土日以外	PM	雨	31	22	22
16	無	有	有	土	夜	晴	37	15	41

表の見方は例えば，NO.16はエンド陳列は無し，値下げ有り，広告有り，土曜日，時間帯は夜，天候は晴で，牛乳は37本，インスタントラーメンは15コ，清涼飲料は41本の売上だったことを示す．順序はNO.1～16をランダムにして実施された．1ヵ月でデータが集まった．分散分析表を求めてどの要因が販売個数を多くするのに効いているかを調べた．販売個数を多くする要因と水準は，牛乳では値下げ・広告有り，インスタントラーメンでは曜日（土日以外）・時間帯（PM），清涼飲料では値下げ・天候(晴)であることがわかった．同時に，効いていない要因もわかった．

（注）清涼飲料を例にとればanovaallの実行は以下のようになる．データは $y<-c$（40, 41, 48, 50, 37, 41, 12, 42, 13, 25, 19, 33, 20, 30, 22, 41)と定義する．計画行列は割り付け表で第1・2・3水準に対応して1，2，3とする．（これをdmatとする）そうするとanovaall（y, dmat,"."）で分散分析表が得られる．分散分析表をみるとB，G以外の要因は効いていないので誤差にプールする．プールするにはanovaall（y, dmat, "B+G"）と指定すればよい．

分散分析表

	Df	Sum of Sq	Mean Sq	F Value	Pr (F)
A	1	36.000	36.000	0.3313	0.5829
B	1	529.000	529.000	4.8680	0.0631
C	1	25.000	25.000	0.2301	0.6461
D	2	9.909	4.954	0.0456	0.9557
E	2	158.342	79.171	0.7286	0.5159
G	1	720.820	720.820	6.6332	0.0367
Res	7	760.679			

3．関数の作成

過去2年間で以下のような関数を作り業務に役立てている．

AICによる分割表解析，非線形回帰，アイテム・カテ

ゴリデータの作成，数量化理論III・IV類，ソフトウェア信頼性成長モデル曲線，回帰分析変数選択規準，判別分析変数選択規準，外れ値の簡易検出法，AR(p)モデルにおけるpの決定，分散分析，AHPモデル，BT（Bradley-Terry)モデル，スーパーなどにおける最適なレジ数決定法，商品の最適仕入れ個数決定法など

例えば，数量化理論Ⅰ～Ⅳ類は20年前，FORTRANで開発した経験がある．数量化理論Ⅳ類は500ステップ位だった．S-PLUSでは数ステップである．ソースを載せておくので参考にしていただきたい．

```
hayashi 4<function(x)          # 93-1-10 作成上田
 {eigen(contc(x))}
 contc<-function(x)            # e_ij要素行列を調整する
 {rsum<-apply(x,1,sum)         #行の和
 diagx<-diag(x)                #対角要素を取り出す
 diagy<-diagx-rsum             #対角要素－行の和
 diag(x)<-diagy                #
 x-max(x)}                     #各要素からデータの
                                 最大値を引く
```

4．所感

筆者はノートパソコンでS-PLUSを使用している．使用した感想を述べる．

①パワフルなソフトウェアである．15年前の汎用機時代ほぼ1カ月かかる解析が1日でできる．豊富な関数がサポートされている．線形代数を体得するのに適している．統計手法の開発に適している．アルゴリズム記述言語の1つの候補である．dimension，go to文，if文から解放された．

②オブジェクト指向ソフトウェアのお手本となる．研究者，実務家にとっては格好の題材になると思う［7］．

③マニュアルは完備しているが，短時間で使いこなせるようになる仕掛けが欲しい．マニュアルの邦訳はS-PLUSの普及のため必須である．（邦訳中とのこと）例題集もあれば親切である．筆者の経験では初心者に，手とり足とりでフォローしたところ2時間で使用可能になった．統計の教科書，Technometrics誌などで引用される著名なデータが組み込まれていて大変便利である．

④プロトタイプ開発に適している．実現可能性の早期確認に有効と思う．非定型なデータ解析業務に威力を発揮する．

⑤各種関数にAICが採り入れられているのは嬉しいが，数量化理論，因子分析，田口メソッドの関数も欲しい．

⑥ユーザサポートは良い．誠意あるサポートに敬意を表します．

謝辞　本稿執筆の機会を提供していただきましたOR学会ORソフトウェア研究部会主査八巻直一氏および当社取締役生産管理部長黒田寿一氏，また原稿を読んでいただき，貴重なコメントをいただきました編集委員会に感謝いたします．

参考文献

［1］　上田太一郎（1993）：“S-PLUSによる回帰分析変数選択の試み”，品質 Vol.23，No.3

［2］　上田太一郎（1993）：“残差を用いた回帰分析変数選択”，情報処理学会第47回（平成5年度後期）全国大会予稿集

［3］　上田太一郎，小林整功（1993）：“システムエンジニアリング企業における入社試験結果と数年後の能力評価との相関について”，情報処理学会第47回（平成5年度後期）全国大会予稿集

［4］　上田太一郎（1993）：“S-PLUSの有効利用”，日本計算機統計学会　第7回シンポジウム予稿集

［5］　上田太一郎，小林整功（1994）：“ソフトウェア業界の特質について－判別分析変数選択規準とその適用例－”，情報処理学会第49回（平成6年度後期）全国大会予稿集

［6］　渋谷政昭，柴田里程著（1992）：「Sによるデータ解析」，共立出版

［7］　J. M. チェンバース＋T. J. ヘイスティ編　柴田里程　訳（1994）「Sと統計モデル」，共立出版

［8］　上田太一郎（1994）：“S-PLUSの有効利用”，94年度第1回（OR学会）ソフトウェア研究部会発表資料

［9］　情報処理振興事業協会（1994）：「先端技術の現状と製品化に関する調査報告書」