

# SAS ユーザのための S-Plus 活用術

高橋 行雄

中外臨床研究センター・バイオメトリックス部

## 要旨

S-Plus は、SAS と同様に製薬会社の臨床開発関連の統計解析に使われている。しかしながら、その 2 つを使いこなしている人たちは、極めて少ない。S-Plus を使ったみたいとチャレンジする SAS のパワーユーザたちもいたのであるが、その努力はなかなか実っていない。S-Plus のパワーユーザにとっても SAS は敷居が高いようで、互いに独立したコミュニティを形成している。S-Plus のグラフの機能は、SAS を凌駕する機能を持っていて、両者を使いこなすことにより統計解析の生産性と質の向上につながると確信することがようやくできた。これまでに目にした S-Plus 教育・研修テキストに、S-Plus に興味を持つ SAS のパワーユーザたちを挫折させる側面があることを見出した。そこで、S-Plus の SAS にない圧倒的なパワーがどこにあるのか、それを SAS とどのように融合させたらよいか、などを主体にした研修テキストを作成しており、S-Plus の何が SAS のパワーユーザたちを魅了する具体例を示したい。これにより、SAS のパワーユーザたちが、P-Plus も自在に操るバイリンガルとなり、臨床試験の関連する統計解析の生産性と品質の両方を同時に向上させることを期待している。

## 1. はじめに

私は、S-Plus のグラフィック関連の機能が、SAS を圧倒する潜在能力をもっていると確信しているのだが、それを引き出すことがなかなかできなかった。SAS のパワーユーザたちは、生データの収集、蓄積、段階的なデータ・レビュー、解析報告書のために SAS プログラムを駆使して、多くの関係者と共に努力を続けている。

中外臨床研究センターには、多数の SAS のパワーユーザたちと少数の S-Plus のパワーユーザたちが存在している。SAS のパワーユーザたちに対して S-Plus を、いつでも自由に使える UNIX の環境と Windows のネットワーク環境が用意してある。S-Plus を使いこなしたいと志した SAS のパワーユーザが何人もいたのであるが、使っている様子はない。SAS のパワーユーザたちは、なぜ S-Plus に価値を見出せなかったのでしょうか。

SAS のパワーユーザたちにとって、さまざまな障害となる要因が S-Plus の学習する過程にある。S-Plus の学習上の問題点を指摘しながら、S-Plus の優位性を引き出す上で、何が必要なのかを述べたい。学習意欲をかき立てるためには、これまでの S 関連の講習会とその資料、出版されている多くの書籍だけでは不十分であり、新たな研修のコンセプトとそのための教材の必要性を強く感

ずる。SAS のパワーユーザたちが、なんとかしたいと思うが、なんともしがたい課題を手品のごとく解決できる事例で迫ることが教材づくりのポイントであり、SAS のパワーユーザたちの日常的な使い方に S-Plus を自ら組み込みたいと思わせるメッセージが必須である。そして、SAS のパワーユーザたちが、S-Plus を日常的に使うようになり、彼ら・彼女らが携わっている臨床試験の統計解析に至るプロセスの生産性と質の向上をはかるためのヒントを述べたい。

私も S-Plus を使う努力を継続的にしてきたが、しばしば意欲が途切れることがあった。Insightful Corp. の M. O'Connell の「*Complementing SAS with S-PLUS 6.2 in clinical and non-clinical environments*」([http://www.insightful.com/news\\_events/webcasts/2004/03sas/0304Webcast\\_SAS.pdf](http://www.insightful.com/news_events/webcasts/2004/03sas/0304Webcast_SAS.pdf)) が、私の意欲を再びかき立てた。なぜ意欲が途切れたのか、それを解決するために必要な知識をどこから得たらよいか、実際の活用事例を示した研修テキストの作成の経験を基に、S-Plus の新たな研修のあり方について論じたい。

## 2. SAS に対する S-Plus の優位性

S-Plus の SAS に対する圧倒的な優位性は、Trellis グラフにある。事例を示そう。臨床試験では、数多くの経時的な検査データを多くの医療施設から収集している。これらのデータには、思いがけないようなはずれ値が混在しがちである。単独では、はずれ値とは思われないが、経時的な推移をみると一目でおかしいと気がつくものもある。すべての検査データを経時的な線グラフで表すことは基本中の基本である。ある程度のデータが集まったときに、予想していたような反応が出ているのか、考えられる予後因子で分類した場合の反応が出ているのか、思わぬ結果が見出された場合に、その原因は何なのか探ったりしたいのである。

SAS のパワーユーザたちもグラフはできるだけコンパクトに表示したいとの思いはある。しかしながら、SAS のグラフに、S-Plus の Trellis に相当する機能が乏しい。そのために SAS のパワーユーザたちは、多くの労力をかけて、コンパクトかつ情報量が多いグラフを作成している。SAS のパワーユーザたちがほしいと願ってはいるが、そのための労力を考えると断念している事例を S-Plus の Trellis グラフで実現してみよう。

サンプルデータは、Milland, Krause (2001), *Applied Statistics in the Pharmaceutical Industry* の第 4 章 *Analysis of Toxicokinetics and Pharmacokinetics Data from Animal Study* を使う。このデータは、O'Connell のプレゼン資料に使われているものである。同じデータを用いてデータのクリーニング・プロセスと探索的な解析を想定して、必要となる様々な線グラフを作成してみよう。

得られたデータ数は、(用量 3) × (性 2) × (検査日 2) × (SESSION 4) × (動物数 4) × (時点 3) = 576 個の薬物濃度データである。投与量 1mg/kg, 360 日目の Male のデータについて測定結果を示す。これは、全データの 12 分の 1 のデータ数である。

薬物濃度データは、576 個と小さなデータセットであるが、ここにいたるまで、多くの実験者、採血者、検体の輸送業者、分析者、分析結果の報告者、データマネージャの手を経て SAS のパワーユーザの手元にくる。得られたデータには、多くの人たちの分業によるものであるので、何か

の勘違いによるおかしなデータが混在することが常である。これを発見するために、各種の経時的な線グラフが必須となってくる。

表 1 投与量 1mg/kg・360 日目の Male の薬物濃度データ

SESSION	ANIMAL	h_0	h_05	h_1	h_2	h_3	h_4	h_5	h_7	h10	h14	h18	h24
1	1	0.00	.	.	.	52.68	.	.	.	14.46	.	.	.
	2	0.00	.	.	.	29.30	.	.	.	22.22	.	.	.
	3	0.00	.	.	.	32.47	.	.	.	6.67	.	.	.
	4	0.00	.	.	.	26.31	.	.	.	6.67	.	.	.
2	5	.	25.73	.	.	.	42.37	.	.	.	6.67	.	.
	6	.	33.77	.	.	.	51.85	.	.	.	6.67	.	.
	7	.	11.92	.	.	.	22.81	.	.	.	6.67	.	.
	8	.	18.00	.	.	.	50.71	.	.	.	14.31	.	.
3	9	.	.	23.53	.	.	.	16.17	.	.	.	0.00	.
	10	.	.	43.95	.	.	.	34.49	.	.	.	0.00	.
	11	.	.	27.61	.	.	.	32.10	.	.	.	0.00	.
	12	.	.	32.69	.	.	.	30.57	.	.	.	0.00	.
4	13	.	.	.	47.80	.	.	.	25.94	.	.	.	0.00
	14	.	.	.	69.19	.	.	.	26.76	.	.	.	0.00
	15	.	.	.	35.05	.	.	.	22.14	.	.	.	0.00
	16	.	.	.	48.21	.	.	.	39.07	.	.	.	0.00
平均		0.00	22.36	31.95	50.06	35.19	41.94	28.33	28.48	12.51	8.58	0.00	0.00

S-Plus で作成した線グラフは、用量 (1, 3, 10mg/kg)・性 (1, 2)・投与時期 (0 日, 360 日) 別の線グラフである。S-Plus ユーザにとってなじみのある格子グラフであろう。用量の増加に伴う経時変化など期待通りなのか、ここに示したデータは、十分な吟味を経たデータであり、不自然な挙動は見受けられない。性別に大きな違いがなければ、合わせたグラフに意味が出てくるし、報告書には、見栄えを重視して 0 日目と 360 日目を別々に 3 用量をまとめた Trellis グラフとしたくなる。

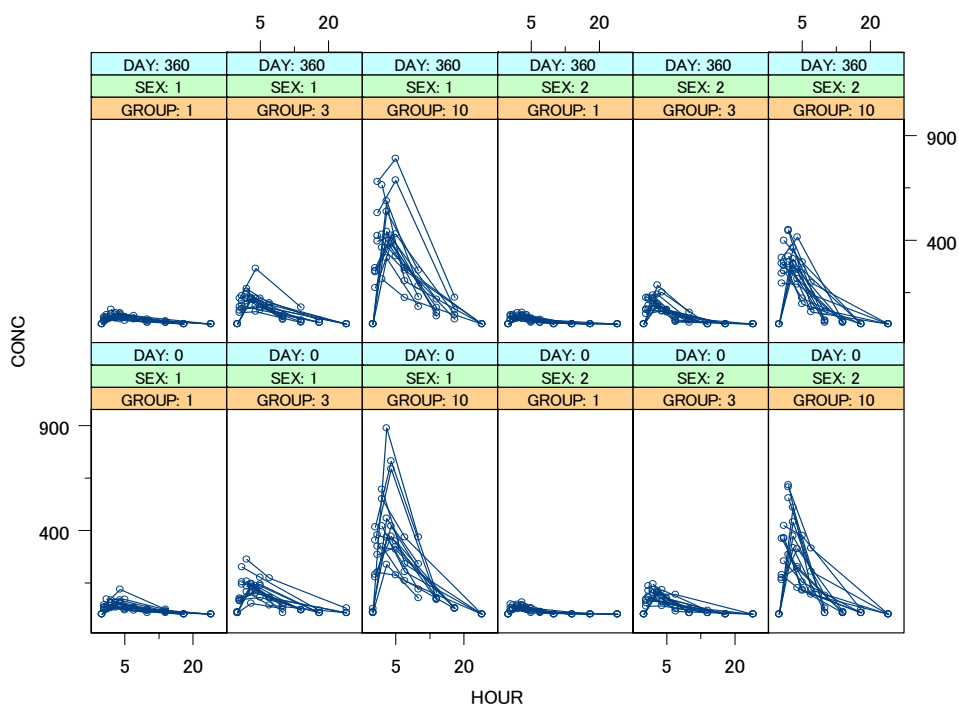


図 1 6 列 2 行に収めた線グラフ

このデータは、0日目と360日目で同じ症例から同じ時点での血中薬物濃度データが得られている。それらに整合性があるのだろうか。全体で96症例あり、できるだけコンパクトにするために12列4行2ページのレイアウトにしたい。そのために、Multipanel タブの画面と作成された Trellis グラフを図2に示す。0日目と360日目の記号を代えるために、Vary Symbols タブでの設定を行っている。

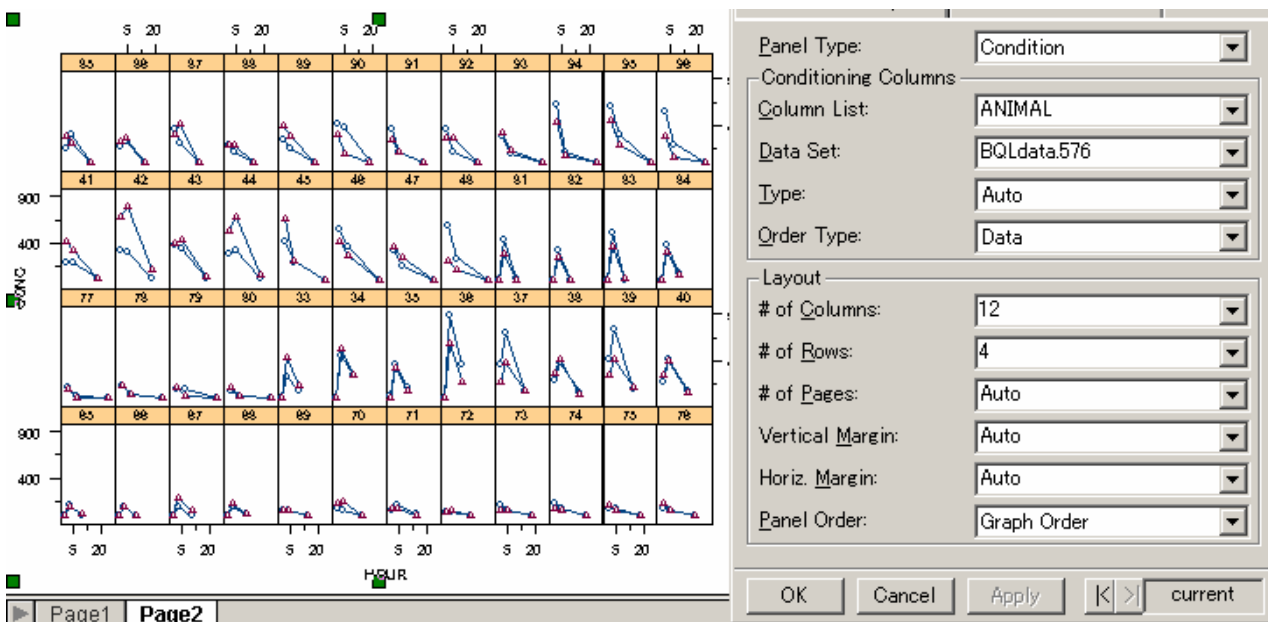


図2 マルチパネルタブの詳細設定での12列4行のレイアウト設定

### 3. S-Plus の標準的な研修スタイル

#### ・コース・ノート

世界的に行われている S-Plus の研修は、3 から 4 日間が標準的である。使用しているテキストは、次の3冊であろう。

S-PLUS Essentials I: The Graphical User Interface, 488 ページ

S-PLUS Essentials II: The Command Line, 493 ページ

S-PLUS Training Manual: S-PLUS for SAS Users, 606 ページ

日本では、S-PLUS for Windows 入門、S-PLUS for Windows 応用、S 言語入門、S-PLUS プログラミング基礎、S-PLUS による探索的データ解析、などの半日コースが開催されている。これらのコースのテキストは、似通った内容になっているので、S-PLUS for SAS Users のテキストを学習する過程で、参考になった点、幻滅を感じた点を示したい。

S-PLUS for SAS Users の最初の事例は、ヒンズークシ地方の地震のデータである。私にとっては、教養を高めてくれた興味深いデータあったが、臨床試験にたずさわる SAS のパワーユーザたちに様々な解析の必要性を類推できる事例として、表1に示した薬物濃度データなどが良いかもしれない。

## ・ Trellis グラフ

変数の drug & drop の手順により連続量を自動的に区分してくれる Trellis グラフは、S-Plus のすばらしい機能と思う。しかし、失敗したら元に戻すのはどうしたらよいのか、自動的な区分しかできないのかとの疑問が常に残り、Trellis グラフを自在に活用する道を遮っているかのようである。Graph Tools パレットの No Conditioning ボタンで元に戻せることは、なかなか気が付かなかった。図 2 に示した Multipanel タブの使い方はテキストに記載がなく、試行錯誤の学習で突き止めたものである。これを使うことにより Trellis グラフを自在にコントロールできるようになった。さらに、変数のデータタイプの指定も Trellis グラフを使いこなすためには、必須である。図 1 では、DAY, SEX, および GROUP を factor 型として、変数名をラベルとして表示させ、図 2 では、ANIMAL をラベルから落とすために文字型としている。Trellis グラフの多彩な機能は、SAS のパワーユーザたちを虜にするに違いない。

## ・異なるグラフを 1 枚のシートへ

S-PLUS for SAS Users テキストでは、シフトキーを使って複数のグラフを 1 枚のシートに描く例示が早めに強調されて出てくる。強調されていることにより、一連の学習の後で記憶が引き出しやすいし、同僚にも特徴を説明しやすい。頻度グラフと箱ひげ図などセットで 1 枚のシートに 5 変数ぐらい入っているような活用事例があれば、SAS では苦手とする機能であるのでさらにインパクトは大きい。

## ・驚きと挫折、そして発見

散布図上で気になる症例ラベルを書き込む機能は、紙に出力することを前提にした SAS グラフの機能を完全に凌駕する。必要なものをグラフに上書きするは S-plus の基本機能であり、プログラムの簡潔さにつながり、SAS に対する優位性になる。

S-Plus で複数の症例の経時変化データを 1 枚のシートにどのようにしたら描けるのか。列方向に展開した複数の変数の散布図を重ね書きした例が S-PLUS for SAS Users テキストに示されている。ある変数の時間経過を行方向に、症例を列方向にとなるようにデータを再構築すれば、すべての症例の経時変化を 1 枚のシートに描く線グラフは作成可能と思われる。しかし、この方法はデータの構造の変換を伴い実用的ではなく、S-Plus は使えないのかと強く感じたのであった。

非線型混合モデルを使い、症例ごとの推定値と生データのプロット図を S-Plus で簡単に重ね書きできるのではないかと、ヒントを求めてテキスト・マニュアル・書籍をあさってみたが見出すことができなかった。しかし、2 年前に偶然に解決した。Line/Scatter Plot ダイアログの Line タブに Break Line ボックスを見出し、その中に Break When: があることに気が付いたのである。症例ごとに時間ごとにデータがソートされていれば、一筆書きされていて使いようもない線グラフが Break When: の X Decreases の選択で突然変化し、図 3 に示すように望んでいた線グラフが完成したのであった。

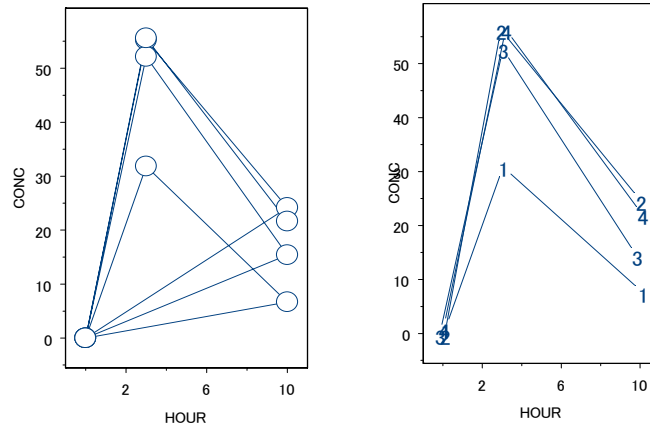


図3 S-Plusの標準グラフ(左)とカスタム化後の線グラフ(右)

何人かのS-Plusのエキスパートに尋ねても、この方法についてのヒントは得られなかった。S-PlusでのGUIによるグラフの作成は、入門者用のおまけの機能だとS-Plusのエキスパートたちは、割り切っているのであろうか。

#### ・GUIの操作をバッチ処理で

S-PLUS for SAS Usersテキストも他のテキストも同じであるが、GUIグラフの使い方の解説の後、コマンド・ラインによるグラフの作成が始まる。S-PlusのSASにない特徴は、GUIでの操作がスクリプトとして履歴ログに残ることである。そのログスクリプトを編集し、別途バッチ処理で運用することにより、グラフの作成効率が向上すると期待していたのであった。しかし、コマンド・ラインによる説明は、GUIで作成したグラフのスクリプトと異なり、これを使うのがS-Plusの標準的なマナーかと思うと学習の意欲が失せてしまった。社内でS-Plusを活用しているパワーユーザに尋ねても、コマンド・ラインによるグラフを作成していて、GUIによるグラフはあまり使っていないとの答えであった。

#### ・対話処理の強調はSASユーザにとって幻滅

S-Plusの特徴に対話処理が強調されている。1990年ごろからSASユーザは、S-Plusのスクリプト・ウィンドウのような環境で育ってきている。コマンド・ラインでの操作の説明で、逐次学習することが、学習のために優れているとS-Plusのエキスパートたちは確信しているのであるか。一連の操作をスクリプトファイルから読み込み、何行かまとめて実行できるような環境があるにもかかわらず、なぜ対話処理を前面に押し出しているのであろうか。これは、SASのパワーユーザたちに幻滅を感じさせている。SASのパワーユーザたちは、SASのプログラム(スクリプト)・ウィンドウで、スクリプトを書き逐次実行しつつバグとりをして、最終的にはバッチ処理用で結果を出している。SASユーザに対するテキストには、SASの標準的な操作をS-Plusで対比して示すことが必須である。

図4にSASユーザにとってなじみのある形式にS-Plusのウィンドウを整えた結果を示す。ここに示している事例は、生物検定法の代表的な手法の一つである平行線検定のグラフ表示

の例である。薬剤を層とする回帰分析をあらかじめ行い、予測値をデータフレームに出力しておく。薬剤の文字コード (S, T) を使った散布を GUI で作成し、その上に予測値の線グラフを上書きし、それぞれの薬剤に平行な回帰直線を引いたものである。履歴ログに記録されているスクリプトを編集して、S-Plus のスクリプトファイルを作成し、そのスクリプトを呼び出した結果が表示されている。最初の guiModify 関数が散布図の作成で、次の guiModify 関数で平行な直線を引いている。この中 BreakOnSlopeTransition = "X Decreases" が、X の値が減少したときに線を区切れとのオプションである。

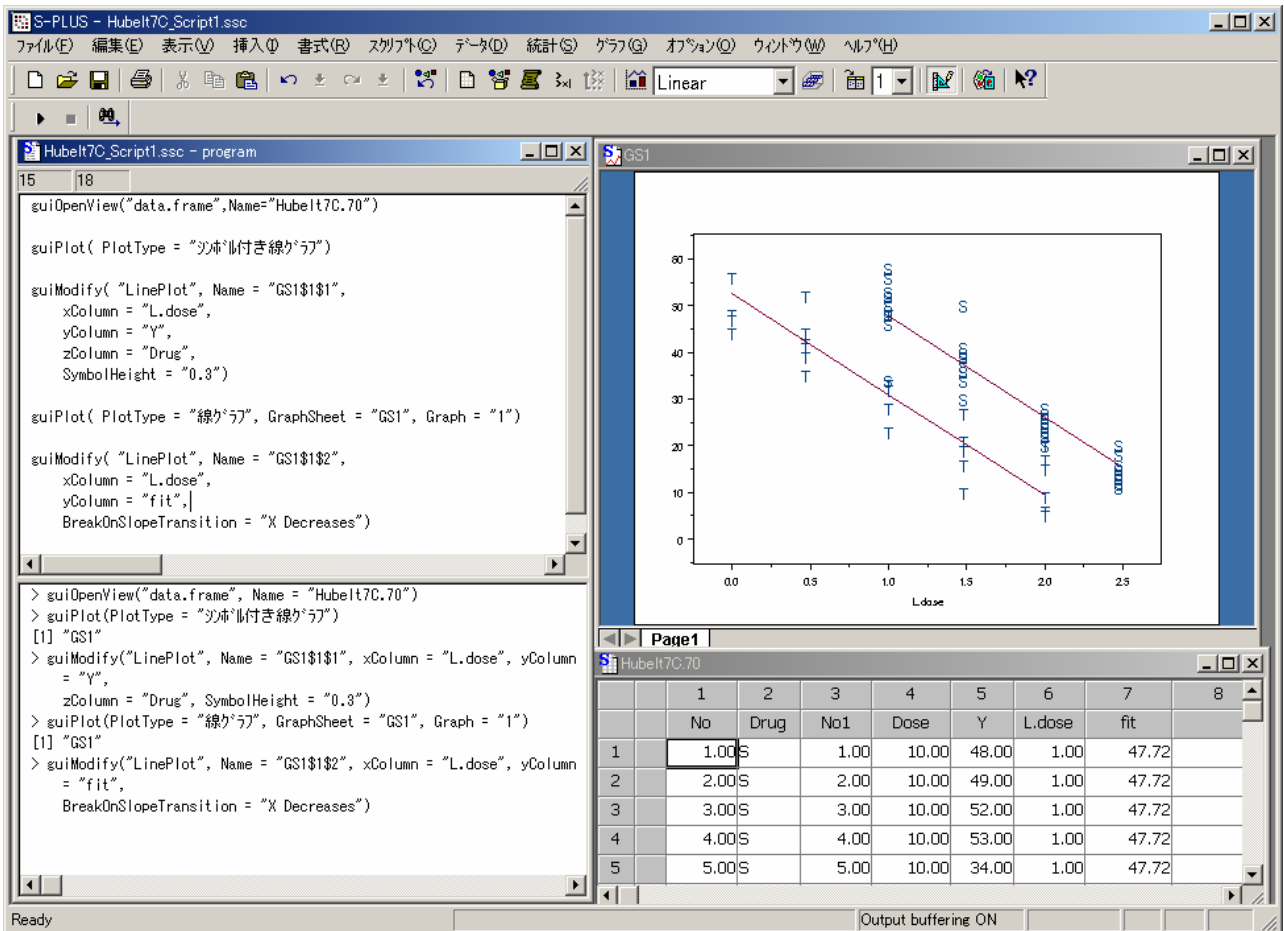


図 4 SAS のワークスタイルに似た S-Plus の画面

#### ・心機一転

数理システムから送られてきたダンボール箱に、GUI の関数を含むすべての関数の仕様書 1 冊 15cm のバインダーで 5 冊が送られてきた。これにより、気を取り直して S-Plus の学習を再開した。自分でも挫折したポイント、SAS のパワーユーザたちも挫折するだろうポイント、何を参照したらよいのかの助言を含めたテキストを作成している。

#### ・SAS のグラフにはない S-Plus のグラフの付加価値

S-Plus のグラフのファイルには、グラフの作成に用いたデータが張り付いている。これは、

グラフの品質を保証するためにすばらしい機能である。論文に示されている散布図を再現しようとしたときに、点の数が合わないことをしばしば経験する。これは、X 軸の設定をした時に、最大値を実際に存在するデータより小さな値とした結果である。グラフにデータが張り付いている場合に、グラフのファイルだけを用いて、おかしいなと思ったときにすぐにグラフの書き直しができる。これは、統計解析の結果の品質の向上にかけながら寄与すると思われる。

#### 4. SAS と S-Plus の共存がもたらす生産性と品質の向上

グローバルな製薬会社において SAS プログラマーは、臨床試験の最終的な図表・症例リストすべてを生産する役割を担っている。試験統計家たちは、解析計画、報告書の作成を通じて当該の臨床試験の説明責任を分担している。実質の解析作業は、統計の素養を持っている SAS プログラマーたちが担っている。

臨床試験は、分業で成り立っているので SAS プログラマーたちに、様々な図表類の作成の要望が殺到する。Excel で作成した図表類は、再現性が保障できないので最終報告書では使えないとの理由も有り、こまごました図表類の作成に SAS プログラマーたちは、忙殺されるのである。

どんなグラフがほしいのか要望を聞き、試作し、さらに手直しして、検討する、少量多品種の試作、生産活動が SAS プログラマーたちの日常である。SAS のグラフは、紙に出力するのが基本である。試作の結果も紙の上である。SAS は繰り返し処理に優れていて、図 1 に示した 1 枚に 2×6 線グラフを 1 枚 1 グラフで 12 ページに打ち出すことは容易である。しかし、この 12 枚の線グラフの相互の比較をするのが、このデータに対する関心事である。図 2 場合に、96 枚の線グラフを出すことはためられる。様々な方法で、コンパクトにする努力は続けてはいるが、忍耐の割には完成度が低い。

S-Plus に期待するのは、1) グラフを要望する人たちと SAS プログラマーたちが、ディスプレイの前で、どのようなグラフが適切なのかを、相互に GUI を使って試作する。2) ログからスクリプトを SAS プログラマーたちが整理し、バッチ処理の準備をする。3) 途中のデータ・レビューで、紙の打ち出しと共に異常変動となった症例番号を、ディスプレイ上のグラフに出して、症例番号を「焼きつけて」打ち出し、症例のフォロー活動の資料とする。4) 作る人たちと利用する人たちの日常的な相互コミュニケーションが、臨床試験全体の品質の向上につながる。5) 最終報告書の図が適切に作られているかを基のデータに戻ることは、検査する人たちには困難であるが、必要なデータだけがグラフに張り付いているならば、そのデータを用いて何らかの確認作が簡便に行える。

多くの期待する事柄を列挙したのであるが、実務で実現したわけではないので、今後、SAS と S-Plus の共存によってもたらされた生産性と品質の向上について、別途報告できるように努力を続けたい。