

STN

SAS Technical News

For Higher
Customer Satisfaction,
We Bridge
the SAS System
Between
Customer's World.



WINTER 2014

特集 01
High Performanceな分析へ
SAS/STAT® 12.3

SAS Partner Network 10
データから価値を生む
富士通の
ビッグデータイニシアティブ

SAS Academic News Digest 15
SAS Certification 16
Q&A 19
リリース/Hot Fix最新情報 23



High Performanceな分析へ SAS/STAT® 12.3

前号につづき、最新リリースSAS® 9.4をご紹介します。
このリリースでは、SAS/STAT®、SAS/ETS®、SAS/OR®、
SAS/IML®、SAS/QC®などの分析系プロダクトはバージョン
12.3となり、2012年夏にリリースされた12.1 (SAS 9.3
TS1M2) のメンテナンス版となります。
12.3では既存プロシジャは大きく機能拡張されていませんが、

SAS® High-Performance Analyticsプロダクトの機能
(一部) を単一サーバーで実行でき、SAS/STATプロシジャ
では対応していない機能もあります。例えば、一般化線形
モデルにおける変数選択があります。本特集記事では、
SAS/STATを含むインストールにて、実行できるSAS
High-Performance Statisticsのプロシジャをご紹介します。

1 SAS High-Performance Analyticsについて

SAS High-Performance Analytics (以下、HPA) は、今日のビックデータに対応し、複数の分散処理オプション(In-Memory、In-Database、およびGrid Computing) にて、モデリング、モデル選択などに対応しているプロダクトとなります。分散処理オプションなどの詳細については、以下のURLをご参照ください。

<http://www.sas.com/jp/high-performance-analytics>

単一サーバー、シングルマシンにて実行できる機能をご紹介します上で、HPAに関する実行モード、およびライセンスについて最初に記述します。

1.1 実行モード

HPA プロシジャと既存のプロシジャにおける基本的な違いは実行環境です。既存のプロシジャは、単一サーバー、シングルマシンにて実行しますが、HPA プロシジャはシングルマシンに加え、分散コンピューティング環境にて実行できます。また、2つ目の違いとしてマルチスレッド処理への対応が挙げられます。いままでは一部のプロシジャ^{*1}のみでしたが、HPA プロシジャはすべてがマルチスレッド処理に対応しています。

※1 マルチスレッド処理対応の従来のプロシジャ

Base SAS : MEANS, REPORT, SORT, SUMMARY, TABULATE, SQL
SAS/STAT : ADAPTIVEREG, FMM, GLM, GLMSELECT, LOESS, MIXED, QUANTLIFE, QUANTREG, QUANTSELECT, ROBUSTREG

1.2 ライセンス

9.4より前のリリースでは、SAS High-Performance Analytics ライセンスにHPA プロシジャが含まれています。このライセンスには、データマイニング、統計モデリング、時系列解析、テキストマイニングが含まれており、個別にはライセンスされていませんでした。さらに、これらのHPA プロシジャは分散コンピューティング環境での実行となり、シングルマシンでは実行できませんでした。

9.4では、SAS High-Performance Analytics ライセンスは分割され、必要とする機能のみを個別にライセンスできます。分割されたライセンスは、表1のように既存のプロダクト(以下、MVA プロダクト) と対応しています。

表1. SAS High-Performance AnalyticsとMVAプロダクトの対応

SAS High-Performance Analytics プロダクト	MVA プロダクト
SAS High-Performance Statistics	SAS/STAT
SAS High-Performance Econometrics	SAS/ETS
SAS High-Performance Optimization	SAS/OR
SAS High-Performance Forecasting	SAS High-Performance Forecasting
SAS High-Performance Data Mining	SAS Enterprise Miner
SAS High-Performance Text Mining	SAS Text Miner

さらに、HPA プロシジャは対応しているMVAプロダクトに含まれ、シングルマシンの実行にHPAライセンスは必要ありません。例えば、SAS High-Performance StatisticsのHPREGプロシジャは、対応しているSAS/STATに含まれ、HPAライセンスがない場合にもシングルマシンで実行できます。ただし、分散処理オプション(In-Memory、In-Database、およびGrid Computing)を用いた環境で実行するときは、HPAライセンスが必要です。

2 SAS High-Performance Statisticsについて

すべてのHPA プロシジャは、分散コンピューティング環境で、分析ができるように設計されています。また、シングルマシンで実行する場合と同じ構文にて分散コンピューティング環境にでも実行できるようになっています。このため、シングルマシンでプログラム、結果の検討などを行ったのち、データ量の増大などにもなう分散コンピューティング環境に移行するときにも、構文をそのまま活用できます。

統計モデリングに対応しているSAS High-Performance Statisticsでは、大量のオブザベーション、非常に多くの変数のデータから、正確な予測モデルを作成することが主な目的となります。このため、可能である限り、説明変数の選択、次元の削減、重要な変数の特定に対応しています。一方、計算上多くのコスト(メモリ、処理時間など)がともなう統計推論の算出について、対応していないものもあります。例えば、混合モデルを推定できるMIXEDプロシジャでは、自由度算出としてSatterthwaite法、KenwardとRoger法などに対応しています。一方、HPAプロシジャの一つHPLMIXEDで混合モデルは推定できますが、これらの自由度算出の手法には対応していません。

この他、HPA 統計モデリングのプロシジャ間にて共通となる機能がある場合には、同じ構文を用いるように設計されています。ただし、同じ機能をもつSAS/STATプロシジャと同じとは限りません。例えば、回帰分析における変数選択は、SAS/STAT REGプロシジャのMODELステートメントにてSELECTION=オプションを用いますが、HPAのHPREGプロシジャではSELECTIONステートメントにて指定します。

HPA統計モデリングにおける プロシジャの説明

HPA 統計モデリングのプロシジャは、シングルマシンで実行できるほか、SAS/STAT プロシジャにおける機能を拡張している部分もあります。以下では、HPA プロシジャにて共通のステートメントを記述したのち、SAS/STAT プロシジャとの比較を含め、HPA プロシジャを説明します。

3.1 共通ステートメント

HPA プロシジャ間にて共通となる機能については、同じ構文となります。このため、各プロシジャの説明の前に、共通であるステートメントのいくつかをご紹介します。

● CLASS ステートメント

カテゴリカルな値をもつ変数、分類変数がモデルに含まれる場合、連続的な変数と区別するためにCLASSステートメントを用います。CLASSステートメントにて指定されている変数については、内部的にダミー変数が作成され、モデルが推定されます。生成されるダミー変数に関しては、SAS/STATのGENMODプロシジャ（デフォルトはGLM法）、LOGISTICプロシジャ（デフォルトは効果法）などと同様にPARAM=オプションで指定します。

ダミー変数の手法に関してですが、例えば、変数CATが3つの水準A、B、Cである場合、表2のような変数が内部的に生成されます。

表2. 分類変数に対するダミーコード: GLM、参照、効果法

GLM法 (GLM)				参照法 (Reference)			効果法 (Effect)		
CAT	D1	D2	D3	CAT	D1	D2	CAT	D1	D2
A	1	0	0	A	1	0	A	1	0
B	0	1	0	B	0	1	B	0	1
C	0	0	1	C	0	0	C	-1	-1

デフォルトでは最後の水準が参照水準となり、GLM法の場合にはパラメータの値は0に設定されます。参照法、効果法における参照水準は、REF=オプションにて指定できます。SAS/STATのプロシジャでは上記の他、さまざまな手法に対応していますが、HPAプロシジャではGLM法（デフォルト）と参照法のみに対応しており、すべての分類変数に対してダミー変数の手法は統一されている必要があります。

分類変数の水準順序については、ORDER=オプションにて指定できます。デフォルトではORDER=FORMATTEDとなりますので、分類変数に適用されている出力形式に基づき、順序が決まります。欠損値がある場合には、MISSINGオプションを追加することで欠損値を一つの水準と扱うことができます。

● MODEL ステートメント

モデルの指定はMODELステートメントで行います。等号 '=' の左側に従属変数、右側に説明変数をリストします。分類変数がモデルに含まれている場合には、MODELステートメントの前にCLASSステートメントを記述しておく必要があります。

非線形モデルでは、等号ではなく、チルダ '~' を用いて、モデル式を記述します。

● SELECTION ステートメント

変数選択に対応しているSAS/STATのプロシジャでは、変数選択の手法をMODELステートメントのSELECTION=オプションで指定します。一方、HPAプロシジャでは、SELECTIONステートメントにて、変数選択の手法、選択基準などの設定を行います。表3は、各プロシジャが対応している手法となります。

表3. 各プロシジャが対応している変数選択手法

手法	REG	GLMSELECT	LOGISTIC	HPREG	HPLOGISTIC	HPGENSELECT
変数増加 (Forward)	✓	✓	✓	✓	✓	✓
変数増加-スワップ (Forwardswap)				✓		
変数減少 (Backward)	✓	✓	✓	✓	✓	✓
変数減少 (高速) (Backward(Fast))					✓	
変数増減 (Stepwise)	✓	✓	✓	✓	✓	✓
LAR		✓		✓		
LASSO		✓		✓		

REGプロシジャにて対応している最大R2乗値改良 (MAXR)、最小R2乗値改良 (MINR)、および総当り法 (R2乗値による選択 (RSquare)、調整済みR2乗値による選択 (ADJRSQ)、MallowsのCP統計量による選択 (CP)) の手法は、HPAプロシジャでは対応していません。また、LOGISTICプロシジャのスコア (Score) 法にもHPAプロシジャは対応していません。

すべてのプロシジャが変数増加、変数減少、変数増減の手法に対応していますが、HPAプロシジャではより汎用的になります。例えば、REGプロシジャではモデルに効果を追加/削除する際、F統計量が判断基準ですが、HPAプロシジャではSELECT=サブオプションで判断基準となる統計量 (ADJRSQ、AIC、AICC、BIC、CP、RSQUARE、SBC、SLなど) を指定できます。この他、STOP=オプションにて変数選択を終了とする判断基準を指定することもできます。また、MAXEFFECTS=オプションにてモデルに含まれるモデル効果の最大数、MAXSTEPS=オプションにて変数選択のステップ数の上限を指定することなどができます。変数減少法の場合には、MINEFFECTS=オプションにてモデルに含まれるモデル効果の最小数を指定できます。モデルの変数選択を行った後、

最適なモデルの選択はCHOOSE=オプションにて指定した基準 (ADJRSQ, AIC, AICC, BIC, CP, PRESS, SBCなど) にて選択されます。CHOOSE=オプションが指定されていない場合には、変数選択のステップにおける最後のモデルが最適なモデルとなります。表4では、変数選択にて指定できるサブオプション、および変数選択の手法との対応を示しています。

表4. 変数選択の手法とサブオプションの対応

サブオプション	FORWARD	BACKWARD	STEPWISE	FORWARDSWAP	LARS	LASSO
ADAPATIVE						✓
CHOOSE=	✓	✓	✓		✓	✓
COMPETITIVE			✓			
CRITERION=	✓	✓	✓	✓		
FAST		✓				
LSCOEFFS					✓	✓
MAXEFFECTS=	✓		✓	✓	✓	✓
MAXSTEPS=	✓	✓	✓	✓	✓	✓
MINEFFECTS=		✓	✓			
SELECT=	✓	✓	✓	✓		
SLENTRY=	✓		✓		✓	✓
SLSTAY=		✓	✓			✓
STOP=	✓	✓	✓	✓	✓	✓

● PARTITION ステートメント

入力データセットを学習データ (TRAIN)、検証データ (VALIDATE)、テスト (TEST) の3つに分割し、変数選択、統計量の算出などに用います。事前に3つのグループに分割されている場合には、グループを示す変数をROLEVAR=オプションにて指定します。またはランダムに分割を行う場合には、FRACTIONオプションで検証データ、テストデータの比率を指定します。PARTITIONステートメントにて分割を行い、SELECTIONステートメントのSELECT=オプションなどでVALIDATEの値を指定している場合には、検証 (VALIDATE) データに対する平均平方誤差 (ASE) が変数選択の判断基準となります。

● FREQ ステートメント

各オブザベーションの度数を示す変数を指定します。値が1より小さい、もしくは欠損である場合には、オブザベーションは分析より除外されます。指定がない場合には、各オブザベーションの度数は1として扱われます。

● WEIGHT ステートメント

各オブザベーションの重みを示す変数を指定します。値が0より小さい、もしくは欠損である場合には、オブザベーションは分析より除外されます。指定がない場合には、各オブザベーションの重みは1として扱われます。

● ID ステートメント

HPA プロシジャにて生成されるデータセットのうち、オブザベーションごとに統計量 (予測値、残差など) が算出される場合において、含める変数を指定します。

● OUTPUT ステートメント

オブザベーションごとに統計量 (予測値、残差など) を算出し、データセットに出力します。デフォルトでは、入力データセットにおける変数は出力データに含まれませんが、IDステートメントにて保持する変数を指定できます。

表5はHPAプロシジャとステートメントの対応となります。

表5. HPAプロシジャとステートメントの対応

	HPREG	HPLOGISTIC	HPGENSELECT	HPLMIXED	HPNLMOD	HPSPPLIT
CLASS	✓	✓	✓	✓		
MODEL	✓	✓	✓	✓	✓	
SELECTION	✓	✓	✓			
PARTITION	✓					✓
FREQ	✓	✓	✓			
WEIGHT	✓	✓	✓			
ID	✓	✓	✓			✓
OUTPUT	✓	✓	✓			✓

この他、すべてのプロシジャはPERFORMANCEステートメントがサポートされており、分散コンピューティング環境にて実行するときの設定を指定できます。また、プロシジャごとにサポートされているステートメントもあります。例えば、混合モデルに対応しているHPLMIXEDプロシジャでは、ランダム効果のRANDOMステートメント、反復効果のREPEATEDステートメントがサポートされています。

3.2 分析手法

● 線形モデル：HPREG プロシジャ

線形モデルには、回帰分析、分散分析などがあり、SAS/STATのプロシジャではREGプロシジャ、GLMプロシジャ、GLMSELECTプロシジャなどがあります。HPAでは、HPREGプロシジャが対応しています。

REGプロシジャはCLASSステートメントがサポートされていないので、分類変数をモデルに含められません。このため、分類変数がある場合には、CLASSステートメントがサポートされているGLMプロシジャ、GLMSELECTプロシジャなどを用います。HPAプロシジャの一つHPREGプロシジャではCLASSステートメントがサポートされています。

REGプロシジャにおける変数選択では、F統計量を判断基準とし、効果が追加／削除されます。この手法は、HPREGプロシジャにて、SELECT=オプションの値をSLとした場合に該当します。変数選択のステップでは、一つの効果としてモデルに追加／削除を検討することがデフォルトとなりますので、分類変数に対するダミー変数の一部のみがモデルに追加／削除されることはありません。ただし、HPREGプロシジャは、CLASSステートメントのSPLITオプションがサポートされており、ダミー変数を個別に扱うことができます。

推定されたモデルを検証する上で、各オブザベーションに対する残差、診断統計量などが用いられます。これらの統計量はGLMプロシジャ、GLMSELECTプロシジャでは算出されませんが、REGプロシジャ、HPREGプロシジャのOUTPUTステートメントで求められます。

● ロジスティック回帰モデル：HPLOGISTIC プロシジャ

線形モデルは従属変数が連続であるケースのモデルとなります。従属変数が2値、もしくは少ない数の水準である場合には、線形モデルではなくロジスティック回帰モデルが一つの分析手法となります。ロジスティック回帰モデルはLOGISTICプロシジャにて対応しており、HPAではHPLOGISTICプロシジャとなります。

HPLOGISTICプロシジャでは、手法をSELECTIONステートメントのMETHOD=オプションで指定し、判断基準をサブオプションSELECT=オプションにて指定できます。ただし、現状サポートしている値はSELECT=SL (有意水準) のみとなります。

● 一般化線形モデル：HPGENSELECT プロシジャ

正規分布を仮定している線形モデル、二項分布(多項分布)を仮定しているロジスティック回帰モデルを含む枠組みとして、一般化線形モデルがあります。例えば、Poisson回帰モデルも一般化線形モデルの一つとなります。一般化線形モデルはGENMODプロシジャにて対応しており、HPAではHPGENSELECTプロシジャが該当します。

GENMODプロシジャ、HPGENSELECTプロシジャのMODELステートメントでは、LINK=オプションにてリンク関数を指定します。このLINK=オプションの指定がない場合、DIST=オプションの値に基づき、リンク関数が選択されます。LINK=オプション、DIST=オプションの指定がない場合には、DIST=NORMAL, LINK=IDENTITYとなります。

DIST=オプション、デフォルトのLINK=オプションの値に関しては、表6をご参照ください。

表6. 対応している分布とリンク関数のリスト

分布	リンク関数 (デフォルト)	GENMOD	HPGENSELECT
二項分布	ロジット	✓	✓
多項分布	累積ロジット	✓	✓
負の二項分布	対数	✓	✓
Poisson分布	対数	✓	✓
ガンマ分布	逆数	✓	✓
幾何分布	対数	✓	✓
正規分布	恒等	✓	✓
逆正規分布	逆2乗	✓	✓
ゼロ強調Poisson分布	ログ/ロジット	✓	✓
ゼロ強調負の二項分布	ログ/ロジット	✓	✓
Tweedie分布	対数		✓

HPGENSELECTプロシジャでは、GENMODプロシジャで対応していないTweedie分布を指定できます。Tweedie分布は、乱数NがPoisson分布に従うとき、N個の独立したガンマ乱数の合計として捉えることができます。

HPGENSELECTプロシジャがGENMODプロシジャと大きく異なる点は、一般化線形モデルの枠組みにおいて変数選択の手法に対応していることです。対応している手法は、変数増加(Forward)、変数減少(Backward)、変数増減(Stepwise)法となります。(表3参照)

HPREGプロシジャ、HPLOGISTICプロシジャとほぼ同じ構文にてHPGENSELECTプロシジャを実行できます。大きく異なる点としては、MODELステートメントにて、DIST=オプション、LINK=オプションを用いた、分布、リンク関数の指定となります。

● 線形混合モデル：HPLMIXED プロシジャ

線形モデルを拡張し、固定効果のみではなくランダム効果を含めたモデルとして、線形混合モデルがあります。データにクラスターが含まれている場合、同じ被験者から繰り返し観測する反復分析などに用いられます。線形混合モデルに関しては、MIXEDプロシジャ、HPMIXEDプロシジャ^{※2}にて対応しています。HPAでは、HPLMIXEDプロシジャが該当します。

※2 SAS/STATのHPMIXEDプロシジャは、プロシジャ名がHPで始まっていますが、HPAプロシジャの枠組みには当てはまりません。このプロシジャはデータがまばらであるケースの手法を用いてハイパフォーマンスな処理を実現していますが、MIXEDプロシジャ、HPLMIXEDプロシジャに対応しているモデルの一部のみ、分析できます。

HPLMIXEDプロシジャでは、ランダム効果を指定するRANDOMステートメント、反復効果を指定するためのREPEATEDステートメントがあり、さまざまな共分散構造を指定した上で、モデルが推定できます。

● 非線形モデル：HPNLMOD プロシジャ

分析によっては、パラメータの線形モデルが適切ではなく、非線形モデルが必要となるケースもあります。このような非線形モデルでは、プログラミングステートメントで式を記述し、モデルの推定を行います。非線形モデルについては、NLIN プロシジャ、NLMIXED プロシジャが対応しており、HPA ではHPNLMOD が該当します。

HPNLMOD プロシジャではNLIN プロシジャと同様、最小2乗法を用いてモデルが推定できます。また、NLIN プロシジャでは対応していませんが、NLMIXED プロシジャと同様、分布の指定を行い、最尤法でのモデル推定にも対応しています。ただし、RANDOM ステートメントはサポートされていないので、ランダム効果には対応していません。

● 決定木モデルリング：HPSPLIT プロシジャ

上述の線形モデル、非線形モデルの他、データマイニングなどで用いられる手法として決定木モデリングがあります。従属変数を説明、予測する上で、説明変数を元にしたルールで入力データセットをセグメント（ノード）に分割していきます。各セグメントの分割を繰り返し、作成される階層が決定木となり、末端のセグメントは葉（leaf）と呼ばれます。各オブザベーションに対する予測値などの統計量は、この葉（leaf）に基づき算出されます。既存プロシジャでは対応しているものはありませんが、HPA プロシジャの一つ、HPSPLIT プロシジャにて実行できます。HPSPLIT プロシジャは、他のプロシジャと異なり、MODEL ステートメントがサポートされていません。このため、従属変数はTARGET ステートメント、説明変数はINPUT ステートメントにて指定します。また、分類変数はCLASS ステートメントではなく、INPUT ステートメントでLEVEL=NOM オプションを追記します。（ただし、説明変数が文字変数である場合には、デフォルトで分類変数と認識されます。）

セグメント（ノード）に分割する際、デフォルトではエントロピー統計量（ENTROPY）が判断基準となります。他の判断基準を用いるときには、CRITERION ステートメントを用いて、GINI、FASTCHAID（評価版）の値を指定します。また、決定木では分割した後の剪定（枝刈り）ステップがあります。一般的に、入力データに基づく決定木ではオーバーフィッティング（過剰適合）している傾向があり、予測モデルとして適さない場合があります。このため、入力データに特化しすぎていないかを検討する上で、剪定ステップがあります。剪定の手法はPRUNE ステートメントで指定します。デフォルトはエントロピー（ENTROPY）となりますが、NONE（剪定なし）、C45（C4.5手法に基づく剪定）を選択できます。この他、平均平方誤差（ASE）、ジニ係数（GINI）、誤分類率（MISC）の統計量を剪定の基準とすることもできます。

3.3 例題

ここでは、HPREG プロシジャで線形モデルにおける変数選択を行う例と、HPSPLIT プロシジャを用いた決定木モデリングの例を取り上げます。ともに、ドキュメントに記載されている例となりますが、オプションの設定など、一部変更しています。

● 線形モデルにおける変数選択：HPREG プロシジャ

1986年のメジャーリーグ選手の成績（nAtBat, nHits, …）と1987年の年棒（Salary）のデータ（BASEBALL）から、1987年の年棒を推定します。ただし、年棒は高額になるほどばらつきが大きくなるので、最初に対数をとって、logSummary = log (Summary) を従属変数とします。HPREG プロシジャを用いた場合、以下のような記述となります。

例

```
PROC HPREG DATA=baseball;
  CLASS league(split) division(split);
  MODEL logSalary = nAtBat nHits nHome nRuns nRBI nBB
                  yrMajor crAtBat crHits crHome crRuns crRbi
                  crBB league division nOuts nAssts nError;
  SELECTION METHOD=stepwise(SELECT=SL CHOOSE=aic STOP=sbcs) STOPHORIZON=5;
RUN;
```

プログラムを実行しますと、ログに以下のメッセージが表示され、実行モードがシングルマシンであることを確認できます。

NOTE:HPREG プロシジャは、single-machine モードで実行されています。

また、結果として表示されるアウトプット（図1）にても‘パフォーマンスの情報’にて確認できます。

図1. 実行モード、モデルの情報、変数選択に関するアウトプット

パフォーマンスの情報	
実行モード	Single-Machine
スレッド数	4

モデルの情報	
データソース	WORK.BASEBALL
従属変数	logSalary
分類パラメータ化	GLM

選択の情報	
選択の方法	ステップワイズ
Select 基準	有意水準
Stop 基準	SBC
Choose 基準	AIC
適用される効果の階層	なし
入力の有意水準 (SLE)	0.05
維持する有意水準 (SLS)	0.05
Stop 区間	5

この他、モデルの情報として、データセット名、従属変数の名称、また、分類変数をCLASSステートメントで指定している場合、ダミー変数の生成方法、'分類パラメータ化'が併せて表示されます。また、変数選択の基準に関する部分では、選択した手法の名称のほか、さまざまな設定が表示されます。

図2. 分類変数に関するアウトプット

分類変数の水準の情報		
分類	水準	値
league	2 *	American National
division	2 *	East West

* Associated Parameters Split

分類変数の水準についても表示されます。変数選択の際、ダミー変数を個別に扱うようにSPLITオプションを指定した場合には、図2のように*と表示されます。

この例題では変数増減法を用いた変数選択を指定しており、結果として変数選択の要約(図3)が出力されます。

図3. 変数選択のステップに関するアウトプット

HPREG プロシジャ						
選択の要約						
ステップ	効果の 入力	効果の 削除	取り込んだ 効果数	AIC	SBC	p 値
0	Intercept		1	204.2238	-57.2041	1.0000
1	crRuns		2	63.5391	-194.3166	<.0001
2	nHits		3	1.7041	-252.5794	<.0001
3	yrMajor		4	-12.0208	-262.7322	<.0001
4	nBB		5	-19.2591	-266.3983	0.0026
5		crRuns	4	-19.0690	-269.7804*	0.1431
6	division_West		5	-21.4702*	-268.6094	0.0379

* Optimal Value of Criterion

SELECT=SLとなっていますので、有意水準を元に変数選択がおこなわれ、crRuns(打点)、nHits(1986年の安打数)、yrMajor(メジャーリーグ在籍年数)、nBB(1986年の四死球数)、division_West(1986年終了時に西地区に所属)が追加され、ステップ5ではcrRunsが削除されています。STOP=SBCとしていますので、変数選択のステップは、有意水準を元に変数の追加、もしくは削除がされない、もしくはSBC統計量が最小となった時点にて終了となります。この例では、有意水準に基づき、終了となり、以下のメッセージが出力されます。

```
Selection stopped because all candidates for removal
are significant at the 0.05 level and no candidate
for entry is significant at the 0.05 level.
```

CHOOSE=AICとしていますので、変数選択のステップにて構築されたモデルのうち、AIC統計量が最小のモデルが最適のモデルとして選択されます。上記では、ステップ6におけるモデルとなり、そのモデルにおける推定結果(図4)が出力されます。

図4. 選択されたモデルに関する統計量のアウトプット

HPREG プロシジャ 選択済みモデル					
Selected Effects: Intercept nHits nBB yrMajor division_West					
分散分析					
要因	自由度	平方和	平均平方	F 値	Pr > F
モデル	4	121.96315	30.49079	92.34	<.0001
誤差	258	85.19059	0.33020		
修正合計	262	207.15373			

Root MSE	0.57463
R2 乗	0.58876
調整済み R2 乗	0.58238
AIC	-21.47020
AICC	-21.14207
SBC	-268.60942
ASE	0.32392

パラメータ推定値					
パラメータ	自由度	推定値	標準誤差	t 値	Pr > t
Intercept	1	4.111662	0.120091	34.24	<.0001
nHits	1	0.007809	0.000989	7.90	<.0001
nBB	1	0.007117	0.002037	3.49	0.0006
yrMajor	1	0.100372	0.007504	13.38	<.0001
division_West	1	-0.148754	0.071291	-2.09	0.0379

分散分析、AICなどの各統計量、また、パラメータ推定値など、REGプロシジャなどの既存プロシジャとほぼ同じ形式にて出力されます。

● 決定木: HPSPLIT プロシジャ

データセットHMEQには、5,960件のローンに関し、債務不履行の有無(BAD)の情報と併せ、職業(JOB)、申込みローン金額(LOAN)、クレジットカード未払い回数(DELIQ)などが含まれています。従属変数は債務不履行の有無であり、決定木を行うため、以下のような記述にてHPSPLITプロシジャを実行します。

例

```
PROC HPSPLIT DATA=hmeq MAXDEPTH=7 MAXBRANCH=2;
TARGET bad;
INPUT delinq derog job ninq reason / LEVEL=nom;
INPUT clno debtinc loan mortdue value yoj / LEVEL=int;
CRITERION entropy;
PRUNE misc / N <=6;
PARTITION FRACTION(validate=0.2);
RULES FILE='hpsplit-rules.txt';
SCORE OUT=scored;
RUN;
```

説明変数には分類変数と連続変数があるため、INPUTステートメントを二度記述しています。決定木を作成する場合には、エントロピーを基準としていますが、剪定の基準は誤分類率 (misc) としています。ただし、追加の条件として $N \leq 6$ とありますので、各葉 (leaf) におけるサイズが6以下となる場合には、剪定のステップは行われません。HPSPLIT プロシジャを実行しますと、実行モードの情報のみが結果として出力されます。

図5. 実行モードに関するアウトプット

HPSPLIT プロシジャ	
パフォーマンスの情報	
実行モード	Single-Machine
スレッド数	4

各ノードにおける統計量はOUTPUTステートメントを用いてデータセットに出力することで確認できます。この例では、OUTPUTステートメントではなく、RULESステートメントを用いて、イングリッシュ・ルールをテキストファイル (図6) と出力しています。

図6. イングリッシュ・ルール (一部) のアウトプット

```

*-----*
NODE = 2
*-----*
DEROG IS ONE OF 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
AND DELINQ IS ONE OF 1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 15
  PREDICTED VALUE IS 1
  PREDICTED 1 = 0.6996( 163/233)
  PREDICTED 0 = 0.3004( 70/233)
*-----*
NODE = 3
*-----*
DEROG IS ONE OF MISSING, 0
AND DELINQ IS ONE OF 1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 15
  PREDICTED VALUE IS 0
  PREDICTED 1 = 0.353( 251/711)
  PREDICTED 0 = 0.647( 460/711)
*-----*
NODE = 7
*-----*
DEROG IS ONE OF 1, 2, 3, 4, 6, 7, 8, 9, 10
AND DEBTINC IS MISSING OR DEBTINC <45.137782
AND CLAGE IS MISSING OR CLAGE <186.91737
AND DELINQ IS ONE OF MISSING, 0
  PREDICTED VALUE IS 0
  PREDICTED 1 = 0.419( 88/210)
  PREDICTED 0 = 0.581( 122/210)
*-----*
NODE = 8
*-----*
DEROG IS ONE OF MISSING, 0, 5
AND DEBTINC IS MISSING OR DEBTINC <45.137782
AND CLAGE IS MISSING OR CLAGE <186.91737
AND DELINQ IS ONE OF MISSING, 0
  PREDICTED VALUE IS 0
  PREDICTED 1 = 0.1576( 316/2005)
  PREDICTED 0 = 0.8424( 1689/2005)

```

各ノードにおける分割のルールとともに、学習データにおける予測値がリストされています。

また、この例題では、SCOREステートメントを用いて、各オブザベーションに対する予測値を算出し、データセットSCORED (図7) に出力しています。

図7. オブザベーションに対する予測確率値 (一部) のアウトプット

OBS	BAD	_NODE_	_LEAF_	P_BAD1	P_BAD0	V_BAD1	V_BAD0
1	1	8	3	0.15761	0.84239	0.13760	0.86240
2	1	3	1	0.35302	0.64698	0.40306	0.59694
3	1	8	3	0.15761	0.84239	0.13760	0.86240
4	1	8	3	0.15761	0.84239	0.13760	0.86240
5	0	8	3	0.15761	0.84239	0.13760	0.86240
6	1	8	3	0.15761	0.84239	0.13760	0.86240
7	1	2	0	0.69957	0.30043	0.67213	0.32787
8	1	8	3	0.15761	0.84239	0.13760	0.86240
9	1	3	1	0.35302	0.64698	0.40306	0.59694
10	1	8	3	0.15761	0.84239	0.13760	0.86240

ここでは、各オブザベーションがどのノードに分類されているかが、変数 _NODE_ に示されています。この値は図6におけるイングリッシュ・ルールのNODEに該当しています。学習データにおける予測値P_BAD1、P_BAD0とともに、検証データにおける従属変数BADの比率がV_BAD1、V_BAD0として算出されています。

4 その他のプロシジャ

HPAプロシジャを含むプロダクト、SAS/STATをインストールした場合、これまでご紹介したプロシジャの他、いくつかのHPAプロシジャを実行できます。

● データセットの要約

データセットにおける変数の情報、要約統計量の算出に、HPDMDBプロシジャ、HPSUMMARYプロシジャがあります。また、変数間の相関を探索、検証するため、HPCORRプロシジャにて相関統計量を求めることができます。

● データセットの加工

元データセットから標本抽出を行う場合、HPSAMPLEプロシジャがあります。このプロシジャでは単純ランダム抽出、層別ランダム抽出の2手法に対応しています。

データモデリングのステップでは、データセットの準備、加工があります。その一つには欠損値の補完があり、HPIMPUTEプロシジャで実行できます。連続変数の欠損値に対し、平均値、中央値、ランダムな値にて補完します。また、他の方法として、連続変数をいくつかの区間 (ビン) に区切り、ビン化することが挙げられます。この加工についてはHPBINプロシジャが対応しています。

5 参考文献

SAS/STAT 12.3のインストールでシングルマシンにて実行できるHPAプロシジャについて、米国にて開催された SAS Global Forum 2013にて紹介されており、以下の文献にて詳細を参照できます。

High-Performance Statistical Modeling

<http://support.sas.com/resources/papers/proceedings13/401-2013.pdf>

また、SAS High-Performance Statisticsのドキュメントは以下にて参照できます。

<http://support.sas.com/documentation/cdl/en/stathpug/66410/HTML/default/viewer.htm>

6 おわりに

HPAプロシジャにて、対応している変数選択の手法が多くなっていること、また、対応しているモデルに一般化線形モデルが加わり、拡張していることをご紹介しました。また、HPREGプロシジャの例題を用いて、既存プロシジャと構文、出力結果などがほぼ同じとなっていることを例示させていただきました。分析手法を検討、選択する上で参考となれば幸いです。

また、この特集では、シングルマシンでの実行としてご紹介しましたが、より大きなデータ、ビッグデータ、よりパワフルな分散コンピューティング環境を考慮する上でも参考になれば幸いです。



SAS SAS パートナーネットワーク Partner Network

Special Interview

データから価値を生む 富士通のビッグデータイニシアティブ

—— 第4回
富士通株式会社

連載第4回目を迎える今回は、東京・浜松町にある富士通トラステッド・クラウド・スクエアに、SAS Technical News 編集部がうかがいました。執筆いただいた方々にこの1年間の連載を振り返っていただくとともに、富士通がSAS Visual Analytics (以下、SAS VA) を活用して行っている先進的な取り組みや導入事例について、お話をお聞きました。



富士通株式会社
鶴田 晃氏 (写真左) 平井 義浩氏 (写真中央)
株式会社富士通ビー・エス・シー
置田 淳夫氏 (写真右)

1

検証で見えてきた SAS Visual Analytics の実際

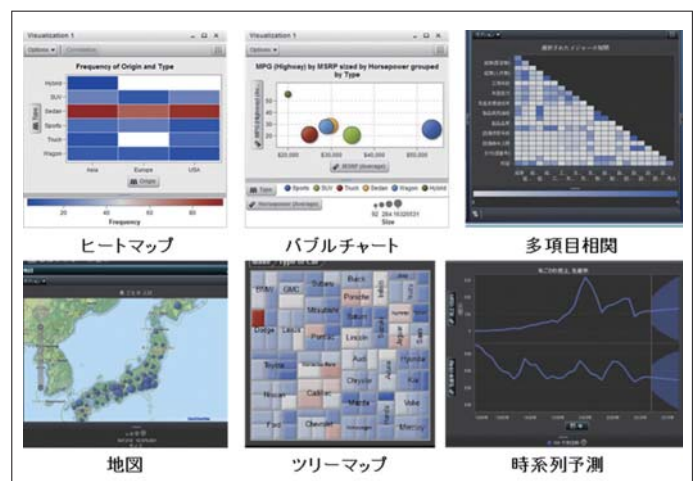
——1年間、本SAS Technical Newsにて、SAS VAのインストールから分析に至るフェーズまでを検証していただき、ありがとうございました。実際にSAS VAを扱ってみた感想はいかがでしたか。

置田 SAS VAは、誰もがとりあえず触ってみることができるので、前提知識がなくても、おそらく何かしら動かすことができる製品です。この使いやすさのおかげで、SASの敷居がグッと下がるのではないかと思います。そういう意味で、これまでのSASとは異なる製品だなと感じました。

平井 SASソリューションを拡販するために、セミナーなどへ出展していますが、お客様に分析についてお話するだけではなかなか足を止めていただけません。ですが、SAS VAのグラフィカルな部分をご紹介しますと、お客様も足を止め、具体的な活用について会話がはずみます。この1年間は、どうやってこのSAS VAをお客様に紹介していったらいいのか、検証を含めて、自分たちなりに整理するための良い機会だったと思います。

鶴田 SAS VAは、イベントでお客様への訴求力はかなりあります。ビジュアルが抜きん出ているので、お客様が、実際に使うシーンをイメージしやすい製品だと思います。

お客様の多くは、いままでのSASに対してハードルが高いというイメージを持っていますから、そこを切り崩すのに最適な製品ですね。弊社の営業やSEも、SASの提案を積極的に盛り込むようになったのも、それが影響しているのではないかと思います。



SAS VA のグラフィカルな図

——では反対に、実際に触れられてみた上で感じた、SAS VAのウィークポイントなどはあるでしょうか。

置田 ■メモリに載る分しかデータが扱えない点でしょうか。分析のためには非圧縮でなければいけないという話はわかりますが、扱えるデータのボリュームが少なくなるのは、お客様も気になるところだと思います。例えば、非圧縮/圧縮の2モードを用意するなどすれば、なお良いのではないのでしょうか。

平井 ■ファイルに関して言えば、SAS VA単体の場合、日本語が混在するファイルが扱いにくいことでしょうか。SAS Enterprise Guide (以下、EG) と組み合わせると違和感なく実行できるので、EGと合わせてご紹介したりもしています。

——日本語ファイルの扱いについては、SAS本社ももう少し洗練されるべき点であると認識しており、SAS VAの開発者に日本人も関わるなどしています。

鶴田 ■SAS VAは、従来のビジネスインテリジェンス (以下、BI) プラスαの「α」部分であると理解すれば分かりやすいのではないのでしょうか。SAS VAのData Explorerは素晴らしい機能ですが、日本でのBIの利用シーンは定型レポートの作成がほとんどだと考えると、そうした機能を使いこなせるような場面はまだ少なく、これから増えていくのではないかと思います。

——たしかに、米国やヨーロッパでは、データ探索が普通のビジネスで行われていますが、日本ではようやくいま、そうした状況になりつつあるということなのかもしれません。

置田 ■ご提案の際にはあえて分析と言わずに、データを理解することの大切さやメリットをご紹介する方が、お客様の検討が進むこともあります。

2

ビッグデータ？ アナリティクス？ お客様の課題はいま

——お客様はいま、ビッグデータやアナリティクスというものに対して、どのようにしたいと考えていらっしゃるのでしょうか。セミナーやイベントで聴こえてくる声はどのようなものですか。

平井 ■セミナーやイベントには多くのお客様にご参加いただいておりますが「ビッグデータを使って、こういうことをやりたい」と具体的な活用方法をお持ちの場合はまだ少ないのが実情です。

アナリティクスというキーワードにしても似た状況です。「そろそろビジネスインテリジェンス (以下、BI) で集計だけじゃなく、一歩先に行きたい」と漠然と思われているような場合でも、手法やツールが先行しているものの、具体的に何をしたいか分からない場合が多いですね。そのため、私たちとしてはまず、お客様が本当に困っていることを聞き出して、弊社の最新製品やSASの製品を組み合わせ、それをどのように具体的に解決していくかを探っていきます。ビッグデータや

アナリティクスというキーワードは、その結果付いてくる言葉に過ぎないのです。

鶴田 ■ビッグデータに関連する技術は、ハードウェアからソフトウェア、そして分析手法に至るまで、多方面にわたります。そのようなお客様のさまざまな課題に対して富士通のビッグデータの取り組み方を明確に

し、一本の窓口で受けるため

に、ビッグデータイニシアティブ

センターを立ち上げています。

お客様が何をやりたいのか悩ん

でいらっしゃる時には、仮説

立案ワークショップを開き、課

題が何かを一緒に考えます。

「課題はない」とおっしゃるお

客様が時折いるのですが、そん

なことはないんですね。「ビッグ

データで何かやりたい」とおっ

しゃっている場合や上席から指

示を受けている場合、いずれの

場合もその背景を丹念に探れ

ば、課題が必ず見つかるのです。



富士通株式会社
総合商品戦略本部
ビッグデータイニシアティブセンター
鶴田 晃 氏

3

データディスカバリーを実践する 山形大学でのチャレンジ

——引き合いがある一方、なかなか具体的な進展がない中で、SAS VAの導入が決定しました。

平井 ■はい。山形大学様に対してですね。導入先は、大学をどのようにマネジメントしていくかを考えるエンロール・マネジメント部門です。ここでは、学生の成績や出身高校などさまざまな情報を入力して、教育改善のためのマネジメントに活かそうとしています。

そこで発見したのは、グラフをアドホックに操作するだけでも、ディスカッションにもものすごいインパクトを与えるということでした。

導入にあたっては、大学のデータをお借りして、SAS VAのデモンストレーションを行いました。年度ごとに学生の出身高校の割合を示す棒グラフやバブルチャートを作成し、ディスカッションに参加することになったのですが「入試制度を変えたら、出身高校はどのように変化するのだろうか」などと、大学運営を知っている方々ならではの視点で、リアルタイムにグラフを変化させて比較・検討を行いました。その結果、思いのほかディスカッションが盛り上がったのです。まさに、データディスカバリーの瞬間を見た思いがしました。

単にグラフを作成するだけなら、お馴染みのスプレッドシートでも可能です。ですが、だんだんそのグラフをつくること自体が目的になったり、完成までにタイムラグが生じてしまったりするなど、本来の目的を

見失いがちです。

業務に詳しい方々が、SAS VAでデータをアドホックに見ていただくだけでも、SAS VAの面白さを十分に分かっていただけるのだなと感じました。

鶴田 ■ 思考のスピードを妨げずに、考えていることをそのまま確認することができるSAS VAのような探索型の製品はそうはありません。そこからさらに、新たな気づきを得ることも可能になります。決まりきった思考ルーチンに沿って、ダッシュボードからレポートを作成するようなBIツールとは対局にありますね。



富士通株式会社
総合商品戦略本部
ビジネスアプリケーション
推進統括部
経営ソリューション推進部
平井 義浩氏

企業の内外にあるビッグデータから情報を収集する技術と、SAS製品によるETLなど、アナリティクスのためのデータ加工などの技術を組み合わせ、どのようなデータ活用ができるかを紐づけてSAS VAをお客様にご紹介することが、私たちのミッションだと思っています。

——まさに、ハードウェアの製品を持っている企業であり、SIerとしてのノウハウを持っている企業ならではのアプローチですね。

置田 ■ SIの部分としてはいま、ビッグデータをパフォーマンスよく扱うための観点に比重が大きくなってきており、サーバやストレージに焦点が当たっています。

以前は、SASをPC 1台で動かしたり、最小限のサーバ構成で運用したりしていましたが、それではパフォーマンスは向上しません。ビッグデータの活用には基盤が疎かにできなくなっているのです。

——お客様からすると、自社のシステムでできるか、新しいシステムはどれだけ性能を出せるのか、実際に試用したいケースがあると思います。その点、富士通トラステッド・クラウド・スクエアにある、ビッグデータイニシアティブラボは、お客様が導入時に検証を行うことができるなど、大変利便性の高い施設ですね。

平井 ■ 担当者レベルでSAS製品を導入したいと考えても、社内で価値を認められなければ話が進みません。「自分たちが思ったことを、こうすればできる」と自分たちで話すことができれば、上申しやすいでしょう。そのような強い思いを抱くお客様にこのデータセンターの環境を提供していますが、大変喜ばれています。

駅に直結しているというアクセスしやすさもさることながら、リモート接続であれば24時間ご利用いただけますので、一晩かかるバッチ処理を半分にするための検証なども柔軟に行うことができます。

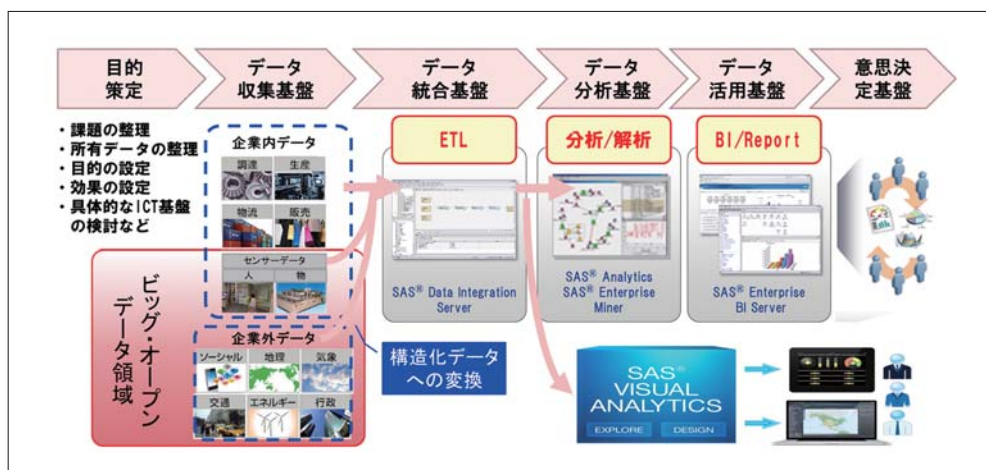
鶴田 ■ セキュリティに関しても、さまざまなレベルのご要望に応じています。リモート接続に抵抗があるお客様には、データをお預かりして検証するためのクローズドな環境もありますし、静脈認証で入室する

4

導入にあたり柔軟に検証できる ビッグデータイニシアティブラボ

——ご提案の際には、SAS VAだけではなく、お客様が既に使われているハードウェアやデータ収集のためのソフトウェアなど、トータルにコーディネーションされているのでしょうか。

平井 ■ 「データをさまざまな軸で分析したいので、SAS VAを導入したい」と漠然とおっしゃるお客様がいらっしゃいますが「さまざまな軸」という自由を確保するには、それを可能にするデータ構成でなければなりません。そこが弊社のノウハウの活きる場所となります。



SAS VA を利用したビッグデータ分析を実際に運用するための全体像

部屋も一部ご用意しています。

また、富士通のハードウェアもひと通り揃えています。最新UNIXサーバであるSPARC M10では、Solarisであれば複数の環境を用意することも可能です。SAS製品の性能を試験したときも、3つのモードで比べました。連載でのSAS VAの検証もこのセンターで行いました。第1回で言及した検証環境の、4ブレード構成パターンと8ブレード構成パターンの比較なども、ここで対応しました。

置田 ■ 専任のSEが常について、ネットワークを組む体制になっております。「ハードディスク、ディスク装置、ストレージが欲しい」とご要請があれば「では付けましょう。どういうふうに繋げばいいですか」とヒアリングしながら、進めています。

購入後にストレージが欲しいと思っても、予算取りや上申から始めなければならず、なかなかスムーズには進みません。このセンターを活用する

ことで、実際に装置を付けたり外したりすることを試しながら考えられる絶好の機会になると思います。おかげさまで、多くのお客様からご相談をいただいています。

—— お客様に対するサポートやハードウェアもかなり充実しているようですね。

鶴田 ■ ハードウェア、ソフトウェア、それらを導入するシステムインテグレーションをトータルで提供し、実運用までサポートするのが富士通の役目と考えています。



株式会社富士通ビー・エス・シー
第二システム本部
第二システム事業部
第三システム部

置田 淳夫 氏

5

人と情報を ICT で捉え、 新たな価値を生むアナリティクスへ

—— 今後、御社ではビッグデータに対してどのような取り組みを進めていく予定でしょうか。

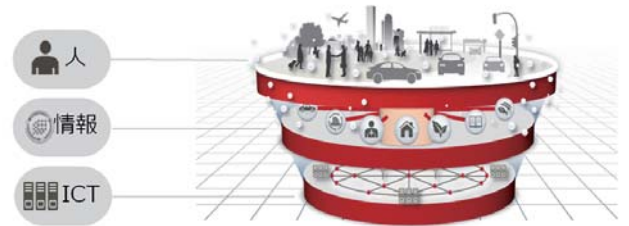
鶴田 ■ 富士通では、ヒューマンセントリック・インテリジェントソサエティというビジョンを掲げています。それは「人」と「情報」と「ICT」という3本の柱に基づいて、世の中を捉えようとするものです。

社会に存在する人々が、可能性を最大限に発揮してイノベーションを生み出し、安心・安全に暮らしながら、情報が新たな価値を生み出していくことで、社会が持続的に成長していく世界。このような世界を富士通のICTが支えます。

平井 ■ 人が社会の中で生きていけば、さまざまな情報が生まれます。

ヒューマンセントリック・インテリジェントソサエティ

「人々が可能性を最大限に発揮してイノベーションを生み出し、安心安全に暮らし、そして情報が新たな価値を生み出し、社会が持続的に成長していく世界」



ヒューマンセントリック・インテリジェントソサエティ

このような大量の情報、つまりビッグデータをICTの部分であるネットワークで捕捉し、新たな価値を生むものとして活用できないかと弊社では考えています。

その中核になるのは、やはり、データ分析や予測モデル構築といったアナリティクスです。特にSAS VAは、これからビッグデータ・アナリティクスを始めようとするお客様に使っていただくのがおすすめです。

お客様をピラミッド型として捉えるとすると、いちばん上には、既に実践されているお客様、次に、具体化されようとしているお客様、そして、課題となるキーワードを漠然と持たれていて、セミナーに出るなど何かしらアクションを起こそうとしているお客様、さらに、まだまだそこまで至っていないお客様が、それぞれいらっしゃるのではないのでしょうか。ビッグデータとSAS VAを組み合わせることで、私たちは3番目の何かしらアクションを起こそうとしているお客様に対して、お客様の課題をご一緒に発見しながら、データの活用をご提案していきたいと思っています。

—— そうしたお客様は、近い将来、ビッグデータをビジネスで扱う可能性が高いというわけですね。

鶴田 ■ 課題の発見には、SAS VAのような探索型の製品が最適です。ビッグデータイニシアティブラボでは、お客様にSAS VAを利用いただける環境をご用意していますので、ご利用いただければと思います。



上/富士通トラステッド・クラウド・スクエア
下/ビッグデータイニシアティブラボ

—— 御社の幅広い取り組みによって、多くのお客様がデータから価値を見つけることが可能になるのでしょうか。

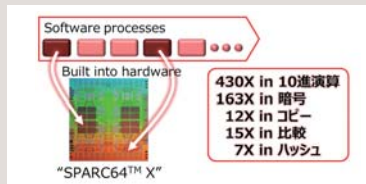
本日は、ありがとうございました。

富士通の次世代サーバ「SPARC M10」が 実現する、超高速データ分析環境

ビッグデータ時代と言われる昨今、爆発的に増加する多種多様なデータを蓄積し分析を行うために、「分析結果をより短時間で得たい」「より多くのパターンでシミュレーションを実施したい」といったお客様の声が増えており、サーバに対する要求はますます高まっています。このような背景の下、富士通は最先端技術により高速性と柔軟性、堅牢性を兼ね備えた最新 UNIX サーバ「SPARC M10」を提供しています。ここでは、ビッグデータ時代に最適な分析環境として、SPARC M10 と SAS の組み合わせをご紹介します。

最新鋭の高性能プロセッサ「SPARC 64 X」を搭載

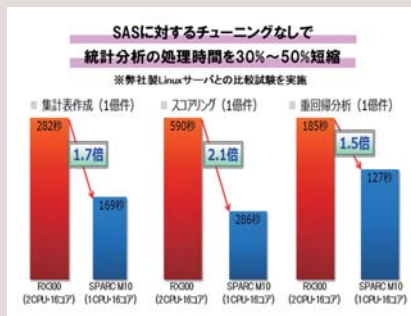
SPARC M10 は、処理高速化技術「Software on Chip」やメモリアクセス高速化技術「System on Chip」など、富士通がスーパーコンピュータで培った独自の技術を活用した高性能プロセッサ「SPARC64™ X」を搭載しています。これにより、従来に比べ飛躍的に高い演算性能や、高信頼性・省電力を実現しています。



SPARC M10 + SAS による大幅な処理性能向上を検証済み

この高性能プロセッサにより、SASの分析環境を大幅に高速化することができます。弊社では、SPARC M10+SAS 環境にて実際にSASプログラムを実行し、その高速性を検証しました。その結果、ほぼ同スペックのCPUと比較し、SASのプログラムやサーバのチューニングを行うことなく性能が1.5倍以上向上しました。

上記の性能向上に加え、実際の運用ではM10の高速メモリ・大量コアを活用するチューニングにより更なる高速化を図ることができます。



超高速データ分析環境により、ビッグデータ分析の課題を解決

SPARC M10+SASによる高速分析環境は、現行の業務の分析やビッグデータ分析にて直面する以下の課題に効果的な解決策となります。

- ベイズ分析などの複雑な分析について、処理に時間がかかり意思決定が遅れる。
- パターンを減らしてシミュレーションを実施しているため、精度が上がらない。
- データ量が膨大なため、夜間バッチ処理のスケジューリングが困難である。

このような課題をお持ちのお客様は、是非弊社までお問い合わせください。

SPARC M10について、詳細は下記のページをご参照ください。

<http://jp.fujitsu.com/platform/server/sparc/>

以上、4回に渡って続いた本連載は今回で終了となります。最後までお読みいただき誠にありがとうございました。

富士通は今後もビッグデータやSASの導入に関する情報を発信してまいりますので、ご興味のある方は是非お問い合わせください。また、富士通主催のBI/BAセミナーやSAS共催セミナーについても詳細が決定しだい下記のサイトでご紹介いたしますので、ご覧いただけますと幸いです。

<http://jp.fujitsu.com/solutions/ba/>

お問い合わせ

富士通株式会社 統合商品戦略本部
ビジネスアプリケーション推進統括部 経営ソリューション推進部
Tel: 03-6424-6648 受付時間 9:00 ~ 17:30
(土・日・祝日・当社指定の休業日を除く)
<http://jp.fujitsu.com/solutions/ba/>

FUJITSU

SAS アカデミック・ニュース ダイジェスト Academic News

Digest

Web版 アカデミック・ニュースのご紹介

随想 「マーケティングとデータ解析」
第6回 六本木のデータ解析サロン



SAS Technical News 2013 Winterまで、SAS Technical NewsにはAcademic Newsのコーナーがありました。

本誌で教育現場におけるSASの活用に関する記事を目にしたご記憶がある方もいらっしゃるのではないのでしょうか？

SAS Technical News 2013 SpringよりAcademic NewsはWeb版へ移行しましたが、読者様の熱いご要望と、記事のなかにはビジネスの現場においてもヒントになる内容が含まれていることから、Academicと並行しTechnical Newsでもご紹介する運びとなりました。

今回の記事では、本年5月からSAS Institute Japan東京本社で開催されたSASマーケティング研究部会の報告をご紹介します。本文のSASマーケティング研究会とは、本年5月に朝野先生を中心として設立した研究会です。研究テーマは、マーケティング活動を効果的に実行することを目的とし開催時毎にテーマを変え議論しています。たとえば、第6回目となる2013年最後の研究会では、ベイズ統計がテーマです。

記事掲載ページ：

随想「マーケティングとデータ解析」

第6回 六本木のデータサロン

<http://www.sas.com/jp/industry/academic/fieldvoice/columAsano/06.html>

Technical Newsでの連載は今号からの掲載となりますが、前号までの記事は、弊社Webページ「教育の現場から」より参照可能となっております。ぜひ一読ください。

<http://www.sas.com/jp/industry/academic/fieldvoice/index.html>



Web版 アカデミック・ニュース 「教育の現場から」

寄稿
随想「マーケティングとデータ解析」

朝野 照彦
中央大学客員教授

【経歴】
千葉大卒業後、千葉大・筑波大講師、専修大・都立大教授を経て中央大学および多摩大学大学院客員教授。学習院マネジメントスクール顧問、日本マーケティング・サイエンス学会論文誌編集委員、日本行動計量学会理事。著書は『ビッグデータの使い方・活かし方』東京図書(編著)、『マーケティング・リサーチ』『最新マーケティング・サイエンスの基礎』『入門共分散構造分析の実際』『入門多変量解析の実際第2版』以上講談社など多数。

SAS Certification

SAS グローバル認定プログラムにチャレンジ! 第3回

『SAS グローバル認定プログラムにチャレンジ!』のコーナーも3回目になりました。

前回の例題は全問正解できたでしょうか?

今回は、少しレベルアップした「SAS Advanced Programmer for SAS® 9」へのチャレンジです。

この認定試験は、前回までご紹介した SAS Base Programmer for SAS® 9 を取得した方向けの、次のレベルの認定プログラムです。

まずは、認定試験の概要をご紹介した後、例題をいくつか出題します。

馬力で勝負!!
知力

《正念馬》

SAS® 認定プロフェッショナル SAS Advanced Programmer for SAS® 9について

概要

SAS® 認定プロフェッショナル SAS Advanced Programmer for SAS® 9 の認定には、SAS を使用したプログラミングとデータ・マネジメントの経験が必要です。

この認定資格の取得には、以下のようなスキルや知識が必要とされます。

- 高度な DATA ステップ・プログラミングと効率化のテクニックを使用した問題解決
- SAS SQL コードの記述と解釈
- SAS マクロ機能の作成と利用
- SAS® 9 で追加された拡張機能や新機能の知識

この試験の制限時間は2時間で、問題は63問出題されます。

正答率65%で合格です。

前回までご紹介した SAS Base Programming for SAS® 9 の試験と、この SAS Advanced Programming for SAS® 9 の試験の両方に合格された方が、『SAS® 認定プロフェッショナル SAS Advanced Programmer for SAS® 9』として認定され、認定証が発行されます。

SAS Advanced Programmer for SAS® 9の例題にチャレンジ!

それでは早速、例題にチャレンジしてみましょう。

解答は、P23 に記載されています。

01 CHALLENGE

SAS Advanced Programmer for SAS® 9

SAS データセット ONE と TWO があります。

SAS データセット ONE			SAS データセット TWO		
YEAR	QTR	BUDGET	YEAR	QTR	SALES
2001	3	500	2001	4	300
2001	4	400	2002	1	600
2002	1	700			

次の SAS プログラムがサブミットされました。

```
proc sql;
  select one.*, sales
         from one right join two
         on one.year = two.year;
quit;
```

生成されるレポートは、次の4つのうちどれになるでしょうか?

A

YEAR	QTR	BUDGET	SALES
2001	3	500	.

B

YEAR	QTR	BUDGET	SALES
2001	4	400	300
2002	1	700	600

C

YEAR	QTR	BUDGET	SALES
2001	3	500	.
2001	4	400	300
2002	1	700	600

D

YEAR	QTR	BUDGET	SALES
2001	3	500	300
2001	4	400	300
2002	1	700	600

02 CHALLENGE

SAS Advanced Programmer for SAS®9

次のプログラムがサブミットされました。

```
%let type = RANCH;
proc sql;
  create view houses as
  select *
  from sasuser.houses
  where style = "&type";
quit;
%let type = CONDO;
proc print data = houses;
run;
```

このプログラムの実行後に生成されるレポートでは、STYLEの値がRANCHのオブザベーションが表示されます。

次のWHERE句の関数のうち、どれを使うと現在のマクロ変数TYPEの値を展開できるでしょうか？

A
GET

B
SYMGET

C
%SYMGET

D
&RETRIEVE

03 CHALLENGE

SAS Advanced Programmer for SAS®9

SASデータセットONEとTWOがあります。

SASデータセットONE

NUM	CHAR1
1	A
2	B
4	D

SASデータセットTWO

NUM	CHAR2
2	X
3	Y
5	V

次のSASプログラムをサブミットして、出力テーブルTHREEを作成しました。

```
data three;
merge one (in = in1) two (in = in2);
  by num;
run;
```

SASデータセットTHREE (出力テーブル)

NUM	CHAR1	CHAR2
1	A	
2	B	X
3		Y
4	D	
5		V

次のSQLプログラムのうち、どれを実行すればSASデータセットTHREEと同じものを作成できるでしょうか？

A

```
proc sql;
create table three as
select *
  from one full join two
  where one.num = two.num;
quit;
```

B

```
proc sql;
create table three as
select coalesce(one.num, two.num)
  as NUM, char1, char2
  from one full join two
  where one.num = two.num;
quit;
```

C

```
proc sql;
create table three as
select one.num, char1, char2
  from one full join two
  on one.num = two.num;
quit;
```

D

```
proc sql;
create table three as
select coalesce(one.num, two.num)
  as NUM, char1, char2
  from one full join two
  on one.num = two.num;
quit;
```

04 CHALLENGE

SAS Advanced Programmer for SAS® 9

変数 DATE に対する単一インデックスと、インデックス化されていない変数 REVENUE が含まれる SAS データセット、SASDATA.SALES があります。

次の SAS プログラムのうち、インデックス DATE が使用されると考えられるのはどれでしょうか？

A

```
proc print data = sasdata.sales;
  by date;
run;
```

B

```
proc print data = sasdata.sales;
  where month(date) = 3;
run;
```

C

```
data march;
  set sasdata.sales;
  if '01mar2002'd < date < '31mar2002'd;
run;
```

D

```
data march;
  set sasdata.sales;
  where date < '31mar2002'd or revenue >
50000;
run;
```

もっと学んでみたい方へ

今回ご紹介した、SAS® 認定プロフェッショナル SAS Advanced Programmer for SAS® 9 はいかがでしたでしょうか？

この認定プログラムについてもっと学んでみたい方、また、受験してみたいという方は、ぜひ弊社で開催しているトレーニングをご活用ください。

SAS グローバル認定プログラムに関する詳細は、次の Web サイトをご確認ください。

<http://www.sas.com/jp/training/certify/index.html>

SAS Advanced Programmer for SAS® 9 を受験する前に、以下のコースの受講をお勧めします。

SAS® 認定プロフェッショナル SAS Base Programming for SAS® 9

SAS マクロ言語 1 :
必須要素

+

SAS SQL1 : 必須要素

+

SAS プログラミング 3 :
上級テクニックと効率化

SAS Advanced
Programmer
for SAS® 9

トレーニングの詳細や関連マニュアルに関する情報は、こちらの Web サイトに記載されています。

<http://www.sas.com/jp/training/certify/benefits/prep.html>

Q&A



- Stored Process Web Application への暗号化パスワード利用
- オブザベーション番号の出力方法
- 実行ログを任意のタイミングで保存する
- ユーザ定義フォーマットのUTF-8セッションエンコーディング環境への移行
- 変数長を変更した際に表示されるWARNINGメッセージを制御する方法
- カイ2乗検定における警告メッセージの出力
- OLAP Server サービスの停止について
- INTCK関数の基準値について
- SASプログラムをEGのプロセスフローへ移行

Q Stored Process Web Applicationを利用する際にURLへそのままパスワードを含めて利用しようとしています。パスワードを平文で利用はしたくありませんのでSAS®で暗号化済みのパスワードを利用したいと考えています。どうすれば暗号化したパスワードを含めたURLで正しく表示させられるでしょうか。

A 標準では暗号化したパスワードをそのままでは利用できません。管理コンソールからの設定にて「Stored Process Web App 9.2 (もしくは9.3)」の拡張設定で「AllowEncodedPassword」属性を有効にし、URL エンコーディングを考慮した文字列へ変換しての利用が必要です。

この件につきましては次のNoteにて紹介しております。
※Noteでは9.2の例が記載されています。

Usage Note 36831: SAS Stored Process Web Application does not accept encoded passwords on the _PASSWORD parameter by default
<http://support.sas.com/kb/36/831.html>

AllowEncodedPasswordオプションについては次のマニュアルに記載がございます。

SAS® Integration Technologies
<http://support.sas.com/documentation/onlineoc/inttech/index.html#inttech93>

SAS 9.3 Stored Processes: Developer's Guide Configuring the SAS Stored Process Web Application
<http://support.sas.com/documentation/cdl/en/stpug/62758/HTML/default/viewer.htm#webappcfg.htm>

Q DATA ステップの SET ステートメントと WHERE ステートメントを使用し、データセットから条件抽出を行います。このとき、条件にマッチして読み込まれたオブザベーションが、データセットの何オブザベーション目だったかを確認できる簡単な方法はありますか。

A SAS® 9.4 より SET ステートメントに追加された新しいオプション CUROBS= を使用すると、読み込まれたオブザベーションの番号を指定の変数に出力することができます。サンプルプログラムを右に示します。

例

```
DATA WOMEN;
  SET sashelp.class CUROBS=cobs;
  WHERE sex = 'F';
  orig_obs = cobs;
RUN;

PROC PRINT DATA=WOMEN;
RUN;
```

以下のような結果となります。

Obs	Name	Sex	Age	Height	Weight	orig_obs
1	Alice	F	13	56.5	84.0	2
2	Barbara	F	13	65.3	98.0	3
3	Carol	F	14	62.8	102.5	4
4	Jane	F	12	59.8	84.5	7
5	Janet	F	15	62.5	112.5	8
6	Joyce	F	11	51.3	50.5	11
7	Judy	F	14	64.3	90.0	12
8	Louise	F	12	56.3	77.0	13
9	Mary	F	15	66.5	112.0	14

Q SAS® の実行ログを任意のタイミングで外部ファイルへ保存するタスクをプログラムで実現することは可能でしょうか。現在は、ログウィンドウの [ファイル] → [名前を付けて保存] からファイル名を指定して保存していますが、これをプログラムに置き換えられれば作業が軽減されます。

A DM ステートメントの中で、ログウィンドウに対して FILE コマンドを実行することで実現可能です。

```
DM "LOG; FILE 'c:%temp%test1.log' ";
```

また、次のように動的なファイル名を設定することも考えられます。

```
DM "LOG; FILE 'C:%temp%test_%SYSFUNC(DATE(),yyymmdd8.)_%TRIM(%SYSFUNC(time()),32..).log' ";
```

Q 今までは、SASをセッションエンコーディングSJISで使用していました。セッションエンコーディングUTF-8の環境で使用することになり、データセットのトランスコードは行っておりますが、ユーザ定義フォーマットのトランスコードはどのようにすればよいでしょうか。

A フォーマットカタログにつきましては、エンコーディングの概念がございせん。他方、PROC FORMATでは、CNTLIN=, CNTLOUT=オプションを用いて、ユーザ定義フォーマットの情報をデータセットに出力することが可能です。一度、データセットに出力した後、トランスコードすることができます。

(1) SAS SJISセッションエンコーディングで、制御データセットの作成

```
LIBNAME sjisfmt "c:%temp%sjisfmt";

                                /** サンプルフォーマットの作成 **/
PROC FORMAT LIB=work;
  VALUE testfmt
    1="男";
RUN;

                                /** CNTLOUT=にて、制御データセット化 **/
PROC FORMAT LIB=work CNTLOUT=sjisfmt.cntlout;
RUN;
```

(2) SAS UTF-8セッションエンコーディングにて、制御データセットのカスタマイズ、再作成

```
LIBNAME sjisfmt "c:%temp%sjisfmt" CVMULT=1.5 INENCODING=sjis;

/** 1) で作成した制御データセットを WORK.FORMATS に取り込む **/
PROC FORMAT CNTLIN=sjisfmt.cntlout LIB=work;
RUN;

                                /** 再度、制御データセットとして、出力 **/
PROC FORMAT CNTLOUT=work.cntlout LIB=work;
RUN;

/** ここで、SJIS->UTF-8に変更する上で、日本語の長さの変更がある **/
/** (今回の例では、"男" が 2->3 バイトになる **/
/** フォーマットの中では、フォーマット値の長さ、フォーマット値を **/
/** 指定しない場合のデフォルトの長さの情報も含まれている **/
/** 制御データセット内では、変数 LENGTH、DEFAULT がそれぞれ該当 **/
/** する。DEFAULT は SJIS で作成した際の長さとなっているため、LENGTH **/
/** と同じ値にする **/

DATA work.cntlout2;
  SET work.cntlout;
  DEFAULT=length;
RUN;

/** カスタマイズした制御データセットをフォーマットカタログに取り込む **/
PROC FORMAT CNTLIN=work.cntlout2 LIB=work;
RUN;

                                /** 結果の確認 **/

DATA _null_;
  a=1;
  PUT a testfmt.;
RUN;
```

Q 下記のサンプルプログラムのように、DATA ステップにおいて、SET ステートメントで指定したデータセット内の変数と同一の変数名の長さを変更した場合、ログに表示される以下のメッセージを出力させない方法はありませんか。

WARNING : 入力データセットによって、変数 TEST に複数の長さが指定されました。データの切り捨てが発生します。

サンプルプログラム

```
DATA test;
  LENGTH name $5;
  SET sashelp.class;
RUN;
```

A 9.2以降のリリースにおいて、変数の長さが変わった場合にその旨をお知らせするWARNINGが表示されるようになりました。詳細に関しては、下記のアナウンスが公開されております。

Problem Note 31850: Using the LENGTH statement might produce a warning and rc=4

<http://support.sas.com/kb/31/850.html>

SAS® 9.2 TS1M0では、上のページの「hotfix」タブをクリックして表示されるリンクから、hotfixをダウンロードいただき、適用することでWARNINGを表示しない動作に戻すことができます。

SAS® 9.2 TS2M0以降では、VARLENCHK=システムオプションを設定することにより、WARNINGを表示するか否かを制御することが可能となりました。SAS® 9.1.3 同様にWARNINGを非表示とするためには、次のように指定します。

```
OPTIONS VARLENCHK=nowarn;
```

なお、VARLENCHK=システムオプションの詳細については、下記URLのオンラインヘルプをご確認いただけますようお願い申し上げます。

VARLENCHK= System Option

<http://support.sas.com/documentation/cdl/en/lrdict/64316/HTML/default/viewer.htm#a003269301.htm>

Q FREQ プロシジャの TABLES ステートメントにて、CHISQ オプションを指定した際、カイ 2 乗検定に関する WARNING メッセージが表示されることがあります。

```
WARNING : セルの xx% において、期待度数が 5 より小さくなっています。
          カイ 2 乗検定は妥当な検定でないと思われる。
```

このメッセージを非表示にできますか。また、WARNING メッセージの有無をデータセットに出力できますか。

A 該当のメッセージは、期待度数が 5 より小さいセルの比率が 20% より大きな場合に表示されます。SAS 9.2 TS2M3 以降のリリースでは、CHISQ オプションのサブオプションを追記することで非表示となります。

例

```
PROC FREQ DATA=test;
  TABLES x1*x2 / CHISQ(WARN=NOPRINT);
RUN;
```

なお、WARN= の値として OUTPUT を追加することによって、WARNING メッセージの有無を出力データセットに含めることができます。

例

```
ODS OUTPUT CHISQ=out1;
PROC FREQ DATA=test;
  TABLES x1*x2 / CHISQ(WARN=(NOPRINT OUTPUT));
  OUTPUT OUT=out2 CHISQ;
RUN;
```

データセット OUT1 では変数 WARNING、データセット OUT2 では変数 WARN_CHI が追加され、値が 1 の場合は WARNING メッセージ有りを示しています。

Q 現在、OLAP Server サービスが稼働しています。OLAP の機能は利用しないため、OLAP Server サービスは停止してもよいでしょうか。

A SAS® Enterprise BI Server をインストールした際、SAS® OLAP Server が構成され、自動的に稼働するように設定されます。

OLAP Server は、SAS® Enterprise Guide®、SAS® Add-in for Microsoft Office、SAS® Web Report Studio 等のデータソースの 1 つとして利用することが可能です。

データソースとして独立したコンポーネントであるため、利用しない場合には停止しておくことができます。

Q INTCK 関数で日付間隔を算出した場合に、開始日を基準値として設定することはできますか。

A SAS® 9.2 以降より追加されました、オプション引数: CONTINUOUS を用いることで開始日を基準値として設定することができます。

INTCK 関数の構文:

```
INTCK(<間隔>, <開始日>, <終了日>, 'オプション引数')
```

オプション引数:

DISCRETE: 間隔の境界が基準値になります。(デフォルト値)
CONTINUOUS: 開始日が基準値となります。

例えば、下記のサンプルプログラムのように、15JAN2010 と 14JAN2011 の日付間隔を求める場合に、DISCRETE は、月末が基準となるため結果は、'12' となります。CONTINUOUS は、開始日が基準となるため結果は、'11' となります。

例

```
DATA _NULL_;

  start_date = '15JAN2010'd;
  end_date   = '14JAN2011'd;

  up  = INTCK('MONTH', start_date, end_date, 'DISCRETE');
  down = INTCK('MONTH', start_date, end_date, 'CONTINUOUS');

  PUT 'DISCRETE   :' up;
  PUT 'CONTINUOUS :' down;
RUN;
```

Q SASの実行環境をSAS FoundationからSAS® Enterprise Guide®に変更しました。既存のSASファイルをEnterprise Guide上で開いた場合、プログラムアイコンがプロセスフロー上に配置されますが、このままではどのような処理が行われるのかわかりにくいです。実行される処理の内容を分かりやすくする方法はありますか？

A Enterprise Guide 4.3以降では、多数のステップからなるSASプログラムを確認し、各ステップ毎のプログラムノードと出力されるデータセットに分けられたプロセスフローを生成する機能が追加されました。

プログラムを開いた状態で、Enterprise Guideのメニューより、「プログラム」->「分析」を選択します。

例えば、右の図では、DATAステップ、SQLプロシジャ、REPORTプロシジャからなる1つのプログラムノードを分析し、各ステップごとのプログラムノードに分割して新たなプロセスフローを作成しています。

The figure illustrates the process of analyzing a SAS program in SAS Enterprise Guide. It consists of four sequential screenshots:

- Screenshot 1:** The SAS Enterprise Guide interface showing a 'Program' icon in the process flow area.
- Screenshot 2:** The 'Program' menu is open, and 'Analyze' is selected. The 'SAS Program Analysis' dialog box is displayed, showing the program steps and data sets.
- Screenshot 3:** The 'SAS Program Analysis' dialog box is shown with a table of program steps and data sets. The table is as follows:

プログラムステップ	データセット	ファイル
DATASTEP	CLASS (SASHELP)SOURC	
SQL	SOURCE1 (WORK)TEMP1	
REPORT	TEMP1 (WORK)	
- Screenshot 4:** The resulting process flow after analysis, showing individual program steps like 'DATA Step', 'PROC SQL', and 'PROC REPORT' in the process flow area.

Latest Releases

最新リリース情報



PCプラットフォーム

Microsoft Windows版 SAS 9.3
Microsoft Windows x64版 SAS 9.3 / 9.4

メインフレームプラットフォーム

IBM版(OS/390, z/OS) SAS 9.3 / 9.4

UNIXプラットフォーム

SunOS/Solaris版 SAS 9.3/ 9.4
x64 Solaris版 SAS 9.3/ 9.4
HP-UX版 SAS 9.3
HP-UX (Itanium) 版 SAS 9.3/ 9.4
AIX版 SAS 9.3/ 9.4
Linux (Intel) 版 SAS 9.3
x64 Linux 版 SAS 9.3/ 9.4

Latest Hot Fix

最新 Hot Fix 情報



SASのHot Fixは、主に特定の不具合に対する修正プログラムであり、緊急度の高い不具合に対応したもから優先的に作成されています。また、SASのHot Fixは累積的な修正であるため、これまで関連する修正をバンドルして提供しています。

Hot Fixを適用するタイミングとしては、問題が発生してから適用する場合と、事前に適用する場合と2通りあります。

前者の場合、適用のタイミングによっては、モジュールのメンテナンスレベルに差がでる可能性があります。

Technical Support Hot Fixes

<http://ftp.sas.com/techsup/download/hotfix/hotfix.html>

※リンク先の情報は英語です。あらかじめご了承ください。

後者の場合、常に最新モジュールを使用することになるため、既知の障害に遭遇する危険性を未然に防ぐことができます。

Hot Fixは、下記の専用ホームページで提供しています。なお、基本的に次期バージョンのリリース時には、Hot Fixで対応したのも含め、現行バージョンのリリース後に見つかった不具合に対応したものが提供されます。

Hot Fixについては、定期的にご確認ください。

SAS Technical News入手

SAS Technical Newsは、右記のURLから入手できます。

<http://www.sas.com/jp/periodicals/technews/index.html>

例題01:D 例題02:B 例題03:D 例題04:A

編集部より

1年間にわたって掲載された、富士通株式会社様の連載記事は今号で終了となります。富士通株式会社鶴田様、平井様、富士通ビー・エス・シー置田様、まことにありがとうございました。記事作成におきましては、ご多忙の中でご対応いただきましたこと、深く感謝申し上げます。本紙面をお借りしお礼申し上げます。

さて、2014年春号からは、弊社アライアンスパートナーであるネイチャーインサイト株式会社から、3回にわたり連載がスタートします。ネイチャーインサイト社は、弊社パートナーとして20年の歴史を持ち、SAS BIシステム開発やデータ分析業務などに携わっていただいているアライアンスパートナーです。次号からの連載にご期待ください。

発行: SAS Institute Japan株式会社



STN
SAS Technical News
WINTER 2014

■ **テクニカルニュースに関するお問い合わせ先**
テクニカルサポートグループ
TEL:03-6434-3680 FAX:03-6434-3681



SAS Institute Japan株式会社

www.sas.com/jp

本社
〒106-6111
東京都港区六本木6-10-1
六本木ヒルズ森タワー 11F
Tel 03 (6434) 3000
Fax 03 (6434) 3001

大阪支店
〒530-0004
大阪市北区堂島浜1-4-16
アクア堂島西館 12F
Tel 06 (6345) 5700
Fax 06 (6345) 5655

このカタログに記載された内容は改良のため、予告なく仕様・性能を変更する場合があります。あらかじめご了承ください。SASロゴ、The Power to Knowは米国SAS Institute Inc.の登録商標です。その他記載のブランド、商品名は、一般の各社の登録商標です。 Copyright©2014, SAS Institute Inc. All rights reserved.