SAS High-Performance Analyticsによる ビッグデータ解析

泉水克之
SAS Institute Japan株式会社
ソリューションコンサルティング第一本部
エンタープライズアナリティクス推進グループ

Katsuyuki Izumi
SAS Institute Japan

SAS High-Performance Analyticsとは？

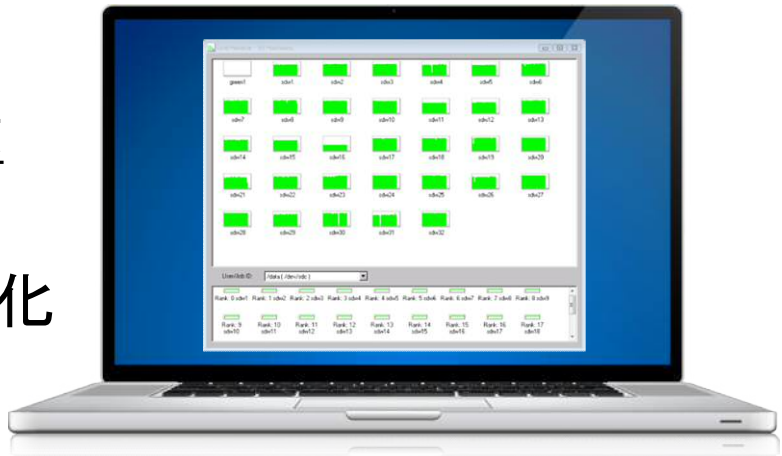
- 従来のSASアドバンスト・アナリティクス製品を分散並列インメモリ環境に対応
- 数十時間、数日かかっていたSASの大規模アナリティクス・バッチ処理を、数分、数秒へと大幅に処理時間を短縮するためのテクノロジー

- <u>既存のSAS環境の多くでも、導入済みSASプロダクトに対応したプロシジャを利用可能</u>

SAS High-Performance Analyticsのメリット

<u>分散環境下</u>

- システム構成を最大限に生かした分析処理の「超」高速化
- 数十時間、数日⇒数分、数秒



<u>多くのSASユーザーの皆様の既存環境（サーバー/PC)</u>

- マルチスレッドへの対応による高速化
- 1時間⇒20分

SAS High-Performance Analyticsのプロダクト群

- SAS® High-Performance Statistics
- SAS® High-Performance Data Mining
- SAS® High-Performance Text Mining
- SAS® High-Performance Optimization
- SAS® High-Performance Econometrics

統計解析	データマイニング	テキストマイニング	最適化	計量経済における統計解析
High-Performance Statistics	**High-Performance Data Mining**	**High-Performance Text Mining**	**High-Performance Optimization**	**High-Performance Econometrics**

• HPLOGISTIC
• HPREG
• HPLMIXED
• HPNLMOD
• HPSPLIT
• HPGENSELECT
• HPFMM
• HPCANDISC
• HPPRINCOMP
• HPQUANTSELECT
• HPPLS
• GAMPL
	• HPREDUCE
	• HPNEURAL
	• HPFOREST
	• HP4SCORE
	• HPDECIDE
	• HPCLUS
	• HPSVM
	• HPBNET
		• HPTMINE
		• HPTMSCORE
		• HPBOOLRULE
			• OPTLSO
			• 以下の最適化エンジンの一部機能
			  • OPTMILP
			  • OPTLP
			  • OPTMODEL
				• HPCOUNTREG
				• HPSEVERITY
				• HPQLIM
				• HPPANEL
				• HPCOPULA
				• HPCDM

共通で利用可能： HPDS2, HPDMDB, HPSAMPLE, HPSUMMARY, HPIMPUTE, HPBIN, HPCORR

2013年7月	12.3 with SAS9.4	2013年12月	13.1 with SAS9.4M1
2014年7月	13.2 with SAS9.4M2	2015年7月	14.1 with SAS9.4M3

対応する既存プロダクトをご利用いただいている方は、これらの機能を原則利用可能です。

SAS/STAT	SAS Enterprise Miner	SAS Text Miner	SAS/OR	SAS/ETS
統計解析	データマイニング	テキストマイニング	最適化	計量経済における統計解析

High-Performance Statistics

- HPLOGISTIC
- HPREG
- HPLMIXED
- HPNLMOD
- HPSPLIT
- HPGENSELECT
- HPFMM
- HPCANDISC
- HPPRINCOMP
- HPQUANTSELECT
- HPPLS
- GAMPL

High-Performance Data Mining

- HPREDUCE
- HPNEURAL
- HPFOREST
- HP4SCORE
- HPDECIDE
- HPCLUS
- HPSVM
- HPBNET

High-Performance Text Mining

- HPTMINE
- HPTMSCORE
- HPBOOLRULE

High-Performance Optimization

- OPTLSO
- 以下の最適化エンジンの一部機能
  - OPTMILP
  - OPTLP
  - OPTMODEL

High-Performance Econometrics

- HPCOUNTREG
- HPSEVERITY
- HPQLIM
- HPPANEL
- HPCOPULA
- HPCDM

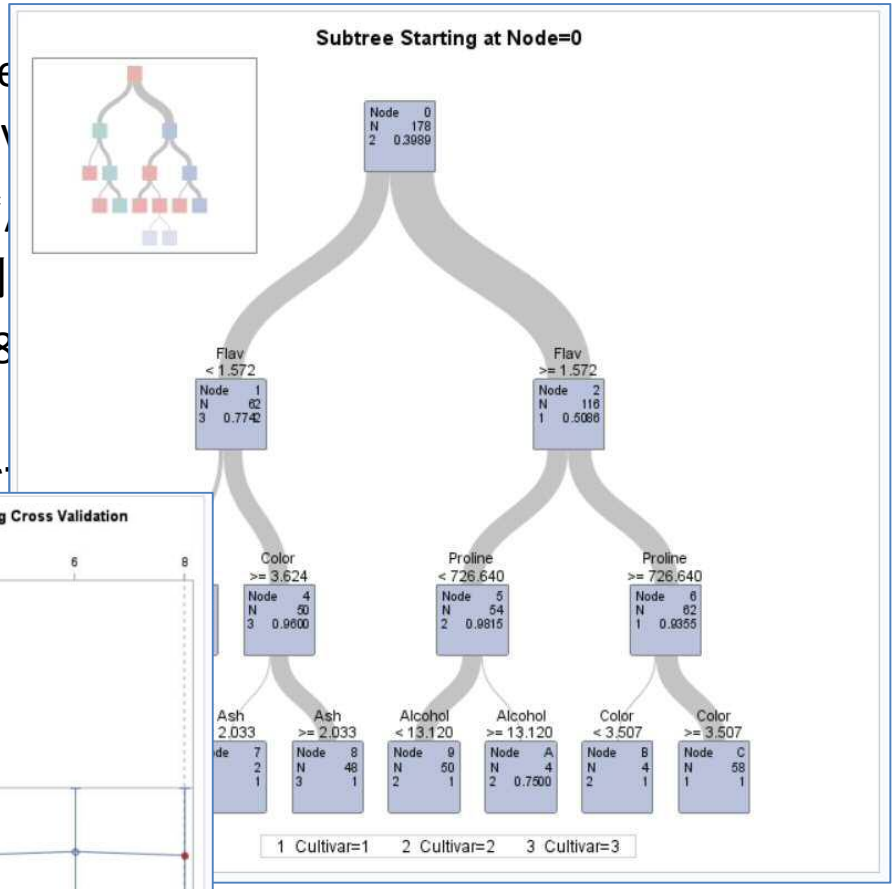
共通で利用可能： HPDS2, HPDMDB, HPSAMPLE, HPSUMMARY, HPIMPUTE, HPBIN, HPCORR

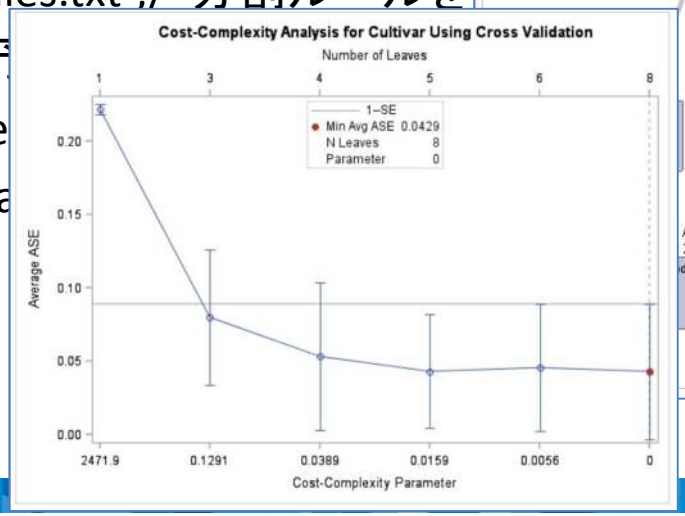
例：SAS® High-Performance Statisticsの機能

HPプロシジャ	対応する統計手法	SAS/STATで対応する既存プロシジャ	HPプロシジャにおける主な利点（高速化、以外）
HPLOGISTIC	ロジスティック回帰	LOGISTIC	変数選択機能の強化
HPREG	線形回帰	REG, GLMSELECT	線形回帰モデルを扱う統一環境
HPLMIXED	線形混合モデル	MIXED	
HPNLMOD	非線形モデル	NLIN	最小2乗法と最尤法
HPSPLIT	決定木モデル	—	決定木！
HPGENSELECT	一般化線形モデル	GENMOD	GLIMにおける変数選択
HPFMM	混合分布への当てはめ	FMM	
HPCANDISC	正準判別分析	CANDISC	
HPPRINCOMP	主成分分析	PRINCOMP	
HPQUANTSELECT	変数選択を伴う分位点回帰	(QUANTREG)	High-Performance Statistics独自機能
HPPLS	PLS回帰	PLS	
GAMPL	罰則付き最尤推定に基づく一般化加法モデル	(GAM)	

例：ツリーモデルを作成するPROC HPSPLIT

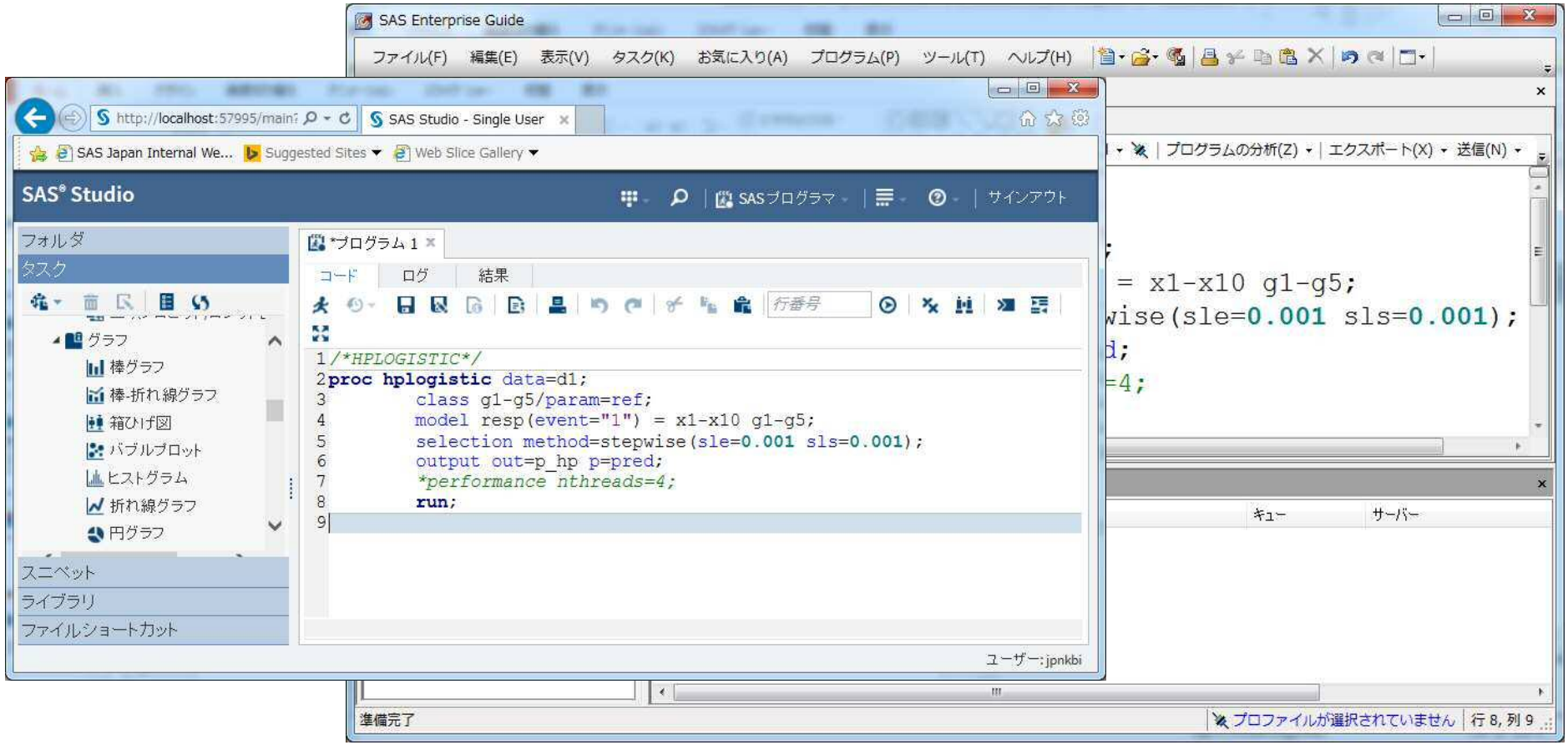
```
proc hpsplit data=temp.hmeq maxdepth=7 maxbranch=2;
 target BAD;          /*予測対象の変数*/
 input DELINQ DEROG JOB NINQ REASON / leve
 input CLAGE CLNO DEBTINC LOAN MORTDUE V
 criterion entropy;/*分割基準はエントロピー*/
 prune misc /N<=10;/*枝刈り時の基準は誤判
 partition fraction(validate=0.2);/*学習データ8
木の作成*/
 rules file='c:¥temp¥rules.txt';/*分割ルールを
 score out=outdata;/*
 output growthsubtree
            nodestats=sta
run;
```





利用方法(1)

- SAS Foundation(DMS)/EG/Stat Studioでプログラミング



利用方法(2)

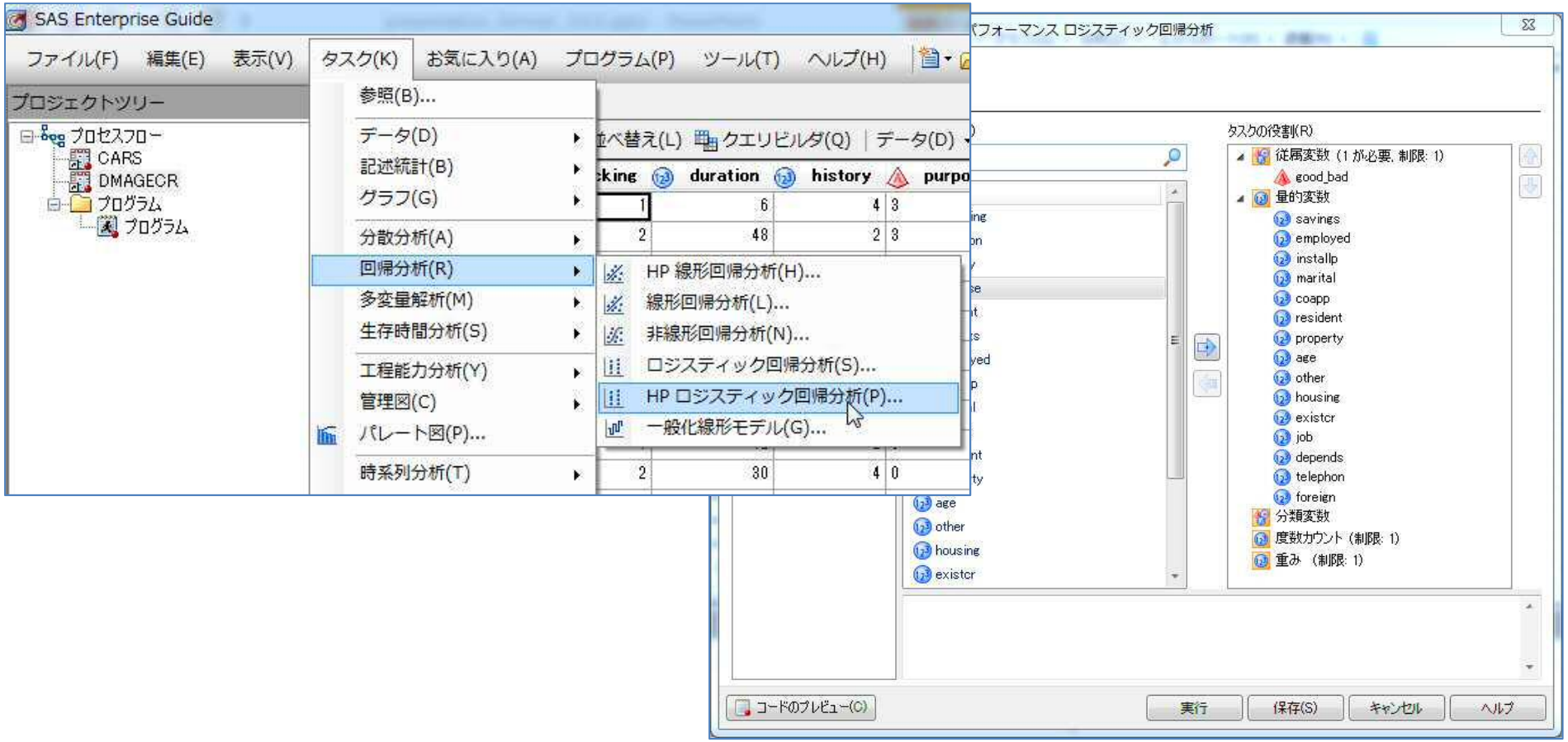
- 構文は、対応する既存プロシジャとほぼ同じ。

```
/*HPLOGISTIC*/
proc hplogistic data=d1;
        class g1-g5/param=ref;
        model resp(event="1") = x1-x10 g1-g5;
        selection method=stepwise(sle=0.001 sls=0.001);
        output out=p_hp p=pred;
        *performance nthreads=4;
        run;

/*LOGISTIC*/
proc logistic data=d1;
        class g:/param=ref;
        model resp(event="1")= x1-x10 g1-g5
                        /selection=stepwise sle=0.001 sls=0.001;
        output out=p p=pred;
        run;
```

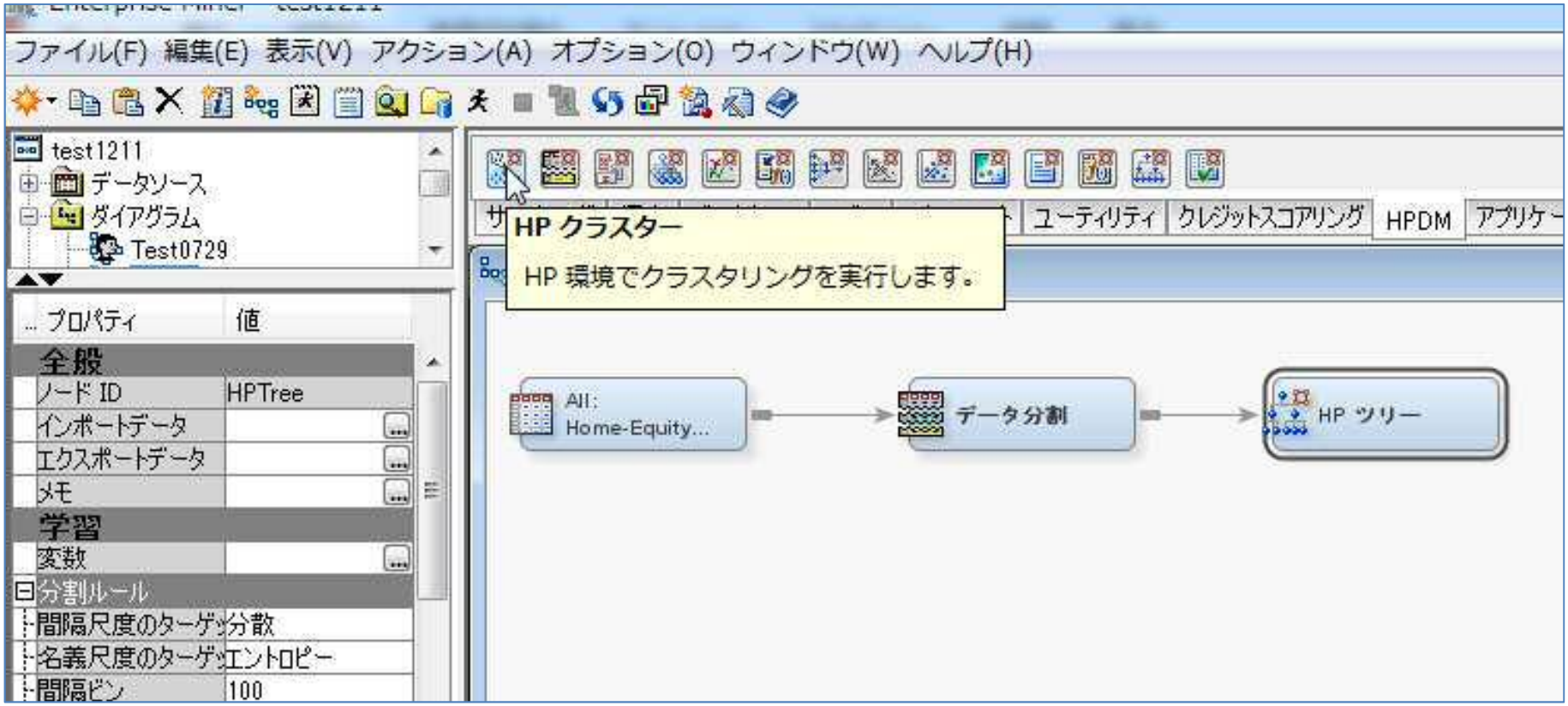
実行方法(2)

SASクライアントツールでGUIベースでの実行（1）：EG

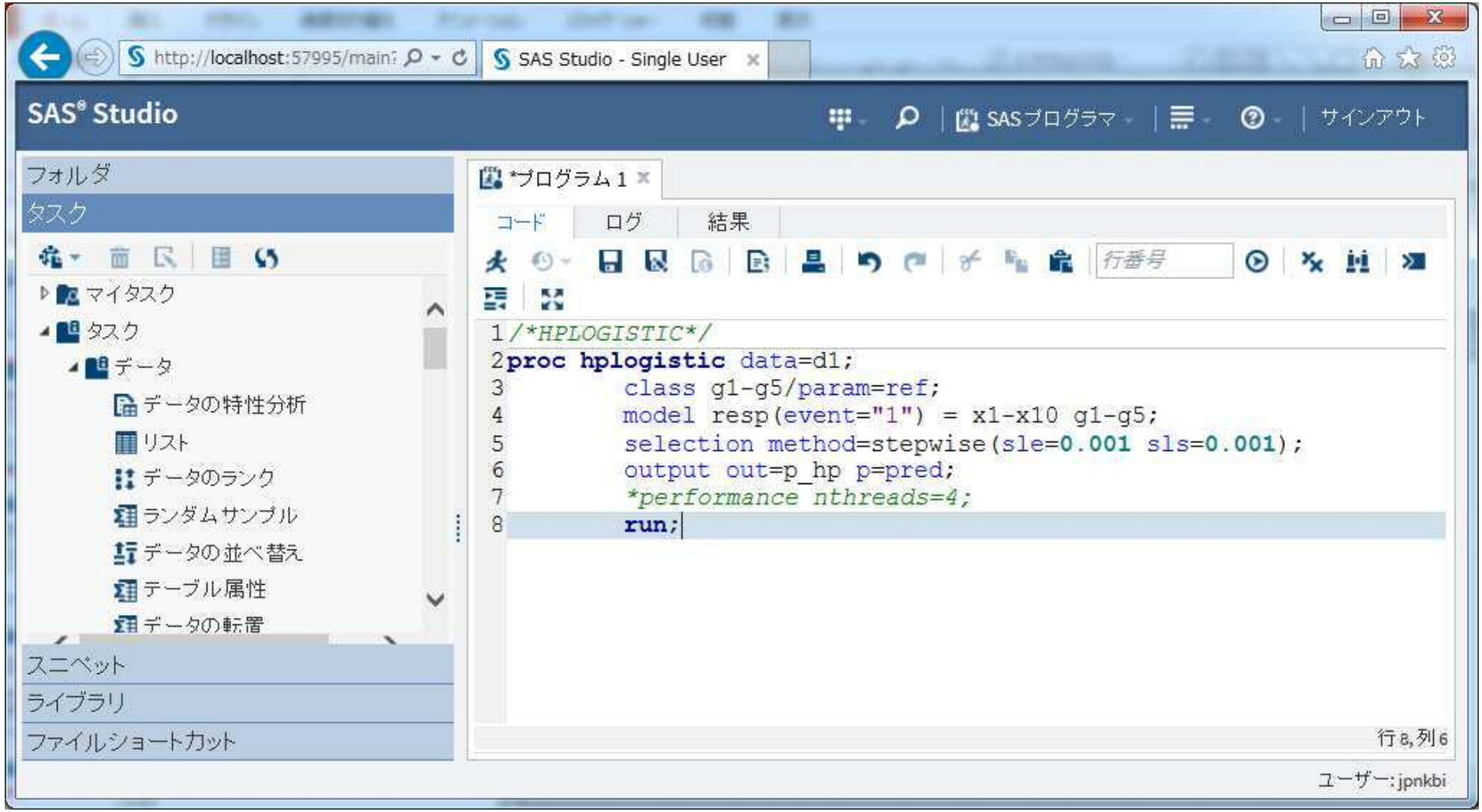


実行方法(2)

SASクライアントツールでGUIベースでの実行（2）：EM



デモンストレーション


http://localhost:57995/main?
SAS Studio - Single User
SAS® Studio
SASプログラマ
サインアウト
フォルダ
タスク
マイタスク
タスク
データ
データの特性分析
リスト
データのランク
ランダムサンプル
データの並べ替え
テーブル属性
データの転置
スニペット
ライブラリ
ファイルショートカット
*プログラム 1
コード
ログ
結果
行番号
/*HPLOGISTIC*/
proc hplogistic data=d1;
class g1-g5/param=ref;
model resp(event="1") = x1-x10 g1-g5;
selection method=stepwise(sle=0.001 sls=0.001);
output out=p_hp p=pred;
*performance nthreads=4;
run;
行 8, 列 6
ユーザー: jpnkbi

SAS High-Performance Analytics(HPA)の未来

- 今後は、Webアプリケーション化と共にますますSASの標準機能になっていきます。
- 分析手法としても、HPAにのみ実装されているものも増えてきています。

- 多くのSAS環境で既に利用可能ですので、是非体感してください。