

Version8における 医薬統計関係の拡張

(株)SASインスティテュートジャパン
テクニカルサポート

小野裕亮

作成:2000年6月23日

修正1:2000年6月28日



ナッシュビル・プロジェクト

- * V8での様々な機能拡張。例えば ...
 - * Base SAS: **ODS**
 - * GRAPH: ActiveX, Java
 - * AF: SAS Component Language
 - * ACCESS: LIBNAME
 - * WebAF,WebEIS : Java
 - * Integration Tech: COM/DCOM,CORBA
 - * CONNECT: 非同期, CPORT不要
 - * Enterprise Guide: thin -client のGUI

Version8, SAS/STATプロダクト の構成

- * 既存42プロシジャ
- * + 新規13プロシジャ
- * + 評価 3プロシジャ
- * -----
- * 合計 58プロシジャ

SAS/STATに 新しく追加されたプロシジャ(1)

- * ノンパラメトリックな回帰 / 確率密度推定
 - LOESS, TPSPLINE(Thin-Plate Spline)
 - GAM(Generalized Additive Model:評価版)
 - KDE (=Kernel Density Estimation)
- * (2次元の)空間統計
 - VARIOGRAM, KRIDGE, SIM2D
- * 層化・集落抽出など および その解析
 - SURVEYSELECT, SURVEYMEANS
 - SURVEYREG

新しく追加されたプロシジャ(2)

- * NLMIXED
 - 非線形 + 一般化 混合モデル
- * PLS
 - PLS回帰
- * MI + MIANALYZE (評価版)
 - Multiple Imputation
- * BOXPLOT (箱ひげ図)
- * STDIZE(変数の様々な方法による標準化)

本発表のトピック

- × NLMIXED →石塚先生が説明。
- * 次のプロシジャを中心に説明。
 - OLOGISTIC, GENMOD
 - MIXED, FREQ
 - (評価版のプロシジャ) MI

LOGISTICプロシジャ

* 数多くの拡張が行われました。

- (1) CLASS → 質的変数
- (2) CONTRAST → 対比
- (3) EXACT → 正確な条件付き...

LOGISTICプロシジャ(1)

* CLASSステートメントの使用例

```
* PROC LOGISTIC DATA=DATA1;  
* CLASS A / PARAM=GLM ;  
* MODEL Y=A B A*B;  
* RUN;  
  - 交互作用もOK。
```

LOGISTICプロシジャ(1)

* 複数のコーディング方法をサポート

– 各水準の効果

- * a)PARAM=EFFECT ... PROC CATMOD流
- * b)PARAM=GLM ... PROC GLM流
- * c)PARAM=REFERENCE ... 最後のセルを0

– 多項式

- * d)PARAM=POLYNOMIAL ... 多項式
- * e)PARAM=ORTHPOLY ... 直交多項式

PARAM=EFFECT (デフォルト)

* A	$\alpha 1$	$\alpha 2$	$\alpha 3$	CATMODプロシジャで採用されているコーディング方法
* 1	1	0	0	パラメータ推定値の和が0になるような制約。
* 2	0	1	0	$\alpha 1 + \alpha 2 + \alpha 3 + \alpha 4 = 0$
* 3	0	0	1	最後の水準における効果は、
* 4	-1	-1	-1	$\alpha 4 = -(\alpha 1 + \alpha 2 + \alpha 3)$ のようにして求める。

PARAM=REFERENCE

* A	$\alpha 1$	$\alpha 2$	$\alpha 3$	最後の水準を0とおく
* -----				コーディング
* 1	1	0	0	最後の水準を基準とした
* 2	0	1	0	時の推定値が算出される。
* 3	0	0	1	
* 4	0	0	0	

PARAM=GLM

* A	$\alpha 1$	$\alpha 2$	$\alpha 3$	$\alpha 4$	PROC GLMの方式
* -----					
* 1	1	0	0	0	
* 2	0	1	0	0	
* 3	0	0	1	0	(注)交互作用がある時、
* 4	0	0	0	1	PARAM=REFERENCE と主効果に対する検定 が異なる。

LOGISTICプロシジャ(2)

- * CONTRASTで、対比を指定できるように。
 - PARMA= オプションに応じて、対比の指定方法は異なるので、注意が必要？
 - expをとった値も計算することができる。
 - * →オッズ比の計算も可能になりました！

LOGISTICプロシジャ(3)

- * EXACTステートメント
 - * 正確な条件付き分布に基づくロジスティック分析
 - * Version8.1から。応答2水準のみ。
- 十分統計量で条件付けた分布のもとで、
 - * 2つの検定のp値および“mid-p”値
 - exact probability test
 - exact conditional scores test
 - * 推定 (ESTIMATE オプション)

最も単純な例(2×2表の解析)

データ		デザイン			計画行列Xの各列 x_i に対応した 十分統計量 t_i
Y	X	y	x0	x1	
Yes	A群	1	1	1	$t_i = y' x_i$
Yes	A群	1	1	1	
No	A群	0	1	1	
Yes	B群	1	1	0	切片x1の十分統計量
No	B群	0	1	0	$t_0 = y' x_0 = 3$
No	B群	0	1	0	効果x1の十分統計量
No	B群	0	1	0	

最も単純な例(2×2表の解析)

	Yes	No	計
A	2	1	3
B	1	4	5
計	3	5	8

* 行和(赤色)だけでなく、列和(水色)も固定した時の(1,1)セルの分布

EXACTステートメントを 手計算すると、、、

* T1 何通り？

* -----

* 0 10 ${}_3C_0 \times {}_5C_3$

* 1 30 ${}_3C_1 \times {}_5C_2$

* 2 15 ${}_3C_2 \times {}_5C_1$

* 3 1 ${}_3C_3 \times {}_5C_0$

2つの考え方=2つの検定方法

* (1) “フィッシャーの直接確率検定” 風

– Prob(T1=u | T0=3)を計算していった、

– Prob (T1=2 | T0=3) 以下となる確率を足していく。

* (2) “コクラン-マンテル-ヘンツェル”風

* $[t1 - E(T1|T0=3)]^2$

* -----

* $Var(T1| T0=3)$ (*単変量の場合)

– よりも大きくなるものが生じる確率を足していく。

Obs	xA	Count	Score	Prob
1	0	10	2.52000	0.17857
2	1	30	0.03111	0.53571
3	2	15	1.52444	0.26786
4	3	1	7.00000	0.01786

0.26786以下の確率を足すと、、、

$$0.17857 + 0.26786 + 0.01786$$

1.52444以上になる確率を求めると、、

$$0.17857 + 0.26786 + 0.01786$$

(* 2つの検定は、少なくとも2 × 2表の時は同じ結果)

Mid-p value

- * 条件付き正確検定
- * → p値の分布は離散
- * → 保守的になる傾向がある。

- * Mid -p ... 「保守的な傾向」を“薄める”ために考え出された方法。

$$-0.17857 + 0.26786 / 2 + 0.01786$$

Test	Statistic	Exact PValue	Mid PValue
Score	1.5244	0.4643	0.3304
Prob	0.2679	0.4643	0.3304

ESTIMATEオプション

- * 条件付き尤度に基づく推定値・信頼区間
 - ESTIMATE= BOTH, ODDS, PARM
- * $(1 - \alpha)\%$ の信頼区間
 - 帰無仮説を α で棄却するような β_0 の値を。
 - $H_0: \beta = \beta_{0-}$ $H_1: \beta > \beta_{0-}$
 - および $H_0: \beta = \beta_{0+}$ $H_1: \beta < \beta_{0+}$
 - (両側の場合は、それぞれ $\alpha/2$ で)。

Exact Odds Ratios (2-tail)					
		Est	95% Confidence Limits		p-Value
x	A	5.784	0.158	586.994	0.5714

Exact Odds Ratios (1-tail)					
		Est	One-sided 95% Confidence Limits		One-sided p-Value
x	A	5.784	0.245	Infinity	0.2857

Cf. 推定値... PHREG, 信頼区間 ...FREQ /EXACT OR
One-sided p-Value ... Fisherの正確検定

GENMODプロシジャ

- * “相関がある”二値応答データの解析
- * 2次のモーメントまでを考慮したGEE推定の他に、Alternating Logistic Reg(交互ロジスティック回帰)が行えるように。
 - Type3検定が行えるように。
 - 「相関構造」ではなく、「(対数)オッズ比」の構造を指定する。
 - 基本は全被験者に対して共通のオッズ比構造

LOGOR= オプション

- * PROC GENMOD;
- * CLASS ...; MODEL
- * / DIST=BIN LINK=LOGIT;
- * REPEATED
- * SUBJECT=id(drug) /
- * WITHIN = time
- * **LOGOR = EXCH ;**
- * RUN;

LOGOR= オプション

- * LOGOR = EXCH
 - ... すべての時点間で等しいオッズ比。
 - (パラメータ数: 1個)
- * LOGOR = FULLCLUS
 - ... すべての時点間でばらばらのオッズ比。
 - (パラメータ数: $t(t-1)/2$ 個)
- * LOGOR= ...ユーザ指定など...

対数オッズの構造を Exchangable にした結果							
標準誤差は Empirical							
Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > Z
Intercept		1.3863	0.7906	-0.1632	2.9358	1.75	0.0795
time	1	-0.0000	1.2500	-2.4500	2.4500	-0.00	1.0000
...	省略...
...
drug	1	0.8109	1.3176	-1.7715	3.3934	0.62	0.5383
drug	2	0.0000	0.0000	0.0000	0.0000	.	.
drug*time	1 1	-1.7918	1.8603	-5.4378	1.8543	-0.96	0.3355
drug*time	1 2	-2.1609	1.6913	-5.4757	1.1540	-1.28	0.2014
drug*time	1 3	-0.2719	1.6962	-3.5964	3.0525	-0.16	0.8726
...
...	省略...
Alpha 1		-1.1206	0.3617	-1.8296	-0.4116	-3.10	0.0019

Score Statistics For Type 3 GEE Analysis			
Source	DF	Chi-Square	Pr > ChiSq
time	4	1.76	0.7800
drug	1	0.01	0.9327
drug*time	4	2.28	0.6842

FREQプロシジャ

- * EXACTステートメント
 - * MCオプションによる乱数シミュレーション
 - * 処理時間の上限を決定するMAXTIME=オプション
 - * NPAR1WAYプロシジャも同様。

- * 二項分布 $\text{Bin}(n,p)$ の確率 p に対する信頼区間が一元表の時にも計算できるように。
 - (方法自体は、RISKDIFFオプションと同じ)

FREQプロシジャの変更点

- * 共通オッズ比と共通相対リスクのマンテル-ヘンツェル流の推定値
 - Version6までは、「帰無仮説のもとでの分散推定値」によって信頼区間を計算していた。
 - → Version8で、ようやく改良。

MIXEDプロシジャ

- * 小標本時の標準誤差・検定
 - DDFM=KENWARDROGER
 - * Kenward & Roger (1997)の論文に基づく方法によって、標準誤差・検定統計量(F値, t値)・p値を計算。
 - * 注)自由度の計算だけでなく、固定パラメータ間の標準誤差なども調整される。

Kenward & Roger(1997)

* REML推定における標準誤差の求め方

- 「 σ の推定値」をプラグインして、
 - 応答変数 Y の分散共分散構造 $\Sigma(\sigma)$ の推定値
 - $\Sigma(\sigma \text{のL推定値})$
- (GLS推定の枠組みに添って、) Σ の推定値を既知のものとして、 β の分散共分散行列 V を推定する。
 - 欠点A) σ の推定に伴う変動を考慮していない。
 - 欠点B) V における小標本に伴うバイアスを考慮していない。

Kenward & Roger(1997)

- * これらの欠点を補うために、近似を高める方法を提案。
- * そして、F分布の近似も良くなるような方法も。既存の検定と対応させて、、、
 - * 自由度1の時: サタースウェイト近似
 - * 正確な検定の時: Hotteling T^2 統計量
- * になるような計算を提案。

- * シミュレーションではカイ2乗近似と比較。

MI & MIANALYZE (V8.1～ ただし、評価版)

- * 3種のMultiple Imputation をサポート
 - PROC MI
 - * 乱数によって欠損値を補ったデータセット (complete dataset)を複数作成。Imputeする方法としては、3種類。
 - PROC MIANALYZE
 - * 様々なプロシジャによる推定結果から標準誤差などを計算。

MIにおける3種類の方法

- * monotoneな欠測パタン
 - 1) 通常の回帰に基づく方法
 - * regression method
 - 2) ロジスティック回帰に基づく方法
 - * propensity score method
- * 多変量正規に基づく
 - 3)MCMC法

Monotoneな欠測値パターン

* ID Y0 Y1 Y2 Y3 Y4

* -----

* 001 ○ ○ ○ ○ ○

* 002 ○ ○ ○ × ×

* 003 ○ ○ ○ ○ ×

* 004 ○ ○ × × ×

–「×の後ろは、必ず×」

MAR

* どの手法も、Missing At Randomでないとは駄目。

– 欠測するかどうかが、その欠測部分の値によっては左右されない。

– 例えば、トービットモデルに従っているようなデータは駄目。→ある以上のものが欠測するようなデータなどは駄目。

方法1: 回帰法

- * 次のモデルで $Y(1)$, $Y(2)$, $Y(3)$... の順番に代入していく。
- * $Y(t)$ の欠測値部分
- * $=b_0 + b_1 Y(1) + \dots + b_{(t-1)} Y(t-1)$
- * $+ s \times \text{乱数}$
- $b_0, b_1, \dots, b_{(t-1)}$ および s は、 $Y(t)$ の非欠測部分から推定

方法2: propensity score method

- * 「欠測」 or 「非欠測」 を応答変数として、ロジスティックモデルを推定。
- * ロジスティックモデルの予測値に近いもの同士で、幾つかのグループに分ける。
- * グループ化されたもののなかで、非欠測値を無作為抽出して、それを欠測値に代入。

方法3: 多変量正規に基づく MCMC法

- * 何らかの方法で、 Σ の初期値を与える。
 - デフォルトはEMアルゴリズムによる推定値
- * 「非欠測部分」および「 Σ の推定値」に基づき、欠測部分に値を代入。
- * Σ を、代入された部分を含めて推定する。
B)に戻る。
 - B), C)の処理は事前に何回も行われる。

MIANALYZE

- * complete datasetsからの推定
 - 各データ毎に推定値 $Q(i)$ およびその分散推定値 $U(i)$ を求める。
 - * 点推定値 = $Q(i)$ の平均
 - within-imputation variance = $U(i)$ の平均
 - between-imputation variance = $Q(i)$ の分散
 - * 点推定値の分散 (total variance)
 - * = within + $(1 + 1/m)$ between

Multiple-Imputation Variance Information						
Variable	Variance			DF	Relative Increase in Variance	Fraction Missing Information
	Between	Within	Total			
x1	0.031937	0.065363	0.098897	165	0.513030	0.346931
x2	0.030456	0.031908	0.063887	76	1.002241	0.513231

GLM, REG, TTEST , UNIVARIATEでも...

- * 信頼区間を出力する機能が追加。
- * GLM: 線形従属の関係を表示できるように。
- * TTEST: 1標本, 対応のあるt検定など。

Base SASでも...

- * UNIVARIATE
 - トリム平均、ウィンザー化平均
 - MADなどによる標準偏差の推定
 - 信頼区間(平均・分散・パーセント点...)
 - ヒストグラム・QQプロット(SAS/QCから移植)
- * MEANS
 - CLASSDATA= オプション
 - パーセント点も計算できるようになった。

その他の統計関連における 拡張 (GUI)

- * Enterprise Guide (クライアント-サーバー環境)
- * SAS/STATプロダクト
 - * Analyst Application (英語)
 - 記述統計・グラフ
 - 1標本・2標本の平均
 - REG・PRINCOMP・GLM・MIXED・LIFETEST・PHREG
 - 1標本や2標本問題における検定・信頼区間・同等性などの検出力計算

補足：ODSに伴う変更点(1)

- * ODSの導入により、outputウィンドウへの出力形式が一部変更されています。
 - “PRINTTOプロシジャ”を利用したV6時代のプログラムが、うまく動かなくなる可能性があります。
 - 対処方法：ODSを使えば、直接データセットに出力できるようになります。

補足：ODSに伴う変更点(2)

- * GENMOD,MIXEDにおける変更
 - Version8との互換性があるステートメント
 - * OMAKEステートメント
 - →しかし、変数名が変更されています。
 - V8で廃止されたステートメントやプロシジャ
 - * × %LET PRINT_=OFF; %LET_DISK_=ON;
 - * × PROC OUTPUT → 将来のバージョンでは、これに相当した機能が正式サポートされる予定。

情報

- * 米国SUGIでのV8に関する論文
 - <http://www.sas.com/service/techsup/news/>
 - からダウンロード可能(PDF形式)
 - * “What’s New in Version8” ...
 - * “What’s New in Version7” ...

- * SUGI-Japan ... *****

以上です。