SAS 9.3 Enhancements for Statistical Analysis

Tor Neilands
UCSF Center for AIDS Prevention Studies
January 27, 2012

Contents

- HMTL output and statistical graphics
- Differences of binomial proportions
- Handling missing data via FIML
- Handling missing data via FCS MI
- EFFECT statement in statistical PROCs
- Post-estimation via PROC PLM
- Bayesian random effects models
- Resources

HTML Output and Graphics

- In SAS 9.3, the default output style is now HTML
- Classical listing output is still available via ODS LISTING ;
- HTML output shows both plots and text as shown in programs *demo1.sas* and *demo2.sas*.
- HTML output accumulates
  - To clear the output window each time you run a program, add ODS HTML CLOSE; ODS HTML ;

    to the top of your program.
  - To selectively exclude specific procedure output, wrap the procedure whose output you want to mask in ODS SELECT NONE ; and ODS SELECT ALL. This is demonstrated in the program *demo3.sas* for excluding PROC LOGISTIC output for each analysis of multiply imputed data sets.

Differences for Binomial Proportions

- Here we are concerned with testing for differences in a two-by-two contingency table
- Various classical methods exist, including:
  - Pearson chi-square
    - Expected cell count should be > 5
  - Fisher's exact test
  - Wald –based CI of the difference of proportions
    - Best for cell counts > 12 (or > 8 with continuity correction).

Differences for Binomial Proportions

- Newer approaches:
  - Farrington-Manning: Invert two one-sided exact score tests and combine
    - Does well, power-wise
    - Interval may be too narrow when $N < 5$ per group
  - Newcombe: Computes a quadratic-based confidence interval for each proportion and integrates them to obtain an overall CI for the difference
    - Continuity correction available for *N*_row < 10
    - Can be somewhat conservative

Smoking Cessation Intervention Example

- Intermittent, non-daily smokers are an increasing percentage of the smoking population.
- This population may be less responsive to existing smoking cessation interventions.
- A pilot study of an intervention targeting non-daily smokers was conducted
- N = 52 who smoked in the past week but not daily were randomized to control or a brief (< 20 minutes) intervention focusing on harm smoking does to them (control) or others (intervention).
- N = 40 completed the study. Measurements were taken at baseline and 3 months following the intervention.

Smoking Cessation Intervention Example

- Primary outcome was quitting.
  - 9.5% (2/21) of the control participants quit
  - 36.8% (7/19) of the intervention participants quit
  - $p = .039$ for the Pearson chi-square test
  - $p = .06$ for the Fisher exact test
- Are there other options for investigating the difference between the groups?
  - A Barnard test, available in R, yielded $p = .07$
    - Metha & Senchaudhuri, 2003: http://www.cytel.com/Papers/twobinomials.pdf
  - Farrington-Manning and Newcombe's method available in SAS. These are shown in *demo1.sas* and described in Stokes & Koch (2011) "Up to Speed with Categorical Data Analysis" from SAS Global Forum.

Missing Data

- Four extant methods for handling missing data under Rubin's Missing at Random (MAR) missingness mechanism assumption
  - Inverse probability weighting
  - Fully Bayesian estimation
  - Full-Information Maximum Likelihood (FIML)
  - Multiple Imputation (MI)
- All perform about equally well and outperform commonly-used ad hoc methods like listwise or pairwise deletion of missing data; or single imputation strategies such as mean substitution.
- Ibrahim et al., 2005, *JASA*, v. 100 (issue 469), pp. 332-346,compares the four methods.

FIML in SAS

- Implemented in the CALIS procedure
- Use METHOD = FIML on the PROC CALIS line
- Available for linear regression, path, and structural equation models with continuous outcomes only (at this time)
- No support for clustered data (at this time)
  - Repeated measures with a fixed number of measurement occasions may be modeled in the wide data structure using the latent growth curve approach, which is similar to multilevel random coefficient models

FIML Example

- Linear regression model example from Paul Allison's Sage publication *Missing Data.*
- Graduation rates from US colleges reported in US News and World Report.
- *N* = 455 cases with complete data for the analysis, a subset of *N* = 1302 cases (35%).
- Outcome: GRADRAT, the ratio of graduating seniors to number enrolled 4 years earlier * 100.
- Explanatory variables:
  - CSAT – Combined mean verbal and math SAT scores
  - LENROLL – Natural logarithm of the number of enrolling freshmen
  - PRIVATE – 0 = public school; 1 = private school

FIML Example

  - STUFAC – Ratio of students to faculty * 100
  - RMBRD – Total annual costs for room and board in thousands of dollars
  - ACT – Mean ACT scores
- Only PRIVATE has complete data. Most data are missing on CSAT (40%), ACT (45%), and RMBRD (40%).
- The SAS program *demo2.sas* fits a linear regression model to the US News data using PROC REG, then PROC CALIS, both with listwise deletion of cases with incomplete data (N = 455).
- The program then demonstrates how to use PROC CALIS to fit the same model using FIML estimation with all 1302 cases.
  - To attain convergence with covariance structure modeling programs, it is sometimes necessary to rescale variables to have similar variances. *Demo2.sas* also demonstrates this.

Multiple Imputation

- For a number of years, SAS featured various imputation methods for monotone missing categorical and continuous variables
- For data sets with arbitrary missingness, SAS features the MCMC data augmentation approach, based on Joe Schafer's text and his freeware NORM program
- This approach assumes multivariate normality for the joint distribution of variables used to generate imputations

Multiple Imputation

- **Experimental** fully-conditional specification (FCS) method in version 9.3 can impute both continuous and categorical variables.
- Works via a chained regression equations approach analogous to MICE in R or categorical imputation options in Stata's -mi- suite (formerly supplied by the user-written program -ice-).
- Requires fewer iterations than the MCMC method.

FCS Imputation Example

- Example drawn from Paul Allison's *Missing Data* text
- *N* = 2992 respondents from the 1994 General Social Survey (GSS)
- Ordinal logistic regression of agreement with the use of SPANKING as a disciplinary technique in child-rearing
  - 1 – Strongly agree
  - 2 – Agree
  - 3 – Disagree
  - 4 – Strongly disagree
- SPANKING question administered to a random 2/3 subset of the sample by design, so there are 1,015 cases missing by design plus 27 more set to missing due to "don't know" or "no answer" responses.

FCS Imputation Example

- Explanatory variables with [number missing]
  - AGE in years [6]
  - EDUC in years of schooling [7]
  - INCOME in thousands of dollars [356]
  - FEMALE: 1 = female; zero otherwise
  - BLACK: 1 = Black; zero otherwise
  - MARITAL: 5 marital status categories [1]
  - REGION: 9 regional categories
  - NOCHILD : 1 = no children; zero otherwise [9]
  - NODOUBT: 1 = no doubt about existence of God; zero otherwise. Missing 1,606 cases by design and another 60 due to "don't know" or no answer responses

FCS Imputation Example

- Most of the missing data appear in various combinations of SPANKING, INCOME, and NODOUBT
- Only 26% of the cases have complete data
- Paul Allison's example drops the one case with missing data on MARITAL to avoid imputation of a multicategory variable. For consistency, I do the same, so the base *N* is 2991.
- Dummy variables created for REGION: EAST, SOUTH, and MIDWEST vs. WEST.

FCS Imputation Example

- Dummy variables created for marital status: NEVMAR (never married), DIVSEP (divorced or separated), and WIDOW (widowed) vs. married.
- To illustrate the FCS implementation in SAS, we will impute values for EDUC, INCOME, NODOUBT, NOCHILD, and SPANKING)
- SAS PROC MI FCS method:
  - Uses linear regression to impute continuous variables (e.g., EDUC, INCOME)
  - Offers a choice for categorical variables
    - Discriminant function analysis for unordered variables
    - Logistic regression for binary and ordinal variables

FCS Imputation Example

- The discriminant method only allows continuous covariates
- The logistic method assumes proportional odds for imputed values
- These features of the imputation process have implications when assessing the proportional odds assumption
- For illustration purposes, we will demonstrate both approaches in *demo03.sas*
  - See Allison's text for further discussion and a SAS macro for assessing the proportional odds assumption across multiple imputed data sets

EFFECT Statement

- Splines and quadratic functions are one approach that may be used to diagnose and handle non-linear relationships of explanatory variables with outcome variables
- Until recently, SAS users had to construct spline variables themselves
- The EFFECT statement now creates such variables internally within procedures
- Supported procedures include GLIMMIX, LOGISTIC, and PHREG

EFFECT Statement Example

- NA Accord study collected repeated measures of CD4 T-cell counts
- Mixed effects model fitted using PROC GLIMMIX (random intercepts for subjects)
- Outcome: SQRTCD4 – square root-transformed CD4 T-cell count
- Explanatory variables (all measured at baseline)
  - CTIME – month of measurement (centered at 4 months to improve convergence)
    - We desire three linear splines: 0-4 months; 4-12 months, and 12-36 months

EFFECT Statement Example

- SREGION – Geographic region
  - 1 = North America (*N* = 23,423)
  - 2 = West Africa (*N* = 520)
  - 3 = East Africa (*N* = 5110)
  - 4 = South Africa (*N* = 41,185)
  - 5 = Asia (*N* = 2,493)
- Csex – Sex of participant (centered)
- Cage – Age of participant (centered)
- EFVBL – On Efavirenz (1 = yes; 0 = no)
- BCD4V – baseline CD4 value
- AZTBL – On AZT (1 = yes; 0 = no)

EFFECT Statement Example

- BVLVC: HIV-RNA viral load ordered categories (WHO)
  - 0 : log10 VL <= 4.0
  - 1 : log10 VL > 4.0 – 4.5
  - 2 : log10 VL > 4.5 – 5.0
  - 3: log10 VL > 5.0 – 5.5
  - 4: log 10 VL > 5.5
- Previous diagnostic work suggests the relationship between baseline CD4 and follow-up CD4 values may not be linear
- One way to address non-linearity is with restricted cubic splines (Harrell, 2001)
- The EFFECT statement can be used in PROC GLIMMIX to construct restricted cubic splines and other spline functions (e.g., B-splines). This is demonstrated in *demo4.sas*.

Post-Estimation

- For many years, Stata users have enjoyed the ability to fit a model and then separately evaluate linear and non-linear contrasts of model parameters using estimation results stored the parameter estimates vector e(b) and the matrix of parameter estimate variances and covariances, e(V).
- Historically, SAS' post-estimation options were procedure-specific

Post-Estimation

- What is suboptimal about procedure-specific post-estimation:
  - Increased developer time investment
  - Users must learn idiosyncratic contrast specifications (e.g., full-rank vs. not-of-full-rank design parameterizations)
  - Some procedures don't support the post-estimation features one wants (if one is lucky, one can switch to another procedure, but that is time consuming)
  - Models must be re-run to obtain post-estimation results (time consuming for computationally demanding models)

Post-Estimation via PLM

- PROC PLM is now available to perform post-estimation following many SAS modeling PROCs.
- Use the STORE statement in the modeling PROCs to create an estimates store file
- Use PROC PLM to access the store file and perform the desired *linear* contrasts
- PROC PLM can also create predicted values for each subject using the SCORE statement
- Saves computing time and also allows sharing of de-identified estimation results with colleagues without having to share the original raw data.

Post-Estimation via PLM

- Supported statements for post-estimation include ESTIMATE, LSMEANS, LSMESTIMATE, SLICE, and TEST.
- WHERE statement is supported for by-group processing (when by-groups are present)
- SHOW statement displays original model specification and results
- Also, various plots are available.
- Using the same model described in the previous example, we created an item store in *demo5a.sas* and performed post-estimation in *demo5b.sas*.

Bayesian Estimation

- SAS continues to expand its Bayesian modeling options.
- Bayesian estimation is available via convenient options in procedures such as GENMOD and PHREG.
- However, there may be times when you want to fit a model with Bayesian estimation that is not supported by SAS procedures with Bayesian estimation options. That is what PROC MCMC is for.
- Version 9.3 includes various Bayesian estimation enhancements, including the addition of the RANDOM statement to PROC MCMC.

Bayesian Estimation

- MCMC syntax is similar to that of PROC NLMIXED
  - User specifies the model parameters and likelihood function
  - Built-in functions are available for commonly-fitted models
  - Various non-normal random effects distributions are supported via the RANDOM statement (e.g., beta, binomial, gamma, and inverse gamma), as well as the familiar and ubiquitous normal distribution.
    - A multivariate normal (MVN) distribution is available.
  - A random coefficients multilevel model fitted with PROC MCMC is demonstrated in *demo6.sas*.

PROC MCMC Example

- Some HIV+ persons do not experience CD4 T-cell recovery as anticipated on treatment, despite suppressed viral load
- Does residual viral replication replenish the latent viral reservoir? Ongoing viral replication could stimulate higher HIV-specific T-cell responses and raltegravir intensification might decrease that response.
  - See Hatano et al., 2011, JID, v. 203 (1 April), pp. 960-968 for details
- *N* = 30 treated participants with CD4 counts < 350 and virologic suppression >= 1 year received raltegravir or a placebo for 24 weeks.
- Outcome: Square root-transformed CD4
- Predictors: Group, Weeks, Group*Weeks, baseline CD4
- Goal: Fit random coefficients model with random intercepts and slopes using PROC MCMC

Other Enhancements

- SURVEYPHREG – Survival analysis using complex survey data.
- FMM – Finite mixture models (useful for fitting zero-inflated, hurdle, and overdispersion models to heavy-tailed data).
- HPMIXED – High performance mixed modeling using sparse matrix algorithms. Suitable for large data sets and models. Addition of a REPEATED statement and more covariance structures.
- Diagnostics for non-linear model fitted in PROC NLIN.
- More – see the SAS "What's New" documentation.

Resources

- **On Deck: SAS/STAT® 9.3**
  - Maura Stokes, Fang Chen, and Ying So, SAS Institute, Cary NC
  - **http://support.sas.com/resources/papers/proceedings11/331-2011.pdf**
- **Up To Speed With Categorical Data Analysis**
  - Maura Stokes, SAS Institute, Inc.
  - Gary Koch, University of North Carolina, Chapel Hill, NC
  - **http://support.sas.com/resources/papers/proceedings11/346-2011.pdf**
- **Making Use of Incomplete Observations in the Analysis of Structural Equation Models: The CALIS Procedure's Full Information Maximum Likelihood Method in SAS/STAT® 9.3**
  - Yiu-Fai Yung and Wei Zhang, SAS Institute Inc.
  - **http://support.sas.com/resources/papers/proceedings11/333-2011.pdf**
  - **See also: http://support.sas.com/rnd/app/papers/stat/imps2011_FIML.pdf**
- **The RANDOM Statement and More: Moving on with PROC MCMC.**
  - Fang Chen, SAS Institute Inc.
  - **http://support.sas.com/resources/papers/proceedings11/334-2011.pdf**

Acknowledgements

- Sharing Data:
  - Stan Glantz – Smoking cessation data
  - Jeff Martin and Elvin Geng – NA Accord data
  - Hiroyu Hatano – Raltegravir CD4 data
- Slide Review: Estie Hudes
- SAS Institute and Staff:
  - Developers of SAS 9.3
    - SAS 9.3 features
    - SGF articles
  - Maura Stokes, SAS R&D
    - Presentations on version 9.3 at WUSS
    - SGF articles
  - Mike Patetta, SAS Education: SAS Bayesian analysis course