Supplement 9.7. SAS commands for evaluations summarized in Table 9.6

Brian R. Gray, US Geological Survey, Upper Midwest Environmental Sciences Center, La Crosse, Wisconsin, USA.

This supplement provides SAS commands for evaluating estimators of variance components (VCs) from log-linear mixed models of negative binomial and Poisson counts derived from two-way cross-classification designs. Corresponding routines were used for all scenarios summarized in Table 9.6. A .sas format file containing the following commands is available on request from gitzenr@missouri.edu.

```
* PURPOSE: estimate variance components (VCs) from loglinear mixed models of negative binomial and Poisson counts derived
from 2-way crossed classification designs;
* CREATED: 11 Jun 2010 by Brian Gray from corresponding binomial/binary VC code;
* REVISIONS: ;


/* THIS ALGORITHM IS PROVIDED WITHOUT WARRANTY, WHETHER EXPRESSED OR IMPLIED.  PLS SEND QUESTIONS AND CRITICISMS TO BRIAN
       GRAY AT brgray@usgs.gov */


* NOTE:
*      - quasilikelihood ('RPQLQ') is appropriate under conditional negative binomial distribution _generating_ models (in
*             which case the variance estimate represents that of the Poisson with a fixed multiplier).
*             RPQLQ output should be ignored when the generating model is that of a Poisson
*             (in which case the population multiplier is identically 1);


* DEFINE PARAMETER/DESIGN VALUES;
%let beta00=1.61;                    * assign mean on log link scale (median on measurement scale);
%let dist=NB;                        * define conditional sampling distribution (valid selections include NB, Poi, P);
%let NBdisp=.5;                             * assign negative binomial dispersion parameter (must be positive);
%let var_site=.3;                           * assign among-site intercept variation on log scale;
%let var_yr=.3;                             * assign among-year intercept variation on log scale;
%let var_sitebyyr=.15;        * assign among-sitexyear intercept variation on log scale;
%let n=5;                                                * assign sample size w/in site-year clusters;
%let sites=20;                              * assign number sites;
%let years=5;                                      * assign number years;
%let simulations=200;         * determine number simulated data sets;

options nocenter;
title1 "2-way factorial design: distribution=&dist, gamma00=&beta00, NBdisp=&NBdisp, sites = &sites, years=&years, n = &n";
title2 "var(yr)=&var_yr, var(site)=&var_site, var(sitebyyr)=&var_sitebyyr, sims=&simulations";


* GENERATE DATA;
* delete old datasets;
proc datasets library=work; save sasgopt gseg; run;

* generate year effects;
data yr_effects;
       do sim = 1 to &simulations;
       do year = 1 to &years;
              yr_effect = sqrt(&var_yr)*rannor(123);
              do site=1 to &sites;
                     output;
              end;
       end;
```

```
        end;
run;
proc sort data=yr_effects out=yr_effects_sort; by sim site year; run;

* generate site effects;
data site_effects;
        do sim = 1 to &simulations;
        do site = 1 to &sites;
                site_effect = sqrt(&var_site)*rannor(234);
                do year = 1 to &years;
                        sitebyyr_effect = sqrt(&var_sitebyyr)*rannor(345);
                output;
                end;
  end;
        end;
run;

* merge site and year effects, and generate outcomes;
data twoway;
        merge yr_effects_sort site_effects;
        by sim site year;
        eta = &beta00 + yr_effect + site_effect + sitebyyr_effect;
        eta_inv = exp(eta);
        n = &n;
        do i = 1 to &n;
                shape = 1/&NBdisp;
                scale = &NBdisp;                                                                        * ensure
E[gamma rv] = 1 and intercept identification by setting scale=1/shape;
                xgammaij = scale*rangam(456,shape);          * generate gamma rv;
                if "&dist" in ("NB","nb") then mean = eta_inv*xgammaij; * generate mean of NB variate;
                        else mean = eta_inv;
                yij = ranpoi(567,mean);                                                 * generate random variate;
                if "&dist" in ("NB","nb") then var_yij = eta_inv*(1+&NBdisp*eta_inv);
                        else var_yij = eta_inv;
                output;
        end;
run;


* ESTIMATE VARIANCE COMPONENTS BY METHOD;
* per SAS email 5/26/10, SAS' ?SPL and ?MPL labels correspond to Breslow and Claytons' PQL1 and MQL1, respectively.  see also
*       4/14/2008 SAS email and p541, Littell et al (2006);
ods listing close;
proc glimmix data=twoway method=mmpl;* hessian;
        class year site;
        model yij = / s d=&dist;
        random int / sub=site;
```

```
        random int / sub=year;
        random int / sub=year*site;
        by sim;
  ods output convergencestatus=convMQL iterhistory=iterMQL parameterestimates=parmestsMQL covparms=covparmsMQL;
run;

* estimate variance components using PQL;
proc glimmix data=twoway method=mspl;
        class year site;
        model yij = / s d=&dist;
        random int / sub=site;
        random int / sub=year;
        random int / sub=year*site;
        by sim;
  ods output convergencestatus=convPQL iterhistory=iterPQL parameterestimates=parmestsPQL covparms=covparmsPQL;
run;

* estimate variance components using RPQL;
proc glimmix data=twoway method=rspl;
        class year site;
        model yij = / s d=&dist;
        random int / sub=site;
        random int / sub=year;
        random int / sub=year*site;
        by sim;
  ods output convergencestatus=convRPQL iterhistory=iterRPQL parameterestimates=parmestsRPQL covparms=covparmsRPQL;
run;

* estimate variance components using RPQL under Poisson assumption with moment-based correction for underdispersion;
proc glimmix data=twoway method=rspl;
        class year site;
        model yij = / s d=Poi;
        random int / sub=site;
        random int / sub=year;
        random int / sub=year*site;
        random _residual_;
        by sim;
  ods output convergencestatus=convRPQLQ iterhistory=iterRPQLQ parameterestimates=parmestsRPQLQ covparms=covparmsRPQLQ;
run;

* estimate variance components using Laplace estimation;
proc glimmix data=twoway method=laplace;
        class year site;
        model yij = / s d=&dist;
        random int / sub=site;
        random int / sub=year;
        random int / sub=year*site;
```

```
        by sim;
  ods output convergencestatus=convLaplace iterhistory=iterLaplace
        parameterestimates=parmestsLaplace covparms=covparmsLaplace;
run;
ods listing;


* COMBINE RESULTS FOR ALL ESTIMATION METHODS;
* create all-method convergence information dataset;
data conv;
        length method $10.;
        set convMQL(in=MQL) convPQL(in=PQL) convRPQL(in=RPQL)/**/ convLaplace(in=Laplace) convRPQLQ(in=RPQLQ);
        if MQL then method="_1MQL"; else if PQL then method="_2PQL"; else if RPQL then method="_3RPQL"; else if Laplace then
method="_4Laplace";
                else if RPQLQ then method="_5RPQLQ"; else method = "_6ERROR";
        if status = 1 or pdg=0 then convpdg=0; else convpdg=1;
        converge=1-status;
run;

* create all-method parameter estimate dataset;
data parmests;
        length method $10.;
        set parmestsMQL(in=MQL) parmestsPQL(in=PQL) parmestsRPQL(in=RPQL) /**/parmestsLaplace(in=Laplace)
                parmestsRPQLQ(in=RPQLQ);
        if MQL then method="_1MQL"; else if PQL then method="_2PQL"; else if RPQL then method="_3RPQL";
                else if Laplace then method="_4Laplace";
                else if RPQLQ then method="_5RPQLQ"; else method = "_6ERROR";
        inthat=estimate;
        SEparm=stderr;
        keep sim method inthat SEparm;
run;

* create all-method estimated covariance matrix dataset;
data covparms;
        length method $10.;
        set covparmsMQL(in=MQL) covparmsPQL(in=PQL) covparmsRPQL(in=RPQL) covparmsLaplace(in=Laplace) covparmsRPQLQ(in=RPQLQ);
        if MQL then method="_1MQL"; else if PQL then method="_2PQL"; else if RPQL then method="_3RPQL";
                else if Laplace then method="_4Laplace";
                else if RPQLQ then method="_5RPQLQ"; else method = "_6ERROR";
        varhat =estimate;
        SDhat =sqrt(estimate);
        SE_varhat=stderr;
        if subject = "" and "&dist" in ("NB","nb") then subject = "NBdisp";  else if subject = "" then subject="QLlocal";
keep sim method subject covparm estimate varhat SDhat SE_varhat;
run;

* merge convergence, parameter and covariance datasets;
```

```
data covparmsplus;
	merge covparms parmests conv;
	by method sim;
run;
proc sort data=covparmsplus out=covparmsplussort;
	by subject method sim;
run;


* CALCULATE MODEL CONVERGENCE PROPORTIONS, AND PREVALENCES OF NON-POSITIVE DEFINITE HESSIAN MATRICES;
title3 "Convergence proportions";
proc means data=conv noprint; var converge; output out=converge mean=probconverge n=n; by method;
proc print data=converge(drop=_TYPE_ _FREQ_) noobs; run;
title3 "Proportion models with positive definite Hessian matrices (of models that converged)";
proc means data=conv noprint; var pdg; output out=pdg mean=probpdg n=n; by method; where converge;
proc print data=pdg(drop=_TYPE_ _FREQ_) noobs; run;


* DETERMINE NUMBER OF 0 OR NEAR-0 (<1e-4 units) VARIANCE ESTIMATES FOR MODEL FITS WITH non-PD G MATRICES;
title3 "% zero estimates by variance component|model convergence";
data minvar; set covparmsplussort; if varhat lt 1e-4 then varhat0=1; else varhat0=0;
proc means data=minvar noprint; var varhat0; by subject method sim; output out=varhat0stats sum=sum0s n=n; where converge;
run;
proc means data=varhat0stats noprint; var sum0s; by subject method ; output out=varhat0statssum mean=mean median=median
min=min max=max n=n; run;
proc print data=varhat0statssum(drop=_TYPE_ _FREQ_) noobs; run;


* CALCULATE VARIANCE STATISTICS (conditional on convergence and, if selected, positive definite G matrix);
title3 "Variance component estimates";
proc means data=covparmsplussort noprint;
	var varhat SDhat SE_varhat inthat SEparm;
	output out=varcompstats mean=meanvarhat meanSDhat meanSEvarhat meaninthat meanSEparm stddev=SDvarhat SDSDhat n=n;
	by subject method;
	where converge;* and pdg;
run;
data varcompstats2;
	set varcompstats(drop=_TYPE_);
	pctconverge=n/&simulations;
	* calculate RMSE(var_hat);
	if subject="site" then MSEvarhat=(meanvarhat-&var_site)**2 + SDvarhat**2;
		else if subject="year" then MSEvarhat=(meanvarhat-&var_yr)**2 + SDvarhat**2;
		else if subject="year*site" then MSEvarhat=(meanvarhat-&var_sitebyyr)**2 + SDvarhat**2;  else if
subject="NBdisp" then MSEvarhat=(meanvarhat-&NBdisp)**2 + SDvarhat**2;
	RMSEvarhat=sqrt(MSEvarhat);
	* calculate RMSE(SD_hat);
	if subject="site" then MSESDhat=(meanSDhat-sqrt(&var_site))**2 + SDSDhat**2;
```

```
               else if subject="year" then MSESDhat=(meanSDhat-sqrt(&var_yr))**2 + SDSDhat**2;
               else if subject="year*site" then MSESDhat=(meanSDhat-sqrt(&var_sitebyyr))**2 + SDSDhat**2;
               else if subject="NBdisp" then MSESDhat=(meanSDhat-sqrt(&NBdisp))**2 + SDSDhat**2;
                      RMSESDhat=sqrt(MSESDhat);     RMSESDhat=sqrt(MSESDhat);
proc print data=varcompstats2 noobs;
        format _numeric_ 8.3 pctconverge 8.2 MSEvarhat MSESDhat 8.4 n;
        var subject method pctconverge meanvarhat SDvarhat meanSEvarhat RMSEvarhat meanSDhat SDSDhat
               RMSESDhat meaninthat meanSEparm;
run;


* PLOT ESTIMATES;
proc boxplot data=covparmsplussort; plot SDhat*method; by subject; where converge and subject ne "year*site"; run;
proc boxplot data=covparmsplussort; plot SDhat*method; by subject; where converge and subject = "year*site"; run;
ods graphics on;
proc sgpanel data=covparmsplussort noautolegend;
        panelby method;
        histogram SDhat;
        density SDhat/type=kernel;
        by subject;
run;
ods graphics off;
```