Paper SD-06

A Simulation Study to Evaluate ANOVA and GEE for Comparing Correlated Proportions With Missing Values

Mark S. Litaker, University of Alabama at Birmingham, Birmingham, AL
Daron G. Ferris, Medical College of Georgia, Augusta, GA

ABSTRACT

The performance of an analysis of variance (ANOVA) model was compared with that of a generalized estimating equations (GEE) model for comparison of correlated proportions using simulated data. The analyses were implemented using SAS® PROC GLM and PROC GENMOD. Simulations were based on a clinical study of three diagnostic techniques, validated with biopsy. Study subjects received two independent evaluations using each of the three methods. Observations represent agreement or disagreement with biopsy results. Correlated observations were generated using observed subject means and hypothesized treatment effects. Simulations included the null and two alternative scenarios, each with complete data and with 5% and 10% missing data. The binomial distribution and logit link function were used in the GEE analysis. Least-squares means and 95% confidence intervals (CI) were calculated for group means using each of the analysis models. Alpha, power and CI coverage probabilities were calculated. Observed alphas were close to the nominal level for both procedures. Power was similar for the two methods, with ANOVA demonstrating a slight power advantage. CI coverage was closer to nominal for ANOVA than for GEE, although the ANOVA CIs may be overadjusted for larger treatment effects. GEE CIs showed higher than nominal coverage levels

INTRODUCTION

Repeated observations are often used in biological research in order to facilitate the removal of between-subject variability from the comparison of treatment effects, and thus to increase power. Correlation among multiple observations made on the same subjects must be incorporated into the estimates of standard error in order to construct valid tests and confidence intervals. Traditional repeated measures ANOVA accomplishes this task for normally-distributed observations. However, an appropriate analysis of non-normally distributed data is often more problematic.

Analysis of variance using subjects as blocks has been shown to perform well for comparison of treatment effects for binary outcome data with multiple observations per subject, as well as for calculating confidence intervals based on least squares means[1]. A problem with this analytic approach, however, is that missing values may cause aberrant results, particularly with respect to point estimates and confidence interval coverage. There are potential problems with the use of the blocked ANOVA approach for estimation of proportions. First, point estimates and confidence intervals obtained from least squares means are not constrained to be in the interval [0,1]. While this is not typically a problem with complete data, the presence of missing values may lead to out-of-range values of estimates. Also, since the F-tests are based on the assumption of normality of the error term, use of ANOVA for binary data has the potential to lead to invalid inference.

Generalized linear models extend ANOVA to independent observations of non-normally-distrubuted data. Use of generalized estimating equations (GEE) extends generalized linear models to the analysis of correlated data. This approach allows linear models to be constructed for correlated observations that are not normally-distributed, including binomial data, and allows parameter estimates to be constrained to specific intervals.

GOALS

The primary goal of this study was to provide a preliminary evaluation and comparison of the performance of ANOVA and GEE models for comparing correlated binary proportions derived from two observations for each of three treatments per subject and to evaluate the performance of each of these methods for estimation of confidence intervals based on least-squares means. Additionally, the goal was to evaluate the effect of randomly missing data on the performance of each method.

THE TELEMEDICINE COLPOSCOPY STUDY

This NCI/AHCPR-funded study was conducted to compare three methods of cervical examination: in-person examination using a colposcope, distant examination using computer imaging, and distant examination using a telemedicine system[2]. Two hundred sixty four women who previously had abnormal Pap smears were examined twice using each of the three examination methods. A diagnosis was assigned for each of the six examinations. Thus, the experimental design was a 3 x 2 doubly repeated measures layout. The outcome measure for this study was the proportion of examinations in which the diagnosis agrees with the biopsy result. The goal of the study was to compare the proportions of examinations which agree with the biopsy result among the three methods, and to calculate 95% confidence intervals on these proportions of agreement, accounting for correlation due to the repeated measurements. The current simulation study was based on examination results for a subset of 145 of the study patients who had diagnostic results by biopsy as well as complete data for all six colposcopic and telemedicine evaluations.

THE SIMULATION STUDY

This study was conducted to compare the performance of a GEE model, implemented with PROC GENMOD and incorporating subjects as blocks, with that of an analogous ANOVA model implemented with PROC GLM, for comparing correlated proportions. The three methods of examination define the treatment groups for the analyses. Correlated data were simulated for six examinations for each of 145 patients for the null situation of no difference among mean agreement levels and for two alternative situations, representing differences of 10% and of 15% between the extreme groups, with the third group having a mean agreement halfway between the other two. These differences were selected based on the size of difference that was judged to be clinically meaningful. Each observation represents either agreement with the biopsy diagnosis or disagreement. Correlations among the observations were induced by incorporating a subject term, consisting of the deviation of the mean agreement observed for each subject from the overall mean agreement, into the probability that a particular observation represents agreement. The probability of agreement for each observation was the sum of the overall mean, the subject term and a fixed treatment effect. Individual observations were then generated as a "0" or "1", with the probability of a "1" being the value determined by the sum of the subject and treatment effects. The observation was constructed by generating a random number from the uniform(0,1) distribution. If the random number was less than the probability defined as the sum of the overall mean and the subject and treatment effects, then the observation was coded as 1, that is, as an agreement. Otherwise, the simulated observation was coded as 0, a disagreement. Using this coding, the crude point estimate of the agreement with biopsy for each treatment is the mean of the observations within that treatment.

Each of the simulated data sets was analyzed using ANOVA, implemented with PROC GLM, and GEE, implemented wth PROC GENMOD. The binomial distribution and the logit link function were used for the GEE analysis. Results of the treatment comparisons were recorded for each analysis. Performance of the confidence intervals was based on whether the observed confidence interval included the "true" value of the parameter, that is, the proportion of agreement that was specified in the simulation.

Three confidence intervals were calculated for each analysis. The first of these was calculated using PROC SUMMARY, and was based on the crude treatment means, assuming independent observations. The second confidence interval was produced using the LSMEANS statement in PROC GLM, and is intended to account for lack of independence of observations made on the same subject. The third confidence interval was calculated using the LSMEANS statement in PROC GENMOD.

Following each complete-data analysis, 5% of the observations were randomly selected and were recoded as missing values. The analyses were repeated using ANOVA and GEE, and the results were recorded. The procedure was repeated with 10% of the data randomly recoded as missing. For this preliminary report, 2000 replicates of each scenario were conducted.

IMPLEMENTING THE SIMULATION STUDY

SPECIFYING THE SUBJECT AND TREATMENT EFFECTS

Subject effects were obtained by calculating mean agreement for each subject and for all subjects combined, using the observed data. A variable, dxbxagree, was coded as 0 or 1 representing disagreement or agreement with biopsy, respectively. This variable was averaged across each subject's six observations, and across all subjects, using PROC SUMMARY. Subject effects were then calculated as the difference between the overall mean agreement and the mean agreement for each subject. Treatment effects were then calculated as the difference between the overall mean and each of the treatment means. Mean overall agreement for the observed data was 0.55287. Treatment means for a 10% difference between the extreme groups were defined as 0.50287, 0.55287 and 0.60287. A 15% difference across the treatment groups was defined as mean agreements of 0.47787, 0.55287, and 0.62787.

The following code illustrates the simulation macro using for complete data. The macro generates simulated data sets, analyzes each of these using PROC GLM and PROC GENMOD, and saves the appropriate p-values and confidence limits to text files.

```
%macro allbx;
  %do i = 1 %to 2000;
```

The macro begins with a data set consisting of six observations per subject that includes subject ID, treatment group, replicate (1 or 2), and the treatment effects that are to be simulated.

```
data rep;
set dims.sim10;
```

A random number is chosen as the "observed probability" for each replicate observation.

```
if rep = 1 then sim = ranuni(-1);
if rep = 2 then sim = ranuni(-1);
```

The "true" probability of agreement is calculated for each subject and treatment.

```
if (method eq 1) then yhat = overall + subject + method1 ;
if (method eq 2) then yhat = overall + subject + method2;
if (method eq 3) then yhat = overall + subject + method3;
```

The observations are coded 1 (agree) if the observed probability is greater than the “true” probability of agreement, and as 0 (disagree) otherwise.

```
if sim le yhat then agree = 1;
if sim gt yhat then agree = 0;
run;
```

Analyze the data using ANOVA and write the p-values and confidence interval endpoints to text files.

```
proc glm data=rep noprint  outstat=pvals;
class id method ;
model agree = id method;
lsmeans method /out=ls;
run;
data pvals;
set pvals;
if _SOURCE_ eq 'method' and _TYPE_ eq 'SS3';
file 'p_simpwr10.txt' mod;
put PROB;
run;

data ls;
set ls;
file 'p_simpwr10ci.txt' mod;
put method lsmean stderr;
```

Calculate confidence intervals on the raw means, using PROC SUMMARY, and write these to text files.

```
proc summary data=rep noprint;
class method;
var agree;
output out=mcl mean=mean lclm=lclm uclm=uclm;
run;
data mcl;
set mcl;
if _type_ eq 1;
file ‘p_pwr10mci.txt' mod;
put method mean lclm uclm;
run;
```

Analyze the data using GENMOD and save the p-value for method and the confidence intervals for the least-squares means.

```
ods listing exclude all;
proc genmod data=rep descending  ;
class  id method ;
model agree =  method /dist=binomial link=logit type3 ;
repeated subject=id ;
lsmeans method /cl;
* ods output Type3 LSMeans;
ods output Type3=type3(keep=ProbChiSq) LSMeans=lsmeans;
ods trace on;
ods show;
run;
ods listing select all;
run;
data dims.gee_pwr10;
set dims.gee_pwr10 type3 lsmeans;
run;
%end;
%mend ;

%allbx;
```

```
run;
```

The next step calculates power for each of the procedures, using PROC MEANS. Power is the proportion of the total number of analyses that show a significant difference among the treatments.

```
data probs;
infile 'p_simpwr10.txt';
input pvalue;
      if pvalue le .05 then power = 1;
      if pvalue gt .05 then power = 0;
run;

proc means n mean data=probs;
title1 'power estimates for GLM analysis';
title2 'treatment effects: 0, +5, -5';
title3 '55.2%, 60.2%, 50.2% agreement';
var power;
run;
```

The endpoints of confidence intervals on the LSMEANS was calculated using the normal approximation. These confidence intervals could also be obtained directly from the LSMEANS statement in PROC GLM

```
data ci;
infile 'p_simpwr10ci.txt';
input method lsmean stderr;
      lower = lsmean - 1.96*stderr;
      upper = lsmean + 1.96*stderr;
```

An indicator variable was used to identify whether the confidence interval included the "true" value of agreement.

```
include  = 0;
if (method eq 1) and (lower le .55287) and (upper ge .55287) then include=1;
if (method eq 2) and (lower le .60287) and (upper  ge .60287) then include=1;
if (method eq 3) and (lower le .50287) and (upper  ge .50287) then include=1;
run;
```

Coverage probabilities of the confidence intervals were calculated for each treatment using PROC MEANS.

```
proc sort data=ci;
by method;
proc means data=ci n mean;
title 'performance of adjusted confidence intervals -- GLM analysis';
var lsmean include;
by method;
run;
```

Evaluation of the performance of the unadjusted confidence intervals and the GEE analysis was conducted in the same manner, with the exception that the endpoints of the confidence intervals must be back-transformed from the logit transformation that was used in the GEE analysis.

```
data geeci;
set gee_pwr10;
if effect = 'method';
      * untransform means and ci estimates from logit;
g_mean = (exp(estimate))/(exp(estimate) + 1);
g_lower = (exp(lowercl))/(exp(lowercl) + 1);
g_upper = (exp(uppercl))/(exp(uppercl) + 1);
      * identify whether ci includes true mean agreement;
include  = 0;
if (method eq 1) and (g_lower  le .55287) and (g_upper  ge .55287) then
      include = 1;
if (method eq 2) and (g_lower  le .60287) and (g_upper  ge .60287) then
      include = 1;
if (method eq 3) and (g_lower  le .50287) and (g_upper  ge .50287) then
      include = 1;
run;
```

```
proc sort data=geeci;
by method;
proc means data=geeci n mean;
title 'inclusion probabilities of gee confidence intervals for LS means';
var  g_mean  include;
by method;
run;
```

For each simulation scenario, the analyses were repeated after randomly deleting 5% of the observations and after 10% deletion. In order to delete the same number of observations for each replicate analysis, deletion was performed by sorting the data set by a random number and specifying the number of observations to keep. The following code randomly deletes 5% of one of the simulated data sets, keeping 826 of 870 observations:

```
data rep5;
set rep;
randelete=ranuni(-1);
run;
proc sort data=rep5;
by randelete;
run;
data rep5;
set rep5(obs=826);
run;
```

RESULTS

For the actual study, the observed overall mean agreement between examination and biopsy diagnosis was 55.3%, and the mean agreement with biopsy diagnoses for the three methods (treatment groups) were 56.9%, 53.4%, and 55.5%.

Observed alpha levels for both ANOVA and GEE analyses of the simulated data were very close to the nominal level of 0.05. ANOVA showed alpha levels of 0.0440, 0.0495 and 0.0530 for complete data, 5% missing data and 10% missing data, respectively. The corresponding alpha levels for GEE analysis were 0.0470, 0.0485 and 0.0525.

ANOVA showed slightly greater power than GEE in all the simulation scenarios that were considered. For the scenario with a 10% difference between the extremes of treatment means, observed power for ANOVA was 66.8% for complete data, 64.5% for 5% missing and 62.0% for 10% missing, versus 65.5%, 61.6% and 59.0% for GEE. With a 15% difference between treatment means, power for the ANOVA analysis was 96.3%, 94.8% and 94.1% for complete data, 5% missing and 10% missing, respectively. Corresponding values of power for GEE were 96.0%, 94.1% and 92.5%.

Coverage probabilities of nominal 95% confidence intervals showed a more marked difference between the methods. As would be expected, the unadjusted confidence intervals were liberal, showing 99.1%, 98.9% and 98.8% coverage of the true agreement for complete data, 5% missing and 10% missing, respectively.

In the null scenario, confidence intervals for least-squares means from GLM showed coverage of 94.9% for complete, 5% missing and 10% missing data. GEE showed confidence intervals with coverage probabilities close to those of unadjusted intervals: 99.7%, 99.7% and 99.6% for complete, 5% missing and 10% missing data, respectively.

In the situation with a 10% treatment difference, GLM produced slightly conservative confidence intervals, with coverage percentages of 93.6%, 93.5% and 93.8% for complete, 5% missing and 10% missing data. Confidence interval coverage for the GEE analyses were 99.6%, 99.7% and 99.5% for this scenario.

For the simulations using a 15% difference across treatments, confidence intervals from GLM showed 91.7%, 91.9% and 91.8% coverage for complete, 5% missing and 10% missing data, respectively. The corresponding coverage percentages for the GEE analysis were 99.1%, 99.0% and 98.9%.

CONCLUSIONS

ANOVA and GEE both provided valid comparisons of correlated proportions, as reflected by alphas for both procedures being approximately equal to the nominal level. Likewise, power was similar for the two procedures, with ANOVA showing a slight power advantage over GEE.

In the null situation, with no treatment difference, coverage of the “true” parameter value for adjusted confidence intervals from ANOVA were closer to the nominal level than were those from GEE. Confidence interval coverage for GEE was similar to that of unadjusted confidence intervals for the null scenario, and consistently showed coverage of over 99% when using a nominal confidence level of 95%. Confidence intervals from GENMOD remained very liberal relative to the nominal confidence level for all scenarios that were considered. Confidence intervals from GLM demonstrated coverage very close to the nominal level for the null scenario. However, with larger differences among the treatment means, the confidence intervals using GLM may be overadjusted, resulting in intervals that are conservative, including the simulation-specified parameter value in 91 – 94% of nominally 95% confidence intervals.

The poor performance of the GEE-based confidence intervals may be due to the fact that the doubly-repeated nature of this data was not represented in the GEE model. Simulation of data with a single repeated measurement might show better results for this method. Also, this simulation study was based on a relatively small number of replications, which may have resulted in unstable estimates of performance. However, the performance of both types of analysis were based on the same set of 2000 replicate simulations, so sample-dependency of results would be expected to be equal for both methods. Based on this study, there seems to be little or no advantage to the use of PROC GENMOD over PROC GLM for comparison of multiple correlated proportions, and PROC GLM appears to be more suitable for the calculation of adjusted confidence intervals.

REFERENCES

1. Litaker MS, Ferris DG. Comparison of Correlated Proportions using SAS® PROC GLM: a Simulation Study. *Proceedings of the SouthEast SAS Users Group Tenth Annual Conference; 2002.*

2. Ferris DG, Bishai DM, Litaker MS, Dickman ED, Miller JA, Macfee MS. Telemedicine network telecolposcopy compared with computer-based telecolposcopy. *Journal of Lower Genital Tract Disease 8(2):94-101; 2004*.

ACKNOWLEDGMENTS

The Telemedicine Study was funded by a grant from the National Cancer Institute and the Agency for Health Care Policy and Research (grant no. R01 HS08814).

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Mark S. Litaker, Ph.D.
Associate Professor / Director of Biostatistics
Department of Diagnostic Sciences
UAB School of Dentistry, SDB 111
1530 3rd Ave. S.
Birmingham, AL 35294-0007
Work Phone: (205) 934-1179
Fax: (205) 975-0603
Email:mlitaker@uab.edu