Conference on Applied Statistics in Agriculture   2001 - 13th Annual Conference Proceedings

A COMPARISON OF SOME METHODS TO ANALYZE REPEATED MEASURES ORDINAL CATEGORICAL DATA

Yaobing Sui

Walter W. Stroup

A COMPARISON OF SOME METHODS TO ANALYZE REPEATED MEASURES ORDINAL CATEGORICAL DATA

by Yaobing Sui and Walter W. Stroup
Department of Biometry, University of Nebraska, Lincoln, NE 68583-0712

Abstract: Recent advances in statistical software made possible by the rapid development of computer technology in the past decade have made many new procedures available to data analysts. We focus in this paper on methods for ordinal categorical data with repeated measures that can be implemented using SAS. These procedures are illustrated using data from an animal health experiment. The responses, measured as severity of symptoms on an ordinal scale, are recorded for test animals over time. The experiment was designed to estimate treatment and time effects on the severity of symptoms. The data were analyzed with various approaches using PROC MIXED, PROC NLMIXED, PROC GENMOD, and the GLIMMIX macro. In this paper, we compare the strengths and weaknesses of these different methods.

1.   Introduction

Consider an experiment in which three treatments are compared. There are r blocks of three animals each, formed using criteria relevant to the experiment. Within each block, one animal is assigned at random to each treatment. Animals are measured at "week 0," the time the treatments first take effect, and again at weeks 4 and 12. The variables measured include weight, presence or absence of disease symptoms, and severity of symptoms, classified as "worse," "no change," or "better." This type of experiment is called a repeated measures experiment. The focus of this paper is on repeated measures analysis of the last two types of data in the above list: categorical data that are either binary or ordinal.

Repeated measures data, also known as longitudinal data, come from experiments in which observations are made on subjects at regular, planned times. These experiments have two or more treatments and are set up using familiar designs: randomized complete or incomplete block designs, if blocking is appropriate, row-column designs such as Latin Squares, when appropriate, or completely randomized assignment of experimental units to treatments when blocking is not required. Repeated measures designs are widely used throughout the life sciences.

Repeated measures analysis is fairly well understood for normally distributed data, but less so for categorical data. However, recent developments in methodology and statistical computing software have greatly increased the number of tools available to categorical data analysts. The purpose of this paper is to review these tools, what we currently know of their advantages and disadvantages, and what we still need to learn about them.

Regardless of whether the observations are normally distributed, or categorical, or have some other distribution, a general approach to repeated measures analysis based on the linear mixed model uses the following general form:

observation = between subject systematic effects + between subjects random variation
        + within subjects systematic effects + within subjects random variation

For non-normal data, a function of the observation, e.g. the link function in a generalized linear

mixed model, often replaces the literal observation in the above model.

In the example that begins this section, the between subjects systematic effects are for block and treatment, the between subjects random effects correspond to block × treatment random effects - i.e. the between subjects model is identical to the model one would use for a randomized complete block analysis of variance. The within subjects systematic effects are the main effects of time and the treatment × time interaction. Within subjects random variation - formally, block × time within treatment variation - is essentially whatever is left unexplained, i.e. variation among the measurements at different times on the same experimental unit not explain by systematic effects already specified.

Formally, for normal errors, the model equation is:

$$y_{ijk} = \mu + \tau_i + r_j + b_{ij} + \omega_k + (\tau\omega)_{ik} + e_{ijk},$$

where y_{ijk} is the observation on the ith treatment, jth block at the kth week (or, more generally, time), μ is the intercept, τ_i is the ith treatment main effect, r_j is the jth block effect, b_{ij} is the ijth block-treatment random effect, assumed i.i.d. $N(0, \sigma_B^2)$, ω_k is the kth time main effect, $(\tau\omega)_{ik}$ is the ikth time-treatment interaction effect, and e_{ijk} is the ijkth within subject error. The e_{ijk} are assumed multivariate normal and, at least potentially, correlated.

There are two main distinguishing features of repeated measures analysis:
1. The primary objective is to see if changes over time are the same for each treatment, i.e. to assess the time × treatment interaction.
2. The errors, e_{ijk}, are correlated. Specifically, let $e_{ij}' = [e_{ij1}, e_{ij2}, \ldots, e_{ijK}]$ be the vector of within subjects errors, where K is the number of time periods observed. Then $e_{ij} \sim MVN(0, \Sigma)$, where Σ is the covariance matrix reflecting the correlation structure.

The vector $e' = [e_{11}', \ldots, e_{1r}', \ldots, e_{a1}', \ldots, e_{ar}']$ is thus distributed with a block-diagonal covariance matrix, i.e. $e \sim MVN(0, I_{ar} \otimes \Sigma)$, where a is the number of treatments.

With normal errors, repeated measures analysis can be implemented with mixed model software such as PROC MIXED. The main issues in using PROC MIXED for repeated measures analysis involve choosing an appropriate covariance model for Σ, realistically approximating the error degrees of freedom for various tests, and adjusting for potential bias of standard errors and test statistics that result from estimating the components of Σ. Readers seeking more detail on the use of PROC MIXED for repeated measures analysis are referred to Littell, et. al. (1996). Carlin and Louis (1996) discussed covariance model selection issues. Kenward and Roger (1997) discussed standard error bias and degree of freedom issues and presented approximations now available with PROC MIXED. Guerin and Stroup (2000) presented an extensive simulation study documenting the small sample behavior of PROC MIXED under various options.

Models with non-normal errors, including categorical data, require some modifications. To make these modifications more understandable, one can re-express the normal errors model in terms that make it more amenable to the required changes. Specifically, define the linear mixed model in terms of the distribution of the random models effects and in terms of the conditional distribution of the observations given the random model effects. Specifically,

$$y|u \sim MVN(X\beta + Zu, R) \text{ and } u \sim MVN(0, G).$$

The linear mixed model is a model of the conditional mean of the observation vector, y, given the random effects, u. For non-normal data, one adapts the generalized linear model approach used for categorical models such as logistic regression and log-linear models. Specifically, drop the

assumption of multivariate normality for $\mathbf{y|u}$ and use $\mathbf{X\beta+Zu}$ to model a function of the conditional mean, $E(\mathbf{y|u})$, called the *link function* in generalized linear models. This results in the *generalized linear mixed model* (GLMM), widely discussed in the statistical literature of the 1980's through the present. See, for example, Breslow and Clayton (1993). The GLMM is thus described as follows:

1. The **distribution of the random effects**: $u \sim \text{MVN}(0,G)$
2. The **conditional distribution of the observations**, y, given the random effects, u. For categorical data, this distribution is typically assumed Poisson (for log-linear models fit to contingency tables), binomial (for logistic models), or multinomial (for extensions of logit models when there are more than two categories). Quasi-likelihood methods allow the use of GLMM-based analysis even when one can only specify the expected value and variance of $y|u$ rather than the distribution *per se.*
3. The *inverse link*, $E(y|u) = h(\mathbf{X\beta+Zu})$. The inverse link may be the inverse of the link function, or the inverse link may be a set of functions, as is the case for some multinomial models. With the latter case, there is no one-to-one relationship between the conditional mean and the link. When a one-to-one relationship does exist, the GLMM can be described in terms of the link function, that is, $\eta=\mathbf{X\beta+Zu}$, where $\eta=g[E(y|u)]$ is the link function.

For the randomized complete block design with repeated measures described above, the GLMM would thus be

$$\eta_{ijk} = \mu + \tau_i + r_j + b_{ij} + \omega_k + (\tau\omega)_{ik}$$

where η_{ijk} is the link function, $g[E(y_{ijk} | b_{ij})]$, and the terms of the right-hand side of the model are defined as they were with the linear mixed model given previously. Alternatively, one can use the inverse link $\quad E(y_{ijk} | b_{ij}) = h[\mu + \tau_i + r_j + b_{ij} + \omega_k + (\tau\omega)_{ik}]$

Several options exist in SAS for fitting categorical repeated measures models. PROC GENMOD can be used to fit log-linear models. For *binomial data only*, GENMOD can also fit certain GLMM's for repeated measures using the method of *generalized estimating equations* (Zeger, et. al. 1988), commonly referred to as GEE's. The GLIMMIX macro can also fit repeated measures GLMM's to binomial data. GLIMMIX uses a *pseudo-likelihood* approach (Wolfinger and O'Connell, 1993) that is similar to the quasi-likelihood approach described by Breslow and Clayton (1993), but somewhat more general. GLIMMIX is not as restrictive as the GENMOD GEE option in terms of the types of covariance models available. PROC NLMIXED, introduced in SAS Version 8, can estimate repeated GLMM's for multinomial data in addition to models for binomial data. It uses a maximum likelihood algorithm based on Gaussian quadrature. With some programming ingenuity, NLMIXED can fit a certain covariance matrices, although convergence can be an issue with more complex structures.

The next section describes in more detail SAS-based methods useful for categorical repeated measures data, with a focus on ordinal data. Section 3 presents an example from an animal health experiment. Section 4 presents some tentative simulation results. These will be pursued in far more detail in work now in progress.

2.   Review of Methods

Table 1 shows the data for the experiment described at the beginning of Section 1 in contingency table form. Each cell contains the number of animals in a given treatment × week × response category combination. This section describes the methods available in SAS to analyze

Applied Statistics in Agriculture

these data.

The simplest categorical data analysis approach is to compute the Cochran-Mantel-Haenszel statistic to test treatment × response category association. A statistically significant result constitutes evidence of a treatment effect, assuming that the association does not change over weeks. SAS PROC FREQ can compute the Cochran-Mantel-Haenszel test. It can also compute the Breslow-Day statistic for no three-way treatment × response category × week association (i.e. no change in treatment × response association over weeks) if the treatment × response table is 2 × 2, but not for the more general case, such as the 3×3 shown here. See Agresti (1996) for a more in depth discussion of the contingency table approach.

Alternatively, the contingency table approach can be implemented using a log-linear model. For the above example, the log-linear model is

$$\log(\lambda_{ijk}) = \mu + \tau_i + \omega_j + (\tau\omega)_{ij} + c_k + (\tau c)_{ik} + (\tau\omega c)_{ijk}$$

where λ_{ijk} is the expected count of the ijk[th] treatment × week × response category combination, and τ, ω, and c refer to treatment, week, and response category effects, respectively. The two effects of primary interest are the three-way association effects and, assuming the three-way effects, $(\tau\omega c)_{ijk}$, are zero, the two-way treatment × response category effects. The test of H_0: all $(\tau\omega c)_{ijk}=0$ is equivalent to the Breslow-Day test, but more general because it is not restricted to 2×2 treatment×response category cases. The test of H_0: all $(\tau c)_{ij}=0$ is equivalent to the Cochran-Mantel-Haenszel test. PROC GENMOD can do all the required computations for the log-linear model.

While the log-linear model is easy to compute, the contingency table approach may not take correlation among repeated measurements on the same experimental unit into account realistically. Agresti (1996) presents the logic of the contingency table approach when there are two times, but the logic does not necessarily extend to three or more times. Approaches using GEE's or other GLMM methods with more flexibility in specifying the covariance structure are, at least in theory, preferable.

In SAS, *for binary data only*, GEE's can be implemented using the REPEATED option in PROC GENMOD. This approach is limited in that it assumes no random model effects. The model thus

$$\eta_{ijk} = \mu + \tau_i + r_j + \omega_k + (\tau\omega)_{ik}$$

where η_{ijk} is usually either the logit or probit link, and τ, r, and ω refer to treatment, block, and week effects, respectively. The logit link is defined as $logit(\pi_{ijk}) = \log\left(\dfrac{\pi_{ijk}}{1-\pi_{ijk}}\right)$, where π_{ijk} is the probability of the outcome of interest occurring for the ijk[th] treatment×block×week combination. The probit link is defined as probit$(\pi_{ijk}) = \Phi^{-1}(\pi_{ijk})$, where Φ^{-1} is the inverse cumulative standard normal distribution. The observations are assumed to have a covariance matrix R=DPD, where D=diag$\left(\dfrac{\pi_{ijk}(1-\pi_{ijk})}{n_{ijk}}\right)$, and n_{ijk} is the number of Bernoulli trials observed on the ijk[th] treatment×block×week combination. The form of D given here is specific to the binomial distribution. In general, D a diagonal matrix whose elements are the variance functions with for each treatment×block×week combination. P is a working correlation matrix. Working correlation matrices are not true correlation matrices, but their structure follows common correlated error

structures in linear mixed models, e.g. compound symmetry, first-order autoregressive, etc. See Diggle, et. al. (1994) for more about working correlation matrices and GEE's. Vonesh and Chinchilli (1997) provide a comprehensive discussion of modeling issues for repeated measures.

Also *for binary data only* the GLIMMIX macro can implement GLMM's using a pseudo-likelihood algorithm, essentially a generalized linear model adaptation of the mixed model algorithm used in PROC MIXED. This approach is similar to the GEE's in the REPEATED option in PROC GENMOD, except that GLIMMIX allows for separate random model effects in addition to the working correlation structure. For example, GEE's allow AR(1) correlation only, but no separate estimate of random between subjects effects, the b_{ij} effects in the above GLMM for the example in Table 1. This is important, because within subject correlation and between subject variation are often distinct effects. Guerin and Stroup (2000) demonstrated, for example, that failing to estimate separate AR(1) correlation and between subject error variance components can seriously impair control over type I error in repeated measures analysis of normally-distributed data. We expect this to be true for non-normal data as well. Hence, GLIMMIX should have an advantage over the GENMOD REPEATED option in these cases.

The logistic and probit capabilities of GLIMMIX and GENMOD are mainly limited by the restriction of binomial data only. For multinomial responses, multi-category logit and probit models exists, but cannot be estimated by GENMOD or GLIMMIX. For ordinal data used in Table 1, one can fit cumulative logit or cumulative probit models. The cumulative logit model defines c-1 logits, where c is the number of response categories. For three categories, "worse," "no change," and "better", code the responses as -1, 0, and 1 respectively and define the two logits as

$$\text{cumulative logit for category "worse (-1)"} = l_{-1} = \log\left(\frac{\pi_{-1}}{1-\pi_{-1}}\right) \text{ and}$$

$$\text{cumulative logit for category "no change (0)"} = l_0 = \log\left(\frac{\pi_{-1}+\pi_0}{1-(\pi_{-1}+\pi_0)}\right)$$

where π_{-1}, π_0, and π_1 are the probabilities of a -1, 0, and 1 (worse, no change, or better) respectively. Then fit the models

$$l_{-1} = \eta_{-1,ijk} = \mu_{-1} + \tau_i + r_j + b_{ij} + \omega_k + (\tau\omega)_{ik} \text{ to the logit } l_{-1}, \text{ and}$$

$$l_0 = \eta_{0,ijk} = \mu_0 + \tau_i + r_j + b_{ij} + \omega_k + (\tau\omega)_{ik} \text{ to the logit } l_0.$$

The parameters of the model are defined as in previous GLMM's, except for the intercepts. The intercept μ_{-1} is for logit l_{-1} and the intercept for logit l_0 is $\mu_0 = \mu_{-1} + \delta$, where $\delta > 0$. The inverse link functions that allow one to compute the probabilities are thus

$$\hat{\pi}_{-1} = \frac{\exp(\hat{\eta}_{-1})}{1+\exp(\hat{\eta}_{-1})}$$

$$\hat{\pi}_0 = \frac{\exp(\hat{\eta}_0)}{1+\exp(\hat{\eta}_0)} - \frac{\exp(\hat{\eta}_{-1})}{1+\exp(\hat{\eta}_{-1})}$$

$$\hat{\pi}_1 = 1 - \left[\frac{\exp(\hat{\eta}_0)}{1+\exp(\hat{\eta}_0)}\right]$$

In standard categorical data analysis, this is also known as a *proportional odds* model. PROC

LOGISTIC can be used to estimate proportional odds models for fixed effects data with no repeated measures structure. However, for the mixed model shown here, PROC NLMIXED must be used.

One can replace the logit link with a probit link. Specifically,

$$\eta_{ijk} = \mu + \tau_i + r_j + b_{ij} + \omega_k + (\tau\omega)_{ik}$$

where the various terms are defined as previously, and the inverse links are thus,

$$\pi_{-1,ijk} = \Phi(-\eta_{ijk})$$

$$\pi_{0,ijk} = \Phi(\delta - \eta_{ijk}) - \Phi(-\eta_{ijk})$$

$$\pi_{1,ijk} = 1 - \Phi(\delta - \eta_{ijk})$$

Details for programming the analysis of this model with SAS PROC NLMIXED are given in the SAS/STAT Online Documentation for Version 8, Example 46.3. In some disciplines, such as animal breeding, the cumulative probit model is called a *threshold model*.

For binomial data, GLIMMIX or NLMIXED can be used to compute GLMM's. The main disadvantage of GLIMMIX is that the pseudo-likelihood algorithm may yield biased estimates when n_{ijk}, the number of Bernoulli trials per experimental unit (treatment×block×week combinations in the Table 1 example), is equal to 1 (see Breslwo and Clayton, 1993). Of interest in this paper is to determine if the computing algorithm used by NLMIXED is less prone to this sort of bias. On the other hand, although NLMIXED *can* accommodate models with correlated errors, such models may require programming heroics. Also, the NLMIXED algorithm is prone to convergence problems. While NLMIXED has a vast array of options to improve convergence, there is no guarantee that these will be effective. Finally, NLMIXED uses maximum likelihood estimation; there is no restricted maximum likelihood option. For normal-errors models, maximum likelihood variance component estimates are known to be biased downward, and the resulting test statistics that use these variance component estimates are biased upward. Type I error rates with maximum likelihood can be atrocious. If this is also true of GLMM's, NLMIXED's usefulness may be seriously limited.

Table 2 shows a summary of the pro's and con's of the various options available in SAS. Table 3 classifies data by number of response categories and by number of times at which repeated measures are taken and list the SAS PROC's that have suitable options.

3. An Example of Results Using the Various Procedures

Following the design format described at the beginning of Section 1, three treatments, a placebo and an experimental treatment applied at a low level (treatment 1) and a high level (treatment 2) were compared. The animals were dogs. A total of 199 dogs were observed. The placebo and treatment 2 were each observed on 67 animals; two dogs were lost from treatment 1 and hence only 65 animals were observed. Each dog was observed at week 0, 4, and 12. Note that dogs were lost as the experiment proceeded, so the frequencies for the tables at weeks 4 and 12 have lower counts. The response variables were house training, orientation, activity, sleep, response, greeting, and health. Table 1 shows the "house training" variable only. Each variable had three response categories: worse (-1), no change (0), improved (1). For week 0, the response was relative to a pre-trial baseline measurement.

The data were analyzed using the four methods described in Section 2. Table 4 shows the results for the variable "house training" of the tests for treatment, time (week), and treatment ×

time effects obtained from the four methods. The house training is used simply to illustrate possible outcomes, similarities, and contradictions that may occur with these methods of analysis.

3.1 Analysis using Log-linear Model

This analysis used the model given in the discussion of log-linear models in Section 2. The following SAS statements generate the analysis:

```
proc genmod data=loglin;
  classes week trt ht:
  model count=week|trt|ht/dist=poisson link=log type1 type3;
```

The tests for the treatment, week, and treatment×week effects are actually tests of associations between these effects and the response category. The tests of primary interest, therefore, are of H_0: all $(\tau c)_{ij}=0$ (no treatment effect) and H_0: all $(\tau\omega c)_{ijk}=0$ (no treatment×week interaction). From Table 4, the likelihood ratio χ^2 for treatment×week interaction is 3.81 with a p-value of 0.7029 and the χ^2 for the treatment main effect is 15.86 with p=0.0032. Note that in theory there should be 4 d.f. for treatment and week (2 d.f. for response category × 2.d.f. for treatment or week) and 8 d.f. for the treatment×week interaction. However, for week and treatment×week only 3 d.f. and 6 d.f., respectively, are shown in Table 4. This is because at week 0, none of the animals on any of the treatments, including the placebo, had a response in category 1 ("better").

The 4 d.f. treatment main effect, i.e. treatment×response category association, can be partitioned into single d.f. contrasts by adding the following statements to the above SAS PROC GENMOD statements:

```
contrast 'ctl vs trt. -1 vs 0&1' trt*ht 4 -2 -2 -2 1 1 -2 1 1;
contrast 'ctl vs trt. 0 vs 1' trt*ht 0 2 -2 0 -1 1 0 -1 1:
contrast 'trt 1 vs 2. -1 vs 0&1' trt*ht 0 0 0 2 -1 -1 -2 1 1;
contrast 'trt 1 vs 2, 0 vs 1' trt*ht 0 0 0 0 1 -1 0 -1 1;
```

For example, the contrast labeled (p v t) × (-1 v ≥0) is the association between the difference between placebo and the average of the two treatments, and response category -1 vs categories 0 and 1. This tests the difference between placebo and the two treatments with respect to the likelihood of dogs getting worse versus not getting worse. The χ^2 value is 9.23 and the p-value is 0.0024. The other contrasts forming an orthogonal set are also shown. They indicate that the only significant component of the treatment × response category association is the contrast just described.

3.2   Analysis using GEE

GEE analysis uses the REPEATED option of GENMOD, which requires binary data. For this example, the "no change" and "better" categories were combined, so the binary response became "worse" versus "not worse." The following SAS statements were used to compute the analysis:

Applied Statistics in Agriculture

```
PROC GENMOD DATA=binarv;
 class animal trt week:
 model ht= trt|week /dist=binomial link=logit type3;
 repeated subject=animal(trt)/type=exch;
```

The model corresponded to the GEE analysis in Section 2 with a logit link and a compound symmetry working correlation matrix.

Obviously, some information was lost relative to not combining these response categories. However, for the treatment effect on worse vs. not worse, the GEE analysis showed results similar to the log-linear model: a non-significant treatment×week interaction (p=0.4085) and a significant treatment main effect (p=0.0174).

3.3 Analysis using GLIMMIX

This analysis used the binary GLMM as given in Section 2 with a logit link and a between subjects error variance. The following SAS statements were used:

```
%include 'c:\Program Files\SAS Institute\SAS\V8\Glmm800.sas';
%Glimmix (data=binarv1, procopt=method=reml, stmts=%str (
    class animal trt week;
     model ht = trt|week:
    random animal(trt);),
    error=binomial,
    link=logit);
```

For normal errors models, compound symmetry and between subjects random model effect result in the same covariance structure, but this is not true for generalized linear mixed models. As with the GEE analysis, the binary worse vs. no worse response variable was used. The results were similar: no significant treatment×week interaction (p=0.3341) and a significant treatment main effect (p=0.0083). Note that GLIMMIX uses an approximate F-test rather than a χ^2 test.

3.4. Proportional odds model using NLMIXED

This analysis used the cumulative logit, or proportional odds, model with a random between subjects error effect as described in Section 2. The following SAS statements were used:

```
    data NLMX binarv: set binarv:
      t1=(trt='1'):  t2=(trt='2'):  t3=(trt='3');
      w1=(week=4):  w2=(week=12);  w3=(week=0);
     proc nlmixed data=week4 12:
      parms b0=0 a1=0 a2=0 b1=0 ab11=0 ab21=0 s2w=1 s2s=1 i1=1;
      bounds i1 > 0:
      eta = b0 + a1*t1 + a2*t2 + b1*w1 + ab11*t1*w1 +ab21*t2*w1 + wpe;
      if (ht=-1) then z = exp(eta)/(1+exp(eta));
      else if (ht=0) then
          z = exp(i1+eta)/(1+exp(i1+eta))- exp(eta)/(1+exp(eta));
```

```
else z = 1 - exp(i1+eta)/(1+exp(i1+eta));
if (z > 1e-8) then ll = log(z);
else ll = -1e100:
model ht ~ general(ll):
random wpe ~ normal(0,s2w)
        subject=animal:
contrast 'trt' a1, a2:
contrast 'week' b1, b2:
contrast 'trt*week' ab11, ab12, ab21, ab22;
```

As with GLIMMIX, NLMIXED computes approximate F-statistics. For these data, the analysis shows no significant treatment×week interaction (p=0.4549) and also no significant treatment effect (p=0.1838). These results are distinctly more conservative than the other three methods.

Note that all four methods showed a highly significant time (week) effect. For these data, the important discrepancy was among the tests of treatment effect: all analyses except the proportional odds model showed a significant treatment effect. Also, a theoretically inelegant, but often used analysis is to ignore the fine points of the multinomial distribution and use -1, 0, and 1 as if they were normally distributed data and compute the analysis directly with PROC MIXED. For these data, the conclusions were the same for treatment×week interaction and treatment main effect as obtained with the log-linear model, GEE, and the binary GLMM using GLIMMIX. The linear mixed model is sufficiently robust that in many cases it leads to the same conclusions as procedures explicitly for categorical data.

4.   A Tentative Simulation Study

A comprehensive simulation study is in progress. This section reports the basic approach to the study and some highly tentative results. Thought tentative, the authors regard the results as sufficiently interesting to those who work with categorical data to warrant reporting.

Three types of simulated binary data were generated. Each data set had two treatments and twenty animals per treatment. The between subjects design was a CRD. For each animal, there were four repeated measures. All data sets assumed no treatment or time effects. The goal was to measure type 1 error rates and the precision of the estimates of treatment and time effects. The types of data were determined by the probability of a "success" (π) and by the variance-working correlation structure. The three types were:
1. $\pi = 0.1$ and covariance structure=CS
2. $\pi = 0.5$ and covariance structure=CS
3. $\pi = 0.5$ and covariance structure=AR(1)
For the types with $\pi = 0.5$, 500 sets of data were simulated. For the type with $\pi=0.1$, 100 sets of data were simulated. Each data set was analyzed using the four methods described in Section 2. For NLMIXED, the model was a generalized linear mixed model with a logit link. Hence, the GLIMMIX and NLMIXED computed analyses on the same model and differ only in the computing algorithm (pseudo-likelihood for GLIMMIX, Gauss-Hermite quadrature for NLMIXED). GEE differed from the GLIMMIX and NLMIXED models in that only a working correlation structure could be used rather than the random between subjects error effect plus the working correlation.

Table 5 shows type I error rate based on type3 tests of fixed effects using the $\pi=0.5$, AR(1)

simulated data. Clearly, treatment×time interaction, treatment effect, and time effect were under-rejected using PROC NLMIXED. Treatment×time interaction and time main effect were also under-rejected using the log-linear model approach implemented by PROC GENMOD. On the other hand, the log-linear model tended to over-reject for the tests treatment effects. These results reflect the log-linear model's failure to adequately account for between-subject versus within-subject variation. Overall, GEE and GLIMMIX did a good job in that their observed rejection rates were close to the nominal α-levels for all tests. For the log-linear model, GEE, and GLIMMIX analyses all 500 sets of data converged. With NLMIXED, 32 of 500 sets of data did not converge.

Table 6 shows the average of parameter estimates and their sample variance. As noted previously the treatment, time, treatment× time parameters were set to zero when the data were simulated. For the $\pi=0.5$ data, the intercept parameter was also 0. The average parameter estimates are very similar for the four methods. The sample variance for each parameter estimate, while large for all methods, was smaller for GEE than for GLIMMIX and NLMIXED.

Simulation results for covariance parameter estimates are only shown for NLMIXED. The AR(1) parameters ρ and σ^2 were set to 0.75 and 1, respectively when simulating the data. The estimated ρ and σ^2 are 0.66 and 1.15, respectively, very close to the values used for simulating the data.

5. Discussion and Conclusions

There is much we do not yet know about the analysis of categorical data analysis with repeated measures. Except for the log-linear model, the procedures discussed here are adaptations of linear models for repeated measures data used for normally-distributed data. The main attraction of this approach is the ability to assess treatment and time effects and to characterize between- and within-subject variability in ways that are familiar with other response variables. GLMM methods place all response variables, categorical or otherwise, under a common methodology.

While the *models* have much appeal, the computational methods are just in their beginning stages of development, and there is much to understand. For the example data set, all of the methods except the proportional odds model computed using NLMIXED would have led to similar conclusions. The tests computed by the proportional odds model were noticeably more conservative. This observation was borne out by the simulation results. Though tentative, NLMIXED yielded a substantially lower rejection rate than either the nominal α-rate or the other procedures. The expected downward bias in variance estimates and resulting upward bias in test statistics seen in linear mixed models was not evident in the simulations discussed in Section 4. The log-linear model analysis, on the other hand, showed substantial departures from the nominal α-rates consistent with its not taking between- versus within-subject error into account. The GEE and GLIMMIX procedures performed rather similarly, although GEE did somewhat better despite the fact that it modeled between-subjects variation indirectly through a working correlation matrix, rather than directly estimating the variance component is GLIMMIX does. This simulation did not include autocorrelated within-subjects effects in conjunction with between-subject variation of substantial magnitude, so we do not know how GEE and GLIMMIX will compare with more complex covariance situations. At this point, we conclude that trying to make a definitive recommendation about how to analyze categorical repeated measures data is problematic.

Clearly, more work it needed. Further simulations will consider a wider variety of covariance situations and include multinomial responses as well. More refined methods of analysis, including GEE2 and various Monte Carlo methods exist and more are being developed. As yet, these are not available in SAS, but software evolves quickly. These methods need to be assessed for their advantages relative to what is currently SAS-accessible, or available in some form equally usable to consultants and subject-matter researchers. Until this happens, questions about how best to analyze categorical repeated measure data will continue to be controversial.

Applied Statistics in Agriculture

6. References

Agresti, A. 1996. *An Introduction to Categorical Data Analysis.* New York: Wiley.

Breslow, N.E. and D.G. Clayton. 1993. "Approximate inference in generalized linear mixed models." *J. Amer. Statist. Assoc.* **88**: 9-25.

Carlin, B.P. and T.A. Louis. 1996. *Bayes and Empirical Bayes Methods for Data Analysis.* Boca Raton, FL: Chapman and Hall.

Diggle, P.J., K.-Y. Liang, and S.L. Zeger. 1994. *Analysis of Longitudinal Data.* Oxford: Oxford University Press.

Guerin, L. and W.W. Stroup. 2000. "A simulation study to evaluate PROC MIXED analysis of repeated measures data." *Proceedings: 12th Annual Conference on Applied Statistics in Agriculture.* Manhattan, KS: Kansas State University Statistics Department.

Kenward, M.G. and J.H. Roger. 1997. "Small sample inference for fixed effects from restricted maximum likelihood." *Biometrics* **53**: 983-997.

Littell, R.C., G.A. Milliken, W.W. Stroup, and R.D. Wolfinger. 1996. *SAS System for Mixed Models.* Cary, NC: SAS Institute.

SAS Institute, Inc. 1999. *SAS Online Doc, Version 8.* Cary, NC: SAS Institute.

Vonesh, E.F. and V.M. Chinchilli. 1997. *Linear and Nonlinear Models for the Analysis of Repeated Measurements.* New York: Marcel-Dekker.

Wolfinger, R.D. and M. O'Connell. 1993. "Generalized linear mixed models: a pseudo-likelihood approach." *J. Statist. Comput. Simul.* **48**: 233-243.

Zeger, S.C., K.-Y. Liang, and P.S. Albert. 1988. "Models for longitudinal data: a generalized estimating equation approach." *Biometrics* **44**: 1049-1060.

Table 1.   Example of Repeated Measures Ordinal Categorical Data in Contingency Table Form

Response category ↓ Trt →	Week 0			Week 4			Week 12		
	Plac	Trt 1	Trt 2	Plac	Trt 1	Trt 2	Plac	Trt 1	Trt 2
Worse	60	59	54	14	5	3	13	10	7
No Change	7	6	13	34	33	38	25	17	21
Better	0	0	0	15	22	17	17	28	21

Table 2. Advantages and Disadvantages of SAS Procedures

Log-linear Model
    Advantages
        1.   East to implement using PROC GENMOD.
        2.   Can be used with any number of response categories
        3.   Response category may be ordinal or nominal

    Disadvantage
        1.   Unrealistic modeling of within subject correlation for ≥ 3 times of measurement.

GEE
    Advantages
        1.   Can be implemented using REPEATED option in PROC GENMOD
        2.   Offers choice of plausible working correlation structures when # times ≥ 3.

    Disadvantages
        1.   Limited to binary (2 response category) case
        2.   Cannot model between subject error variance and working correlation separately

GLIMMIX
    Advantages
        1.   Very flexible in modeling variation, e.g. permits separate estimates of between subject error variance and working correlation
        2.   Ease of syntax: similar to PROC MIXED

    Disadvantages
        1.   Limited to binary case.
        2.   Pseudo-likelihood algorithm vulnerable to bias when N (number of Bernoulli trials per experimental unit) is small, especially when N=1.

NLMIXED
    Advantages
        1.   Can be used for binary or multinomial data
        2.   Can be programmed to account for between subject error variance and working correlation separately
        3.   Gauss-Hermite quadrature less susceptible to bias when N small

    Disadvantages
        1.   Maximum likelihood only. ML estimates of variance components known to be downward biased and hence test statistics biased upward.
        2.   Prone to convergence problems.
        3   No CLASS statement makes analysis of treatment and time effects less convenient.

Table 3. Applicability of SAS Procedures by Response Type and Number of Measurement
Times

Response Categories →

Number of Measurement
Times ↓

	2	≥3
2	ALL	Log-linear Model (GENMOD) NLMIXED
≥3	GEE (GENMOD) GLIMMIX NLMIXED	NLMIXED

Applied Statistics in Agriculture

Table 4: Type 3 Tests of Fixed Effects for Analysis of Example Data, House Training Response

Method	Source	DF		Chi-Square	Pr > ChiSq
Log-Linear	trt	4		15.86	0.0032
	(p v t) × (-1 v ≥0)		1	9.23	0.0024
	(p v t) × (0 v 1)		1	2.66	0.1028
	(t1 v t2) × (-1 v ≥0)		1	1.76	0.1821
	(t1 v t2) × (0 v 1)		1	2.64	0.1042
	week	3		200.71	<.0001
	week*trt	6		3.81	0.7029
Working Corr (GEE)	trt	2		8.10	0.0174
	week	2		147.29	<.0001
	trt*week	4		3.98	0.4085

Method	Source	NDF	DDF	F Value	Pr>F
BSE + W Corr (GLIMMIX)	trt	2	196	4.92	0.0083
	week	2	334	184.02	<.0001
	trt*week	4	334	1.15	0.3341
BSE + W Corr (NLMIXED)	trt	2	198	1.71	0.1838
	week	2	198	25.04	<.0001
	trt*week	4	198	0.92	0.4549

Table 5. Observed Rejection Rates for Time and Treatment Main Effects and Treatment
×Time Interaction at α=0.01,0.05,and 0.1
AR(1) data, π=0.5

Analysis*	Nominal alpha**	N	time	trt	time*trt
			Observed Rejection Rate		
Log-Linear	reject_10	500	0.070	0.218	0.044
(GENMOD)	reject_05	500	0.024	0.144	0.016
	reject_01	500	0.002	0.054	0.002
GEE	reject_10	500	0.116	0.114	0.102
(GENMOD)	reject_05	500	0.068	0.052	0.046
	reject_01	500	0.010	0.010	0.002
Random BSE	reject_10	500	0.144	0.098	0.128
+AR(1)	reject_05	500	0.086	0.040	0.086
_(GLIMMIX)	reject_01	500	0.022	0.006	0.012
Random BSE	reject_10	468	0.036	0.034	0.038
+ AR(1)	reject_05	468	0.021	0.015	0.015
(NLMIXED)	reject_01	468	0.011	0.006	0.009

* Type of Analysis as described in Section 2
  BSE denotes "between subjects error"
  GEE denotes "Geneeralized estimating equations" with AR(1) working correlation matrix

** "Reject_10" denotes nominal α-level=0.10. Nominal α levels are 0.10, 0.05, and 0.01 respectively

Table 6. Average of Parameter Estimates, AR(1) data, $\pi=0.5$
$\mu=0$, trt effects=0, time effects=0, trt×time interaction effects=0, $\rho=0.75$, $\sigma^2=1$

method	parameter	N	Mean	Sample Variance
GEE - GENMOD	μ	500	0.0033085	0.2197468
	TRT	500	-0.0514023	0.4127234
	TIME	500	0.0087798	0.3293952
	TRT*TIME	500	0.0381633	0.6515009
GLIMMIX	μ	500	0.0038056	0.2822495
	TRT	500	-0.0576805	0.5290325
	TIME	500	0.0018012	0.4432861
	TRT*TIME	500	0.0439123	0.8719174
NLMIXED	μ	468	-0.0081427	0.2932152
	TRT	468	-0.0524051	0.5110301
	TIME	468	0.0223616	0.4292667
	TRT*TIME	468	0.0187420	0.7501278
	ρ	468	0.6560738	0.1988518
	σ^2	468	1.1496017	0.7700970