Kansas State University Libraries

New Prairie Press

Conference on Applied Statistics in Agriculture 1997 - 9th Annual Conference Proceedings

ANALYSING BINARY DATA IN A REPEATED MEASUREMENTS SETTING USING SAS

Eleanor F. Allan

Follow this and additional works at: https://newprairiepress.org/agstatconference

 Part of the Agriculture Commons, and the Applied Statistics Commons

This work is licensed under a Creative Commons Attribution-Noncommercial-No Derivative Works 4.0 License.

Recommended Citation

Allan, Eleanor F. (1997). "ANALYSING BINARY DATA IN A REPEATED MEASUREMENTS SETTING USING SAS," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1293

This is brought to you for free and open access by the Conferences at New Prairie Press. It has been accepted for inclusion in Conference on Applied Statistics in Agriculture by an authorized administrator of New Prairie Press. For more information, please contact cads@k-state.edu.

ANALYSING BINARY DATA IN A REPEATED MEASUREMENTS SETTING USING SAS

Eleanor F Allan
Statistical Services Centre
Department of Applied Statistics
The University of Reading
Reading RG6 6FN
UK

Abstract

Whilst the repeated measurements methods appropriate for the analysis of normally distributed data are well established, methods for handling binary and categorical data in a repeated measurements context are not so commonly known or used. The application of population averaged models and subject effects models to repeated binary data are discussed and their implementation with the aid of SAS are illustrated by example. Comparisons with other approaches are also considered.

1. Introduction

Experiments to compare treatments often involve taking repeated measurements of a response variable (or response variables) for each experimental unit. Examples might be comparisons of different insecticides applied to crops, or different forage treatments fed to cattle, where interest is in the response to treatment over time. Since treatments are randomly allocated to the experimental units, with several units per treatment, whilst the repeated measurements are made on the experimental units, the data have two different levels of variability : within and between unit variability. When comparing treatments in such a setting, one is usually interested in how the response over time is influenced by treatment. This comparison of treatment profiles can be made by interpreting the treatment main effect and the treatment-by-time interaction simultaneously.

When the response of interest is a normally distributed variable (or can be transformed to normality), there are many well-developed methods of analysis available for comparing treatments, which deal with the repeated measurements structure. One of these is the split-plot-in-time analysis of variance where the treatment main effect is assessed relative to the between unit variability and the treatment by time interaction relative to the within unit variability.

Response variables in agricultural experiments however are not always normally distributed, but can be binary or even categorical. For instance, the presence or absence of aphids (yes, no) may be of interest in the crop experiment to compare insecticides or the condition score of a cow (1=normal, 2=subdued, 3=dull, 4=very dull) in a livestock study. These responses can be

recorded on several occasions throughout the experiment, in which case the objective is to assess the treatment effects over time. Methods of analysis which are available are based on the underlying assumption of a multinomial distribution. A full discussion of the different approaches to analysing binary and categorical repeated measurement data may be found in Kenward (1992). Agresti (1989) also gives an account of methods available when the data are of an ordered categorical structure. Two methods of analysis which can be used for binary repeated measurement data will be discussed here: *the marginal probability model* and *the subject effects model*. In the former a model is fitted to the probability of the level of a response at each time, irrespective of the observation recorded at other times. The subject effects model on the other hand, is the categorical equivalent of the split-plot analysis, including an effect for the experimental unit in the model. How these two analyses for binary data can be implemented in SAS and the results interpreted will be described here using an example. The details of the methodologies are described only briefly.

2. Example

Consider the following example looking at swine fever in pigs. Three treatments were randomly allocated to the pigs; treatment A is a currently used vaccine, B and C are new vaccines. The pigs were vaccinated on day -5, then challenged with the virus on day 0. Respiratory rate was recorded daily for 7 days as either normal (0) or increased (1). The important times post-challenge were days 1, 3 and 7. The data recorded at these times are summarised in Table 1.

From the joint frequencies presented in this table observed marginal proportions, i.e. proportions responding with a 0 or 1 on each day, can be calculated (Table 2). Inspection of these data suggests that the probability of having a normal respiratory rate increases with time. The effect of treatment is less clear, as is the treatment-by-time interaction. A marginal probability model would attempt to model the probability of a response on a particular day for a particular treatment in terms of effects due to treatment, time etc.

3. Marginal Probability Model

3.1 Method

One method of fitting marginal probability models is the *empirical generalised least squares* approach as discussed by Koch and Landis (1977). This can be implemented in SAS using PROC CATMOD.

The method fits a linear model to functions of the marginal probabilities :

$$E\left\{\underset{\sim}{f}(\underset{\sim}{p})\right\} = X\underset{\sim}{\beta}$$

where $\underset{\sim}{p}$ is the vector of observed marginal probabilities

$\underset{\sim}{f}(p)$ is a $(n \times 1)$ vector of functions of these marginal probabilities

X is the design matrix

$\underset{\sim}{\beta}$ is a $(p \times 1)$ vector of unknown parameters

In the case of binary data, the most frequently-used function is the logit of the "success" probability, and in our example this is

$$\log\left\{\frac{P(\text{response} = 0)}{P(\text{response} = 1)}\right\}$$

since a normal respiratory rate can be regarded as a "success". There is one observed response function for each margin (day) for each treatment.

The covariance matrix of these functions, W, is block diagonal, with blocks for each treatment group. The functions are neither independent, nor do they have constant variance; hence the ordinary least squares method of model fitting does not apply. Parameters are estimated instead using generalised least squares :

$$\underset{\sim}{\tilde{\beta}} = (X^T \tilde{W}^{-1} X)^{-1} X^T \tilde{W}^{-1} \underset{\sim}{f}$$

with estimated covariance matrix

$$\tilde{\psi} = (X^T \tilde{W}^{-1} X)^{-1}$$

where $\tilde{W}$ is the sample estimate of W.

To test for example the treatment effect, one can formulate a hypothesis of the form

$$H_0 : \underset{\sim}{\beta}_1 = \underset{\sim}{0} \qquad \text{(a subset of the parameters)}$$

and carry out a Wald test

$$X^2 = \underset{\sim}{\tilde{\beta}}_1^T \tilde{\psi}_1^{-1} \underset{\sim}{\tilde{\beta}}_1$$

which has an asymptotic χ^2 distribution with p_1 degrees of freedom (where p_1 is the number of independent parameters, and $\tilde{\psi}_1$, is the appropriate $(p_1 \times p_1)$ subset of $\tilde{\psi}$).

The residual sum of squares, $(\underset{\sim}{f} - X\underset{\sim}{\tilde{\beta}})^T \tilde{W}^{-1} (\underset{\sim}{f} - X\underset{\sim}{\tilde{\beta}})$ has an asymptotic χ^2 distribution with (n-p) degrees of freedom, and can be used to assess the goodness-of-fit of the model.

3.2 Analysis

Appendix 1 contains a SAS program for fitting a marginal probability model with effects for time (period) and treatment to the data in the example, using a logit transformation of the probability of a normal response (score=0). The resulting output is also presented.

A model of the form: $f_{ij} = \mu + t_i + p_j$ is fitted to the data where f_{ij} is the response function for treatment i in period j; i = 1,2,3; j = 1,2,3

μ is overall mean
t_i is the effect associated with the ith treatment
p_j is the effect associated with the jth period

The residual chi-square on 4 degrees of freedom in the CATMOD output assesses the goodness-of-fit of the model, or the significance of the treatment-by-period interaction; and is clearly suggesting a model with just period and treatment effects is an adequate representation of the data. The Wald χ^2-tests of 18.05 on 2 d.f. for periods and 1.29 on 2 d.f. for treatments, show no significant difference between treatments A, B and C, but a significant difference between days 1, 3 and 7.

PROC CATMOD employs the constraint that the effect parameters sum to zero. Hence from the output, the estimates of treatment effects are : treatment 1 (A) = −0.2025, treatment 2 (B) = −0.0535; and so treatment 3 (C) = 0.2560. Period effects are similarly estimated.

Estimates of treatment differences, and period differences, can be derived from differences of the individual effect parameters and, using their variance-covariance matrix, standard errors of these differences can be calculated. Examples are illustrated in Table 3.

Since a logit transformation was used in modelling the response probabilities, these differences can be interpreted as log odds ratios of having a normal respiratory rate (response = 0). Consequently estimates of odds ratios, and approximate 95% confidence intervals for odds ratios, can be derived. For instance the odds of having a normal respiratory rate as opposed to an increased one for day 1 as opposed to day 3 is 0.69. The period effects are therefore suggesting an improvement in respiratory rate over time.

3.3 Comments on Approach

One main disadvantage of the generalised least squares approach is a sample size consideration. The method is based on large sample approximations and requires a non-singular estimated covariance matrix for the response functions, $\tilde{W}$, for each "population"-x-period combination. Thus if there are several populations (e.g. combinations of treatment and other explanatory factors) then the table may be too sparse for this. The SAS manual suggests an effective sample

size of 25-30 for each response function. Koch and Landis in their paper also suggested that a sample size of ≥25 is necessary. The method therefore does not seem inappropriate in the swine fever example. The other disadvantage of the approach is that it cannot handle continuous covariate information, other than by categorising it. This in turn could lead to a sparse data representation.

Instead of using a least squares approach to fit a model to the observed marginal proportions, the modelling could have been carried out using the Liang and Zeger (1986) method of generalised estimating equations. This approach fits, using a 'working' correlation matrix, a model to data which otherwise would be analysed by a generalised linear model, but which have some correlation structure. Recent versions of SAS (version 6.12 for example) now contain a facility within PROC GENMOD for using generalised estimating equations. In the swine fever example the unknown marginal probabilities of respiratory rate response on the three different days in the study in the three groups were modelled, using a 'working' correlation matrix of the repeated measurements. To be consistent with the generalised least squares analysis, a model with only effects for treatment and period was fitted. The SAS output however only gives parameter estimates and standard errors; no significance tests are produced. Results for treatment effect differences obtained from PROC GENMOD are presented in Table 4.

These estimates are slightly different from the ones obtained with the least squares approach, though when interpreted in terms of odds ratios (and approximate 95% confidence intervals for the odds ratios) they are fairly similar. The standard errors obtained with the generalised estimating equations were very slightly smaller.

This method will work in situations where the generalised least squares assumptions might not hold e.g. when the data being analysed are from a sparse table.

4. Subject Effects Model

4.1 Method

The second method of analysis for binary data is the subject effects model. Here the model incorporates an effect for the subject, just as in the split-plot analysis for normally distributed data. To investigate the effects of treatment and time on binary data the following model could be used :

$$f\{P(\text{response} = 0)\} = \mu + t_i + p_j + (tp)_{ij} + \delta_{i\ell}$$

where μ is the overall mean

t_i is the effect associated with treatment i

p_j is the effect associated with period j

$(tp)_{ij}$ is the treatment-period interaction effect

and $\delta_{i\ell}$ is a subject effect for ℓ th unit on the ith treatment.

As with other binary data situations the function being modelled is a logit of the probability of "success".

However, it is not possible to estimate effects for each subject and then base inferences on the estimates of the treatment effects, period effects etc.; the chi-square asymptotic approximations in the hypothesis testing will not hold, since the number of model parameters now increases as the number of observations increases. Two approaches are possible to address this problem. One is to assume that the subject effects are random effects from some particular distribution (such as the normal). Analysis is then possible, using maximum likelihood estimation, as with the package EGRET.

A second approach is to carry out a conditional analysis, using a method suggested by Blackwood (1988) where the problem reduces to a conventional log-linear model. With this method the subject totals (i.e. sum of responses over the times) form a sufficient statistic for the subject effects. Conditioning on these eliminates the subject effects and reduces the problem to a log-linear model on a multidimensional contingency table classified by treatment and response at each separate time point i.e. a $t \times \underbrace{2 \times 2 \times \cdots \times 2}_{\text{q times}}$ contingency table where t is the number of treatments. Log-linear modelling can be carried out in SAS using PROC GENMOD.

The main disadvantage of this analysis is that all between subject information, including treatment effects, is now lost. It is however possible to investigate a treatment by time interaction.

To implement the conditioning, a model with effects for treatment, subject total score (S) and their interaction must be fitted in the model. This is the minimal model [Model (1)]. In the swine fever example the factor S takes values as follows:

S = 1 if responses are (0,0,0) on days 1, 3 and 7
2 if responses are (0,0,1) or (0,1,0) or (1,0,0) on days 1, 3 and 7
3 if responses are (0,1,1) or (1,0,1) or (1,1,0) on days 1, 3 and 7
4 if responses are (1,1,1) on days 1, 3 and 7

Time effects are added using 2-level factors (say P1, P2, ...Pq) associated with response in each period. The time main effect has only (q – 1) df, and hence only (q – 1) of the factors are needed for the model with time effects [Model (2)]. The treatment-by-time interaction is addressed similarly by incorporating into the model terms for the interaction between treatment and these period factors [Model (3)]. Significance testing is then carried out using conventional deviance methods for log-linear modelling. Parameter estimates from the model are interpreted as log odds ratios.

4.2 Analysis

The SAS program for the analysis of the swine fever data using this approach, and fitting the three models just discussed is presented in Appendix 2, along with the output from the model with no treatment-by-time interaction. Table 5 summaries the results of the model fitting process in terms of deviances and deviance differences.

From this table both models (2) and (3) seem to be reasonable fits for the data (deviances were non-significant), and the change in deviance between the two shows that the treatment by period interaction is non-significant (χ^2 test on 4 df). Comparison of models (1) and (2) shows period effects to be significant (χ^2 test on 2 df; $p<0.001$). In the absence of a treatment by time interaction, estimates of period effects can be extracted from model (2). The GENMOD syntax which was used in this model incorporated factors for period 1 and 2 only (days 1 and 3), and consequently the model parameters are estimates relative to period 3 (day 7). Furthermore they are estimates of log odds ratios of a normal respiratory rate (score = 0) as opposed to an increased rate (score = 1) for each of these days relative to day 7. Estimates of log odds ratios for some period comparisons are presented in Table 6 along with (for interest) their EGRET counterparts.

4.3 Comments on Approach

The results of this conditional analysis are very similar to the result obtained using EGRET, where the change in deviance associated with the treatment-by-period interaction was 5.60 on 4 d.f. Parameter estimates were slightly different, though when interpreted in terms of odds ratios, and confidence intervals for odds ratios, the results seem fairly similar.

The main disadvantage with this conditional subject effects analysis is that all between subject information is lost in the conditioning, and consequently it is not possible to investigate treatment main effects. If this is of interest, then a summary statistics analysis could be performed. In the swine fever example the average score for an individual pig was taken as a suitable summary statistic, and a Kruskal-Wallis non-parametric analysis of variance used to compare the groups. This yielded a χ^2 value of 1.33, on 2 d.f., which is clearly not significant.

5. Summary

Two different types of models have been fitted to a repeated measurements binary data example using SAS. A marginal probability model was fitted via empirical generalised least squares in PROC CATMOD. Treatment, time and treatment by time effects could all be investigated. The method is based on some large sample approximations which appeared to hold in this particular instance. With smaller samples a generalised estimating equations approach may be more appropriate, but as yet this approach using PROC GENMOD in SAS only yields parameter estimates.

A subject effects model was also fitted to the same example. This can only be done in SAS via a conditional analysis, conditioning on subject total response, which reduces the problem to a conventional log-linear model for contingency table data. PROC GENMOD can be used for this. The analysis allows investigation of period effects and the treatment-by-period interaction, but because of the nature of the conditioning the treatment main effect cannot be investigated. A summary statistics analysis would be required to explore this.

Finally, it should be noted that the interpretation of the parameters in the subject effects model and the marginal probability model is different. In the subject effects case, the model is describing how an individual's probability is modified over time. The marginal probability model, on the other hand is a population averaged model, and the parameter estimates relate to the probability of an individual chosen at random.

References:

Agresti, A. (1989). A survey of models for repeated ordered categorical response data. *Statistics in Medicine*, **8**, 1209-1224.

Blackwood, L (1988). Latent variable models for the analysis of medical data with repeated measures of binary variables. *Statistics in Medicine*, 7, 975-981.

Kenward, M.G and Jones, B (1992). Alternative approaches to the analysis of binary and categorical repeated measurements. *Journal of Biopharmaceutical Statistics*, **2(2),** 137-170.

Koch, G.G, Landis, J.R *et al* (1977). A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics*, **33**, 133-158.

Liang, K and Zeger S.L (1986). Longitudinal data analysis using generalised linear models. *Biometrika*, **73**, 13-22.

Table 1: Swine fever in pigs. Frequency of respiratory rate responses on the three days for each treatment group

Treatment	Response (day 1, day 3, day 7)								
	(0,0,0)	(0,0,1)	(0,1,0)	(0,1,1)	(1,0,0)	(1,0,1)	(1,1,0)	(1,1,1)	Total
A	4	1	2	2	1	1	6	13	30
B	3	0	1	2	3	4	8	8	29
C	4	0	2	0	6	4	5	9	30

Table 2: Swine fever in pigs. Proportion of animals responding 0 or 1 in each treatment group, by day

	Day 1		Day 3		Day 7	
	0	1	0	1	0	1
A	0.300	0.700	0. 233	0.767	0.433	0.567
B	0.207	0.793	0.345	0.655	0.517	0.483
C	0.200	0.800	0.467	0.533	0.567	0.433

Table 3: Some estimates of treatment effect differences and period differences using PROC CATMOD in SAS for marginal probability modelling

	Treatments			Periods	
	Estimate	s.e.		Estimate	s.e.
$t_1 - t_3$	−0.4585	0.4159	$p_1 - p_2$	−0.3769	0.2805
$t_2 - t_3$	−0.3095	0.3870	$p_1 - p_3$	−1.1189	0.2702

Table 4: Some estimates of treatment effect differences using PROC GENMOD in SAS for fitting generalised estimating equations.

	Estimate	s.e.
$t_1 - t_3$	−0.4059	0.4123
$t_2 - t_3$	−0.2454	0.3757

Table 5: Summary of models fitted in PROC GENMOD in SAS for the subject effects analysis

Model		Deviance	df	Change in: Deviance	df
(1)	Minimal	29.25	12		
(2)	+ period effects	10.29	10	18.96	2
(3)	+ treatment*period	4.90	6	5.39	4

Table 6: Some estimated period differences (log odds ratios) for the subject effects model

	Genmod Analysis		EGRET	
	estimate	s.e.	estimate	s.e.
$p_1 - p_2$	−0.8008	0.4146	−0.746	0.396
$p_1 - p_3$	−1.6627	0.4190	−1.653	0.414

Appendix 1: Marginal effects model for the analysis of respiratory rate in pigs with swine fever.

SAS Program

```
/*-------------------------------------
| Marginal effects model : EGLS       |
-------------------------------------*/

data a;
do treat = 1 to 3;
do p1 = 0,1;
do p2 = 0,1;
do p3 = 0,1;
input resp @@;
output;
end;
end;
end;
end;

cards;
4 1 2 2 1 1 6 13
3 0 1 2 3 4 8 8
4 0 2 0 6 4 5 9
;


proc catmod;
population treat;
weight resp;
response logit;
model p1*p2*p3 = _response_ treat  / covb;
repeated period;
run;
```

Note: Variables p1, p2, p3 refer to the response recorded on days 1, 3 and 7.

Treatments 1, 2 and 3 refer to treatments A, B and C.

The data are being read in contingency table form.

Output

ANALYSIS-OF-VARIANCE TABLE

Source	DF	Chi-Square	Prob
INTERCEPT	1	11.29	0.0008
PERIOD	2	18.05	0.0001
TREAT	2	1.29	0.5234
RESIDUAL	4	6.53	0.1630

ANALYSIS OF WEIGHTED-LEAST-SQUARES ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
INTERCEPT	1	-0.5531	0.1646	11.29	0.0008
PERIOD	2	-0.4986	0.1600	9.70	0.0018
	3	-0.1217	0.1596	0.58	0.4458
TREAT	4	-0.2025	0.2395	0.72	0.3976
	5	-0.0535	0.2227	0.06	0.8100

COVARIANCE MATRIX OF THE PARAMETER ESTIMATES

	1	2	3	4	5
1	0.02710199	0.00345396	-.00041280	0.00314991	-.00344501
2	0.00345396	0.02561443	-.01372379	-.00300155	0.00060463
3	-.00041280	-.01372379	0.02546448	0.00506764	-.00084245
4	0.00314991	-.00300155	0.00506764	0.05733676	-.02648665
5	-.00344501	0.00060463	-.00084245	-.02648665	0.04958796

Appendix 2: Subject effects model for the analysis of the respiratory rate in pigs with swine fever.

SAS Program

```
/*-----------------------------------------------
| Subjects effects model : log-lin approach      |
-----------------------------------------------*/

data a;
do treat = 1 to 3;
do p1 = 0,1;
do p2 = 0,1;
do p3 = 0,1;
input count @@;
output;
end;
end;
end;
end;
cards;
4 1 2 2 1 1 6 13
3 0 1 2 3 4 8 8
4 0 2 0 6 4 5 9
;

data b; set a;
s = p1+p2+p3 +1 ;

proc genmod;
class treat p1 p2 p3 s;
model count = treat s s*treat / d=poisson;
run;

proc genmod;
class treat p1 p2 p3 s;
model count = treat s s*treat p1 p2 / d=poisson;
run;

proc genmod;
class treat p1 p2 p3 s;
model count = treat s s*treat p1 p2 treat*p1 treat*p2
                          / d=poisson;
run;
```

Output

Model : TREAT S TREAT*S P1 P2

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	10	10.2949	1.0295
Scaled Deviance	10	10.2949	1.0295
Pearson Chi-Square	10	8.0902	0.8090
Scaled Pearson X2	10	8.0902	0.8090
Log Likelihood	.	56.1089	.

Analysis Of Parameter Estimates

Parameter			DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT			1	2.1972	0.3333	43.4502	0.0001
TREAT	1		1	0.3677	0.4336	0.7191	0.3964
TREAT	2		1	-0.1178	0.4859	0.0588	0.8085
TREAT	3		0	0.0000	0.0000	.	.
S	1		1	1.7137	0.8986	3.6373	0.0565
S	2		1	0.2503	0.6099	0.1684	0.6815
S	3		1	-0.4775	0.4879	0.9576	0.3278
S	4		0	0.0000	0.0000	.	.
TREAT*S	1	1	1	-0.3677	0.8295	0.1965	0.6575
TREAT*S	1	2	1	-1.0609	0.7504	1.9989	0.1574
TREAT*S	1	3	1	-0.3677	0.6405	0.3296	0.5659
TREAT*S	1	4	0	0.0000	0.0000	.	.
TREAT*S	2	1	1	-0.1699	0.9052	0.0352	0.8511
TREAT*S	2	2	1	-0.5754	0.7817	0.5417	0.4617
TREAT*S	2	3	1	0.5596	0.6470	0.7480	0.3871
TREAT*S	2	4	0	0.0000	0.0000	.	.
TREAT*S	3	1	0	0.0000	0.0000	.	.
TREAT*S	3	2	0	0.0000	0.0000	.	.
TREAT*S	3	3	0	0.0000	0.0000	.	.
TREAT*S	3	4	0	0.0000	0.0000	.	.
P1	0		1	-1.6627	0.4190	15.7471	0.0001
P1	1		0	0.0000	0.0000	.	.
P2	0		1	-0.8619	0.3654	5.5639	0.0183
P2	1		0	0.0000	0.0000	.	.
SCALE			0	1.0000	0.0000	.	.