

Kansas State University Libraries

**New Prairie Press**

---

Conference on Applied Statistics in Agriculture

1996 - 8th Annual Conference Proceedings

---

## **ANALYSIS OF UNBALANCED MIXED MODEL DATA: Traditional ANOVA Versus Contemporary Methods**

Ramon C. Littell

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**

Littell, Ramon C. (1996). "ANALYSIS OF UNBALANCED MIXED MODEL DATA: Traditional ANOVA Versus Contemporary Methods," *Conference on Applied Statistics in Agriculture*. <https://doi.org/10.4148/2475-7772.1318>

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](mailto:cads@k-state.edu).

## **ANALYSIS OF UNBALANCED MIXED MODEL DATA Traditional ANOVA Versus Contemporary Methods**

Ramon C. Littell  
Department of Statistics  
University of Florida

### **ABSTRACT**

Analysis of unbalanced data and analysis of mixed model data are important topics of statistical discussion. Analysis of unbalanced data with fixed effects gives rise to the different types of sums of squares in analysis of variance. Mixed model data raises issues of determining appropriate error terms for test statistics and standard errors of estimates. The situation is even more difficult when the two topics occur together, resulting in unbalanced mixed model data. These problems have plagued users of PROC GLM in the SAS System. Now, with PROC MIXED available, some of the problems are resolved while others remain. This paper gives an overview of two areas of difficulty in analysis of variance using PROC GLM, and describes which problems carry over to PROC MIXED, and which are essentially solved with PROC MIXED.

### **1. INTRODUCTION**

Most unbalanced mixed model applications were analyzed with PROC GLM prior to the introduction of PROC MIXED. The basic testing and estimation procedures in PROC GLM are based on ordinary least squares for a fixed effects model. Mixed model applications with GLM are adaptations of fixed effect methods. Many of the adaptations work satisfactorily well with balanced mixed model data. But serious problems exist with these methods when applied to unbalanced mixed model data, due to the lack of theoretical basis. Most applications use ad hoc and subjective decisions to choose analysis methods. Littell, Wilcox, and Van Horn (1995) discussed related problems pertaining to estimation.

Analysis of variance of unbalanced mixed model data presents problems of choosing quadratic forms for testing hypotheses about fixed effects. This is related to the four types of sums of squares in PROC GLM. In simple terms, the problems are to decide whether sums of squares for fixed effects to be tested should be "adjusted" for other effects in the model. However, the problem of whether to adjust a fixed effect for other fixed effects is quite different from the problem of whether to adjust a fixed effect for a random effect. Since GLM essentially treats all effects as fixed, this distinction is difficult to resolve. When using PROC MIXED, fixed effects and random effects are dealt with separately, which solves part to the problem.

In this paper a dairy cattle feeding trial described by Tomlinson, Van Horn, Wilcox and Harris (1994) is used to illustrate some of the issues encountered when using PROC GLM and PROC MIXED for unbalanced mixed model data analysis.

## 2. MIXED MODEL FOR UNBALANCED DATA

The example data set to be used was from a change-over design with twelve dietary treatments (TRT = 1-12) in four lactation periods (PER = 1-4). Thirty-three cows were used, of which seventeen were in their first lactation (GRP = 1) and sixteen were in a second or later lactation (GRP = 2). Four treatments were applied to each cow in a sequence over the four periods. Also, cows were fed one of the treatments in a period 0, prior to data collection. This yielded a variable (PRETRT = 1-12) which, for any cow in any period, is the treatment she received in the previous period. The design was such that all six two-way tables for the four variables (GRP, TRT, PER and PRETRT) have no empty cells, except the 12×12 table for TRT\*PRETRT, which is connected (Searle, 1987, pp 139-144).

A mathematical model for the usual type of changeover design (in which no treatment is given in period zero) with a between-animal grouping variable and within-animal treatment variable is

$$Y_{ijklm} = \mu + \alpha_i + d_{ij} + \beta_k + \tau_l + \gamma_{m,k-1} + e_{ijklm} \quad (2.1)$$

where  $\mu$  = reference mean,  $\alpha_i$  = effect of  $i^{\text{th}}$  group,  $d_{ij}$  = effect of  $j^{\text{th}}$  animal in  $i^{\text{th}}$  group,  $\beta_k$  = effect of  $k^{\text{th}}$  period,  $\tau_l$  = direct effect of  $l^{\text{th}}$  treatment,  $\gamma_{m,k-1}$  = residual effect of treatment  $m$  having been applied in period  $k-1$ , and  $e_{ijklm}$  = residual error. All effects are considered fixed except  $d_{ij}$  and  $e_{ijklm}$ , which are taken to be normally and independently distributed with variances  $\sigma_d^2$  and  $\sigma_e^2$ , respectively. It is assumed that  $\gamma_{m,0} = 0$ , due to there being no treatment in period 0. In terms of model (2.1), the direct effect mean for treatment  $l$  in group  $i$ , as implicitly defined by Cochran and Cox (1957, p 137), is given by

$$\mu_{il} = \mu + \alpha_i + \beta + \tau_l + \gamma, \quad (2.2)$$

where  $\beta$  stands for the average period effect and  $\gamma$  stands for the average residual effect. In some applications it may be more meaningful to define  $\beta$  to be a random effect, such as when periods serve as replication over time. But in the present context, periods represent specific times in a lactation, and thus are meaningfully defined as fixed.

The model with equation (2.1) is an example of the general linear mixed model (GLMM). The designs discussed by Cochran and Cox (1957, section 4.6a) are balanced, and analyses are amenable to hand computation. The example analysis in section 4.62a of Cochran and Cox (1957) was reproduced using PROC GLM of the SAS System by Littell, Freund and Spector (1991, pg. 205). The CLASS statement in PROC GLM sets up appropriate definitions of indicator variables in model (2.1) except for the residual effects  $\gamma_{m,k-1}$ , for which indicator variables can be constructed explicitly. The methods given by Littell et al (1991) do not require balance; only estimability of parameters of interest.

A model for a changeover design in which treatments were administered in period 0 also has equation (2.1), but  $\gamma_{m,0}$  is not assumed to be zero, and thus can be represented simply as  $\gamma_m$ .

This also is a GLMM and can be fitted with least squares or maximum likelihood. However, there are numerous versions of least squares and maximum likelihood represented in various computer programs, and all give somewhat different results. The central issue is the manner in which to deal with random effects. In the remainder of this section some of the possibilities are described.

To illustrate the problems concerned with adjusting fixed effects for other model terms, some two-way interactions will be added to model equation (2.2). The new model equation is

$$Y_{ijklm} = \mu + \alpha_i + d_{ij} + \beta_k + \tau_l + (\beta\tau)_{kl} + \gamma_m + (\alpha\gamma)_{im} + e_{ijklm} \quad (2.3)$$

Statements for fitting the model with PROC GLM are

```
proc glm; class grp cow per trt pretrt;
  model my=grp cow(grp) per trt per*trt pretrt grp*pretrt / ss1 ss2 ss3;
```

Analysis of variance results are shown in Table 1. Values for the sums of squares for a given effect vary from Type I to Type II to Type III because each was computationally "adjusted" for different other effects in the model. The term "adjusted" is used here without precise definition, but is intended in the sense used by Searle (1987, Chapter 4) and Damon and Harvey (1987, p 114). Generally, the Type I SS for an effect is adjusted for terms which precede it in the model. Type II SS for an effect is adjusted for all other terms which do not symbolically "contain" it. Thus, Type II SS for PER is adjusted for TRT, but not for PER\*TRT. Type III SS for an effect is adjusted for all other terms in the model. One statement is categorically true: Random effects should be adjusted for all fixed effects. Thus, the (common value of) Type II or Type III SS for COW(GRP) should be used.

All F tests shown in Table 1 are those computed by default by PROC GLM, and have  $MS(ERROR) = 132870.4$  in their denominators. None of the three tests for GRP use a valid error term, which would be a combination of  $MS(COW(GRP))$  and  $MS(ERROR)$ . But  $MS(GRP)$  is so large that it would be highly significant even with a proper error term.

The "adjustment" question for GRP pertains to which type of SS should be used for the numerator of the F test. Again, since GRP effect is so large, all three types of tests would be highly significant. But Type I SS for GRP (46584637) is three times larger than Type III SS (13762274), and this could be an important difference in other applications. Type II SS is between Type I SS and Type III SS in value. A problem that has not been resolved in situations such as this is whether GRP should be adjusted for COW(GRP); that is, whether a fixed effect should be adjusted for a random effect that "contains" the fixed effect. Few authors have stated

Table 1. Analysis of Variance Tests from PROC GLM

<u>Source</u>	<u>DF</u>	<u>Type I SS</u>	<u>Mean Square</u>	<u>F Value</u>	<u>Pr &gt; F</u>
GRP	1	46584637.4	46584637.4	350.60	0.0001
COW(GRP)	31	29713608.0	958503.5	7.21	0.0001
PER	3	8503833.5	2834611.2	21.33	0.0001
TRT	11	3956742.5	359703.9	2.71	0.0156
PER*TRT	33	5189561.2	157259.4	1.18	0.3244
PRETRT	11	1574581.1	143143.7	1.08	0.4117
GRP*PRETRT	11	1256056.2	114186.9	0.86	0.5868

  

<u>Source</u>	<u>DF</u>	<u>Type II SS</u>	<u>Mean Square</u>	<u>F Value</u>	<u>Pr &gt; F</u>
GRP	1	26673059.3	26673059.3	200.75	0.0001
COW(GRP)	31	12882561.7	415566.5	3.13	0.0014
PER	3	7023738.8	2341246.3	17.62	0.0001
TRT	11	2967809.7	269800.9	2.03	0.0624
PER*TRT	33	5585607.3	169260.8	1.27	0.2557
PRETRT	11	1574581.1	143143.7	1.08	0.4117
GRP*PRETRT	11	1256056.2	114186.9	0.86	0.5868

  

<u>Source</u>	<u>DF</u>	<u>Type III SS</u>	<u>Mean Square</u>	<u>F Value</u>	<u>Pr &gt; F</u>
GRP	1	13762274.5	13762274.5	103.58	0.0001
COW(GRP)	31	12882561.7	415566.5	3.13	0.0014
PER	3	1978938.5	659646.2	4.96	0.0067
TRT	11	2836554.6	257868.6	1.94	0.0752
PER*TRT	33	5585607.3	169260.8	1.27	0.2557
PRETRT	11	1560343.4	141849.4	1.07	0.4187
GRP*PRETRT	11	1256056.2	114186.9	0.86	0.5868
ERROR	29	3853240.3	132870.4		

their preference. However, Milliken and Johnson (1984, p 395) indicate a Type III test, while Damon and Harvey (1987, p 141) indicate a Type II test. Type III SS for GRP is “adjusted” for COW(GRP), but Types I and II SS for GRP are not. The  $\sigma_d^2$  variance component cannot be removed from the GRP expected MS by adjustment for COW(GRP) because GRP is “contained” in COW(GRP). Thus, the benefits of adjusting GRP for COW(GRP) are not clear.

The tests for PER present more dramatic results. All three F tests on PER in Table 2 use the correct error term (MS(ERROR)) and are valid in that sense, because all three types of SS's for PER are adjusted for COW(GRP) and thereby the COW(GRP) variance component is removed from the expected mean squares for PER. But Type I MS for PER is four times larger than

Table 2. Mixed Model Tests from PROC MIXED  
Tests of Fixed Effects

<u>Source</u>	<u>NDF</u>	<u>DDF</u>	<u>Type I F</u>	<u>Pr &gt; F</u>
GRP	1	31	63.81	0.0001
PER	3	20	20.40	0.0001
TRT	11	20	3.07	0.0077
PER*TRT	33	20	1.34	0.2152
PRETRT	11	20	0.59	0.8206
GRP*PRETRT	11	20	1.06	0.4237

  

<u>Source</u>	<u>NDF</u>	<u>DDF</u>	<u>Type III F</u>	<u>Pr &gt; F</u>
GRP	1	31	55.72	0.0001
PER	3	20	5.40	0.0045
TRT	11	20	2.31	0.0348
PER*TRT	33	20	1.38	0.4237
PRETRT	11	20	0.64	0.7840
GRP*PRETRT	11	20	1.06	0.4237

Type III MS for PER, and the levels of significance range from .0001 for Type I and Type II, to .0067 for Type III. These differences result from PER being adjusted for different sets of fixed effects. This type of problem has been extensively discussed (see Searle, 1987, Chapter 4) but not resolved. At issue is the hypothesis tested using a particular SS Type. It is usually possible to identify the ideal hypotheses to be tested, but the test can suffer from low power with highly unbalanced data.

Statements for fitting the model with PROC MIXED are

```
proc mixed; class grp cow per trt pretrt;
  model my = grp per trt per*trt pretrt grp*pretrt / htype=1, htype=3;
  random cow(grp);
run;
```

Results are shown in Table 2. The output from PROC MIXED is based on generalized least squares (GLS) methodology for the GLMM. Specifically, the model in matrix notation is

$$Y = X\beta + Zu + e \quad (2.4)$$

where  $X\beta$  models all fixed effects,  $Zu$  models random effects, and  $e$  is residual variation. For

the present example, between-cow variation ( $d_{ij}$ ) is modeled in  $Z\mathbf{u}$ , and within-cow variation ( $e_{ijklm}$ ) is contained in  $\mathbf{e}$ . The covariance matrix of  $\mathbf{Y}$  is

$$\mathbf{V} = \mathbf{V}(\mathbf{Y}) = \mathbf{ZGZ}' + \mathbf{R}, \quad (2.5)$$

where  $\mathbf{G} = \mathbf{V}(\mathbf{u})$  and  $\mathbf{R} = \mathbf{V}(\mathbf{e})$ . For the example,  $\mathbf{G} = \sigma_d^2 \mathbf{I}$  and  $\mathbf{R} = \sigma_e^2 \mathbf{I}$ . The GLS estimate of  $\beta$  is

$$\tilde{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y} \quad (2.6)$$

and

$$\mathbf{V}(\tilde{\beta}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}. \quad (2.7)$$

Tests of hypothesis about linear combinations of  $\beta$ ,  $\mathbf{K}'\beta$ , are based on the fact that

$$\tilde{\beta}'\mathbf{K}(\mathbf{K}'\mathbf{V}^{-1}\mathbf{K})^{-1}\mathbf{K}'\tilde{\beta} \sim \chi^2(r(\mathbf{K})). \quad (2.8)$$

However, the  $\mathbf{V}$  matrix is not known, so it must be replaced by  $\hat{\mathbf{V}}$ , which contains estimates of  $\sigma_d^2$  and  $\sigma_e^2$ . Then  $\hat{\mathbf{V}}$  is used in place of  $\mathbf{V}$  in (2.6), (2.7), and (2.8). This results in (2.8) having an approximate F distribution. But the concept of the methodology is on a firm mixed model theoretical basis. Therefore, the tests for GRP are valid in the sense that random variation is properly accommodated in the F statistics. In this sense, the problem in GLM of whether a fixed effect should be adjusted for a random effect is resolved in MIXED. But the question remains of whether GRP should be adjusted for other fixed effects. This problem persists also in the Type I and Type III tests for PER, with results similar to those from GLM. Type II tests are not presently available in MIXED.

### 3. SUMMARY AND CONCLUSION

Data from an unbalanced change-over dairy feeding trial was used to illustrate problems with analysis of unbalanced mixed model data using PROC GLM and PROC MIXED. Most of the problems with using PROC GLM are due to the fact that GLM was written on the basis of fixed effect methodology. Therefore, adjusting a fixed effect for a random effect is not distinguished

from adjusting a fixed effect for another fixed effect. PROC MIXED is based on mixed model methodology, with fixed and random effects dealt with as separate issues. Therefore, adjusting a fixed effect for a random effect is not a problem in MIXED. But concerns about adjusting a fixed effect for another fixed effect are as much a problem in MIXED as in GLM.

## REFERENCES

- Cochran, W.G., and G.M. Cox. 1962. *Experimental Designs*. (2nd edition). John Wiley and Sons, New York.
- Damon, R.A. Jr., and W.R. Harvey. 1987. *Experimental Design, ANOVA, and Regression*. Harper and Row, New York.
- Hartley, H.O. 1967. Expectations, variances and covariances of ANOVA mean squares by "synthesis." *Biometrics* 23:105; Corrigenda, 853.
- Littell, R.C., R.J. Freund, and P.C. Spector, P.C. 1991. *SAS System for Linear Models*. SAS Institute Inc., Cary, NC. p. 205-209.
- Milliken, G.A., and D.E. Johnson. 1984. *Analysis of Messy Data*. Lifetime Learning. Belmont, CA.
- SAS Institute Inc. 1982. *SAS Users Guide*. Statistics, Cary, North Carolina. p. 205-209.
- Searle, S.R. 1987. *Linear Models for Unbalanced Data*. John Wiley & Sons, New York.
- Tomlinson, A.P., H.H. Van Horn, C.J. Wilcox, and B. Harris, Jr. 1994. Effects of undegradable protein and supplemental fat on milk yield and composition and physiological responses of cows. *J. Dairy Sci.* 77:145.