Conference on Applied Statistics in Agriculture 2003 - 15th Annual Conference Proceedings

DATA STRUCTURE WITH RESPECT TO THE MAIN EFFECTS MODEL: A DISCUSSION MOTIVATED BY A META-ANALYSIS DATA SET

Dawn M. VanLeeuwen

David S. Birkes

Cynda Clary

Chadelle Robinson

See next page for additional authors

Recommended Citation

VanLeeuwen, Dawn M.; Birkes, David S.; Clary, Cynda; and Robinson, Chadelle (2003). "DATA STRUCTURE WITH RESPECT TO THE MAIN EFFECTS MODEL: A DISCUSSION MOTIVATED BY A META-ANALYSIS DATA SET," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1171

Author Information

Dawn M. VanLeeuwen, David S. Birkes, Cynda Clary, and Chadelle Robinson

DATA STRUCTURE WITH RESPECT TO THE MAIN EFFECTS MODEL: A DISCUSSION MOTIVATED BY A META-ANALYSIS DATA SET

Dawn M. VanLeeuwen[1], David S. Birkes[2], Cynda Clary[1], and Chadelle Robinson[3]

[1]New Mexico State University, Las Cruces
[2]Oregon State University, Corvallis
[3]New Mexico Department of Agriculture, Las Cruces

Abstract

A discussion on data structure relative to the main effects model is motivated by a severely unbalanced meta-analysis data set. This data set is used to highlight the difficulty of assessing data structure when multiple factor data sets are severely unbalanced. Both theoretical results and numerical examples are used to establish that simple approaches to examining data structure using two-way tables provide easily assimilated information about the effect of data unbalance on main effect contrast variances. In addition, notions of balance, proportionality, unbalance, and missing cells with respect to the main effects model are defined in terms of the two-way tables and are related to main effect contrast estimate variances as assessed using the D-optimality criterion.

1. Introduction

While much has already been written on unbalanced and missing cells data, analyzing such data continues to present difficulties to statistical practitioners and researchers in other disciplines. Unbalanced data are the norm in some disciplines where most research is observational. Furthermore, quantitative, model-based approaches to research are being applied to more settings than ever before; many of these settings are likely to produce unbalanced data. For example, the quantitative literature review, or meta-analysis, is used increasingly in subject areas such as economics, marketing, medicine, and education. Often these data are analyzed using a main effects ANOVA-type model on an extremely sparse unbalanced data structure (Farley and Lehmann, 1986). Concerning the unbalance likely to exist in these data sets, practitioners often are advised to use the condition index or some other diagnostic borrowed from regression to assess whether a problematic degree of collinearity exists. While such diagnostics may provide some insight, they and the guidelines for using them have been developed for regression models with quantitative variables. Other approaches to examining data structure tailored to the qualitative factors used in many meta-analyses may be useful to the data analyst.

Examining data structure and understanding the relationships among independent variables is an important phase in data analysis. For multiple regression with quantitative regressors, Ramsey and Schafer (1997, pp.242-246) recommend obtaining scatter plots for each pair of explanatory variables. It also is common to calculate correlation coefficients among the explanatory variables (Myers, 1990, pp.124-125). An analogous set of descriptive statistics for

qualitative factors under the main effects model would be the set of two-way tables corresponding to every factor pair. However, this simple descriptive approach to examining data structure often is overlooked even in problematic data sets.

In this paper, we borrow from the optimal experimental design literature to gain a perspective on examining observational data set structure under the main effects model. In particular, we explore the degree to which two-way tables for pairs of independent variables inform the data analyst about the data structure's impact on effect estimates. We begin with a motivating example and then present an optimal design result. Based on the optimality result, basic linear models theory, and an extensive consideration of designs for the $3 \times 3 \times 2$ main effects model with $n = 36$, the degree to which the two-way tables inform the data analyst about the data structure's impact on the variance of main effect contrast estimates is examined. We recommend that, in the early stages of data analysis, practitioners use two-way tables as a tool in the process of understanding data structure.

2. Motivating Example

The motivating example is a meta-analysis data set from a masters' project in agricultural economics. The researchers (the student and their advisor) had compiled data from an extensive literature review of generic advertising's effectiveness and intended to use the main effects ANOVA approach to meta-analysis outlined by Farley and Lehmann (1986) to evaluate the effects of various study attributes on the magnitude of advertising elasticity estimates. They sought input from a statistician when estimates of the "shared" variance were negative. The shared variance is computed by first summing the partial sums of squares for each effect in the model (i.e., the type II sums of squares) then subtracting that sum from the full model sum of squares. A common interpretation of this difference is that it is, indeed, shared variance or the amount of variability in the dependent variable that can be accounted for by more than one of the independent variables. Consequently, a negative value is counterintuitive. However, a negative value for this shared variance is mathematically possible.

Extremely simple examination of the data suggested that the problem was likely due to severe collinearity. The meta-analysis data set consisted of 156 observations. There were 14 explanatory variables: eight with two levels, and six with three levels. Only 38 of the 186,624 possible data cells were observed to contain any data. Even though all main effects were estimable (20 model degrees of freedom), the data structure was very sparse. Furthermore, the 156 observations were distributed into the 38 cells in a very uneven fashion. There were 55 observations in a single cell, while 18 cells contained only a single observation. Numbers of observations in the remaining 19 cells ranged from two to 14.

The researchers had not conducted the cursory examination of the data structure summarized in the above paragraph. In fact, while Farley and Lehmann (1986) (henceforth referred to as FL) indicated that meta-analysis data sets were likely to be severely unbalanced and had emphasized the need to examine the data structure, they provided few details on tools that might be used beyond one-way frequencies and the condition number. Consequently, prior to attempting the main effects analysis, the researchers had followed all steps outlined by FL and

believed that all data structure problems had been resolved by collapsing factor categories (levels) on the basis of one-way frequencies. Additionally, FL suggested using the condition number, a tool typically used in regression, to assess whether a problematic degree of collinearity exists in the data. According to FL, problematic degrees of collinearity would be indicated by a condition number of around 1000 or more. The researchers obtained a condition number of 42 using SAS® Proc Reg. However, the condition number may be based on a singular value decomposition of either the design matrix or the design matrix premultiplied by its transpose. FL assumed use of the design matrix premultiplied by its transpose while Proc Reg has implemented the collinearity diagnostic proposed by Belsley, Kuh and Welsch (1980) (henceforth referred to as BKW) based on a singular value decomposition of the design matrix. Consequently, the value of 42 would need to be squared to compare it to the cutoff of 1000 suggested by FL. While the different scales for computing and reporting the condition number did lead to a misinterpretation, a subset of the variables obtained by dropping just one explanatory variable resulted in a condition number below 30 (below 900 on the FL scale) but did not resolve the problem with the negative shared variance estimate.

BKW's diagnostic uses the entire set of condition indices obtained by dividing the maximum singular value (i.e., eigenvalue) from the decomposition of the design matrix by each other singular value. BKW indicate that when accompanied by high variance proportions for two or more variables, condition indices greater than 30 may suggest a strong linear dependency. However, BKW's empirical explorations suggested that, for quantitative variables, weak dependencies may be identified by condition indices of around 10 and even as low as 5 (p.153). Indices from 15 to 30 are considered borderline and may reflect borderline tight dependencies.

As already noted, when applied to the meta-analysis data set, the BKW diagnostic identified a dominant dependency associated with a condition index of about 42. However, several other condition indices were suggestive of additional dependencies. Four more condition indices fell between 15 and 23, and an additional three indices fell between 15 and 10. Furthermore, the diagnostic indicated that all but two of 21 regression coefficients had more than 50% of their variance associated with condition indices greater than 10. Of the two coefficients that were not implicated in possible near linear dependencies, one corresponded to a three-level variable so that all but one variable were implicated. When a single or relatively few dependencies exist, the diagnostic can provide a great deal of guidance in isolating the dependencies. However, when several exist, the diagnostic may provide little assistance beyond suggesting the number of dependencies that may exist.

In an attempt to gain insight into the dependencies suggested by the BKW diagnostic, two-way tables for every pair of explanatory variables were obtained. Of these 91 two-way tables, 60 had one or more cells with counts of zero. Nearly all were highly unbalanced (i.e., not only were zero cell counts present in most tables, but among cells with positive counts, the ratio of the largest to smallest count typically exceeded 10). Had the researchers begun with this simple, descriptive, data structure assessment, they would have realized that their data structure barely supported models incorporating only two explanatory variables. Additionally, they would have come to this realization before becoming invested in an analysis that the data structure did

not support using and would have begun considering alternative approaches to analyzing and reporting the data earlier in the process.

Examining the two-way tables led the researchers to realize that one level of one variable was completely confounded with a single research group that had conducted multiple studies. It is likely that this insight into the data structure would have been overlooked if, as suggested by FL, a canned stepwise regression procedure had been used to simply eliminate some of the factors from the model. As already noted, when the analysis was rerun with this variable removed, the condition indices all fell below 30 but the negative shared variance did not go away. As the two-way tables suggest, a problematic degree of collinearity was still at the root of the negative shared variance even though the condition index no longer exceeded 30.

While these data represent a fairly extreme case, the seeming correspondence between insight provided by simple two-way tables versus a sophisticated and difficult to interpret regression diagnostic raises interesting questions about the amount of information contained in the two-way tables. For this rather extreme case, there appeared to be a correspondence between the information provided by a fairly quick perusal of the two-way tables and by the condition indices and variance proportions. That is, those variables involved in more tables having missing cells tended to be variables that had high variance proportions associated with the condition indices over 10. Additionally, for those variables with three levels, having two-way tables with more missing cells seemed also to be associated with having high variance proportions associated with higher condition indices.

3. Examining Data Structure Using Two-Way Tables

In the example, simple examination of the two-way tables was a useful way to look at the data and to help the researchers appreciate concerns with their data structure. Because the example suggests that the two-way tables may be useful for helping both statisticians and nonstatisticians appreciate problems with data structure, the authors resolved to explore the usefulness of the two-way tables under the main effects model. The results of this exploration are presented in the remainder of this paper. The optimal design literature, while aimed at designed experiments, provided the tools that were used for comparing the relative performance of differing designs. We refer to universal optimality but use the D-optimality criterion as the basis for comparing designs. We establish that, under the main effects model, optimal designs can be defined in terms of the two-way tables. A numerical experiment then uses the D-optimality criterion to evaluate and compare designs generated for a $3 \times 3 \times 2$ main effects model with n=36.

3.1. Two-Way Tables and Optimality for the Main Effects Model

This section summarizes notation, provides a brief overview of optimal design theory, presents an optimal design result, and applies the result to the main effects model.

Notation. We use s to denote the number of effects in the model, $t_1, ...t_s$ to denote the numbers of levels for each of the s effects, and n to denote the total number of observations. Unless implied by context, we use d to index designs belonging to $\mathbb{D}$, the class of all allowable designs. N or N^d represents the $t_1 \times t_2 \times ... \times t_s$ incidence matrix of $\mathrm{n}_{i_1 i_2 ... i_s}$'s. N^{fg} (N_d^{fg}) is the marginal incidence matrix for the effects indexed by f and g and contains the $t_f \times t_g$ values n_{ij}^{fg}. The terms marginal incidence matrices and two-way tables will be used interchangeably to refer to the set of N^{fg} matrices. The fixed main effects model for $Y_{i_1 i_2 ... i_s k}$ can be written

$$Y_{i_1 i_2 ... i_s k} = \mu + \alpha_{1 i_1} + \alpha_{2 i_2} + \ldots \alpha_{s i_s} + e_{i_1 i_2 ... i_s k} \qquad (3.1.1)$$

where the $e_{i_1 i_2 ... i_s k}$ are independent random error terms from the normal distribution with mean 0 and variance σ^2. In matrix notation, the model can be written

$$Y = J\mu + A_1\alpha_1 + A_2\alpha_2 + \ldots A_s\alpha_s + e \qquad (3.1.2)$$

where α_h is the vector of fixed effects corresponding to the hth factor and $e \sim N(0, \sigma^2 I)$. The A_h are sometimes called classification matrices (also, incidence matrices or design matrices) and are composed of 0's and 1's with at most a single 1 in each row and n_i^h 1's corresponding to the position of α_{hi} in the ith column (see Tjur, 1984). Here I represents the identity matrix with dimensions implied by context; similarly J represents a matrix of ones.

The range (or column space) of a matrix A is denoted by $\mathcal{R}(A)$, and the unique orthogonal projection matrix whose range is $\mathcal{R}(A)$ is denoted by $\mathrm{P}_A = A(A'A)^- A'$ where A' denotes the transpose of the matrix A and $(A'A)^-$ denotes the generalized inverse of the matrix $(A'A)$. Factors f and g are defined to be orthogonal if their respective projection matrices commute, that is, if $\mathrm{P}_f \mathrm{P}_g = \mathrm{P}_g \mathrm{P}_f$, where P_f represents the orthogonal projection matrix on A_f. Matrices A and B are ordered $A \geq B$ if $A - B$ is nonnegative definite.

Overview of Optimal Design. For effect matrix A_h, define $X_h = (J, A_1, \ldots, A_{h-1}, A_{h+1}, \ldots, A_s)$ and also define $M_h = I - P_{X_h}$. Then the information matrix, C_h (or C_{dh}), for the hth factor treatment effects is

$$C_h = A_h' M_h A_h \qquad (3.1.3)$$

(Morgan and Bailey, 2000).

Typically, an optimal design is defined to be any design, $d \in \mathbb{D}$, which minimizes the value of some optimality criterion, ϕ, defined as a function of C_h. In general, functions chosen as optimality criteria relate in some way to a notion of minimizing the variance of contrast estimates for the hth treatment effects (Shah and Sinha, 1989). The D-optimality criterion is defined as $\phi_{\mathrm{D}}(C_h) = \prod \frac{1}{\lambda_k}$ where, $\lambda_1, \ldots, \lambda_{r_h}$ are the nonzero eigenvalues of C_h.

Kiefer's (1975) proposition 1 provides conditions that can be used to establish a design's universal optimality. Designs that are universally optimal are D-optimal. His proposition states that if there exists $C_{d^*h} \in \mathbb{C}_h$ such that C_{d^*h} is completely symmetric (i.e., $C_{d^*h}$is of the form $aI + bJ$) and trace$(C_{d^*h}) = \max_{d \in \mathbb{D}}$ trace C_{dh}, then d^* is universally optimal for estimating the h-effects.

Two-Way Tables and Optimality. The main effects model (3.1.1) is a special case of the following fixed effects model:

$$Y = J\mu + A\alpha + B\beta + e. \qquad (3.1.4)$$

In this setting, α contains the t parameters of interest, while $(\mu, \beta')'$ is a vector of nuisance parameters. The following result is related to results in Kunert (1993) and Shah and Sinha (1993) but allows for a greater number of factors.

Optimal Design Result: For the model (3.1.4), suppose that there exists a design d^* with n observations that satisfies the following two conditions:

1) $A'A = mI$,
2) $\mathrm{P}_A\mathrm{P}_B = \mathrm{P}_J$.

Then among all designs for this model with n observations, d^* is universally optimal with respect to estimation of the α-effects.

proof: For d^*, $C^* = A^{*\prime}(I - \mathrm{P}_{J,B})A^* = A^{*\prime}A^* - A^{*\prime}\mathrm{P}_{J,B}A^* =$
$A^{*\prime}A^* - A^{*\prime}\mathrm{P}_J A^* = mI - \frac{m^2}{\mathrm{n}}J.$

Hence, C^* is completely symmetric and satisfies Kiefer's first condition. It remains only to establish that trace$(C^*) \geq$ trace(C) for any other design having n observations.
Let n_i^1 represent the number of observations on the ith level of α. From the above, it follows that trace$(C^*) = \mathrm{n} - \frac{tm^2}{\mathrm{n}}$.
For any design, $C = A'M_A A \leq A'(I - P_J)A$. So that trace$(C) \leq A'(I - P_J)A =$
trace$(A'A - A'P_J A) = \mathrm{n} - \frac{(\mathrm{n}_1^1)^2 + \ldots + (\mathrm{n}_t^1)^2}{\mathrm{n}} \leq \mathrm{n} - \frac{tm^2}{\mathrm{n}}$ since $(\mathrm{n}_1^1)^2 + \ldots + (\mathrm{n}_t^1)^2$ is minimized by taking $\mathrm{n}_i^1 = \frac{\mathrm{n}}{t} = m$. □

The optimal design result implies that a main effects model design will be optimal for estimating the α_1-contrasts, if the following two conditions hold:

1) $\mathrm{n}_i^1 = \mathrm{m}$ for all $i = 1, \ldots, t_1$ and
2) $\mathrm{n}_{ij}^{1h} = \frac{m\,\mathrm{n}_j^h}{\mathrm{n}} = \frac{\mathrm{n}_j^h}{t_1}$ for $h = 2, \ldots, s$.

While the optimal design result only applies when $\mathrm{n} = \mathrm{m}t_1$, where n is the total sample size, it provides insight into the relationship between the two-way tables and optimality criterion values.

Many of the existing results in optimal design are predicated on the assumption that the blocking factors, which are considered to correspond to "nuisance" parameters, form a statistically orthogonal arrangement (Morgan and Bailey, 2000; Shah and Sinha, 1989). However, the optimal design result implies that the structure among the blocking or nuisance factors is immaterial. As long as each level of the effect of interest is replicated the same number of times and each level of the effect of interest appears the same number of times within each level of each nuisance factor, the design will be optimal for the effect of interest. This implies that a design may be poor for estimating some contrast sets while it is quite good for estimating other contrast sets.

The second optimal design result condition is a proportionality condition and implies statistical orthogonality between the first effect and each of the other effects (Tjur, 1984). When this condition is satisfied, the usual two-way table χ^2-test for independence would yield a test statistic value of zero.

Example 1. Consider a $3 \times 3 \times 2$ three factor main effects model. For convenience, use a, b, and c to denote the three effect vectors so that the model becomes:

$$Y = J\mu + Aa + Bb + Cc + e$$

Throughout this paper, we use this model with n $= 36$. Consider the design:

		n_{ij1}			n_{ij2}		
$j =$		1	2	3	1	2	3
	1	2	2	2	2	2	2
i	2	1	4	2	4	1	2
	3	3	0	2	0	3	2

Then:

$n_{ij.}$			$n_{i.k}$		$n_{.jk}$	
4	4	4	6	6	6	6
5	5	4	7	7	6	6
3	3	4	5	5	6	6

This design is optimal for estimating the c-contrast, $c_1 - c_2$; both N^{ac} and N^{bc} satisfy the proportionality condition and $n_k^c = 18$ for $k = 1, 2$. Because N^{bc} is a constant multiple of the J matrix, it is obvious that it satisfies the proportionality condition. N^{ac} also satisfies the condition because, within each row (i.e., within each level of the a-effect), each level of the c-effect is replicated the same number of times. Note that estimates for a and b effects are not optimal, since N^{ab} does not satisfy the proportionality condition (and different levels of a have different numbers of replication). □

The optimal design result implies that if all of the two-way tables are multiples of a matrix of 1's (i.e., $N^{fg} \propto J$ for all $f \neq g$), then the design will be universally optimal for all s factors. Consequently, while a completely balanced design with the same number of replications in every cell will be universally optimal, designs containing missing cells also may be optimal.

Example 2. Consider the model introduced in Example 1. Then the design with $n_{ijk} = 2$ will be completely balanced and therefore universally optimal with $N^{ab} = 4J$, and $N^{ac} = N^{bc} = 6J$. The following two designs have the same marginal incidence matrices as the completely balanced design.

Optimal Design 1 (OD1):

		n_{ij1}			n_{ij2}		
$j =$		1	2	3	1	2	3
	1	2	2	2	2	2	2
i	2	1	3	2	3	1	2
	3	3	1	2	1	3	2

Optimal Design 2 (OD2) (from VanLeeuwen, Birkes, and Seely, 1999):

		n_{ij1}			n_{ij2}		
$j =$		1	2	3	1	2	3
	1	2	0	4	2	4	0
i	2	4	2	0	0	2	4
	3	0	4	2	4	0	2

Under the main effects model, the completely balanced design, OD1 and OD2 are universally optimal for estimating all three main effect contrast sets. □

For a multiple factor main effects model, a design will be optimal if all possible two-factor submodels are balanced. This, in a sense, defines a type of balance with respect to the main effects model. The example of a $3 \times 3 \times 2$ model with $n = 36$ is continued in the next section to explore the relationship between simple unbalance assessments and degradation to effect estimates as measured by the D-optimality criterion.

3.2. Two-Way Tables and Data Unbalance: A Numerical Experiment

This section summarizes a numerical experiment that examines more generally the behavior of designs for the $3 \times 3 \times 2$ model with $n = 36$. The experiment's goal was to explore the relationship between the D-optimality criterion and simple summaries of pairwise unbalance for this model.

To ensure that the sample contained a broad range of design attributes, random samples of 100 designs from each of 13 configurations were generated using SAS Proc IML (SAS Institute, 2000). All 1300 designs in the combined sample were chosen so that all effect contrasts were estimable. The first sample corresponded to designs having no missing cells (i.e., designs having no zeros in N). The second sample consisted of designs with one missing cell, the third sample consisted of designs with two missing cells, and so on up to the 13th sample, which contained designs with 12 missing cells. Note for example, that the design in Example 1 is a design with two missing cells in N. Example 2 OD 1 has no missing cells in N while OD 2

has 6 missing cells. While these example designs may or may not have been chosen in the experiment, all effect contrasts are estimable with all three designs. Consequently, they would have been candidate designs for random generation in the following samples. The Example 1 design would have been a candidate for sample 3; OD 1 would have been a candidate for sample 1; and OD 2 a candidate for sample 7.

For each design, the D-optimality criterion (D-opt) values were calculated for the A-effect. For each two-way table, the numbers of missing cells (AB_miss, AC_miss, BC_miss) and the variance of the two-way table cell counts (AB_imb, AC_imb, BC_imb) were calculated. In addition, similar overall quantities for N were calculated. Each sample had a value for the number of missing cells in N (miss) associated with it and the variance of the n_{ijk} (imb) was calculated for each design.

All 13 samples were pooled. While an initial regression analysis found substantial associations between D-opt and the quantitative two-way table descriptions, diagnostics indicated that as the predicted optimality criterion value increased, so did the residual variance. Log transformations of D-opt greatly reduced the severity of the variance's observed heterogeneity. Consequently, the reported analysis focuses on the log transformed optimality criterion (log(D-opt)).

For the optimal designs in Example 2, the value of log(D-opt) for the A-effect was -4.9762. Experimental values of log(D-opt) ranged from -4.9546 to -1.3863. Fairly substantial associations (as measured by the Pearson correlation coefficient) were observed to exist between log(D-opt) and both the missing cell counts and the cell count variances. The correlation between miss and log(D-opt) was 0.82. The correlations between log(D-opt) and the pairwise missing counts also were substantial: AB_miss (0.86), AC_miss (0.77), BC_miss (0.54). Observed correlations between log(D-opt) and the cell count variances also were substantial: imb (0.88), AB_imb (0.84), AC_imb (0.76), and BC_imb (0.46). Correlations among miss, imb, AB_miss, AC_miss, BC_miss, AB_imb, AC_imb, BC_imb were all substantial as well and ranged in value from 0.50 (between BC_imb and AC_imb) to 0.94 (between miss and imb).

Four regression models were fit (Table 1) using log(D-opt) as the dependent variable. The first model used only two-way table missing cell counts as explanatory variables; the second used only two-way table cell count variances; the third used both two-way table missing cell counts and two-way table cell count variances; and the fourth included all model three explanatory variables, as well as miss and imb.

For log(D-opt), the lowest value of R^2 is 0.83, which corresponds to models 1 and 2. In both models, all coefficients differ significantly from 0. Not surprizingly, since the optimality criterion is assessed for the A-effect, the coefficients associated with measures on the AB and AC two-way tables are positive and larger in magnitude than the coefficient associated with the measure on the BC marginal matrix. Furthermore, for both models, the coefficients associated with the BC matrix are negative. Model 3 explains 88% of the variability observed in log(D-opt). Only the coefficient associated with the BC missing cell count is not significantly different from 0 ($P = 0.0630$). Model 4 establishes that some information can still be retrieved from the

measures on N. When miss and imb are added to the explanatory variables, R^2 increases to 0.92. As with model 3, only the coefficient for BC_miss is not significant ($P = 0.6088$).

This examination of the $3 \times 3 \times 2$ model with 36 observations supports the notion that examining the two-way tables may be a useful step in examining data structure. For this particular case, simple two-way table summaries, such as missing cell counts and marginal cell count variances, explain much of the variability observed in the log D-optimality criterion. Furthermore, as our earlier discussion would suggest, the associations between logged optimality criterion and measures on the marginal incidence matrices for A are stronger than the associations with marginal incidence matrices that do not include A as a factor. Furthermore, while simple associations between log(D-opt) and BC_miss and BC_imb are positive, their regression coefficients are negative. This suggests that the partial correlations, after adjusting for the associations with the AB and AC two-way tables, are negative.

4. Discussion and Summary

As suggested by the motivating example, two-way tables between main effect model factors are useful tools for gaining insight into the data structure's likely impact on effect estimate variances. In the motivating example, simple examination of two-way tables revealed both missing cells and high cell count variability which seemed to be associated with problematic degrees of collinearity. Subsequent exploration of simple measures, such as numbers of missing cells and variances of cell counts, on the two-way tables for the $3 \times 3 \times 2$ model with $n = 36$ suggests that these measures provide substantial insight into changes in D-optimality criterion values. The experiment confirms what seems to be intuitively obvious: That an increasing degree of unbalance in the two-way tables corresponds to an increasing degree of confounding and partial confounding of effects.

While the optimal design result is simple and applies to only limited model-resource combinations, which preclude incomplete blocks, it is instructive when thinking conceptually about unbalance in observational experiments. (Here "model" refers to the number of factors and numbers of levels corresponding to each factor in a main effects model and "resource" refers to the total number of observations, n.) The result emphasizes the role of the pairwise relationships between factors in the model and is consistent with many of the optimal design results that essentially require as much balance as possible between a treatment factor and each of the blocking factors (Shah and Sinha, 1989). Readers are, however, cautioned against overgeneralizing for two reasons. First, most optimal design literature assumes orthogonal block structure. Second, because of the discrete nature of the mathematics involved in obtaining optimal designs, certain model-resource combinations may behave pathologically.

Urquhart and Weeks (1978) noted that even when all effects are estimable, missing cells are problematic in the two-way main effects model. They provided a numerical example to illustrate the degradation to estimates that begins to occur as missing cells appear. Consequently, it is not surprising that D-optimality criterion values were strongly associated with measures as simple as the missing cell counts in the two-way tables.

While Urquhart and Weeks (1978) noted that "models excluding many possible, perhaps even all, interactions are the *rule*, not the exception, particularly for large data sets," they recommended that with messy data, "the researcher . . . should define the relevant parametric functions in terms of cell means." This is good advice in some cases. However, model-based analyses are common in many disciplines and often are used when analyzing both observational data and data from designed experiments. Ultimately, both the research goals and the data structure should be used to inform the approach to analyzing the data. Designed experiments using Latin Squares and fractional factorials are common in industrial settings. Both Latin Squares and some fractional factorial designs are used widely and accepted for use under an assumed main effects model, despite the missing cells in the incidence matrix, N. These designs do not, however, have missing cells in the two-way tables. This research suggests that what constitutes a meaningful degree of missingness may depend, in part, on the model. Under the main effects model, missing cells in the two-way tables may, in some senses, be the natural extension of missing cells in a design for a main effects model having only two factors.

It is not surprising that simple measures on two-way table cell counts are associated strongly with optimality criteria. Both the criteria and the two-way cell counts are functions of the $X'X$ matrix where $X = (J,\ A_1, \ldots\ , A_s)$. The $X'X$ matrix consists of blocks with the N^{fg} making up the off-diagonal blocks and diagonal blocks consisting of off-diagonal zeros with single-factor marginal totals along the diagonal. Consequently, for a main effects model, information criteria are a function of the two-way tables. However, given the complexity of these functions, it is interesting that simple two-way table measures provide as much insight into criteria values as this study suggests. More research is need to examine the association between criteria and these simple measures for other main effect model-resource combinations.

The observations presented lead to notions of four (overlapping) design categories defined in terms of the two-way tables. The first are balanced designs with respect to the main effects model. These are the optimal designs with all two-way tables being multiples of matrices of ones. Another category is suggested by Tjur (1984). Designs that are proportional with respect to the main effects model have statistically orthogonal factors. These designs are characterized by $\mathrm{n}_{ij}^{fg} = (\mathrm{n}_i^f \mathrm{n}_j^g)/\mathrm{n}$ for all $f \neq g$. Balanced designs are a special case of proportional designs. The remaining two classes are designs that are unbalanced or have missing cells with respect to the main effects model. Unbalanced designs simply include at least one unbalanced two-way table, while missing cells designs have at least one missing cell in at least one two-way table. While these design categories do not define a clear hierarchy with respect to optimality criteria, they are instructive for clarifying conceptual thinking about design structure.

Completely balanced designs are not defined in terms of the two-way tables but are the traditional equal replication designs. Using criteria that include design optimality, orthogonality, and robustness to departures from the main effects model, it can be argued that completely balanced designs are the "best" designs. As long as the main effects model holds, however, designs that are balanced with respect to the main effects model are just as good; they will be both optimal and orthogonal. Proportional designs are orthogonal and so possess some desirable analytical properties. However, some designs that are unbalanced with respect to the main

effects model may have better (smaller) D-optimality criterion values than some proportional designs. Similarly, some unbalanced designs may have larger D-optimality criterion values than some designs that have missing cells with respect to the main effects model. Nonetheless, as the experiment demonstrated, there is a strong relationship between D-optimality criterion values and simple measures on the two-way tables. This empirical study suggests that designs with missing cells generally have higher criterion values, and the more missing cells in the two-way tables, the higher the criterion. Similarly, the more variability in the cell counts for two-way tables involving the factor on which the criterion is measured, the higher the criterion. Because such simple measures were used, a practitioner can obtain the two-way tables and, by inspection, quickly assess whether there are missing cells in the two-way tables and whether the two-way table cell count variances are relatively large. While not a substitute for estimating variances of estimate effects and other traditional data analysis details, such inspections may contribute early in the analysis to the practitioner's understanding of the data structure.

5. Recommendations

Practitioners should use two-way tables to examine observational study data structure. Particularly in cases of severe unbalance, the two-way tables may provide insight into partial confounding among factors. Furthermore, we recommend that practitioners use two-way tables early in the process of analyzing observational data. In some cases, this will allow them to assess whether the data structure supports the intended analysis before too much effort is invested in that analysis. In some disciplines, a common procedure is to collapse factor categories prior to analysis. This often is done on the basis of one-way frequencies. If a main effects model is to be used, it may be better to consider both the one-way and all two-way frequencies involving the factor when making decisions about collapsing categories. In some cases, practitioners also should be aware that a design may be quite good (even optimal) for estimating one set of effect contrasts but poor for estimating others.

Acknowledgements. This research was supported, in part, by the New Mexico Agricultural Experiment Station.

REFERENCES

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, NY: John Wiley & Sons, Inc.

Farley, J. U., and Lehmann, D. R. (1986), *Meta-Analysis in Marketing: Generalization of Response Models*, Lexington, MA: D. C. Heath and Company.

Kiefer, J. (1975), "Construction and Optimality of Generalized Youden Designs," In *A Survey of Statistical Decision and Linear Models*, ed. J. N. Srivastava, Amsterdam: North-Holland Publishing, pp. 333-353.

Kunert, J. (1993), "A Note on Optimal Designs with a Non-Orthogonal Row-Column-Structure," *Journal of Statistical Planning and Inference*, 37, 265-270.

Morgan, J. P., and Bailey, R. A. (2000), "Optimal Design with Many Blocking Factors," *The Annals of Statistics*, 28, 553-577.

Myers R. H. (1990), *Classical and Modern Regression with Applications*, (2nd ed.), Boston: PWS-Kent Publishing Company.

Ramsey, F. L., and Schafer, D. W. (1997), *The Statistical Sleuth: A Course in Methods of Data Analysis*, Belmont, CA: Wadsworth Publishing Company.

SAS Institute, Inc. (2000), *SAS OnlineDoc®, Version 8*, Cary, NC: SAS Institute, Inc.

Shah, K. R., and Sinha, B. K. (1993), "Optimality Aspects of Row-Column Designs with Nonorthogonal Structure," *Journal of Statistical Planning and Inference*, 36, 331-346.

Shah, K. R., and Sinha, B. K. (1989), *Theory of Optimal Designs*, NY: Springer-Verlag.

Tjur, T. (1984), "Analysis of Variance Models in Orthogonal Designs," *International Statistical Review*, 52, 33-81.

Urquhart, N. S., and Weeks, D. L. (1978), "Linear Models in Messy Data: Some Problems and Alternatives," *Biometrics*, 34, 696-705.

VanLeeuwen, D. M., Birkes, D. S., and Seely, J. F. (1999), "Balance and Orthogonality in Designs for Mixed Classification Models," *The Annals of Statistics*, 27(6), 1927-1947.

Table 1. Summary of regression coefficients and R^2 values for regressions on the log transformed D-optimality criterion values for the A-effect contrast set.

Model	Intercept	AB_miss	AC_miss	BC_miss	AB_imb	AC_imb	BC_imb	miss	imb	R^2
1	− 4.7643	0.3441	0.3934	− 0.1445	-	-	-	-	-	0.83
2	− 5.0470	-	-	-	0.0832	0.0330	− 0.0111	-	-	0.83
3	− 4.8766	0.1897	0.2310	0.0357	0.0445	0.0166	− 0.0183	-	-	0.88
4	− 4.9787	0.1193	0.1815	− 0.0084	0.0268	0.0091	− 0.0256	− 0.0156	0.1530	0.92

NOTE: Miss, AB_Miss, AC_miss, and BC_miss are missing cell counts. Imb, AB_imb, AC_imb, and BC_imb are variances of N, N^{ab}, N^{ac}, and N^{bc}cell counts, respectively.