SAS® Technical Report R-101

Tests of Hypotheses in Fixed-Effects Linear Models

SAS Institute Inc.

5980

SAS® Technical Report R-101 Tests of the Hypotheses in Fixed-Effects Linear Models

SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS® Technical Report R-101, Tests of Hypotheses in Fixed-Effects Linear Models*, Cary, NC: SAS Institute Inc., 1978. 11 pp.

SAS® Technical Report R-101, Tests of Hypotheses in Fixed-Effects Linear Models

ISBN 1-55544-970-0

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st printing, August 1978
2nd printing, October 1996
3rd printing, June 1997

Tests of Hypotheses in Fixed Effects Linear Models

ABSTRACT

Using the concept of estimability, tests of hypotheses in multifactor fixed-effects linear models are developed without resorting to the usual assumptions. Three types of estimable functions that are usable in SAS® are defined. Each type handles unequal n's, missing cells, and any degree of confounding for any fixed-effects linear model.

INTRODUCTION

The long-smoldering issue of which are the correct sums of squares in a two-way, unbalanced model with interaction resurfaced with Kutner's (1974) article. In numerous letters to the editor in subsequent issues of The American Statistician, Bryce (1975), Hinkelman (1975), Carlson (1975), Gianola (1975), Kutner (1975), and Nelder (1975) took sides and split into what can be referred to as the two camps of linear modelers. In one camp are the R notationers, who do not reparameterize the model and who compute the various SS in the model using the R notation (Searle 1971). The R notationers compute the reduction in SS due to a specific effect given all effects except those containing the specified effect.

In the other camp are the R* notationers who reparameterize the model using the usual assumptions (Searle 1976). Since the reparameterized columns of the design matrix no longer have the property that the columns associated with lower order effects are linear functions of the higher level effect columns, the R* notationers compute the reduction in SS due to a specific effect given all other effects.

Urquhart et al. (1973) and Hocking and Speed (1975) have suggested the formation of another camp that can be called the μ notationers. The μ notationers abandon the overparameterized model and use instead a μ_{ij} model subject to a set of side conditions on the μ_{ij}'s.

Hocking and Speed (1975) show for a particular model what the R, R*, and μ notationers are actually testing in terms of the overparameterized model. As they point out, the usual assumptions and the extension of the usual assumption concept into the nonestimable functions concept of linear model theory have caused considerable confusion among practitioners.

When missing cells occur in a two-way crossed model, for example, the R* notationers in order to produce SS must not only assume the usual assumptions but must also assume that certain interaction parameters or estimates must be zero. This is typical of the problems that face reparameterizers in a general unbalanced, missing cells, and confounded design. To properly reparameterize requires a complete analysis of the design matrix, X, to determine exactly what can be estimated.

THE FIXED EFFECTS MODEL

The fixed effects model can be represented as:

$$Y=X\beta + \varepsilon \qquad (1)$$

where Y is an **n**-vector of observations, X is a known **n** x **k** matrix, β is a **k** x 1 vector of unknown parameters, and ε is an **n** x 1 vector of random variables, distributed normally with mean 0 and variance $I\sigma^2$.

In today's higher level statistical computing languages, model (1) is represented by specifying the various effects that make up the β parameter. A few examples of possible computer language representation of models are given below. The variables A, B, and C are classification variables, each having one or more levels. The variables X, Y, and Z are continuous variables. An intercept term is assumed in all models.

Model	**Represents**
Y=X	linear regression
Y=X Z	multiple regression
Y=A B C	main-effects model
Y=A B A*B	two-factor crossed model
Y=A X	main effect with covariable
Y=A X(A)	separate slopes for each level of A
Y=A X X*A	X*A used to test homogeneity of slopes
Y=X Z X*X Z*Z X*Z	response surface model
Y=A B(A) C(A B)	nested effects
Y=A B(A) C A*C B*C(A)	mixture of crossed and nested effects

In general, an effect can be represented as a sequence of continuous variables each separated by an *, followed by a sequence of class variables also separated by an *, followed by a sequence of class variables inside parentheses.

ADVANTAGES OF NOT REPARAMETERIZING

Some of the advantages of not reparameterizing have already been mentioned. These as well as others are listed below:

- The X'X matrix can usually be constructed using additions instead of multiplications.

- No distinction needs to be made between interaction and nested effects. (The columns of X are the same.)

- The preprocessing annoyance of reparameterization is avoided, including decisions about how to reparameterize when missing cells occur, or when effects are confounded.

- All sums of squares may be computed for the testable hypotheses $L\beta=0$ using:

  $$SS(HO{:}L\beta=0) = (Lb)'(L(X'X)^{-}L')^{-1}(Lb) \qquad (2)$$

  where b is any solution to the normal equations ($X'Xb=X'Y$) and $(X'X)^{-}$ is any generalized inverse of X'X (Searle 1971).

- The Ls mentioned above are easy to construct and can be printed to show exactly what is being tested. The difference in the hypotheses tested by the R and R* camps can be readily seen.

- Linear model theory and methodology are unified.

- Extension of a fixed effects program to handle mixed and random models is greatly simplified.

ESTIMABILITY

The concept of estimability has always been associated with existence. That is, for any L we say that $L\beta$ is estimable if a linear combination of the Ys exists that has an expected value of $L\beta$. Since any linear combination of the Ys, say KY, has $E(KY) = KX\beta$, then for $L\beta$ to be estimable, L must be a linear combination of the rows of the X matrix. This leads to the definition given by most authors: that $L\beta$ is estimable if there exists a K such that $L=KX$.

From the above definitions of estimability, it is clear that if an L is to be constructed such that $L\beta$ is estimable, then only linear combinations of the rows of X need be considered. In fact any linear combination of the rows of X yields an L such that $L\beta$ is estimable. Thus the X matrix or any matrix (with the same row rank of X) constructed from the rows of X may be used as a generating set for constructing any and all Ls. Some possible generating sets for L are:

- X
- X'X
- the Forward Doolittle or Cholesky matrix of X'X
- $(X'X)^{-}X'X$
- the Hermite canonical form of X'X
- any matrix produced from row operations on any of the above if the row rank is preserved.

Obviously an infinite number of Ls can be constructed from above generating sets. For a given model, an entire set of Ls must be generated, one L for each effect in the model. It is convenient to categorize a set of Ls generated for the effects in a model based on the method of generation. Three types of estimable functions (Ls) are described. Type I and Type II are for the R camp, and Type III is for the R* camp. The usual assumptions and the concept of nonestimable functions are not used for any of the three types. The theoretical development relies only on the concept of estimability.

Type I--Estimable Functions

One of the simplest sets of Ls is generated by computing the Forward Doolittle matrix from the X'X matrix (Goodnight 1978) and letting the Ls for each effect be the nonzero rows associated with that effect. For example, let

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}$$

with i,j = 1,2 and k=1,..,nij. If $n_{11}=2$ and $n_{12}= n_{21}=n_{22}=1$, then the X'X matrix is:

The $X'X$ Matrix

	μ	$\alpha 1$	$\alpha 2$	$\beta 1$	$\beta 2$	$\alpha\beta 11$	$\alpha\beta 12$	$\alpha\beta 21$	$\alpha\beta 22$
μ	5	3	2	3	2	2	1	1	1
$\alpha 1$	3	3	0	2	1	2	1	0	0
$\alpha 2$	2	0	2	1	1	0	0	1	1
$\beta 1$	3	2	1	3	0	2	0	1	0
$\beta 2$	2	1	1	0	1	0	1	0	1
$\alpha\beta 11$	2	2	0	2	0	2	0	0	0
$\alpha\beta 12$	1	1	0	0	1	0	1	0	0
$\alpha\beta 21$	1	0	1	1	0	0	0	1	0
$\alpha\beta 22$	1	0	1	0	1	0	0	0	1

The Forward Doolittle matrix (with each nonzero row divided by its diagonal) is:

The Forward Doolittle

	μ	$\alpha 1$	$\alpha 2$	$\beta 1$	$\beta 2$	$\alpha\beta 11$	$\alpha\beta 12$	$\alpha\beta 21$	$\alpha\beta 22$
μ	1	3/5	2/5	3/5	2/5	2/5	2/5	1/5	1/5
$\alpha 1$	0	1	−1	1/6	−1/6	2/3	1/3	−1/2	−1/2
$\alpha 2$	0	0	0	0	0	0	0	0	0
$\beta 1$	0	0	0	1	−1	4/7	−4/7	3/7	−3/7
$\beta 2$	0	0	0	0	0	0	0	0	0
$\alpha\beta 11$	0	0	0	0	0	1	−1	−1	1
$\alpha\beta 12$	0	0	0	0	0	0	0	0	0
$\alpha\beta 21$	0	0	0	0	0	0	0	0	0
$\alpha\beta 22$	0	0	0	0	0	0	0	0	0

The Type I Ls for each effect are:

$$L_{\alpha} = \{0 \quad 1 \quad -1 \quad 1/6 \quad -1/6 \quad 2/3 \quad 1/3 \quad -1/2 \quad -1/2\}$$

$$L_{\beta} = \{0 \quad 0 \quad 0 \quad 1 \quad -1 \quad 4/7 \quad -4/7 \quad 3/7 \quad -3/7\}$$

$$L_{\alpha\beta} = \{0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad -1 \quad -1 \quad 1\}$$

Using the above Ls the Type I SS may be computed using (2). Furthermore, the SS computed in this manner are equivalent to:

$$SS(H_o\colon\ L_{\alpha}\ \beta = 0) = R(\alpha|\mu)$$

$$SS(H_o\colon\ L_{\beta}\ \beta = 0) = R(\beta|\mu,\alpha)$$

$$SS(h_o\colon\ L_{\alpha\beta}\ \beta = 0) = R(\alpha\beta|\mu,\alpha,\beta)$$

The Type I Ls via (2) produce orthogonal quadratic forms (additive SS) which for $I\sigma^2$ error structures are thus independent. However, the Type I hypotheses have several drawbacks. First, they are not invariant with respect to the ordering of effects in the model. Second, for unbalanced designs, each hypotheses generally involves the parameters of the effect being tested plus all remaining parameters that follow in the model statement. In the above example, $L_\alpha\beta$ involves α, β, and $\alpha\beta$ parameters. Third, as can be demonstrated by varying the number of observations per cell, some of the Type I hypotheses are dependent on the cell frequencies.

Type II--Estimable Functions

An excellent use of Type I estimable functions is for the study of the nature of hypotheses being tested in balanced designs. The Type I Ls for a balanced design are unique (providing no higher level effects precede associated lower-order effects in the model). If the preceding example had been balanced, then L_α would involve only α and $\alpha\beta$ parameters. L_β would involve only β and $\alpha\beta$ parameters, and $L_{\alpha\beta}$ would involve only $\alpha\beta$ parameters. For balanced designs, the hypotheses normally computed for an effect are a function only of the parameters of that effect and the parameters of effects which contain that effect. For balanced as well as unbalanced designs, containment may be defined as follows: an effect E_2 is said to contain the effect E_1 if it is known, by observing only the model statement, that all of the columns of the X matrix associated with E_1 can be represented as linear combinations of the columns associated with E_2.

For example, in the model Y = A B A*B:

μ is contained in A, B, A*B
A is contained in A*B
B is contained in A*B
A*B is not contained in any other effect.

Furthermore, in the unreparameterized model, if E_2 contains E_1 then $R(E_1|E_2) = 0$. If E_2 appears in the model statement before E_1, then the Type I L associated with E_1 is zero. It is clear then that to generate Ls, which for each effect involves only the parameters of that effect and the parameters of effects containing that effect, several orderings of the model could be made and a Doolittle performed.

The Type II Ls for an effect E_1 are the associated rows of the Doolittle for the model that has been arranged to put all effects which do not contain E_1 before E_1. The columns of each L, once generated, are naturally rearranged to reflect the original order of the model. Goodnight (1978) describes a reversible Sweep Operator which allows the computation of Type II Ls without actually rearranging the terms in the model.

For the previous example, the Type II Ls are given below:

$$L_\alpha = \{0 \quad 1 \quad -1 \quad 0 \quad 0 \quad 4/7 \quad 3/7 \quad -4/7 \quad -3/7\}$$

$$L_\beta = \{0 \quad 0 \quad 0 \quad 1 \quad -1 \quad 4/7 \quad -4/7 \quad 3/7 \quad -3/7\}$$

$$L_{\alpha\beta} = \{0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad -1 \quad -1 \quad 1\}$$

Type II Ls, in general, no longer produce orthogonal quadratic forms. However, the hypotheses they test are more in line with hypotheses being tested in the balanced case, at least in the sense that only an effects parameters and parameters of effects containing that effect are involved in the hypothesis. If an effect E_1 is contained in no other effect, then the Type II L generated for that effect is such that $L\beta$ is a maximum rank hypothesis involving only the parameters of E_1. However, if an effect E_1 is contained in higher-level effects, then the elements of L associated with those higher-level parameters are functions of the cell frequencies; this can be demonstrated by computing the Type II Ls for the previous example under differing cell frequencies. Additional discussion on the related R notation may be found in Hocking and Speed (1975) and Speed and Hocking (1976).

Type III--Estimable Functions

For most unbalanced designs it is usually possible to test the same set of hypotheses (estimable functions) that would have been tested if the design had been balanced. For those designs which started out balanced, but for which observations were lost due to external forces, there is no reason to alter the hypotheses. Type II hypotheses, for an effect that is contained in other effects do vary depending on the cell frequencies. The Type III hypotheses developed here do not vary, but designs with no missing cells do correspond to the estimable function used in the balanced case.

Had the 2x2 example been balanced, the Type II Ls would have been:

$$L_\alpha = \{0 \quad 1 \quad -1 \quad 0 \quad 0 \quad 1/2 \quad 1/2 \quad -1/2 \quad -1/2\}$$
$$L_\beta = \{0 \quad 0 \quad 0 \quad 1 \quad -1 \quad 1/2 \quad -1/2 \quad 1/2 \quad -1/2\}$$
$$L_{\alpha\beta} = \{0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad -1 \quad -1 \quad 1\}$$

Note that in each of the above Ls only the effect being tested and any effect that contains it are involved. Also, note that both L_α and L_β are orthogonal to $L_{\alpha\beta}$. In fact for any balanced design the Type II L for any effect is orthogonal to the Ls of effects that contain it. This leads to the following definition: a set of Ls, one for each effect in the model, is Type III if each L is a maximum rank hypothesis involving only the parameters of the effect in question and parameters of effects that contain it; and each L is orthogonal to all Ls of effects that contain the effect in question.

The above definition implies that Type II Ls may be converted to Type III Ls by simply making each lower-order L orthogonal to the Ls of all effects that contain the lower-order effect. Furthermore if an effect is not contained in any other effect, then the Type II and III Ls for that effect are the same.

Type III Ls can be computed directly from any generating set (Goodnight 1976) and need not be computed from the Type II Ls. For designs with no missing cells, the sums of squares generated by the Type III hypotheses correspond to the SS computed in the reparameterized model for which the usual assumptions were made.

SYMBOLIC REPRESENTATION OF A GENERATING SET

Since all Ls can be computed from linear combinations of the rows of any generating set, the general form of all Ls can be represented symbolically by multiplying each row of the generating set by a symbolic constant and then adding the resultant rows together.

The following matrix is a generating set for the 2x2 example:

	μ	α_1	α_2	β_1	β_2	$\alpha\beta_{11}$	$\alpha\beta_{12}$	$\alpha\beta_{21}$	$\alpha\beta_{22}$
L1	−1	0	1	0	1	0	0	0	1
L2	0	1	−1	0	0	0	1	0	−1
L3	0	0	0	0	0	0	0	0	0
L4	0	0	0	1	−1	0	0	1	−1
L5	0	0	0	0	0	0	0	0	0
L6	0	0	0	0	0	1	−1	−1	1
L7	0	0	0	0	0	0	0	0	0
L8	0	0	0	0	0	0	0	0	0
L9	0	0	0	0	0	0	0	0	0
SUM	L1	L2	L1 − L2	L4	L1 − L4	L6	L2 − L6	L4 − L6	L1 − L2 − L4 + L6

The symbolic sum, representing the general form of all estimable functions, is best written beside the parameters as follows:

μ	L1
$\alpha 1$	L2
$\alpha 2$	L1-L2
$\beta 1$	L4
$\beta 2$	L1-L4
$\alpha\beta 11$	L6
$\alpha\beta 12$	L2-L6
$\alpha\beta 21$	L4-L6
$\alpha\beta 22$	L1-L2-L4+L6

Note that any values for L1, L2, L4, and L6 produce an L for which Lβ is estimable.

The symbolic notation can be manipulated. For example, by letting L1 and L4 represent zero, the general form of all estimable functions not involving μ, $\beta 1$, and $\beta 2$ can be seen as follows:

μ	0
$\alpha 1$	L2
$\alpha 2$	-L2
$\beta 1$	0
$\beta 2$	0
$\alpha\beta 11$	L6
$\alpha\beta 12$	L2-L6
$\alpha\beta 21$	-L6
$\alpha\beta 22$	-L2+L6

For a one degree-of-freedom contrast that involves the comparison $\alpha 1$ vs $\alpha 2$, a logical necessity is determining the value of L6. This, in essence, is the difference between Type II and Type III. Estimable functions involving only interaction parameters can be achieved by setting L1=L2=L4=0.

The generating set used to represent the general form of estimable functions should, for simplicity, be as sparse as possible with the number of nonzero rows equaling the rank of X'X. The Hermite canonical form of X'X (as used in the previous example) meets the simplicity requirements as well as any other matrix does. The Hermite canonical form of X'X can be computed by pivoting on each nonzero diagonal in sequence of X'X, the Forward Doolittle, or Cholesky matrix (Goodnight 1978).

CONCLUSION

Each of the three types of estimable functions defined handles unequal n's, missing cells, and any degree of confounding. Unlike Type I and Type II estimable functions, the general philosophy behind the Type III estimable functions is that tests of hypotheses made for any given effect should be the same for all designs that have the same general form of estimable functions.

Using a unified approach to define the different types of estimable functions allows for meaningful comparisons to be made between the R and R* camps. It also eliminates the need to try to justify a particular computing approach through the use of buzz words and ambiguous jargon. Although an infinitive number of types of estimable functions exists for a given set of data, only three were defined here. A fourth type of estimable function is given in Barr et al. (1976) and Goodnight (1976). This fourth type of estimable function deals primarily with designs involving missing cells, and provides alternative tests of hypotheses to the three types defined here.

REFERENCES

Barr, A.J., Goodnight, J.H., Sall, J.P., and Helwig, J.T. (1976), A User's Guide to SAS 76, Cary, N.C.: SAS Institute Inc.

Bryce, G.R. (1975), "The One-Way Model," The American Statistician, 29, 69.

Carlson, J.E. (1975), "Hypotheses Tests in Linear Models," The American Statistician, 29, 133.

Gianola, D. (1975), "Hypothesis Tests in Linear Models," The American Statistician, 29, 133.

Goodnight, J.H. (1976), "The General Linear Models Procedure," Proceedings of the First International SAS User's Group, Cary, N.C.: SAS Institute Inc.

Goodnight, J.H. (1978), "The Sweep Operator: Its Importance to Statistical Computing," Proceedings of the Eleventh Interface of Statistics and Computer Science, Institute of Statistics, N.C. State University, Raleigh, N.C.

Hinkelmann, Klaus (1975), "Hypothesis Testing in Linear Models," The American Statistician, 29, 110.

Hocking, R.R. and Speed (1975), "A Full Rank Analysis of Some Linear Model Problems," Journal of the American Statistical Association, 70, 706-712.

Kutner, M.H. (1974), "Hypothesis Testing in Linear Models (Eisenhardt Model I)," The American Statistician, 29, 98-100, 133-134.

Kutner, M.H. (1975), (letter to the editor), The American Statistician, 29, 133-134.

Nelder, J.A. (1974), "Analysis of Various Programs, Least Squares and Two-Way Tables," Applied Statistics, 23, 101.

Nelder, J.A. (1976), "Hypothesis Testing in Linear Models," The American Statistician, 30, 103.

Searle, S.R. (1971), Linear Models, New York: John Wiley & Sons.

Searle, S.R. (1976), "Comments on Anova Calculations for Messy Data," Proceedings of the First International SAS Users Group, Cary, N.C.: SAS Institute Inc.

Speed, F.M., and Hocking, R.R. (1976), "The Use of the R()-Notation with Unbalanced Data," The American Statistician, 30, 30-33.

Urquhart, N.S., Weeks, D.L., and Henderson, C.R. (1973), "Estimation Associated with Linear Models: A Revisitation," Communications in Statistics, I, 303-330.

Index

E

F

L

SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513

ISBN 1-55544-970-0

9 781555 449704