A Pitfall in Linear Models: The Misuse of "Sequential" F Statistics

by

Myra L. Samuels
Purdue University

Technical Report #92-29

Department of Statistics
Purdue University

July 1992

A PITFALL IN LINEAR MODELS: THE MISUSE OF "SEQUENTIAL" *F* STATISTICS

by

Myra L. Samuels
Purdue University

ABSTRACT

This article calls attention to a pitfall in the use of ordinary fixed-effects linear models. Many textbooks and computing package manuals describe several types of *F* statistic, including the type sometimes called the "sequential" or "variables-added-in-order" or "Type I" *F* statistic. In the common situation where the regressor variables are random rather than fixed by design, the sequential *F* statistic is meaningless and should never be used.

KEY WORDS: Analysis of variance; *F* test; Nonorthogonality; Random regressors; Regression; Type I sum of squares; Unbalanced data.

1. Introduction

A question frequently asked by users of statistical computing packages is "How do I know which type of sum of squares to use in analyzing a linear model?" This article points out a conceptual error that can arise concerning this question. Some textbooks propagate the error by explicit mistaken assertions; some others do so by vague, incomplete, or missing discussions of the topic.

The erroneous belief is that the choice between two types of F test (sometimes called "partial" and "sequential") is a matter of taste — the choice to be made according to which of two hypotheses the user really wants to test, or according to which of two variance estimates the user prefers. This view is seriously misleading. If the partial and sequential F statistics differ, it is very often because the values of the regressor variables are random rather than fixed by design. It will be argued that in this case the sequential F test should *never* be used.

2. The Issue

Consider the standard linear model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + \epsilon_i \tag{1}$$

$i = 1, \cdots, n$, where the βs are constants (fixed-effects model). In analyzing the model (1), there are two ways to define a sum of squares (SS) corresponding to each regressor: (1) the *"partial"* SSs, with the SS corresponding to each regressor generated by adding that regressor to a model already containing the other $p-1$ regressors, and (2) the *"sequential"* SS, with regressors entered in the order $X_1, X_2, \ldots, X_p$. In standard notation (Neter, Wasserman and Kutner 1990, p. 277) the partial SSs are $\mathrm{SSR}(X_1|X_2, X_3, X_4, \ldots, X_p)$, $\mathrm{SSR}(X_2|X_1, X_3, X_4, \ldots, X_p)$, $\mathrm{SSR}(X_3|X_1, X_2, X_4, \ldots, X_p)$, and so on, and the sequential SSs are $\mathrm{SSR}(X_1)$, $\mathrm{SSR}(X_2|X_1)$, $\mathrm{SSR}(X_3|X_2, X_1)$, and so on. More generally, partial and sequential SSs can be defined for batches of regressors, corresponding to factor main effects and interactions in an analysis of variance (ANOVA) model.

Some computer packages (see, for instance, SAS Institute Inc 1990, pp. 934-936, p. 1368, and SPSS 1990, pp. 64-65, p. 377) produce both partial and sequential SSs and also corresponding F statistics. These F statistics, which use as denominator the residual MS from the model (1), will be called here partial and sequential F statistics, respectively. Some textbooks discuss both types of SS and F statistic (see, for instance, Draper and Smith 1981, pp. 208-209; Kirk 1982, pp. 412-413; Kleinbaum, Kupper, and Muller 1988, pp. 134-137, Maxwell and Delaney 1990, pp. 286-290, Milliken and Johnson 1984, pp. 146-151; Searle 1987, p. 112, p. 121; note that not all of these authors use the terms "partial" and "sequential" in the same sense as defined here).

The point at issue is whether the sequential F statistics and their corresponding P values have any meaning. We will argue that they do *not*, in the important case where the X_{ji} are regarded as random variables, rather than fixed by design. (If the X_{ji} are fixed by an orthogonal design, then the two types of SSs are identical, so that the issue does not arise. The case where the X_{ji} are fixed by a nonorthogonal design is considered briefly in Section 5.)

For simplicity, we discuss in detail only the case $p = 2$. The model (1) is then

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i \tag{2}$$

Table 1 shows, in standard notation (Neter, Wasserman and Kutner 1990, p. 277) the two sequential decompositions of SS(total) $= \sum(Y_i - \bar{Y})^2$.

****

Table 1 goes about here

****

For the regressor X_1, the partial SS is SSR$(X_1|X_2)$ and the sequential SS is SSR(X_1). The corresponding partial and sequential F statistics are:

$$F_{\text{part}} = \frac{\text{MSR}(X_1|X_2)}{\text{MSE}} \tag{3}$$

$$F_{\text{seq}} = \frac{\text{MSR}(X_1)}{\text{MSE}} \tag{4}$$

where each mean square (MS) is obtained as MS = SS/df, and in each case consider the test obtained by referring the statistic to an $\mathcal{F}_{1,n-3}$ distribution.

The test based on F_{part} is the standard normal-theory test of the hypothesis $H_0 : \beta_1 = 0$ within the model (2). Our concern here is with the test based on F_{seq}, which will be called the sequential F test.

The following examples illustrate the context of the conceptual error that can arise with respect to the sequential F test.

Example 1. ANOVA context. In an observational study, the variables of interest are Y = income of employees in a corporation, X_1 = gender (0 = female, 1 = male), and X_2 = educational level (0 = low, 1 = high). The statistic F_{part} tests the effect of gender on income, adjusted for education, and would (perhaps) be relevant to the question of gender discrimination in hiring. But suppose that for some reason the analyst wants to compare the income of males and females *without* adjusting for education. Is F_{seq} appropriate for this purpose?

Example 2. Regression context. In an observational study, the variables of interest are Y = blood pressure of men, X_1 = amount of smoking, and X_2 = age. The test based on F_{part} assesses the significance of adding 'smoking' to a model already containing 'age.' " But suppose the analyst wishes to assess the significance of 'smoking' alone as a predictor. Is F_{seq} appropriate for this purpose?

Contrary to the assertion or implication of some textbooks and computer package manuals, the answer to the questions in these examples is an unequivocal "No." In fact, in settings like those illustrated in the examples, F_{seq} is not appropriate for any purpose at all.

3. A Counterexample

Before proceeding to a formal analysis, we demonstrate the flaw in F_{seq} by an example that makes it intuitively clear.

Example 3. A scientist has measured the weight Y of n_1 males and n_2 females.

She wants to test the null hypothesis

$$H_0 : \mu^{(\mathrm{M})} = \mu^{(\mathrm{F})} \tag{5}$$

where $\mu^{(\mathrm{M})}$ and $\mu^{(\mathrm{F})}$ denote the population mean weights for males and females, respectively. She proposes to use the t test she learned in STAT 101, based on the test statistic

$$t = \frac{\bar{Y}^{(\mathrm{M})} - \bar{Y}^{(\mathrm{F})}}{s^{\star}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{6}$$

where $\bar{Y}^{(\mathrm{M})}$ and $\bar{Y}^{(\mathrm{F})}$ are, respectively, the sample mean weights for males and females and $s^{\star}$ is the pooled standard deviation from the two samples. Now the scientist has a friend, a statistical consultant, who notices that data on height is also available and so suggests a fancier analysis based on fitting the model (2) with Y = weight, X_1 = gender, and X_2 = height. The test based on F_{part} would then be a comparison of men's and women's weights after adjusting for their difference in height (analysis of covariance). But our scientist wants the *unadjusted* comparison (5), so her friend advises the use of F_{seq}.

To see what happens, note that the t test (6) is equivalent to the F test based on the reduced model

$$Y_i = \beta_0^{\star} + \beta_1^{\star} X_{1i} + \epsilon_i^{\star} \tag{7}$$

with test statistic

$$F_{\mathrm{red}} = \frac{\mathrm{MSR}(X_1)}{\mathrm{MSE}^{\star}} \tag{8}$$

where $\mathrm{MSR}(X_1)$ is the same as in Table 1 but $\mathrm{MSE}^{\star}$ is the residual mean square from the reduced model (7). The statistics (6) and (8) are linked by the relation $t^2 = F_{\mathrm{red}}$.

Since F_{seq} and F_{red} have the same numerator, and since the full model (2) fits the data much more closely than the reduced model (7), resulting in a much smaller denominator for F_{seq}, our delighted scientist finds that she obtains a much smaller P value from F_{seq} than she did from (6). Catching on quickly, she realizes that she can make it even smaller by incorporating data on other variables related to weight, such as X_3 = shoulder breadth, X_4 = waist size, and so on. She is tormented, though, by a sneaking feeling that she is getting something for nothing.

4. Formal Analysis

To see what is going on, let us specify the regression models more precisely, assuming now that the regressors are random variables. The full model (2) can be specified by requiring that the triples (Y_i, X_{1i}, X_{2i}) be independent and identically distributed (IID) with

$$\begin{aligned} \mathrm{E}(Y_i|X_{1i}, X_{2i}) &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} && (9) \\ \mathrm{Var}(Y_i|X_{1i}, X_{2i}) &= \sigma^2 && (10) \end{aligned}$$

The conditions (9) and (10) imply that (2) holds with $E(\epsilon_i) = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2$, and ϵ_i uncorrelated with X_{1i} and X_{2i}.

Similarly, the reduced model (7) can be specified by requiring that the pairs (Y_i, X_{1i}) be IID with

$$\begin{aligned} \mathrm{E}(Y_i|X_{1i}) &= \beta_0^\star + \beta_1^\star X_{1i} && (11) \\ \mathrm{Var}(Y_i|X_{1i}) &= \sigma^{\star 2} && (12) \end{aligned}$$

The conditions (11) and (12) imply that (7) holds with $E(\epsilon_i^\star) = 0$, $\mathrm{Var}(\epsilon_i^\star) = \sigma^{\star 2}$, and $\epsilon_i^\star$ uncorrelated with X_{1i}. To begin with, let us assume that (9), (10), (11) and (12) are simultaneously valid. We also assume that $\beta_2 \neq 0$, since otherwise the full and reduced models are equivalent and the partial and sequential F statistics are virtually equivalent.

To investigate the statistics F_{seq} and F_{red}, let us consider the expectations of their numerator and denominators. Straightforward calculation (see Appendix) yields

$$\begin{aligned} E[\mathrm{MSR}(X_1)] &= (n-1)\sigma_1^2\beta_1^{\star 2} + \sigma^{\star 2} & (13) \\ E[\mathrm{MSE}^\star] &= \sigma^{\star 2} & (14) \\ E[\mathrm{MSE}] &= \sigma^2 & (15) \end{aligned}$$

where $\sigma_1^2 = \mathrm{Var}(X_{1i})$.

In all three examples given above, the implicit erroneous claim is that the hypothesis tested by F_{seq} is

$$H_0 : \beta_1^\star = 0 \tag{16}$$

where $\beta_1^\star$ is defined by the reduced model (11). It is clear from (13) and (14) that the numerator and denominator of F_{red} have the same expectation under (16); indeed, F_{red} is the standard normal-theory test statistic for (16). Turning now to F_{seq}, it follows from (13) and (15) that under (16) the ratio of expectations of its numerator and denominator is $\sigma^{\star 2}/\sigma^2$, which in turn can be written (see Appendix for proof) as

$$\frac{E[\mathrm{MSR}(X_1)]}{E[\mathrm{MSE}]} = \frac{1}{1 - \rho^2_{YX_2 \cdot X_1}} \tag{17}$$

where $\rho_{YX_2 \cdot X_1}$ is the partial correlation of Y_i with X_{2i}, eliminating X_{1i}. The relation (17) shows that the statistic F_{seq}, viewed as a test of (16), will tend to be inflated, leading to inappropriately small P values. The distortion can be very large, if the partial correlation is large.

For another perspective on the flaw in F_{seq} as a test of (16), note that the least squares estimates $\hat{\beta}_1^\star$, $\hat{\beta}_1$, and $\hat{\beta}_2$ satisfy the relation

$$\hat{\beta}_1^\star = \hat{\beta}_1 + \hat{\gamma}\hat{\beta}_2 \tag{18}$$

where $\hat{\gamma}$ is the sample regression coefficient of X_2 upon X_1, that is, $\hat{\gamma} = \sum(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)/\sum(X_{1i} - \bar{X}_1)^2$, with $\bar{X}_j = n^{-1}\sum_i X_{ji}$. The statistic F_{seq} is faulty as a test of (16) because it takes into account the sampling error in $\hat{\beta}_1$ and $\hat{\beta}_2$ but *ignores* the sampling error in $\hat{\gamma}$, which arises from the randomness of the X_{ji}.

The relation (18) is particularly clear in the context of Example 3, where it becomes

$$\bar{Y}^{(M)} - \bar{Y}^{(F)} = \hat{\beta}_1 + \hat{\beta}_2(\bar{X}_2^{(M)} - \bar{X}_2^{(F)}), \tag{19}$$

where $\bar{X}_2^{(M)}$ and $\bar{X}_2^{(F)}$ are the sample mean heights of males and females, respectively. In assessing the sampling error in $(\bar{Y}^{(M)} - \bar{Y}^{(F)})$, the faulty analysis ignores sampling error in $\bar{X}_2^{(M)}$ and $\bar{X}_2^{(F)}$. Thus, the misguided scientist of Example 3, in using F_{seq}, is trying to reduce uncertainty about weight by pretending there is no uncertainty about height.

The above argument shows that if we assume (9), (10), (11) and (12) then F_{seq} is not valid as a test of (16); rather, F_{red} is the correct test for this hypothesis. This leaves open the question of whether, perhaps under weaker assumptions, F_{seq} might be a valid test of some other hypothesis. In fact, however, even if we assume only (9) and (10) , then (15) still holds and it is not hard to show (see Appendix 2) that if $\beta_2 \neq 0$ and degenerate cases are excluded, then $E[\mathrm{MSR}(X_1)] > \sigma^2$, so that F_{seq} cannot possibly have a central $\mathcal{F}$ distribution under *any* hypothesis.

5. Discussion

Although the preceding exposition has concentrated on the case $p = 2$, the sequential F statistics are equally meaningless (when the X_{ji} are random) in the general model (1). Moreover, the critique applies equally to F tests involving batches of regressors, such as those encountered in factorial ANOVA when the factors have more than two levels.

Note that the difficulties with the sequential F statistics are not due to the presence of interaction (our examples use additive models), nor are they due to measurement error in the X_{ji}. Rather, they are due solely to the randomness of the X_{ji}.

If the X_{ji} are fixed rather than random, then the sequential F statistics can be meaningful for certain special purposes, such as testing orthogonalized components in a polynomial regression (Freund, Littell and Spector 1986, pp. 28-30). Moreover, the sequential SSs (rather than the F statistics) can be useful in a variety of ways, for instance, in constructing a split-plot ANOVA.

Textbooks that cover linear models usually include some discussion of nonorthogonal design matrices under such names as "collinearity" and "unbalanced ANOVA." In practice a common source of nonorthogonality is the presence of random regressors. Nevertheless, it is not common for textbooks to give a careful discussion of the interpretation of linear models when the regressors are random (even though the problems and examples often involve random regressors).

In discussing factorial ANOVA, some textbooks display prominently the "hypotheses" tested by sequential F statistics and caution the student that with nonorthogonal data these "hypotheses" will depend on the cell frequencies. While these statements are not, strictly speaking, incorrect, they may be misleading because nonorthogonality is often the result of random X_{ji}, in which case the cell frequencies are random and (as shown in Section 4) the sequential F test does not test any hypothesis at all.

It would seem advisable to alert students and users of statistics to the pitfall in the use of sequential F tests. In addition, it would be desirable to disseminate more widely the following facts: (1) in observational, as opposed to experimental, research it is often appropriate to regard the regressors as random; (2) much, but not all, of the theory developed for fixed regressors carries over to the case of random regressors; (3) when regressors are random, the parameters being estimated by a full model and a reduced model have different meanings, and the decision as to which parameter is to be preferred should depend on the question the research is intended to answer; (4) when regressors are random, nonorthogonality is the usual case, rather than the exceptional one; (5) the situation of random regressors is conceptually distinct from the problem of measurement error in the regressors.

References

Draper, N.R. and Smith, H. (1981), *Applied Regression Analysis, Second Edition*, New York: John Wiley.

Freund, R.J., Littell, R.C., and Spector, P.C. (1986), *SAS System for Linear Models*, Cary, NC: SAS Institute Inc.

Johnson, R.A. and Wichern, D.W. (1982), *Applied Multivariate Statistical Analysis*, Englewood Cliffs, NJ: Prentice-Hall.

Kirk, R.E. (1982), *Experimental Design: Procedures for the Behavioral Sciences*, Belmont, CA: Brooks-Cole.

Kleinbaum, D.G., Kupper, L.L., and Muller, K.E. (1988), *Applied Regression Analysis and Other Multivariable Methods*, Boston, MA: PWS-Kent.

Maxwell, S.E. and Delaney, H.D. (1990), *Designing Experiments and Analyzing Data*, Belmont, CA: Wadsworth.

Milliken, G.A. and Johnson, D.E. (1984), *Analysis of Messy Data, Volume 1*, New York: Van Nostrand Reinhold.

Neter, J., Wasserman, W., and Kutner, M.H. (1990), *Applied Linear Statistical Models*, Homewood, IL: Irwin.

SAS Institute Inc (1990), *SAS/STAT User's Guide, Version 6, Fourth Edition*, Cary, NC: Sas Institute Inc.

Searle, S.R. (1987), *Linear Models for Unbalanced Data*, New York: John Wiley.

SPSS Inc (1990), *SPSS Reference Guide*, Chicago, IL: SPSS Inc.

Appendix. Proofs

To ease the notation, let $V_1 = \mathrm{MSR}(X_1)$, $V_2 = \mathrm{MSE}^\star$, $V_3 = \mathrm{MSE}$, and, for $j = 1, 2$, let $\mathbf{X}_j$ denote the vector $(X_{j1}, X_{j2}, \cdots, X_{jn})$. Assuming (9), (10), (11), and (12), standard linear model theory (Neter, Wasserman and Kutner 1990, pp. 94-95, 240) yields

$$\begin{aligned} V_1 &= (n-1)s_1^2\hat{\beta}_1^{\star 2} && (20) \\ E[V_1|\mathbf{X}_1] &= (n-1)s_1^2\beta_1^{\star 2} + \sigma^{\star 2} && (21) \\ E[V_2|\mathbf{X}_1] &= \sigma^{\star 2} && (22) \\ E[V_3|\mathbf{X}_1, \mathbf{X}_2] &= \sigma^2 && (23) \end{aligned}$$

where $s_1^2 = (n-1)^{-1}\sum_i (X_{1i} - \bar{X}_1)^2$.

To prove (13), (14), and (15), one simply takes expectations in (21), (22), and (23).

To prove (17), let (Y, X_1, X_2) be distributed as (Y, X_{1i}, X_{2i}) and let $\epsilon^\star$ denote the residual $\epsilon^\star = Y - \beta_0^\star - \beta_1^\star X_1$ from the best linear prediction of Y based on X_1. The partial correlation $\rho_{YX_2 \cdot X_1}$ is defined (Johnson and Wichern 1982, p. 342) as the simple correlation between $\epsilon^\star$ and δ, where δ is the residual of X_2 from its best linear predictor based on X_1. Consequently we can write

$$1 - \rho^2_{YX_2 \cdot X_1} = \frac{\sigma_A^2}{\mathrm{Var}(\epsilon^\star)} \qquad (24)$$

where σ_A^2 is the residual variance in $\epsilon^\star$ after linear prediction of $\epsilon^\star$ from δ. Now (9), (10), (11) and (12) imply that $\mathrm{Var}(\epsilon^\star) = \sigma^{\star 2}$ and $\sigma_A^2 = \sigma^2$, and (17) then follows from (24).

Turning to the claim made in the final paragraph of Section 4, we now assume only (9) and (10). Since $\hat{\beta}_1^\star = (n-1)^{-1}s_1^{-2}\sum_i Y_i(X_{1i} - \bar{X}_1)$, it follows that

$$\text{Var}[\hat{\beta}_1^\star|\mathbf{X}_1, \mathbf{X}_2] = \sigma^2/(n-1)s_1^2 \tag{25}$$

From the relation (18) and standard linear model theory it follows that

$$E[\hat{\beta}_1^\star|\mathbf{X}_1, \mathbf{X}_2] = \beta_1 + \hat{\gamma}\beta_2 \tag{26}$$

From (20), (25) and (26) we have

$$E[V_1|\mathbf{X}_1, \mathbf{X}_2] = \sigma^2 + (n-1)s_1^2(\beta_1 + \hat{\gamma}\beta_2)^2 \tag{27}$$

Taking expectations in (27) yields

$$E[V_1] = \sigma^2 + (n-1)\sigma_1^2 E[(\beta_1 + \hat{\gamma}\beta_2)^2] \tag{28}$$

Recalling that $\hat{\gamma}$ is the sample regression coefficient of $\mathbf{X}_2$ upon $\mathbf{X}_1$, it is clear that $\text{Var}(\hat{\gamma}) > 0$ except in the degenerate cases that either X_1 or X_2 is a constant or that X_2 is a linear function of X_1. Consequently, if $\beta_2 > 0$ it follows from (28) that $E(V_1) > \sigma^2$, as was to be shown.

Table 1. Sequential Sums of Squares

		Sum of Squares	
Source	df	X_1 entered first	X_2 entered first
X_1	1	SSR(X_1)	SSR(X_2)
X_2	1	SSR($X_2 \| X_1$)	SSR($X_1 \| X_2$)
Residual	$n-3$	SSE	SSE
Total	$n-1$	SS(total)	SS(total)