Statistica Sinica **8**(1998), 101-117

FORWARD SELECTION ERROR CONTROL IN THE ANALYSIS OF SUPERSATURATED DESIGNS

Peter H. Westfall, S. Stanley Young and Dennis K. J. Lin

Texas Tech University, Glaxo Wellcome Inc. and Penn State University

Abstract: Supersaturated designs are designed to assess the effects of many factors simultaneously. The assumption of "effect sparsity" is often needed to justify the selection of these designs. However, when effect sparsity holds, Type I errors can easily occur. Forward-selection multiple test procedures are proposed to address and solve this problem.

Key words and phrases: Adjusted p-values, control variates, multiplicity adjustment, resampling, variable selection.

1. Introduction

In industry the first phase of experimentation often begins with a screening experiment, where many questions are addressed with few experimental units. It is often found that many of the factors are unimportant, and further experiments are carried out using only those "significant" factors. The situation where many effects are unimportant is called "effect sparsity" (Box and Meyer (1986)). Saturated designs have been extended to supersaturated designs (SSDs) by Booth and Cox (1962); recent advances are given by Lin (1993, 1995), Wu (1993), and Nguyen (1996). In these designs, there are more experimental factors under consideration than there are experimental units. Here, effect sparsity is relied upon to make analysis feasible; if the few important factors are fit to a model, the residual error can be used to test the significance of those terms included in the model.

As with any decision problem, errors of various types must be balanced against costs. In screening designs, there are costs of declaring an inactive factor to be active (Type I error), and costs of declaring an active effect to be inactive (Type II error). Type II errors are troublesome as addressed in Lin (1995). However, Type I errors are also troublesome, as they can cause unnecessary cost in the follow-up experiments, and they can cause detrimental actions if the experiment has immediate impact on practice. Unfortunately, Type I errors are very likely when "effect sparsity" holds.

Lenth (1989) recognized the problem of Type I errors and provided an approximate multiplicity adjustment procedure for detecting effects in saturated designs, using effect sparsity and independence of estimated effects. The assumption of independence fails in the more complex SSDs, where effects are confounded, even in main-effects only models. For these reasons, generic variable selection methods commonly employed in regression, such as forward selection, have been suggested and performed by Lin (1993). The data analyzed by Lin consisted of a half fraction of a Plackett-Burman design. Wang (1995) points out the problem with forward selection analysis of SSDs by replicating Lin's analysis on the other half, and observing that four of the five "important" factors found in one half fraction were not found in the other. The problem of so little agreement between the two analyses could easily be attributable to the multiplicity problem, as we will show.

One appeal of stepwise-type variable selection methods is that nominal Type I error significance levels are used, providing the analyst with some measure of confidence that the variables selected are real. However, if effect sparsity holds in an SSD, we find that Type I error rates can be quite high—easily in the 70% range for $\alpha = .05$ forward selection with 20 or so variables, and nearly certain for $\alpha = .15$ and higher forward selection cutoff values. These error rates are derived using a particular example of an SSD, but can be much worse, depending upon the number of variables investigated. Even more troublesome is that a *large number* of false significances is probable under effect sparsity.

As in Draperet et al. (1971), Aitken (1974), Butler (1982), and Grechanovsky and Pinsker (1995), we consider forward selection methods based on significance tests, and attempt to provide actual significance levels to correct for this problem. The method uses the exact distribution of the maximum forward-selection F statistic to judge significance of a variable to be entered, under the assumption that the entry of the first variables has been pre-selected and not data-steered. This distribution is virtually intractable when there are more variables than data points, as occurs in SSDs, but its significance levels can be estimated consistently and efficiently using a Bonferroni control variates resampling method. Draper et al. (1971), obtained this distribution for uncorrelated predictors; Forsythe et al. (1972), describe a variation of the method based on permutation tests; Butler (1982, 1984) provides upper and lower bounds for the significance level of an entering variable using this distribution; Miller (1990), p. 89 describes a variation for use when there are more observations than variables; and Grechanovsky and Pinsker (1995) calculate the significance level analytically as the content of multidimensional parallelepipeds.

An attractive feature of this approach is that the significance level is *exact* at the first stage of the analysis (modulo Monte Carlo error, which can be controlled). It is also exact at any stage if the first variables' entry are forced *a priori.* Generally, significance levels at later entry stages are only approximate since the procedure is conditional on the observed ordering of variables entered. Nevertheless, theoretical and simulation results suggest that this method controls overall Type I errors well and conservatively.

Interestingly, the ordinary Bonferroni method (without resampling) provides a very accurate approximation to the true critical level of the multiple significance tests in the designs we consider. The accuracy of the Bonferroni approximation in this instance is perhaps surprising. Since there are perfect linear dependencies among the numerators of the forward-selection F-statistics, as well as dependence induced by correlation among the denominator MSE's, one might expect the Bonferroni approximation to be inaccurate. The fact that the Bonferroni approximation is so accurate in these designs is very useful for the practitioner, since it is easy to implement.

In Section 2 the model is described. The forward selection method and analysis of its error rates are discussed in Section 3. Procedures that adjust for selection of maximal effects are presented in Section 4, and their properties are explored both analytically and via a large simulation study. Examples are discussed in Section 5 and concluding remarks are given in Section 6.

2. The Model

Assume there are n experimental runs and q factors under study, of which k are active. Let $\mathbf{A} = \{i_1, \ldots, i_k\}$ and $\mathbf{N} = \{i_{k+1}, \ldots, i_q\}$ denote indexes of active and inert factors, respectively, so $\mathbf{N} \cup \mathbf{A} = \{1, \ldots, q\} = \mathbf{S}$. If X denotes the $(n \times q)$ design matrix (without intercept column), our model is $Y = \mu\mathbf{1} + X\beta + \epsilon$, where Y is the $(n \times 1)$ observable data vector, μ is the intercept term and $\mathbf{1}$ is an n-vector of 1's, β is a $(q \times 1)$ fixed and unknown vector of factor effects and ϵ is a vector assumed to be distributed as $N_n(0, \sigma^2 I)$. In the multiple hypothesis-testing framework, we have null and alternative pairs $H_j : \beta_j = 0$ and $H_j^c : \beta_j \neq 0$, with H_j true for $j \in \mathbf{N}$, and H_j^c true for $j \in \mathbf{A}$. Under effect sparsity, we assume that k is small relative to q.

3. Forward Selection and Type I Errors

With forward selection, one identifies the maximum F-statistic at successive stages. Let $F_j^{(s)}$ denote the F-statistic for testing H_j at stage s, $s = 1, 2 \ldots$ Sequentially define $j_1 = arg\max_{j\, \in\, \mathbf{S}} F_j^{(1)}$, $j_2 = arg\max_{j\, \in\, \mathbf{S}-\{j_1\}} F_j^{(2)}$, $j_3 = arg\max_{j\, \in\, \mathbf{S}-\{j_1, j_2\}} F_j^{(3)}$, etc., where $F_j^{(s)}{=}RSS(j \,|\, j_1, \ldots, j_{s-1})/\, MSE(j, j_1, \ldots, j_{s-1})$. Letting $F^{(s)} = \max_{j\, \in\, \mathbf{S}-\{j_1, \ldots j_{s-1}\}} F_j^{(s)}$, the forward selection procedure is defined by selecting variables $j_1, \ldots, j_f$, where $F^{(f)} \leq \alpha$ and $F^{(f+1)} > \alpha$. If $F^{(1)} > \alpha$ then no variables are selected.

If effect sparsity holds, then a particular effect is more likely to be inert than active. At particular stages of the selection process it is possible that no remaining effects are active; we find the distribution of the maximal F statistic under this condition. The situation where no effects are active is called the *complete null hypothesis*; when some (but not all) effects are inactive we have a *partial null hypothesis.*

Lin (1993) gives as an example a SSD consisting of a half fraction of the 28-run Plackett and Burman design used by Williams (1968). Since columns 13 and 16 are identical, column 16 was removed, leaving 23 columns. Using the resulting $X_{14\times 23}$ design matrix, the complete null hypothesis may be created by associating $Y^* \sim N_{14}(0, I)$ with the given X. Letting V denote the number of variables selected in a given sample, Figure 1 displays estimates of $P(V \geq k)$, for $k = 0, \ldots, 6$ based on 10,000 simulations, for forward selection using $\alpha = .05$, .10, and .15. These are fairly conservative entry levels—Kennedy and Bancroft (1971) recommend $\alpha = .25$, and some software packages use $\alpha = .50$ as a default. Clearly, there is a high probability that *one or more* inert effects are declared significant; and for $\alpha = .15$, *many* inert effects will be declared significant. With $\alpha = .5$ (not shown in the graph), *all* 10,000 samples had 6 or more selected variables!

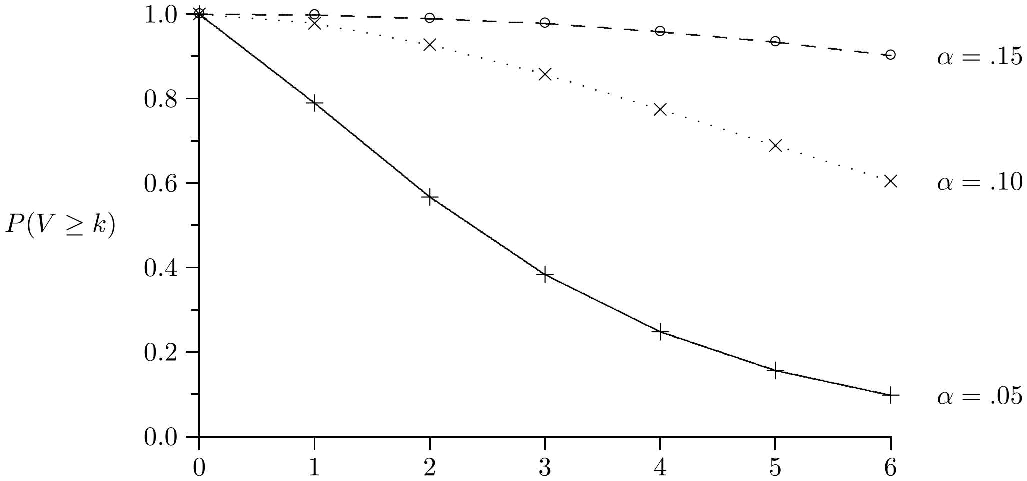

Figure 1. Probability that the number of selected variables (V) is at least k, for $\alpha = .05$, .10, or .15.

The *number* of false significances can be very large for this design for two reasons: (i) the F-statistics are *maxima* at each step, implying that the ordinary F-critical values are much too liberal; and (ii) the denominator estimates of the residual variance are biased downward, causing larger F-statistics (Copas and Long (1991)). Thus, a *large* number of selected variables need not indicate that *any* effects are active.

4. A Resampling Method to Control Type I Errors

The alternative forward selection procedure is developed in this section. We attempt to control the Type I errors at each stage of the forward selection process. The distribution of the maximal F statistic (conditional on a forced entry of the first selected variables) is invariant to all parameters, and can be calculated. In particular, the distribution is invariant to the parameters associated with the first selected variables. According to the simulations and analytical work of this section, we find that the method controls Type I errors conservatively, despite the fact that it attempts to control Type I errors only at each stage of the selection process.

4.1. Adjusted p-values

Suppose all H_j are true. Since the statistics $F_j^{(1)}$ are invariant to location (μ) and scale (σ^2), the distribution of $F^{(1)}$ is completely determined by the known design X, and is therefore known in principle. Letting $f_\alpha^{(1)}$ denote the upper $1-\alpha$ quantile of this distribution, the rule $F^{(1)} > f_\alpha^{(1)}$ defines an *exact* α-level test of the complete null hypothesis. Letting $f^{(1)}$ denote the observed value of $F^{(1)}$, the adjusted p-value is given by

$$p^{(1)} = P(F^{(1)} > f^{(1)} \,|\, \text{all } H_i \text{ true}). \tag{1}$$

At later stages, we calculate the adjusted p-values as if the order of entry of the first $(s-1)$ variables are forced. Thus, $p^{(s)} = P(F^{(s)} > f^{(s)} \,|\, \text{all } H_i,\ i \in \mathbf{S} - \{j_1, \ldots, j_{(s-1)}\}$ are true). Perfect linear relationships involving columns of the X matrix induce complex dependence structures among the numerators of the F_j statistics; thus, the distribution of $F^{(1)}$ appears intractable, and Monte Carlo methods must be used to estimate (1). What follows is a description of a control variate method for doing so.

To simplify notation, the method is described for calculating $p^{(1)}$, but an identical method is used for each $p^{(s)}$, with the exception that the maximal F_i is computed for all $i \in \mathbf{S} - \{j_1, \ldots, j_{(s-1)}\}$. Since the complete null distribution of $F^{(1)}$ is invariant to location and scale, random variables $F^{*(1)}$ having its distribution may be simulated by generating $Y^* \sim N_n(0, I)$, computing the statistics $F_j^{*(1)}$ from Y^* and X, and letting $F^{*(1)} = \max_{j \in \mathbf{S}} F_j^{*(1)}$. While the so-called "uniform resampling" method works directly with the $F^{*(1)}$, the control variate method approximates the p-value using the Bonferroni inequality, then uses simulation to estimate the remainder. The adjusted p-value is then the sum of the analytic Bonferroni estimate and the Monte Carlo estimate of the remainder. Related versions of this procedure are discussed in Heyse and Rom (1988) and Naiman and Wynn (1992).

Letting $\delta(\cdot)$ denote the indicator function, write $p^{(1)} = E\{\sum_{j\in\mathbf{S}} \delta(F_j^{(1)} > f^{(1)})\} - E\{\sum_{j\in\mathbf{S}} \delta(F_j^{(1)} > f^{(1)}) - \delta(\max_{j\in\mathbf{S}} F_j^{(1)} > f^{(1)})\}$. Noting that each $F_j^{(1)}$ is distributed as $F_{1,n-2}$, the first summand is easily obtained as $E\{\sum_{j\in\mathbf{S}} \delta(F_j^{(1)} > f^{(1)})\} = \sum_{j\in\mathbf{S}} E\{\delta(F_j^{(1)} > f^{(1)})\} = q\mathrm{P}(F_{1,n-2} > f^{(1)})$, the Bonferroni approximation. We then estimate $E\{\sum_{j\in\mathbf{S}} \delta(F_j^{(1)} > f^{(1)}) - \delta(\max_{j\in\mathbf{S}} F_j^{(1)} > f^{(1)})\}$ via Monte Carlo and subtract the result from $q\mathrm{P}(F_{1,n-2} > f^{(1)})$. Specifically, generate $Y^* \sim N_n(0, I)$ as with uniform resampling, compute the difference $\Delta^* = \sum_{j\in\mathbf{S}} \delta(F_j^{*(1)} > f^{(1)}) - \delta(\max_{j\in\mathbf{S}} F_j^{*(1)} > f^{(1)})$, and average the values Δ^* over many (say M) resampled data sets. The estimated p-value is then $q\mathrm{P}(F_{1,n-2} > f^{(1)}) - \bar{\Delta}_M^*$, and the standard error of the estimate is $s.d.(\Delta^*)/M^{1/2}$, where $s.d.(\Delta^*)$ is the standard deviation of the M Δ^*'s.

We are most interested in estimating $p^{(1)}$ when $f^{(1)}$ is relatively large; this is also the case where the Bonferroni approximation is best (Miller (1977)). In this case, the standard error of the Bonferroni control variate method is much smaller than that of the uniform resampling method, as described in Westfall (1997).

Using the $X_{14\times 23}$ design matrix of Lin (1993) and $Y^* \sim N_{14}(0, I)$, Table 1 compares the unadjusted, Bonferroni-adjusted, and Control-Variate estimated adjusted p-values for the first selected variable. Note that the Bonferroni approximation is quite accurate in the tail of the distribution; Draper et al. (1971) also noticed this for the forward selection statistic with uncorrelated predictors. However, it is surprising that the Bonferroni multiplier $k = 23$ remains appropriate for large f despite the perfect linear dependencies among the contrasts defining the numerators of the F statistics. One might think that the multiplier would be only 14, since this is the maximum number of linearly independent columns defining the contrasts. This result is yet another caution in the use of SSDs: with a larger number of factors, determining significance becomes increasingly difficult. Incorporating the perfect linear dependencies into the multiplicity adjustment does not improve matters when f is large, since the Bonferroni adjustment (which ignores correlations) is an adequate approximation.

On the other hand, variable selection procedures are often recommended with relaxed variable entry significance levels such as $\alpha = .50$ instead of $\alpha = .05$. If this criterion is applied to the adjusted p-value, Table 1 shows potentially more significances can be attained when the properly adjusted p-values are used, since the Bonferroni adjusted p-values are larger than the Control Variate adjusted p-values.

Table 1. Comparison of unadjusted p-values with Bonferroni and Control Variate (CV) resampling-based ($M = 50,000$) estimates of $p^{(1)} = P(F^{(1)} > f)$, with $k = 23$.

	Unadjusted	Adjusted	
f	$P(F_{1,12} > f)$	Bonferroni	CV
4.0	0.068655	1.579065	0.90337
5.0	0.045115	1.037653	0.74893
6.0	0.030622	0.704301	0.58220
7.0	0.021346	0.490955	0.43950
8.0	0.015220	0.350062	0.32772
9.0	0.011067	0.254534	0.24413
10.0	0.008186	0.188282	0.18358
11.0	0.006149	0.141415	0.13916
12.0	0.004682	0.107677	0.10628
13.0	0.003609	0.083006	0.08239
14.0	0.002813	0.064709	0.06421
15.0	0.002216	0.050965	0.05069

4.2. An alternative stepwise method and error rate control

Type I errors indicated by Figure 1 may be controlled using the adjusted p-values. Algorithmically, at step j, if $p^{(j)} > \alpha$, then stop; otherwise, enter X_j and continue. This procedure controls the Type I error rate *exactly* at level α under the complete null hypothesis (the "weak" sense described by Hochberg and Tamhane (1987), p. 3), since $P(\text{Reject at least one } H_i \,|\, \text{all } H_i \text{ true}) = P(F^{(1)} \geq f_\alpha^{(1)}) = \alpha$. In addition, if the first s variables are *forced*, and the test is used to evaluate the significance of the next entering variable (of the remaining $q - s$), the procedure is again exact under the hypothesis of no effects among the $q - s$ remaining variables. The exactness disappears with simulated p-values, but the errors can be made very small.

What can be said about control of Type I errors under partial nulls? This is a difficult problem, but one can argue heuristically that the procedure should be conservative in most cases. Ignore for the moment the randomness in the variable entry, and assume that the first s variables with indices $j_1, \ldots, j_s$ are forced to enter. Assume also that the set of null effects, $\mathbf{N}$, is a subset of the remaining indices $\mathbf{R}_s = \mathbf{S} - \{j_1, \ldots, j_s\}$. For a false significance to occur at this stage, it is necessary and sufficient that $\max_{j\in\mathbf{N}} F_j^{(s+1)} \geq f_\alpha^{(s+1)}$ $\{arg \max_{j\in\mathbf{S}-\{j_1,\ldots,j_s\}} F_j^{(s+1)}\} \in \mathbf{N}$, where $f_\alpha^{(s+1)}$ is the $1 - \alpha$ quantile of the distribution of $\max_{j\in\mathbf{R}_s} F_j^{(s+1)}$ under the complete null hypothesis that all effects in $\mathbf{R}_s$ are null. Thus, the probability of incorrectly declaring a variable significant at step s is bounded above by

$$P\Big(\{\max_{j\in\mathbf{N}} F_j^{(s+1)} \geq f_\alpha^{(s+1)}\}\Big). \tag{2}$$

The distribution of each $F_j^{(s+1)}$ is the doubly non-central $F_{1,n-s-2,\delta_1,\delta_2}$, where

$$\delta_1 = \|(P_{j,s} - P_s)X_{\mathbf{R}_s(-j)}\beta_{\mathbf{R}_s(-j)} + (P_{j,s} - P_s)X_j\beta_j\|^2 \tag{3}$$

and

$$\delta_2 = \|(I - P_{j,s})X_{\mathbf{R}_s(-j)}\beta_{\mathbf{R}_s(-j)}\|^2. \tag{4}$$

The notation in (3) and (4) follows: $P_{j,s}$ is the projection matrix for the column space of $(\mathbf{1} : X_{j_1} : \cdots : X_{j_s} : X_j)$; P_s is the projection matrix for the column space of $(\mathbf{1} : X_{j_1} : \cdots : X_{j_s})$; $X_{\mathbf{R}_s(-j)}$ is the design matrix whose rows are comprised by indices in $\mathbf{R}_s - \{\mathbf{N} \cup \{j\}\}$; and $\beta_{\mathbf{R}_s(-j)}$ is the vector comprising all non-null effects, removing j, not yet selected.

Note that the terms involving β_j drop out in the case H_j is true. What follows is an heuristic argument (not a proof) that the probability in (2) is less than α in many situations, as will be shown in the simulation studies.

Point 1. In (2) the maximum of the $F_j^{(s+1)}$ is considered only over $j \in \mathbf{N}$, and not over the entire set of remaining indices $j \in \mathbf{R}_s$. When the complete null is true, all F-statistics would be central, and the probability in (2) is bounded above by α.

Point 2. When the complete null is not true, the noncentrality parameters change the situation. If the numerator noncentrality parameters were zero, then the noncentral denominators would make the F-statistics stochastically smaller, again suggesting that the probability in (2) should be less than α.

Point 3. The only problem is that the numerators are also noncentral, acting stochastically in the opposite direction as described in Points 1 and 2. However, we note that in the case of saturated and supersaturated designs, the design vectors are chosen to be orthogonal (for saturated) or nearly orthogonal (for supersaturated), so the numerator noncentrality vector $(P_{j,s}-P_s)X_{\mathbf{R}_s(-j)}\beta_{\mathbf{R}_s(-j)}$ should be "small", in some sense, relative to the denominator noncentrality vector $(I - P_{j,s})X_{\mathbf{R}_s(-j)}\beta_{\mathbf{R}_s(-j)}$. Specifically, assuming the appropriate inverses exist,

$$\delta_1/\delta_2 = \frac{\beta'_{\mathbf{R}_s(-j)}X'_{\mathbf{R}_s(-j)}(I - P_s)P_{j,s}(I - P_s)X_{\mathbf{R}_s(-j)}\beta_{\mathbf{R}_s(-j)}}{\beta'_{\mathbf{R}_s(-j)}X'_{\mathbf{R}_s(-j)}(I - P_s)(I - P_{j,s})(I - P_s)X_{\mathbf{R}_s(-j)}\beta_{\mathbf{R}_s(-j)}} \leq \frac{\rho}{1-\rho},$$

where $\rho = \lambda_1\{(X'_{\mathbf{R}_s(-j)}(I-P_s)X_{\mathbf{R}_s(-j)})^{-1}X'_{\mathbf{R}_s(-j)}(I-P_s)X_{j,s}(X'_{j,s}X_{j,s})^{-1}X'_{j,s}(I-P_s)X_{\mathbf{R}_s(-j)}\}$, which is the maximal canonical correlation between the selected variables (those in $X_{j,s}$) and the residualized remaining active variables $(I - P_s)X_{\mathbf{R}_s(-j)}$. This is in turn bounded (more crudely) by the maximal canonical correlation between the selected variables and the non-residualized remaining active variables. This bound illustrates one important reason to have the columns of SSDs as orthogonal as possible: false significances might arise due to confounding alone, as suggested by the numerator noncentrality parameter of the F statistic. A good SSD from this perspective will require the projection matrix P_s to be as orthogonal as possible, for all possible projections.

These arguments suggest, but do not prove that Type I error rates are less than nominal levels: it may be possible to select parameters to maximize the canonical correlation referred to above, so that inert effects have enough noncentrality to be selected often. Additionally, the arguments do not account for the randomness of the variable entry; Berk (1978) notes that F-values at later steps of the variable selection procedure have a downward bias since important dimensions have been "selected out" at earlier stages, which provides further rationale for the conservativeness of the procedure from the familywise Type I error standpoint.

The effect of the term $(P_{j,s} - P_s)X_{\mathbf{R}_s(-j)}\beta_{\mathbf{R}_s(-j)}$ in the numerator noncentrality δ_1 can be to increase or decrease the Type II error probability. With an orthogonal array, the term is identically zero. SSDs are not orthogonal, and the effects of this nonorthogonality on Type II error probabilities are unpredictable in practice, depending upon the actual values of the parameters in $\beta_{\mathbf{R}_s(-j)}$. When the error variance is known, we find that the power can be increased dramatically, but the Type I errors also can be dramatically inflated since there is no denominator noncentrality to counterbalance the numerator noncentrality.

4.3. Simulation and analytical studies

The design used for our simulation study is again that of Lin (1993), p. 30. Data were simulated under various null and alternative configurations by generating random Y-vectors and associating them with the given design matrix. The simulation required two loops, an outer loop, where the data were generated, the ordinary forward selection procedures were used, and the summary statistics were tabulated; and an inner loop, where the resampling estimates were tabulated and passed to the outer loop. For the forward selection and Bonferroni probabilities, the simulation size was taken to be 10,000, insuring a maximum 95% margin of error less than ±.01. The control variate entries require more extensive simulations: the outer loop was chosen to be $NSIM = 800$ and the simulation size for the inner loop was $NRES = 400$. Results from Westfall and Young (1993), p. 41 suggest that the inner loop should be made smaller than the outer loop, that the margin of error from using 400 instead of infinite resamples is ±1/401, and that the maximum 95% margin error from using 800 outer loop simulations is $\pm 1.96\{(.5)(.5)/800\}^{1/2}$, yielding a maximum 95% margin of error for the control variate simulations of ±.037; the value is much smaller for proportions that are near 1.0 and 0.0. In practice, there would be no "outer loop", and one would take $NRES$ much larger than 400, say 10,000 or more. Further, one would examine the simulation standard error to determine whether the sample size is adequate.

The methods compared are (i) ordinary forward selection, (ii) Bonferroni forward selection, and (iii) resampling (using control variates). Consistent with the notion of "effect sparsity", the number of active effects considered are 0,1,. . . ,5. In all cases the effects are assumed to have the same size, $\beta/\sigma = 5.0$. Due to the confounding structure in the SSD, the sign of the effect can be very important in determining the operating characteristics of the method. Thus the signs of the effects are allowed to be either + or −.

To simplify matters, it is assumed that the active effects occur only for the first five variables. Given the near symmetry of the SSD, one would expect that the results would be similar for most other locations for active effects. Note that in this particular SSD, the correlations for any specific column to all others has 20 ±1/7's and 3 ±3/7's. The correlations among first five columns, have all ±1/7, excepting columns (1,2), (2,4) and (4,5). Thus we have covered columns with 0, 1, and 2 ± 3/7 (Columns 3, 1, 2 respectively). Column 5 has one 3/7 correlation pair, and thus represents the typical situation.

When "−" signs were used, they were assumed to occur only in the last active variables. For example, in Table 2, the notation "3 + +−" indicates $(\beta_1, \beta_2, \beta_3) = (5, 5, -5)$. For each parameter configuration, the FWE, the probability of declaring at least one effect significant, and the probability of declaring all effects significant are tabulated. Forward selection is used in all cases, with $\alpha = .05$, .15, and .50 entry criteria.

Table 2 displays the results. While the FWE of the forward selection procedure is uniformly high, the Bonferroni and resampling estimates control the FWE generally at or below the nominal level. This suggests that when a variable is declared significant by either of these methods, it is "honestly significant", and not an artifact of excessive testing.

One could argue against multiplicity-adjusted procedures on the basis of increased Type II error probabilities. In some cases, the adjusted procedures have virtually no power to detect *any* effect, whereas ordinary forward selection has high power (e.g., the 2 + − configuration). However, with ordinary forward selection, the error rates are so completely uncontrolled that one cannot expect selected effects to be real. At least with the resampling method, one is able to state a prescribed significance level. The price of familywise Type I error control is the reduced power shown in Table 2.

Table 2. Familywise Type I error rates and power functions.

Number of	$\alpha = .05$			$\alpha = .15$			$\alpha = .50$		
Effects	FS	BON	RES	FS	BON	RES	FS	BON	RES
				Familywise Error Rate					
0	0.80	0.05	0.04	1.00	0.15	0.15	1.00	0.46	0.52
1	0.77	0.05	0.04	1.00	0.15	0.16	1.00	0.46	0.51
2	0.76	0.05	0.04	1.00	0.15	0.15	1.00	0.46	0.50
2+−	0.86	0.00	0.00	1.00	0.00	0.00	1.00	0.11	0.13
3	0.73	0.00	0.00	1.00	0.01	0.01	1.00	0.45	0.52
3+ + −	0.73	0.00	0.00	1.00	0.02	0.00	1.00	0.44	0.52
4	0.71	0.00	0.00	1.00	0.00	0.00	1.00	0.28	0.32
4+ + +−	0.71	0.00	0.00	0.99	0.00	0.00	1.00	0.20	0.21
4+ + −−	0.76	0.00	0.00	1.00	0.00	0.00	1.00	0.22	0.28
5	0.25	0.00	0.00	1.00	0.00	0.00	1.00	0.12	0.13
5+ + + + −	0.96	0.00	0.00	1.00	0.00	0.00	1.00	0.27	0.34
5+ + + − −	0.51	0.00	0.00	1.00	0.00	0.00	1.00	0.43	0.44
				Power for at least one effect					
1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2+−	0.89	0.00	0.00	0.98	0.00	0.00	0.98	0.09	0.14
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3+ + −	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4+ + +−	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.92	0.95
4+ + −−	1.00	0.03	0.03	1.00	0.90	0.90	1.00	1.00	1.00
5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5+ + + + −	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5+ + + − −	1.00	0.54	0.52	1.00	1.00	1.00	1.00	1.00	1.00
				Power for all effects					
2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2+−	0.89	0.00	0.00	0.98	0.00	0.00	0.98	0.07	0.14
3	1.00	0.00	0.00	1.00	0.07	0.07	1.00	0.92	0.99
3+ + −	1.00	0.00	0.00	1.00	0.07	0.07	1.00	0.90	1.00
4	0.99	0.00	0.00	0.99	0.00	0.00	0.99	0.38	0.64
4+ + +−	0.87	0.00	0.00	0.97	0.00	0.00	0.97	0.17	0.31
4+ + −−	0.78	0.00	0.00	0.98	0.00	0.00	0.98	0.19	0.28
5	0.01	0.00	0.00	0.29	0.00	0.00	0.29	0.00	0.00
5+ + + + −	0.49	0.00	0.00	0.73	0.00	0.00	0.73	0.04	0.07
5+ + + − −	0.08	0.00	0.00	0.64	0.00	0.00	0.65	0.01	0.02

The assumption that some effects are active and other are completely inactive is artificial. Rather, it is likely that some effects are important, and others are relatively less important. This situation may be simulated by generating active effects exactly as in Table 2, but by generating the remaining effects independently as $N(0,1)$. A $N(0,1)$ variable can rarely be expected to exceed 3.0, and the active effects are always 5.0, so it is reasonable to call the remaining effects "inactive". This model is related to the Bayesian formulation of Box and Meyer (1986), who assumed that the all parameters were normally distributed, with a smaller variance for the inactive factors. These simulations (not shown) showed similar control of Type I errors for the Control Variate and Bonferroni adjustments, excessive Type I errors for the FS method, and generally lower power for all methods.

The unpredictable effect of nonorthogonality on Type II error probabilities is seen by comparing the "$2 + -$" entries ($\beta_1/\sigma = 5$, $\beta_2/\sigma = -5$) with the "2" entries ($\beta_1/\sigma = \beta_2/\sigma = 5$) in Table 2. For "$2 + -$", $\delta_1/\delta_2 = 4.80$, while under "2", $\delta_1/\delta_2 = 30.00$, for the first variable entered (either X_1 or X_2). Noting that $F = \delta_1/\delta_2$ when $\sigma = 0$, this disparity explains the large differences in power for these configurations. For situation "$2 + -$", the first-entering variable is *barely significant*, even *before* multiplicity adjustment, in cases where there is *no* experimental error! The "2" case is "highly significant" in this case, even after multiplicity adjustment: the Bonferroni adjustment is $23 \times P(F_{1,12} > 30.00) = .0033$. The disparity between the "$2 + -$" and "2" cases is attributable to the fairly large correlation (.43) between X_1 and X_2, which is the maximal correlation in the given SSD, shared by many variable pairs.

To further examine the effects of numerator noncentrality, we considered the case where the variance is known (in this example, $\sigma = 1.0$). The algorithm then changes by substituting 1.0 for the denominator MSE in all cases, and changing the F probabilities to χ_1^2. In this case there is no denominator noncentrality, and we can isolate the effects of numerator noncentrality. For the $2 + -$ parameter configuration of Table 2, we find familywise error rates of .50, .53, and .63 for Bonferroni forward selection at $\alpha = .05$, .15, and .50 respectively, using 10,000 simulations. Examination of the simulations resulting in Type I errors showed that the null effects that are confounded with the non-null effects were entered at one of the first two steps. In the unknown variance case represented by Table 2, there was sufficient denominator noncentrality to prevent these occurrences. On the other hand, power is predictably higher when variances are known: the power for detecting both effects in this simulation was identically .98 for $\alpha = .05$, .15, and .50.

5. Applications

Table 3 displays the analysis of the actual response data (Lin (1993), p. 30). Simulated values are estimated using $NSIM = 200,000$ resampled data sets. A

SAS/IML$^{\circledR}$ file was used for this analysis, which took approximately 3.0 hours on a 2100 Alpha VAX EV-5 computer. The probability of observing a maximum F-statistic as large as 20.5859 at the first step is estimated as 0.0155662, highly significant. We therefore claim that X_{15} has a real effect, and continue. At the second stage, the probability of observing an F-statistic as large as 4.5883, in models including only the variable X_{15}, is estimated as 0.816161, which is insignificant at any reasonable level. Our analysis of this data will stop at this point, declaring X_{15} to be the only significant variable.

Table 3. Forward selection results for Williams (1968) data. Monte Carlo standard errors (based on 200,000 simulations) in parentheses.

				Adjusted p-Value	
			Unadjusted		
Step	Variable	F	p-value	Bonferroni	CV
1	15	20.5859	0.000681	0.015667	0.015662 (0.000005)
2	12	4.5883	0.055410	1.219016	0.816161 (0.001328)
3	20	10.0744	0.009920	0.208313	0.200448 (0.000199)
4	4	16.7527	0.002705	0.054097	0.053782 (0.000040)
5	10	5.4188	0.048325	0.918169	0.691004 (0.001041)
6	11	7.1906	0.031469	0.566449	0.486729 (0.000635)

Note that variables 20 and 4 appear marginally significant (adjusted p-levels of 0.200448 and 0.053782, respectively); however, if the second stage variable X_{12} is a Type I error, then the denominator of the F-statistics at these later stages is biased low, resulting in inflated F-statistics. Thus it is reasonable to stop the selection process at stage 2. The remaining steps are shown to indicate the adequacy of the Bonferroni approximation, which requires no simulation.

The second example is a cast fatigue experiment analyzed by Hamada and Wu (1992), with data reported in their Table 2, p. 132. There were seven factors, labelled A–F, in an orthogonal array with 12 runs. Hamada and Wu considered an analysis strategy to identify significant main effects and interactions. We re-analyze their data, with special attention to the multiplicity problem. Using all seven main effects and $7!/(5!2!) = 21$ interactions, we create a single SSD with $q = 28$ factors and $n = 12$ observations. Table 4 displays the results of the forward selection process using multiplicity-adjusted p-values, as described in Section 4.2, using 10,000 samples. Using a FWE = .50 entry-level criterion, variables FG, F, and AE enter. Hamada and Wu found similar results without multiplicity adjustments, but the AE term would not have been considered using their "effect heredity" principle.

Table 4. Forward selection results for cast fatigue data. Monte Carlo standard errors (based on 10,000 simulations) in parentheses.

			Unadjusted	Adjusted p-Value	
Step	Variable	F	p-value	Bonferroni	CV
1	FG	8.0963	0.017387	0.486825	0.440825 (0.002138)
2	F	37.2770	0.000178	0.004808	0.004808 (0.000000)†
3	AE	10.1568	0.012862	0.334409	0.320209 (0.001192)
4	EF	3.5719	0.100684	2.517090	0.986190 (0.009815)

† - $\Delta^* \equiv 0$ for all 10,000 samples.

6. Concluding Remarks

The general message is that identification of significant variables in SSDs is very tricky. *Many* Type I and Type II errors are expected using forward variable selection. Type I errors can be alleviated by using adjusted p-values, at the expense of increasing Type II errors. If Type I errors are considered important, then we recommend using adjusted p-values, with an entry criterion no higher than FWE = .50. The justification for recommending an FWE of .50 is that the procedure is conservative, and we would like to compensate for this conservativeness by allowing more terms in the model. The value .50 is the largest value of a FWE that seems reasonable: with this value, our results suggest that the probability of claiming that an inactive effect is significant is no more than .50. Any value larger than .50 would imply that we expect (or, it is more likely than not) that some of the effects that we declare as active will in fact be inactive. Further, while the simulations should be carried out on a wider range of designs, our limited simulation study (shown in Table 2) suggests that the true FWE is likely to be much less than .50 in practice.

It is wise to combine several data analysis methods to evaluate the inferences. The proposed methods of this article should be included in the analyst's toolkit, particularly to highlight the potential pitfalls of Type I errors. At the very least, we recommend that ordinary forward selection p-values and adjusted p-values be displayed side-by-side in the analysis, to help the investigator gauge the likelihood that a given effect is real. The assumption of effect sparsity that is implicit in the experimenter's choice of an SSD implies that Type I errors will be likely, should ordinary forward selection be used. When using forward selection with multiplicity-adjusted p-values, the experimenter can be reasonably confident (with prescribed familywise Type I error level) that effects declared significant using this approach are not multiple testing artifacts.

It is surprising how well the Bonferroni adjustment compares to the actual forward-selection adjusted p-values for the SSDs we considered. While we recommend the use of actual adjusted levels formally, we can recommend the informal use of the Bonferroni adjusted p-values as crude but effective upper bounds, for those analysts who are concerned with the possibility of Type I errors in the analysis of data arising from SSDs. The fact the appropriate multiplicity adjustment acts like the Bonferroni multiplier reinforces the fact that it is (and should be) difficult to claim significances with SSDs.

Statistics deals with managing uncertainty, but one might question whether control of FWE is appropriate for the analysis of SSDs. An alternative to controlling the FWE is to control the "false discovery rate", described in an unrelated application by Benjamini and Hochberg (1995). Perhaps this is a useful avenue for further research.

There are several alternatives to forward selection, and it may be possible to derive methods related to those of this paper for FWE control for those methods as well. One may supplement forward stepping with backstepping at each stage to see whether all remaining terms are significant, commonly called stepwise regression. For example, if there are 3 variables selected, we might then test whether all of the selected variables remain significant at the $\alpha/3$ level (assuming that the Bonferroni correction is reasonable; and all indications of the present paper suggest that it is). Now, the three variables that have had to enter the model required significance at much more stringent levels, α/k, $\alpha/(k-1)$, and $\alpha/(k-2)$. This fact, coupled with the fact that the collinearity is minor by nature of SSD construction, makes it unlikely that any of the variables will become insignificant at the backstep operation. To check this conjecture, we performed a simulation of the results of the forward selection using the "$5+++--$" parameter configuration of Table 2, and noted at the final step whether any parameters became insignificant. Out of 10,000 simulations, only 5 times was a variable found insignificant at the final step. Thus, it appears reasonable to restrict attention to forward selection only, and not consider backstepping, when performing Bonferroni-corrected adjustments with SSDs.

Another possibility, suggested by a referee, is to perform all-subsets style regression, using a criterion such as AIC or its finite-sample variants, to select a subset of variables, then to evaluate the significance of the resulting coefficients while allowing for multiplicity and selection effects. One could calculate via resampling the distribution of the minimum p-value for tests of $H_0 : \beta_{i_1} = \cdots = \beta_{i_j} = 0$ over all subsets $\{i_1, \ldots, i_j\}$ having cardinality, say, six or less. (Note, the min p-value must be used here, not maximum F, since the distributions of the F's for different subset sizes are not comparable). The data could be simulated from the complete null hypothesis, and the minimum p-value could be computed from an all-subsets regression procedure. This adjusted p-value then would be a reasonable indicator of whether at least one of the selected β's is non-zero. The procedure seems computationally laborious, particularly since all subsets must be evaluated for each resampled data set. Additionally, the operating characteristics would need to be explored via simulation analysis, which would make the computational aspect virtually impossible. Nevertheless, this could be a promising avenue worth further exploration.

Finally, when choosing a SSD, it is particularly important that correlations be low, since large correlations can dramatically affect the likelihoods of both Type I and Type II errors. Consequently, the popular "$E(s^2)$" criterion which averages all correlations, will make sense for comparing two SSDs only when the two SSDs share the same maximal correlations.

Acknowledgement

The authors thank the reviewers for many suggestions which have been incorporated in this article.

References

Aitken, M. A. (1974). Simultaneous inference and the choice of variable subsets in multiple regression. *Technometrics* **16**, 221-227.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57**, 289-300.

Berk, K. N. (1978). Comparing subset regression procedures. *Technometrics* **20**, 1-6.

Booth, K. H. V. and Cox, D. R. (1962). Some systematic supersaturated designs. *Technometrics* **4**, 489-495.

Box, G. E. P and Meyer, R. D. (1986). An analysis for unreplicated fractional factorials. *Technometrics* **28**, 11-18.

Butler, R. W. (1982). Bounds on the significance attained by the best-fitting regressor variables. *Applied Statistics* **31**, 290-292.

Butler, R. W. (1984). The significance attained by the best-fitting regressor variable. *J. Amer. Statist. Assoc.* **79**, 341-348.

Copas, J. B. and Long, T. (1991). Estimating the residual variance in orthogonal regression with variable selection. *The Statistician* **40**, 51-59.

Draper, N. R., Guttman, I. and Kanemasu, H. (1971). The distribution of certain regression statistics. *Biometrika* **58**, 295-298.

Forsythe, A. B., Engleman, L., Jennrich, R. and May, P. R. A. (1973). A stopping rule for variable selection in multiple regression. *J. Amer. Statist. Assoc.* **68**, 75-77.

Grechanovsky, E. and Pinsker, I. (1995). Conditional p-values for the F-statistic in a forward selection procedure. *Comput. Statist. Data Anal.* **20**, 239-263.

Hamada, M. and Wu, C. F. J. (1992). Analysis of designed experiments with complex aliasing. *J. Quality Technology* **24**, 130-137.

Heyse, J. F. and Rom, D. (1988). Adjusting for multiplicity of statistical tests in the analysis of carcinogenicity studies. *Biometrical J.* **30**, 883-896.

Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures.* John Wiley, New York.

Kennedy, W. J. and Bancroft, T. A. (1971). Model building for prediction in regression based upon repeated significance tests. *Ann. Math. Statist.* **42**, 1273-1284.

Lenth, R. V. (1989). Quick and easy analysis of unreplicated factorials. *Technometrics* **31**, 469-473.

Lin, D. K. J. (1993). A new class of supersaturated designs. *Technometrics* **35**, 28-31.

Lin, D. K. J. (1995). Generating systematic supersaturated designs. *Technometrics* **37**, 213-225.

Miller, A. J. (1990). *Subset Selection in Regression.* Chapman and Hall, New York.

Miller, R. G. (1977). Developments in multiple comparisons 1966-1976. *J. Amer. Statist. Assoc.* **72**, 779-788.

Naiman, D. Q. and Wynn, H. P. (1992). Inclusion-exclusion-Bonferroni identities and inequalities for discrete tube-like problems via Euler characteristics. *Ann. Statist.* **20**, 43-76.

Nguyen, N.-K. (1996). An algorithmic approach to constructing supersaturated designs. *Technometrics* **38**, 69-73.

Srivastava, J. N. (1975). Designs for searching non-negligible effects. In *A Survey of Statistical Design and Linear Models* (Edited by J. N. Srivastava), 507-519. North-Holland, Amsterdam.

Wang, P. C. (1995). Comments on Lin. *Technometrics* **37**, 358-359.

Westfall, P. H. (1997). Multiple testing of general contrasts using logical constraints and correlations. *J. Amer. Statist. Assoc.* **92**, 299-306.

Westfall, P. H. and Young, S. S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-value Adjustment.* John Wiley, New York.

Williams, K. R. (1968). Designed experiments. *Rubber Age* **100**, 65-71.

Wu, C. F. J. (1993). Construction of supersaturated designs through partially aliased interactions. *Biometrika* **80**, 661-669.

Department of Information Systems and Quantitative Sciences, Texas Tech University, Lubbock, TX 79409, U.S.A.

E-mail: westfall@ttu.edu

Glaxo Wellcome Inc., Research Triangle Park, NC 27709, U.S.A.

E-mail: ssy0487@glaxo.com

Department of Management Science and Information Systems, Penn State University, University Park, PA 16802, U.S.A.

E-mail: dkl5@psu.edu

(Received September 1995; accepted June 1997)