Supersaturated Designs: Are Our Results Significant?

David J. Edwards
Department of Statistical Sciences and Operations Research
Virginia Commonwealth University
Richmond, VA 23284
(dedwards7@vcu.edu)

Robert W. Mee
Department of Statistics, Operations, and Management Science
University of Tennessee
Knoxville, TN 37996
(rmee@utk.edu)

Abstract

Two-level supersaturated designs (SSDs) are designs that examine more than $n-1$ factors in n runs. Although SSD literature for both construction and analysis is plentiful, the dearth of actual applications suggests that SSDs are still an unproven tool. Whether using forward selection or all-subsets regression, it is easy to select simple models from SSDs that explain a very large percentage of the total variation. Hence, naive p-values can persuade the user that included factors are indeed active. We propose the use of a global model randomization test in conjunction with all-subsets to more appropriately select candidate models of interest. For settings where the large number of factors makes repeated use of all-subsets expensive, we propose a short-cut approximation for the p-values. Finally, we propose a randomization test for reducing the number of terms in candidate models with small global p-values.

Keywords: Adjusted p-value; All-subsets regression; Effect sparsity; Forward selection; Global model test; Randomization test.

1 Introduction and Motivating Example

The initial stage of experimentation often consists of an experiment involving many factors. If the number of factors is very large and/or experimental runs are very expensive, then even resolution III fractional factorials and strength 2 orthogonal arrays accommodating all the factors become impractical. Two-level supersaturated designs (SSDs) were introduced to handle such situations. Naturally, SSDs have too few runs to support estimating main effects for all the factors, which is a source of ambiguity in any analysis. Consider the following example.

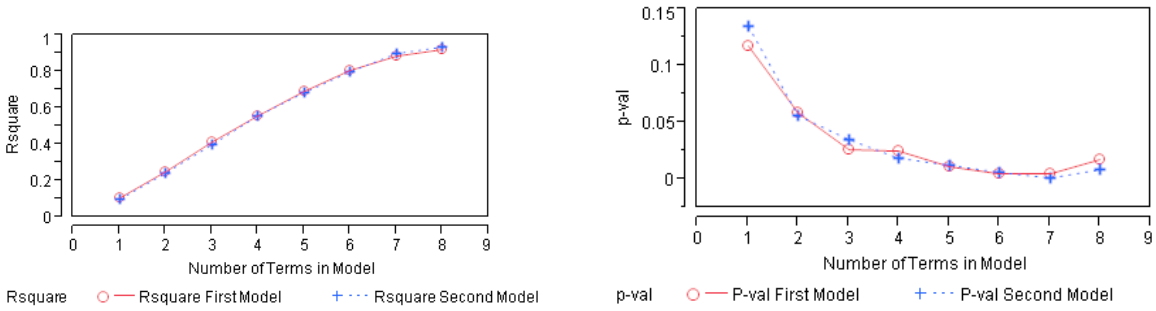
Lin (1995) describes a SSD with 24 runs and 138 factors (denoted as $X_1 - X_{138}$) based on a case study for testing and validating an acquired immune deficiency syndrome (AIDS) model. The response is the AIDS incidence rate per 100,000 persons. The SSD was constructed using a 24-run Hadamard design for $X_1 - X_{23}$, with columns for the remaining 115 factors generated by two-factor interactions. Because the 24 observed incidence rates were highly skewed, with the largest value more than triple all the rest, use of ln(Incidence Rate) or some other transformation would have been advisable.

The success of SSDs depends on having a few dominant and essentially additive effects. That is, a first-order model with just a few terms must explain most of the variation. So which are the main drivers for the AIDS incidence model? Lin (1995) used forward selection to identify a model with 11 factors and $R^2 = 0.99$; the first eight of these factors (with $R^2 = 0.92$) were selected for further study. However, is it true that just (11/138 =) 8% or fewer of the 138 factors affect incidence rate?

It is possible to replicate Lin's results through the first seven steps. However, the eighth factor to enter is X_{71} rather than X_{76} as reported by Lin; see Table 1. Whereas this model has the appearance of being useful, one can exclude the eight factors shown in Table 1, repeat forward selection using the remaining 130 factors, and obtain comparable R^2 and (naive) p-values (see Figure 1). In particular, when 7 or 8 factors are included in the model,

Table 1: Forward Selection for Lin (1995) AIDS Data

Step	Factor	naive p-value	R^2
1	X_{118}	0.1171	0.1079
2	X_{25}	0.0587	0.2505
3	X_{129}	0.0265	0.4176
4	X_{13}	0.0254	0.5554
5	X_{91}	0.0109	0.6928
6	X_{93}	0.0055	0.8072
7	X_{86}	0.0055	0.8825
8	X_{71}	0.0175	0.9204



(a) R^2 (b) Naive p-value

Figure 1: Forward Selection R^2 and Naive p-Value Comparisons

we see higher R^2 and smaller p-values for the second forward selection model. A similar result occurs if we exclude the 16 factors in the first and second models and rerun forward selection with the remaining 122 factors! Clearly we must rely on more than naive p-values in assessing the importance of factors in these model.

Lin (1995) ran a confirmatory 2^{8-4} experiment for the first seven factors in Table 1 plus X_{76}. As a result, factors X_{13}, X_{118}, and X_{129} (with coefficients b_{13}, b_{118}, and b_{129}, respectively) are declared statistically significant. However, b_{13} changed sign, being positive in the forward selection model but negative for the model fitted to the confirmation data, and b_{118} is $1/100^{th}$ of its former estimate. Certainly, one cannot hope to learn much from such SSD data unless a very small subset of the factors truly explains most of the variation.

Westfall et al. (1997) point out that nominal p-values offered by software provide analysts with a misleading sense of confidence about the effects in the forward selection regression model being real. To control Type I error rates, they recommend Monte Carlo simulation using standard normal responses to determine valid p-values for each step. More conveniently, they found that the Bonferroni inequality approximates well these adjusted p-values; one simply has to multiply each nominal p-value by the number of eligible variables at that step. Thus, the Bonferroni adjusted p-values for Table 1 are 138(.1171), 137(.0587), etc. Since the first p-value is 1, by Westfall et al. (1997)'s approach, no terms are added regardless the choice of α. One difficulty with forward selection when we actually control the risk of overfitting is a substantial loss of power due to premature termination. That is, controlling the risk of Type I error for SSDs fit using forward selection leads to very sparse models.

Abraham et al. (1999) highlight forward selection's propensity to miss the real active factors and to select inactive ones instead due to the bias of regression estimates in simple models caused by correlations among the factor columns. Since the aliasing structure inherent in a SSD can hide real effects or encourage identifying nonactive effects as active, it is common for forward selection to be led astray by the entry of a nonactive effect. Furthermore, it is difficult or impossible for the forward selection procedure to recover from such errors. Abraham et al. (1999) utilized eight different supersaturated subsets of the data of Williams (1968), a 28-run Plackett-Burman design in 23 factors, to illustrate this deficiency and found that all-subsets faired somewhat better.

Since models obtained by forward selection's early steps are biased by omitted active factors (see Miller (2002), Chapter 6), we concur with Abraham et al. (1999) that all-subsets provides a better means to obtain candidate models. Table 2 provides models found using all-subsets for ln(Incidence rate) for up to $m = 6$ terms in the model. Running one all-subsets regression for $m = 6$ took approximately 7 hours to perform and is estimated

Table 2: All-Subsets for Lin (1995) AIDS Data, Y = ln(Incidence rate)

Model Size	Subset						R^2	Global p-Value*
1	118						0.366	0.164 (0.003)
1	63						0.295	0.534 (0.004)
1	39						0.274	0.675 (0.003)
2	87	118					0.587	0.158 (0.003)
2	66	118					0.523	0.529 (0.004)
2	74	118					0.519	0.546 (0.004)
3	55	87	118				0.709	0.36 (0.05)
3	6	87	118				0.708	0.39 (0.05)
3	85	87	118				0.691	0.56 (0.05)
4	18	58	63	105			0.811	0.52 (0.05)
4	3	75	85	87			0.809	0.54 (0.05)
4	6	87	118	121			0.801	0.63 (0.05)
5	3	48	75	85	87		0.880	0.61 (0.05)
5	6	63	66	118	130		0.878	0.65 (0.05)
5	18	58	63	105	120		0.876	0.66 (0.05)
6	10	25	87	103	118	133	0.932	> 0.50
6	66	72	77	101	118	121	0.932	> 0.50
6	31	67	87	118	122	123	0.931	> 0.52

*Parentheses indicate standard error, $[\hat{p}(1-\hat{p})/B]^{1/2}$, for estimated p-value. Section 2 explains these p-value computations.

that even one all-subsets for $m = 7$ would take several days. However, since R^2 values above 93% have already been obtained, larger models are not expected to offer anything useful. Similarity of R^2 values among the best models of a given size is generally indicative of models containing factors that are unimportant. This feature is evident for $m \geq 3$.

The only criticism of all-subsets regression is the computational challenge. Beattie et al. (2002) consider all-subsets to be impractical even when a moderate number of factors are active. For instance, a SSD with 23 factors and at most six active factors leads to consideration of 145,498 models, which was deemed to be a formidable comparison by those authors. However, we see that conducting 500 all-subsets regressions for a SSD with 23 factors and at most seven active factors ($\sum_{i=1}^{7} \binom{23}{i}$ = 390,655 models) took less than 5 minutes (i.e. less than 1 second for each all-subsets regression) using SAS or R software

packages on a 1.1GHz Pentium. (As will be described shortly, repetition with all-subsets will be necessary to compute p-values.) Therefore, the ever increasing rise in computing abilities makes all-subsets regression more practical. Kelly and Voelkel (2000) suggested examining all-subsets of effects up to size m, where m is chosen to be at least as large as the maximum number of effects expected. We add the further condition that once models with $R^2 > 0.90$ have been obtained, there is rarely any need to consider larger models.

Bonferroni adjusted p-values provide a simple means of controlling the risk of Type I errors for models obtained by forward selection. How can we control this risk when selecting models using all-subsets? Beginning with Section 2, we answer this question by proposing a permutation-based global test for all-subsets models as a method for evaluating model significance. Realizing that all-subsets can indeed pose considerable computational challenges, Section 3 considers an approximation to the p-values for the global model test introduced in Section 2. Once we determine overall model significance, one should test for the statistical significance of individual terms; in Section 4, we propose a permutation test for reducing the number of terms in candidate models with small global p-values. Section 5 concludes the article with discussion and suggestions for future research.

2 Global Model Test for All-Subsets

The best models determined via all-subsets should be evaluated in terms of a global model test. Permutation tests calculate the probability of getting a test statistic value equal to or more extreme than the observed value under a specified null hypothesis by recalculating the test statistic after random shuffling of the data. Recent work in this area for linear regression models can be found in Anderson and Legendre (1999) and Anderson and Robinson (2001). Anderson (2001) provides a thorough review of permutation test procedures, consolidates recent findings, and provides practical recommendations for practitioners. For a book length treatment, see Manly (1997), among others.

Consider the model,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_q X_q + \varepsilon$$

and suppose we want to test the global null hypothesis $H_0 : \beta_1 = \beta_2 = ... = \beta_q = 0$. An appropriate test statistic, among others, is the usual coefficient of determination (R^2). In order to conduct a permutation test, one needs to consider what is exchangeable under a true null hypothesis (i.e. $Y = \beta_0 + \varepsilon$). Under the assumption that the errors, ε, are i.i.d., the observations are exchangeable, which means that if Y has no relationship with any of the explanatory variables, $X_1, ..., X_q$, then the values obtained for Y could have been observed in any order. Thus, an exact p-value for the above hypothesis test, conditional on the observations, is obtained by fitting every permutation of Y. An estimate $\hat{p}$ is obtained by randomly permuting Y, leaving $X_1, ..., X_q$ fixed, and recalculating R^2 for each of B permutations (denoted by $R^{2(b)}$). That is, calculate

$$\hat{p} = \frac{\#(R^{2(b)} \geq R^2)}{B}, \tag{2.1}$$

where # means 'number of'.

The analysis strategy proposed thus far can be summarized as follows:

1. *All-Subsets Regression*. Perform all-subsets regression and retain the best few models of each size under consideration. The user has considerable freedom in this step with regards to the maximum model size, m, as well as the number of candidate models retained for further exploration.

2. *Global Model Test*. For each model under consideration, perform a test of the global null hypothesis ($H_0 : \beta_1 = \beta_2 = ... = \beta_q = 0$). Any model failing this test need not be examined further. A permutation test for the global null hypothesis of a model with q variables is conducted as follows:

(a) Compute R^2 for a model of interest.

(b) For each of B permutations (shuffling) of the response, Y, perform all-subsets regression for models of size q and select the model with largest R^2. Denote this R^2 as $R^{2(b)}$.

(c) Compute (2.1) and its standard error, $[\hat{p}(1-\hat{p})/B]^{1/2}$. Thus, we estimate the probability (under the null hypothesis) of finding an R^2 greater than that observed for the model fit to the observed data. A small p-value is evidence that one or more terms is accounting for systematic variation in the data.

Once the permutation distribution of $R^{2(b)}$ is obtained for models of a given size, the global p-value for all models of that size are easily estimated. Thus, there is no added computational burden to considering several models of size q rather than just the best one. If possible, we recommend using at least B = 1000 for each size model. However, this choice is not absolute, as we illustrate for the AIDS data.

2.1 Example 1: AIDS Data Global Test p-Values

Estimated p-values (and their standard errors) for the models in Table 2 were obtained as described above. For q = 1 and 2, where the number of possible models is small, we used B = 20,000 and so have small standard errors. For q = 3, 4, and 5, we used B = 100; because these estimated p-values all exceed 0.35, greater precision is not required to identify that these models do no better in explaining ln(Incidence rate) than all-subsets would typically do in explaining random error. For p-values in the range 0.05 to 0.10, which are relevant values for α in the SSD context, the standard errors for the estimated p-value with B = 100 are $[p(1-p)/B]^{1/2}$ = 0.022 and 0.030, respectively. If this does not provide sufficient precision, then either increase B or use the approximation discussed in the next section. For $q = 6$ where even B = 10 repetitions of all-subsets is a challenge, one may adopt a

hybrid strategy of using all subsets for models of size $q_1 < q$; then for the best (50–100) models of size q_1 obtained via permutation of the response, we add $q - q_1$ terms using forward selection. The best R^2 for each permutation serves as a lower bound for the $R^{2(b)}$ that would have been obtained using all-subsets for models of size q. The lower bounds for $q = 6$ in Table 2 were obtained using $q_1 = 5$ and $B = 100$.

The models shown in Table 2 show little evidence for explaining systematic variation in ln(Incidence rate). The single variable model consisting of X_{118} and the two-factor model that adds X_{87} each have a p-value of approximately 0.16. Since these are the smallest p-values for all the models considered, we surmise that the extreme effect sparsity and effect additivity required for SSDs to be informative likely do not hold for the AIDS model.

2.2 Example 2: 5th Fraction of Williams Data

Abraham et al. (1999) analyzed eight different half-fractions of Williams' rubber data (see p. 136). Here we provide global p-values for all-subsets models fitted to their fractions 3 and 5. Lenth's analysis of the full 28 run Plackett-Burman design identifies factor 15 as active, but nothing else. Using additional experimentation, Williams concluded that factors 17 and 20 also had a major effect. Lin (1993) constructed a SSD which corresponds to Abraham et al. (1999)'s fraction 5. Table 3 shows the best three models of each size obtained by all-subsets for up to $m = 7$ terms. Permutation test p-values were obtained using B = 20,000 for models with five or fewer terms and B = 4,000 for larger models. From Table 3's 4th column, we note twelve models that appear "remarkable" and worthy of further study based on a significance level of 10%. Based on our earlier rule, we might have stopped at $m = 4$, since this produced models with $R^2 > 0.9$. Note the similarity of R^2 values among the best models of each size for q = 5, 6, and 7. We will see later whether it was useful to fit these larger models.

Westfall et al. (1997) found p-values by simulating data from a standard normal distri-

bution. One could compute the null distribution of $R^{2(b)}$ by simulating standard normal Y's rather than permuting the data; see the last column of Table 3 for the outcome. When the assumptions of normal theory tests are satisfied, the permutation test and normal theory test agree (Manly (1997)). On the other hand, if the data contains one or more anomalous values, the tests may not agree. Clearly, there are marked differences among the global model p-values based on the permutation test and those based on simulating from a $N(0,1)$, especially for $q = 2$ here. A histogram of the response (not shown) is right skewed with one extreme value, providing some justification for the discepancies. As randomization tests are conditional on the observed data, their use has more relevance than the normal test.

Table 3: All-Subsets and Global Model p-values for Williams' Fraction 5

Model Size	Subset	R^2	$P(R^{2(b)} \geq R^2)$ (Permutation)*	$P(R^{2(b)} \geq R^2)$ (Normal)*
1	15	0.6317	0.013	0.016
1	17	0.3209	0.537	0.643
1	2	0.1202	1.000	1.000
2	12 15	0.7401	0.005	0.075
2	15 20	0.7225	0.011	0.102
2	15 17	0.6942	0.020	0.160
3	12 15 20	0.8705	0.027	0.055
3	4 15 20	0.8192	0.134	0.201
3	12 15 23	0.8120	0.153	0.229
4	4 12 15 20	0.9548	0.011	0.014
4	12 13 15 20	0.9011	0.118	0.227
4	10 12 15 20	0.9004	0.122	0.232
5	4 10 12 15 20	0.9730	0.025	0.044
5	1 4 12 15 20	0.9697	0.036	0.063
5	4 12 15 20 21	0.9688	0.040	0.068
6	4 10 11 12 15 20	0.9867	0.07	0.09
6	4 10 12 15 20 21	0.9826	0.14	0.17
6	1 4 10 12 15 20	0.9817	0.17	0.19
7	4 7 10 11 12 15 20	0.9982	0.01	0.01
7	2 4 5 12 15 20 21	0.9953	0.09	0.11
7	1 4 10 11 12 15 20	0.9935	0.20	0.23

* P-values based on $B = 20{,}000$ permutations for $m = 1$–5 and $B = 4{,}000$ for $m = 6, 7$.

2.3 Example 3: 3rd Fraction of Williams Data

Table 4 shows the results of all-subsets as well as the global model p-values for Abraham et al. (1999)'s 3rd fraction of the Williams data. For this subset, eight models have global p-values that are less than 0.1. Note that none of the models chosen by the global model test in this supersaturated subset are the same as those for Example 2! However every statistically significant model but one in Tables 3 and 4 contains factor 15.

Table 4: All-Subsets and Global Model p-values for Williams' Fraction 3

Model Size	Subset	R^2	Global P-value (Permutation)*
1	15	0.5558	0.012
1	8	0.4021	0.289
1	17	0.1487	0.999
2	15 20	0.7287	0.018
2	8 15	0.6770	0.134
2	15 17	0.6353	0.315
3	5 8 15	0.8544	0.086
3	3 15 20	0.8020	0.318
3	1 8 15	0.8020	0.318
4	1 5 8 15	0.9568	0.016
4	5 8 15 20	0.9166	0.115
4	5 8 12 15	0.8844	0.334
5	1 5 8 15 21	0.9741	0.043
5	1 5 8 12 15	0.9710	0.058
5	1 5 8 15 22	0.9695	0.068
6	1 5 8 11 15 22	0.9851	0.15
6	1 5 8 12 15 21	0.9839	0.17
6	5 6 8 10 18 21	0.9830	0.19
7	1 5 6 8 10 18 21	0.9976	0.03
7	1 4 5 8 11 15 22	0.9945	0.18
7	1 5 8 11 12 15 22	0.9943	0.20

* P-values based on B = 20,000 permutations for m = 1–5 and B = 4,000 for m = 6, 7.

3 Approximating p-Values when Repeating All-Subsets Is Infeasible

Here we provide a simple means of approximating the p-value for cases where q is too large for repetition for all subsets to be feasible. It is well known that the null distribution of R^2 for a particular subset follows a beta distribution (Miller (2002), Chapter 4). Thus, we propose an approximation for the global p-value as

$$\tilde{p} = 1 - P[X(\alpha, \beta) < R^2]^M, \tag{3.1}$$

where $\alpha = q/2$, $\beta = (n-q-1)/2$, $X(\alpha,\beta)$ is a beta random variable, and M is a function of $\binom{k}{q}$, the number of models explored by all-subsets. If the R^2 values for the $\binom{k}{q}$ subsets were independently distributed (and the errors were normally distributed), then M would equal $\binom{k}{q}$. However, we can observe empirically that a smaller value for M is needed.

Using the normal-based distributions for $R^{2(b)}$ from Example 2, we estimated the 50th, 80th, and 90th percentiles of the null distribution for the best all-subsets R^2. Let X_r denote the rth quantile. Then

$$\tilde{M} = \ln(r)/(\ln P[X(\alpha,\beta) < X_r])$$

is the value of M for which the approximation (3.1) matches the true probability for a p-value of $1-r$. Figure 2 shows a plot of $\tilde{M}$ on a logarithmic scale as a function of q. Three facts are evident from this plot. First, the appropriate value for $\tilde{M}$ is less than $\binom{k}{q}$. Second, the value of $\tilde{M}$ is similar for various quantiles, with upper tail probabilities requiring a slightly larger $\tilde{M}$ than for the median. Third, the logarithm of $\tilde{M}$ is very nearly a linear function of q, especially for the median.

The relationship between $\tilde{M}$ and q suggests the following approximation. For small q, it is presumed that one can perform all subsets regression repeatedly and so estimate the

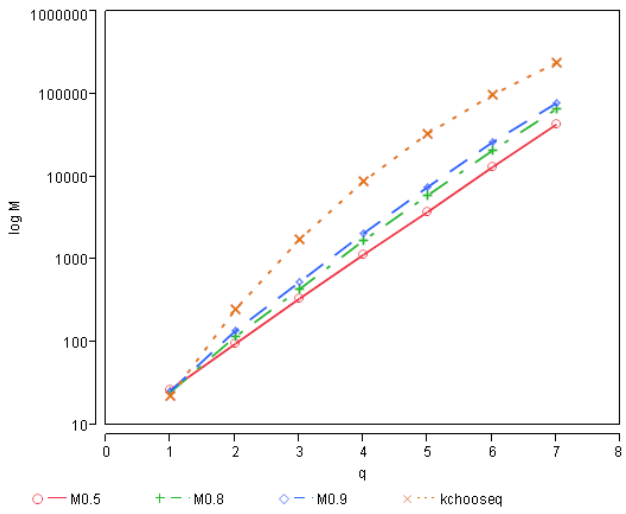


Figure 2: Plot of $\log(\tilde{M})$ vs. q

required quantiles of $R^{2(b)}$. Using these data, a model is fitted for $\ln(\tilde{M})$ as a function of q and fitted values for $\ln(\tilde{M})$ at larger q are obtained. Then (3.1) can be used to approximate p-values where repeating all-subsets is not feasible. We now illustrate this calculation for the AIDS data.

Table 5 shows the medians for $R^{2(b)}$ for q = 1, 2, ..., 5, each based on B = 100. Using these medians from the permutation $R^{2(b)}$ distributions, we obtain $\tilde{M}$. For instance, for q = 1, $\tilde{M}$ = ln(0.5)/ln($P[X(0.5,11) < 0.295]$) = 113.8. We regress $\ln(\tilde{M})$ on q; the linear model

$$\widehat{\ln(\tilde{M})} = 1.787746 + 2.890922q$$

explains 99.975% of the variation in $\ln(\tilde{M})$. Now substitute $q = 6$ and exponentiate to obtain $\tilde{M} = exp(19.1333) = 2.03928 \times 10^7$. The approximate p-value for R^2 = 0.932 is

$$1 - P[X(6/2,17/2) < 0.932]^{20392800} = 0.657.$$

Thus, none of the six-factor models show any evidence of explaining systematic variation.

Table 5: Median $R^{2(b)}$ Values for AIDS Data Based on Permutations of Y = ln(Incidence Rate

q	α	β	$\binom{k}{q}$	median R^2	$\tilde{M}$
1	0.5	11.0	138	0.295	113.8
2	1.0	10.5	9453	0.525	1738.1
3	1.5	10.0	428536	0.699	35578.6
4	2.0	9.5	14463090	0.813	671661.4
5	2.5	9.0	387610812	0.887	10968847
6	3.0	8.5	8592039666		

4 Choosing Factors of Interest

Given a model with a small global test p-value, one would then want to test $H_{0,i} : \beta_i = 0$ for individual $i = 1, 2, ..., q$. In order to handle the multiple testing problem in this situation, individual tests are made more conservative by adjusting naive p-values. Procedures developed to do so are often designed to control the experimentwise error rate (EER).

Bonferroni adjusted p-values are computed by multiplying the naive p-value times the number of eligible factors not in the model, plus one for the factor under consideration. That is, $p_{i,Bonf} = \min((k-q+1)p_i, 1)$. While suitable for models obtained by forward selection (Westfall et al. 1997), Table 7 shows that these Bonferroni p-values are not an adequate adjustment in the context of models chosen by all-subsets. Clearly, too many terms are identified as significant using both the naive and Bonferroni adjusted p-values.

Westfall and Young (1993) discuss resampling-based procedures for computing adjusted p-values that are more conservative than the Bonferroni adjustment. The single-step maxT procedure adjusts all p-values according to the maximum t-statistic distribution. Improved power over this single-step approach can be achieved by adopting a step-down procedure that adjusts only the minimum p-value according to the maximum t-statistic distribution. The remaining p-values are then adjusted based on smaller and smaller sets of t-statistics. Let the ordered naive p-values, $p_{(1)} < p_{(2)} < ... < p_{(q)}$, have fixed indices

$i_1, i_2, ..., i_q$ based on the order of the original hypotheses. The step-down resampling-based adjusted p-values are defined sequentially as follows:

$$\begin{aligned}
\tilde{p}_{(1)} &= P\left(\min_{\ell \in \{i_1,\ldots,i_q\}} p^*_\ell \leq p_{(1)} | \text{all } H_{0,j} \text{ are true}\right) \\
&\vdots \\
\tilde{p}_{(j)} &= \max\left[\tilde{p}_{(j-1)}, P\left(\min_{\ell \in \{i_j,\ldots,i_q\}} p^*_\ell \leq p_{(j)} | \text{all } H_{0,j} \text{ are true}\right)\right] \qquad (4.1) \\
&\vdots \\
\tilde{p}_{(q)} &= \max\left[\tilde{p}_{(q-1)}, P(p^*_{i_q} \leq p_{(j)} | \text{all } H_{0,j} \text{ are true})\right]
\end{aligned}$$

where p^*_j represents a resampling-based p-value. Use of "max" in (4.1) enforces monotonicity of the adjusted p-values. We slightly modify the step-down procedure in order to incorporate the variable selection procedure into the computation of adjusted p-values. The procedure is as follows:

1. For a candidate model of interest of size q (denoted by $\mathcal{M}_q$), compute the t-statistic for each term in the model (denoted by t_j).

2. Compute the residuals of $\mathcal{M}_q$, which we denote as $\hat{\varepsilon}_{\mathcal{M}_q}$. The distribution of the test statistics depends only on $\hat{\varepsilon}_{\mathcal{M}_q}$ under the complete null hypothesis (i. e., all $H_{0,j}$ are true). Therefore, this distribution is estimated by permuting $\hat{\varepsilon}_{\mathcal{M}_q}$.

3. For each of the B permutations of $\hat{\varepsilon}_{\mathcal{M}_q}$,

   (a) Perform all-subsets regression and locate the best model of size q and compute the t-statistic for each term in the model (denoted by $t^{(b)}_j$).

(b) Compute

$$
\begin{aligned}
u_q^{(b)} &= t_{i_q}^{(b)} \\
u_{q-1}^{(b)} &= \max(u_q^{(b)}, t_{i_{q-1}}^{(b)}) \\
u_{q-2}^{(b)} &= \max(u_{q-1}^{(b)}, t_{i_{q-2}}^{(b)}) \\
&\vdots \\
u_1^{(b)} &= \max(u_2^{(b)}, t_{i_1}^{(b)})
\end{aligned}
$$

4. For each $j = 1, 2, ..., q$, compute the adjusted p-value

$$\tilde{p}_{(j)} = \#\left(u_j^{(b)} \geq t_{i_j}\right)/B. \tag{4.2}$$

5. Enforce monotonicity of the adjusted p-values using successive maximization.

Note that step 3 does require repeated use of all-subsets. Further research is suggested for an approximation to this resampling procedure when using all-subsets.

For orthogonal designs with effect sparsity, the p-values from (4.2) are similar to Lenth's experimentwise p-values. For the full n = 28 Plackett-Burman design of Williams (1968), Lenth's method implemented in JMP using the 27 othogonal columns provides (simultaneous) p-values of 0.042 and 0.418 for X_{15} and X_{20}, respectively. Table 6 shows the largest five estimates. Experimentwise error rate bounds as large as α=0.2 or even α=0.5 are typical (Daniel (1959)). Thus, Lenth's method points to either one or two effects. Implementing the step-down Lenth method of Ye et al. (2001) yielded p-values identical to those found in Table 6. This is due to the Lenth PSE remaining the same for each successive step.

Performing the modified step-down procedure for $m = 5$ produces p-values of 0.022, 0.537, 0.647, 0.756, and 0.756 for X_{15}, X_{20}, X_{17}, X_4, and X_{22}, respectively. This analysis would lead one to conclude that X_{15} is active, but not necessarily any other factors. In

Table 6: Lenth Analysis of Full Williams Data, $n = 28$

Term	Estimate	Lenth t	Individual p-Value	Experimentwise p-Value
X_{15}	-43.179	-4.13	0.002	0.042
X_{20}	-24.393	-2.34	0.032	0.418
X_{17}	-21.393	-2.05	0.054	0.607
X_4	18.250	1.75	0.091	0.816
X_{22}	-16.107	-1.54	0.126	0.920
$\vdots$				

practice, the least significant factor would then be dropped, and the reduced model tested. For the reduced model containing only X_{15} and X_{20}, the adjusted p-values are 0.005 and 0.455, respectively. Thus, the adjusted p-values also indicate one and possibly two active effects for this orthogonal design.

As seen in Table 7, the modified step-down procedure produces more conservative p-values than the Bonferroni adjusted p-values as desired. Adjusted p-values for the best model of each size in Table 3 are shown in Table 7. Note that all models are nested within the best seven variable model, which clearly has unimportant terms. Removing the least significant term (X_7), fitting the reduced model, and computing adjusted p-values also suggests that further reduction is required. Successive model reduction indicates that likely the only active factor is X_{15}, which agrees with our previous analysis of the full data.

A small simulation study (as outlined in Westfall et al. (1997), section 4.3) is performed to investigate power and experimentwise error rate control of the modified step-down procedure. Using the 14-run design of Example 2, data were generated from the model $Y = X\beta + \varepsilon$, where X is the design matrix plus a column of 1's and $\varepsilon \sim N(0,1)$. The number of active effects considered are 0, 1, ..., 5 and are assumed to have the same size, $\beta/\sigma = 5$ with varying signs. As in Westfall et al. (1997), active effects occur only for the first five variables given the near symmetry of this SSD. Furthermore, we follow their notation to indicate the number of and signs of active effects. For example, "2+-" indicates

$(\beta_1, \beta_2) = (5, -5)$. For each scenario, the EER, power for at least one effect, and power for all effects are tabulated.

Two loops were required: an outer loop generated 1000 response vectors while an inner loop performed 500 all-subsets up to $m = 5$ terms in the model in order to tabulate the adjusted p-values. The choice of $m = 5$ for each case is consistent with the effect sparsity assumption in which one would expect at most 20% of the factorial effects to be active (i.e. $\lceil 0.2 * 23 \rceil = 5$). Results are shown in Table 8 for α=0.05, 0.1, 0.2, 0.5 and indicate very favorable performance of the modified step-down procedure. Control of the EER is at or below the nominal level while power is at or above 0.843 for all scenarios considered.

A large decrease in EER among cases "0", "1", and "2" is evident. Empirical investigation revealed that the absolute t-statistics for non-active effects become smaller as active effects are included in the model. For example, it was determined via simulation that $\mathbb{E}(|t_i| \,|H_{0,i}, i = 1, ..., 5$ is true)=5.17 while for case "1", $\mathbb{E}(|t_i| \,|H_{0,i}, i = 2, ..., 5$ is true)=4.46. This naturally lends itself to larger adjusted p-values for non-active effects and hence, fewer Type I errors. This occurrence appears to be more pronounced for SSDs due to the correlations among effect estimates. That is, performing the same simulation study using the full 28-run Plackett-Burman design indicates a less prominent decrease in EER as active effects are added to the model. For instance, with case "1", we observe EERs of 0.03, 0.08, 0.17, and 0.46 for $\alpha = 0.05, 0.1, 0.2$ and 0.5, respectively.

5 Discussion

5.1 Some Advice Regarding Use of SSDs

Why was the AIDS model SSD so unsuccessful? While Lin's forward selection analysis of a SSD identified 11 active factors out of 138, randomization procedures suggest that one cannot be sure that anything useful was learned. This example provides an indication that

Table 7: Naive and Adjusted p-Values for Best Table 3 Models

		P-values		
Term	t-ratio	Naïve	Bonferroni	Permutation
12	-2.14	0.055	1.000	0.877
15	-5.42	0.000	0.004	0.001
12	-3.38	0.002	0.042	0.696
15	-7.75	0.000	0.002	0.001
20	-3.17	0.001	0.021	0.696
4	4.09	0.003	0.054	0.455
12	-5.19	0.001	0.012	0.376
15	-12.96	0.000	0.002	0.002
20	-5.86	0.000	0.004	0.338
4	4.64	0.002	0.032	0.760
10	-2.33	0.048	0.918	0.972
12	-6.63	0.000	0.004	0.535
15	-15.96	0.000	0.002	0.011
20	-6.80	0.000	0.002	0.535
4	6.46	0.000	0.054	0.818
10	-3.29	0.013	0.239	0.988
11	2.68	0.032	0.567	0.988
12	-8.35	0.000	0.002	0.638
15	-21.01	0.000	0.002	0.022
20	-9.31	0.000	0.002	0.567
4	14.91	0.000	0.017	0.390
7	-6.27	0.001	0.014	0.813
10	-9.16	0.000	0.002	0.812
11	7.95	0.000	0.003	0.813
12	-20.04	0.000	0.002	0.206
15	-50.52	0.000	0.002	0.001
20	-24.04	0.000	0.002	0.098

SSDs will not be useful unless interactions can be ignored and only a very small subset of the factors truly explains most of the systematic variation.

For unreplicated full factorials and fractional factorials of strength two, the effect sparsity assumption plays an important role in the analysis of these experiments in order to obtain an estimate of the error variance, σ^2 (see, for example, Lenth (1989)). In the case of SSDs, however, the assumption is vital for obtaining useful estimates for the factor ef-

Table 8: Experimentwise Type I Error Rates and Power

Number of Effects	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.5$
Experimentwise Error Rate				
0	0.054	0.100	0.210	0.510
1	0.016	0.038	0.078	0.283
2	0.003	0.013	0.031	0.136
2+-	0.003	0.009	0.024	0.129
3	0.000	0.003	0.015	0.069
3++-	0.001	0.004	0.012	0.069
4	0.000	0.000	0.002	0.027
4+++-	0.000	0.000	0.002	0.028
4++–	0.002	0.002	0.005	0.029
5	0.000	0.000	0.000	0.002
5++++-	0.000	0.000	0.000	0.000
5+++–	0.000	0.000	0.000	0.000
Power for at least one effect				
1	1.000	1.000	1.000	1.000
2	1.000	1.000	1.000	1.000
2+-	1.000	1.000	1.000	1.000
3	1.000	1.000	1.000	1.000
3++-	1.000	1.000	1.000	1.000
4	0.998	1.000	1.000	1.000
4+++-	0.999	1.000	1.000	1.000
4++–	0.998	1.000	1.000	1.000
5	0.947	0.985	0.999	1.000
5++++-	0.910	0.952	0.976	1.000
5+++–	0.920	0.968	0.992	1.000
Power for all effects				
2	0.990	0.990	1.000	1.000
2+-	0.994	0.998	1.000	1.000
3	0.996	0.999	1.000	1.000
3++-	0.995	1.000	1.000	1.000
4	0.986	0.999	1.000	1.000
4+++-	0.989	1.000	1.000	1.000
4++–	0.985	0.998	0.999	1.000
5	0.881	0.960	0.979	0.982
5++++-	0.849	0.923	0.953	0.955
5+++–	0.848	0.934	0.961	0.963

fects. Since the complications with SSDs has to do with obtaining proper estimates for the effects (rather than estimation of the error variance), adding replication to a SSD would accomplish little. Furthermore, the rule of thumb set forth by Box and Meyer (1986) which defines effect sparsity as 20% or fewer of the factorial effects are active does not apply for SSDs. For instance, 20% of $k = 138$ exceeds the sample size for the AIDS data. Correlations among the contrast estimates makes clear identification of even 5% of the factors infeasible. Unless a very few large effects truly dominate, SSDs will not be informative.

Additivity is also a requirement for successful interpretation of SSDs. We suspect that the AIDS model interactions cannot be ignored. If this is the case, it is even less surprising that the SSD, with the purpose of identifying important main effects, failed to be effective. One might consider the use of group screening methods (see Morris (2006) for an overview) as an alternative to SSDs when interaction effects are plausible.

5.2 Analysis of SSDs

Naive p-values, which are unable to account for the multiple comparison aspect of model fitting with SSDs, do not control against over-fitting. Furthermore, any criterion that ignores the number of factors under consideration (which includes information criteria such as AIC_c) is flawed in the context of SSDs.

To illustrate this point, consider augmenting the 28-run Plackett-Burman design in 27 factors with two-factor interaction columns. In particular, one could construct a 28-run SSD with up to 378 factors using only the main effect and two-factor interaction columns. Following this construction method, we investigate five designs: the 27-factor Plackett-Burman design and four SSDs with $k = 50, 100, 150, 200$. For each design, we simulate 100 response vectors with each $y_1, ..., y_{28}$ from a $N(0, 1)$ distribution and fit models of size $q = 1, 2, ..., 10$ found via forward selection. For each fitted model, R^2 is computed as a measure of model adequacy. Figure 3 displays the average R^2 for each design and model

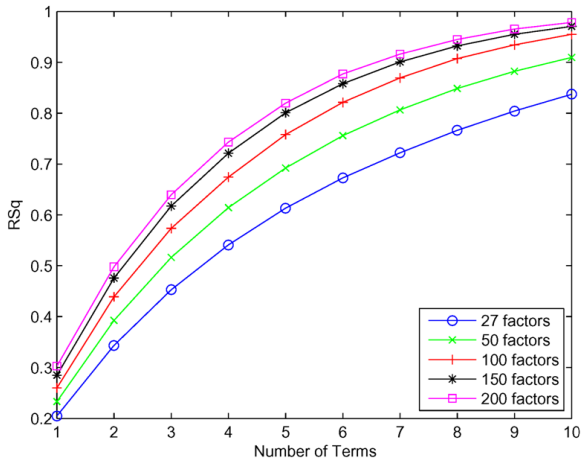


Figure 3: R^2 vs. Number of Factors - Null Model

size. Note that forward selection was utilized here for ease in simulation; the mean R^2s from all-subsets would be even larger.

Simulating from the null model, Figure 3 clearly indicates that as k, the number of factors under consideration increases, the perceived systematic variation explained by the fitted model of a given size q also increases. Therefore, true effects may potentially be masked in SSDs with a large factors/run size ratio. That is, if on average the best s-variable model explains $>80\%$ of the variation when there are no true effects, our ability to identify s (or more) true effects is clearly diminished. This observation is pertinent both for the choice of SSDs and their analysis.

Sections 2 and 3 introduced a means to construct (approximate) p-values for the global randomization test based on all-subsets. This, in conjunction with adjusted p-values for individual terms as proposed in Section 4, provides a simple strategy for ascertaining significance in SSDs and thus, a starting point for follow-up experimentation. The ability to approximate the all-subsets randomization test means that this procedure can be applied to any all-subsets analysis of SSDs. This latter statement is true provided that it is feasible to perform all-subsets regression once up to a specified (and reasonable) number of terms in the model.

Throughout, we have assumed that n is sufficiently larger that one must sample from the $n!$ permutations of Y. However, for $n = 6$, rather than sampling, one should examine all 720 permutations to compute p-values, which would then be exact rather than estimated.

What about really large cases? Li (2008) reports an application of a SSD with 120 runs and nearly 500 factors. In such a case, all-subsets for even four factor models requires comparison of 2.57 billion models, which is almost certainly infeasible by most modern means. With 500 factors, however, surely one or two dozen factors would have systematic effects large enough to warrant further study. In such a case, the model fitting strategy must change to another method, such as a Bayesian variable selection (Chipman et al. (1997)), the Dantzig Selector (Phoa et al. (2009)), or a genetic algorithm. Then, the randomization tests we have recommended could easily be applied to models obtained by any of these methods. Even though we suggest the use of all-subsets regression, we believe randomization tests should be used regardless of the model fitting procedure employed for SSDs.

References

Abraham, B., Chipman, H., and Vijayan, K. (1999), "Some Risks in the Construction and Analysis of Supersaturated Designs," *Technometrics*, 41, 135–141.

Anderson, M.J. (2001), "Permutation Tests for Univariate or Multivariate Analysis of Variance and Regression," *Canadian Journal of Fisheries and Aquatic Science*, 58, 626–639.

Anderson, M.J. and Legendre, P. (1999), "An Empirical Comparison of Permutation Test of Partial Regression Coefficients in a Linear Model," *Journal of Statistical Computation and Simulation*, 62, 271–303.

Anderson, M.J. and Robinson, J. (2001), "Permutation Tests for Linear Models," *Australian and New Zealand Journal of Statistics*, 43, 75–88.

Beattie, S.D., Gong, D.K.H., and Lin, D.K.J. (2002), "A Two-Stage Bayesian Model Selection Strategy for Supersaturated Designs," *Technometrics*, 44, 55–63.

Box, G.E.P. and Meyer, R.D. (1986), "An Anaylsis of Unreplicated Fractional Factorials," *Technometrics*, 28, 11–18.

Chipman, H., Hamada, M., and Wu, C.F.J. (1997), "A Bayesian Variable Selection Approach for Analyzing Designed Experiments with Complex Aliasing," *Technometrics*, 39, 372–381.

Daniel, C. (1959), "Use of Half-Normal Plots in Interpreting Factorial Two-Level Experiments," *Technometrics*, 1, 123–133.

Kelly, H.W. and Voelkel, J.O. (2000), "Asymptotic-Power Problems in the Analysis of Supersaturated Designs," *Statistics and Probability Letters*, 47, 317–324.

Lenth, R.V. (1989), "Quick and Easy Analysis of Unreplicated Factorials," *Technometrics*, 31, 469–473.

Li, W. (2008), "Analyzing Supersaturated Designs via Model Selection Methods - Do They Work?" in *Proceedings of the 2008 Spring Research Conference*, Atlanta, GA.

Lin, D.K.J. (1993), "A New Class of Supersaturated Designs," *Technometrics*, 35, 28–31.

— (1995), "Generating Systematic Supersaturated Designs," *Technometrics*, 37, 213–225.

Manly, B.F.J. (1997), *Randomization, Bootstrap, and Monte Carlo Methods in Biology*, London: Chapman and Hall.

Miller, A.J. (2002), *Subset Selection in Regression*, Boca Raton, FL: CRC Press.

Morris, M. (2006), *Screening: Methods for Experimentation in Industry, Drug Discovery, and Genetics*, New York, NY: Springer, chap. An Overview of Group Factor Screening, pp. 191–206.

Phoa, F.K.H., Pan, Y.H., and Xu, H. (2009), "Analysis of Supersaturated Designs via Dantzig Selector," *Journal of Statistical Planning and Inference*, 139, 2362–2372.

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

Westfall, P.H., Young, S.S., and Lin, D.K.J. (1997), "Forward Selection Error Control in the Analysis of Supersaturated Designs," *Statistica Sinica*, 8, 101–117.

Williams, K.R. (1968), "Designed Experiments," *Rubber Age*, 100, 65–71.

Ye, K.Q., Hamada, M., and Wu, C.F.J. (2001), "Step-Down Lenth Method for Analyzing Unreplicated Factorial Designs," *Journal of Quality Technology*, 33, 140–152.