Analysis of Definitive Screening Designs

Bradley Jones
JMP Discovery Conference
March 2016

Outline

1. Introduction & Motivation
2. Three Ideas for Analysis
3. Simulation Studies
4. Summary

Notation and terminology

m factors, n runs

Linear main effect model (ME)

$$y_i = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} + \varepsilon_i \qquad i = 1, \ldots, n$$

Full second order model

$$y_i = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} + \sum_{j=1}^{m-1} \sum_{k=j+1}^{m} \beta_{jk} x_{ij} x_{ik} + \sum_{j=1}^{m} \beta_{jj} x_{ij}^2 + \varepsilon_i \quad i = 1, \ldots, n$$

Two-factor interactions (2FIs)

Quadratic effects (Q)

The full second order (RSM) model

The response surface model (RSM) is the model consisting of:

1. The intercept term.
2. All main linear effects (for *m* factors, there are m of these)
3. All main quadratic (curvature) effects (*m* of these)
4. All two-factor interactions [there are $m(m-1)/2$ of these]

Number of terms in the full RSM:

$$1 + 2m + m(m-1)/2 = (m+1)(m+2)/2$$

Example: Six Factor RSM (m = 6)

1. Constant term
2. m = 6 main linear effects: X_1, X_2, X_3, X_4, X_5, X_6
3. m = 6 main quadratic effects: X_1^2, X_2^2, X_3^2, X_4^2, X_5^2, X_6^2
4. m = m(m-1)/2 = 15 two-factor interactions:

X_1X_2	X_1X_3	X_1X_4	X_1X_5	X_1X_6
	X_2X_3	X_2X_4	X_2X_5	X_2X_6
		X_3X_4	X_3X_5	X_3X_6
			X_4X_5	X_4X_6
				X_5X_6

Total is 1 + 6 + 6 + 15 = 28 model terms

Definitive Screening Design – minimum runs

Foldover Pair	Run (i)	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$\cdots$	$x_{i,m}$
1	1	0	± 1	± 1	$\cdots$	± 1
	2	0	∓ 1	∓ 1	$\cdots$	∓ 1
2	3	± 1	0	± 1	$\cdots$	± 1
	4	∓ 1	0	∓ 1	$\cdots$	∓ 1
3	5	± 1	± 1	0	$\cdots$	± 1
	6	∓ 1	∓ 1	0	$\cdots$	∓ 1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
m	$2m-1$	± 1	± 1	± 1	$\cdots$	0
	$2m$	∓ 1	∓ 1	∓ 1	$\cdots$	0
Centerpoint	$2m+1$	0	0	0	$\cdots$	0

(Column group heading over $x_{i,1}$ … $x_{i,m}$: Factor Levels)

Minimum design is saturated for the ME + Q effects.

Conference Matrix Definition

A conference matrix is an *mxm* matrix, C, with *0* for each diagonal element and *+1* or *–1* for each off diagonal element such that

$$\mathbf{C}^{\mathrm{T}}\mathbf{C} = (m - 1)\mathbf{I}_{m\times m}$$

The columns of a conference matrix are orthogonal to each other.

A 6x6 conference matrix →

$$\begin{pmatrix} 0 & +1 & +1 & +1 & +1 & +1 \\ +1 & 0 & +1 & -1 & -1 & +1 \\ +1 & +1 & 0 & +1 & -1 & -1 \\ +1 & -1 & +1 & 0 & +1 & -1 \\ +1 & -1 & -1 & +1 & 0 & +1 \\ +1 & +1 & -1 & -1 & +1 & 0 \end{pmatrix}$$

Conference Matrix Construction

Let C be a conference matrix with *m* rows and *m* columns, then

$$\mathbf{D}_m = \begin{bmatrix} \mathbf{C}_m \\ -\mathbf{C}_m \\ \mathbf{0}' \end{bmatrix}$$

where D_m is a DSD with *m* factors and *2m+1* runs.

To construct a DSD with more than the minimal number of runs, use a conference matrix with $c > m$ columns and do not assign the last $c - m$ columns to factors.

Design Properties

1. Small number of runs – 2m + 1 at a minimum
2. Orthogonal main effects (MEs)
3. MEs orthogonal to 2FIs
4. 2FIs not confounded with other 2FIs
5. All the MEs and pure quadratic effects are estimable
6. DSDs with more than 5 factors project onto any 3 factors to allow fitting the full quadratic model

Model selection for unreplicated factorial designs

Any orthogonal main effects plan works for factor screening if:

1. Main effects are >> σ
2. No 2FIs are active
3. The number of active factors < *n/2*

If the number of active effects is greater than n/2, automated model selection procedures tend to break down.

This suggests that for DSDs, identifying models having more than *m* active effects may be problematic.

This talk investigates whether it possible to do better than this.

Outline

1. Introduction & Motivation
2. Three Ideas for Analysis
3. Simulation Studies
4. Summary

Simplest idea – Fit the main (linear) effects model

Advantages:

MEs unbiased – you can believe the coefficient estimates

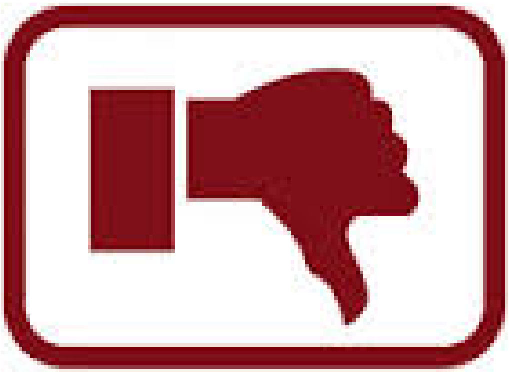
Fit the Main Effects Model

Disadvantages:

Estimate of σ inflated with strong 2FIs or quadratic effects

May make active MEs appear not statistically significant.

You cannot believe the coefficient standard errors.



Example: Laser Etch Experiment

Main Effects Model

Parameter Estimates

Term	Estimate	Std Error	Prob>|t|
Intercept	9.45	1.27	0.00
Speed(8,15)	1.14	1.44	0.46
Frequency(1,5)	1.61	1.44	0.31
Power(15,55)	-2.16	1.44	0.19
Repetitions(1,5)	0.29	1.44	0.85
Humidity(5,15)	0.48	1.44	0.75
Plastic(1,3)	-0.46	1.44	0.76

Best Model

Parameter Estimates

Term	Estimate	Std Error	Prob>|t|
Intercept	4.66	0.71	0.0003*
Speed(8,15)	1.14	0.35	0.0145*
Frequency(1,5)	1.61	0.35	0.0027*
Power(15,55)	-2.16	0.35	0.0005*
Power*Power	6.23	0.83	0.0001*
Frequency*Speed	0.98	0.45	0.0655

The main effects have the same coefficients in each model but the standard errors are inflated for the main effects model.

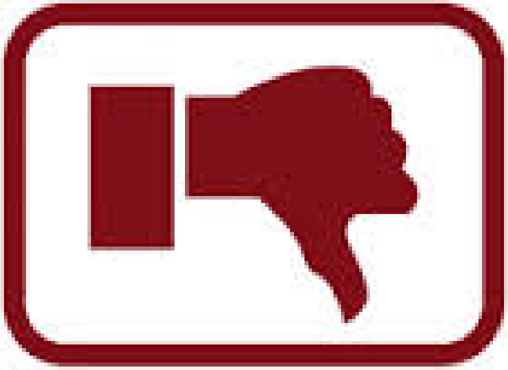
Why did the analysis fail to identify the active main effects?

There are active 2nd order effects included in the estimate of the error standard deviation (RMSE)

This results in an insignificant overall F-test.

Because the denominator in the F statistic is too big

Just fitting the main effects model is not enough.



Analysis Idea #2 – Use Stepwise Regression

Advantages:

Easy to do

Available in most software

Recommended Procedure

Specify a response surface model

Use forward variable selection with stopping based on the AICc criterion

Problem!

The number of model terms (28) is greater than the number of runs in the DSD (13).

Consequences of $n = 13$:

We cannot fit the full RSM

We can estimate at most 13 model terms

One solution: Use Forward Stepwise Selection

Sparsity assumption: not all effects are active

We hope that the number of active effects is substantially fewer than the number of runs.

OK, which effects are active, which are not?

Use a forward stepwise procedure to find out

I prefer minimizing the AICc criterion to decide when to stop. Simulation studies show that it does a better job of finding the active effects when analyzing data with small numbers of runs.

Limitations of DSDs

No design can do everything in one shot. DSDs are no exception.

Limitations include:

1. Stepwise breaks down if there are more than about n/2 active terms in the model

   For example, for six factors, m = 13, if there are more than about 6 active terms, stepwise has difficulty finding the correct model.

   Generally only good if there are just a few two-factor interactions and/or quadratic effects

2. Power is low for finding moderate quadratic effects. The quadratic effect must be large (3 sigma) to have high (>0.9) power.

Addressing the limitations

1. If many terms appear to be active:

   Augment the DSD to identify interactions and quadratic terms.
2. Run a DSD with more than the minimum run size (next)

Analysis Idea #3 – New Method

Since main effects and 2nd order effects are orthogonal to each other you can split the response column (Y) into two new columns.

- One column for identifying main effects – call it YME
- One column for identifying 2nd order effects – call it Y2nd
- The two columns are orthogonal to each other and their sum is Y.

Computing the New Responses

1. Fit the main effects model (No Intercept) and save the predicted values (YME). These are the responses for the main effects model.
2. Save the residuals from the fit above – these residuals are the responses for the 2nd order effects (Y2nd).

Digression: Benefits of “Fake” Factors

Adding Fake Factors (factors you don’t use) provides a way to estimate variance without repeating center runs!

Why?

- Fake factors are orthogonal to the real factors
- Fake factors are orthogonal to all the 2^{nd} order effects
- Assuming the 3^{rd} and higher order effects are negligible, we can use the fake factor degrees of freedom to create an unbiased estimate of the error variance!

Note: Use both the real and fake factors when fitting the main effects model in step 1 of the previous slide.

Example: Six real factors and two fake factors

A	B	C	D	E	F	Fake 1	Fake 2	Y	Y2nd	YME
0	1	1	1	1	1	1	1	94.51	101.04	-6.53
0	-1	-1	-1	-1	-1	-1	-1	107.57	101.04	6.53
1	0	1	1	-1	1	-1	-1	94.36	101.175	-6.815
-1	0	-1	-1	1	-1	1	1	107.99	101.175	6.815
1	-1	0	1	1	-1	1	-1	91.80	90.525	1.275
-1	1	0	-1	-1	1	-1	1	89.25	90.525	-1.275
1	-1	-1	0	1	1	-1	1	93.70	94.485	-0.785
-1	1	1	0	-1	-1	1	-1	95.27	94.485	0.785
1	1	-1	-1	0	1	1	-1	89.55	88.71	0.84
-1	-1	1	1	0	-1	-1	1	87.87	88.71	-0.84
1	-1	1	-1	-1	0	1	1	94.58	95.235	-0.655
-1	1	-1	1	1	0	-1	-1	95.89	95.235	0.655
1	1	-1	1	-1	-1	0	1	93.23	89.58	3.65
-1	-1	1	-1	1	1	0	-1	85.93	89.58	-3.65
1	1	1	-1	1	-1	-1	0	98.11	95.815	2.295
-1	-1	-1	1	-1	1	1	0	93.52	95.815	-2.295
0	0	0	0	0	0	0	0	99.75	99.75	0

Adds 4 runs – 2 error df

Example Column Correlations

Correlations

	Y	Y2nd	YME
Y	1.0000	0.7828	0.6223
Y2nd	0.7828	1.0000	0.0000
YME	0.6223	0.0000	1.0000

Note that the two new responses are orthogonal to each other.

YME
-6.53
6.53
-6.815
6.815
1.275
-1.275
-0.785
0.785
0.84
-0.84
-0.655
0.655
3.65
-3.65
2.295
-2.295
0

Examining the Main Effects Response (YME)

Note responses for each foldover pair sum to zero.

The response for the center run is zero.

There are 17 rows but only 8 independent values

(degrees of freedom – df)

There are 6 real factors but 8 df, so there are

8 – 6 = 2 df for estimating σ^2

Y2nd
101.04
101.04
101.175
101.175
90.525
90.525
94.485
94.485
88.71
88.71
95.235
95.235
89.58
89.58
95.815
95.815
99.75

Examining the 2nd Order Response (Y2nd)

Responses for each foldover pair are the same.

There are 17 rows but only 9 independent values

(degrees of freedom – df)

After estimating the Intercept, there are 8 df left for estimating 2nd order effects.

Analysis – Identify Active Main Effects

1. Recall that the residuals from fitting the Main Effects data to the real factors have 2 degrees of freedom.
2. To estimate σ^2, sum the squared residuals from this fit and divide the result by 2.
3. Using this estimate, do t-tests of each coefficient
4. If the resulting p-value for an effect is small (<0.05 say), conclude that effect is active.

Digression: Model Heredity Assumption

The heredity assumption stipulates that 2nd order effects only occur when the associated main effects are active.

Example 1: If main effects A and B are in the model you can consider the two-factor interaction AB

Example 2: B must be in the model before considering the quadratic effect B^2

While there is no physical law requiring that models exhibit heredity, there is empirical evidence that such models are much more probable in real systems.

Advantage of the Heredity Assumption

The set of possible models using the heredity assumption may be much smaller than allowing any 2nd order effect to appear in the model

Example: Suppose your main effects analysis yields 3 active main effects (C, D, F say). Then the allowable 2nd order terms are CD, CF, DF, C^2, D^2, F^2

We have 8 degrees of freedom and only 6 effects, so it is possible to identify all 6 if they are active.

If we allow consideration all 2nd order effects, there are 15 two-factor interactions and 6 quadratic terms – or 21 terms in all.

There are 2^{21} or more 2 million possible models – a much harder model selection problem.

Analysis – Identifying 2nd Order Effects

Form all the 2nd order terms involving the active main effects

Do all subsets regression up to the point where the MSE of the best 2nd order model for a given number of terms is not significantly larger than your estimate of σ^2

Outline

1. Introduction & Motivation
2. Three Ideas for Analysis
3. Simulation Studies
4. Summary

DSD with 4 factors and 13 runs

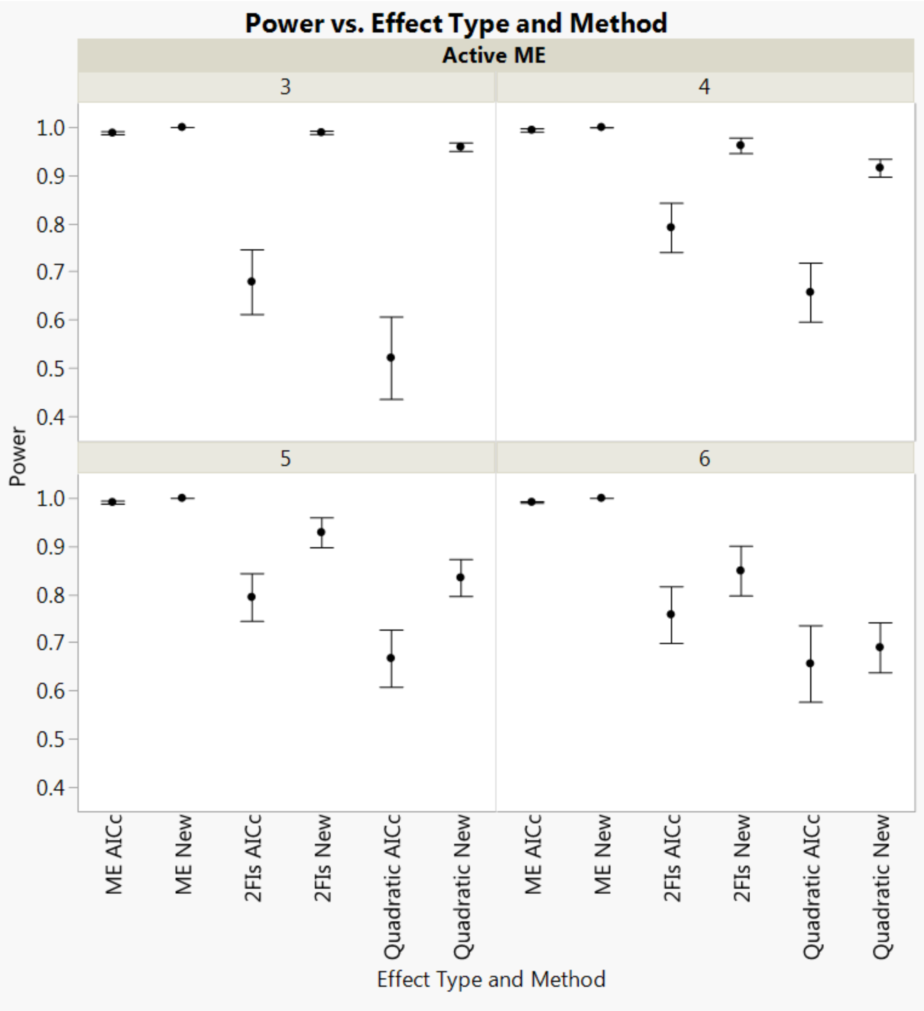
Number of Active ME	Number of Active 2FIs	Number of Active Quadratic Effects	Power ME	Power 2FIs	Power Quadratic
3	1	1	0.9997	0.998	0.958
3	1	2	0.9987	0.989	0.955
3	1	3	0.9983	0.977	0.952
3	2	1	0.9987	0.995	0.947
3	2	2	0.9983	0.973	0.940
3	2	3	0.9980	0.967	0.949
3	3	1	0.9993	0.979	0.926
3	3	2	0.9990	0.979	0.959
3	3	3	0.9980	0.972	0.964
4	1	1	0.9993	0.990	0.884
4	1	2	0.9995	0.867	0.755
4	2	1	0.9983	0.926	0.793

Active coefficients generated by adding 2 to an exponentially distributed random number.

Model heredity assumed.

Two “fake” factors.

Simulation Comparisons New Method vs. Stepwise



Comparison for DSD with 6 factors and 17 runs (i.e. 2 fake factors)

Power for detecting 2FIs and Quadratic effects is much higher for the new method especially when fewer MEs are active

Analyzing DSDs Conclusion

Stepwise with AICc works adequately if there are few active 2nd order effects (one or two 2FIs and/or one quadratic effect)

The new method is not in commercial software but performs better than any existing alternative analytical procedure I know.

Recommendations

Prefer using fake factors to repeated center runs.

Assume model heredity unless there is substantial scientific evidence to the contrary.

Model main effects separately from 2nd order effects by breaking the response into two responses.

And one last thing…

You can use the two response decomposition idea for any foldover design.

References

Jones, Bradley and Nachtsheim, C. J. (2011) “A Class of Three-Level Designs for Definitive Screening in the Presence of Second-Order Effects” *Journal of Quality Technology*, **43**. 1-15.

Jones, Bradley**,** and Nachtsheim, C. J. (2013) “Definitive Screening Designs with Added Two-Level Categorical Factors”, *Journal of Quality Technology*, **45:2**, 120-129.

Jones, Bradley, and Nachtsheim, C. J. (2015): Blocking Schemes for Definitive Screening Designs, *Technometrics*, DOI: 10.1080/00401706.2015.1013777

Miller, A., and Sitter, R. R. (2005). “Using Folded-Over Nonorthogonal Designs,” *Technometrics*, **47:4**, 502-513.

Xiao, L, Lin, D. K.J., and Fengshan, B. (2012), Constructing Definitive Screening Designs Using Conference Matrices, *Journal of Quality Technology*, 44, 1-7.