Model Selection Strategies for Definitive Screening Designs Using JMP Pro and R

Philip J. Ramsey, Ph.D.
University of New Hampshire
& North Haven Group
Durham, NH, USA
philip.ramsey@unh.edu

Maria Weese, Ph.D.
Department of Information Systems and Analytics
Miami University
Oxford, OH, USA
weeseml@miamioh.edu

Douglas Montgomery, Ph.D.
Department of Industrial Engineering
Tempe, AZ, USA
doug.montgomery@asu.edu

Outline

- Introduction
- Simulation Protocol
- Model Selection Strategies
- The Dantzig Selector
- Simulation Results for DSDs
- Conclusions from Simulations
- Case Study: The Glycoprofiling DSD Experiment
- Conclusions from Case Study

Introduction

Definitive Screening Designs (DSD), are receiving a lot of interest due in part due their high efficiency, which makes them attractive for experimenters where experimental resources are limited.

DSDs have especially attracted a great deal of attention in the Biopharmaceutical Industries for process characterization and optimization due in part to the Quality by Design (QbD) initiatives of the FDA and the European Medicines Association; see Elliot et al. (2013).

Although DSDs have enjoyed well deserved popularity as experimental designs, there has only recently been much research on how to best analyze these designs.

In this talk we will discuss a number of approaches to the analysis of DSDs and make recommendations.

Explanatory vs. Predictive Modeling

An important consideration in the analysis of any experiment is the objective of the analysts.

Schmueli (2010) points out that most statistical modeling can be divided into two major categories:

1. **Explanatory Modeling**
2. **Predictive Modeling**

She defines explanatory modeling as when the analysis goal is to discover a causal relationship between the factors and the response and predictive modeling when the analysis goal is to predict the response from the factors.

A good model for explanation is frequently not a good model for prediction and this point has been well demonstrated by examples from predictive analytics and machine learning.

Explanatory Modeling Simulation Details

Four separate simulation scenarios were employed.

	Main Effects		Quadratic Effects		Interaction Effects		Heredity
Scenario	a	τ	a	τ	a	τ	
1	n/3	6	1	3	1	3	Strong
2	3	6	3	6	3	6	Strong
3	n/4	3	2	9	2	9	Strong
4	3	6	3	6	3	6	Weak

In the table *a* is the number of effects of each type randomly selected from a full quadratic model for *k* experimental factors and τ represents the magnitude of the selected effects.

The signs of the coefficients were randomly assigned.

Inactive effect coefficients were randomly assigned from a N(0, 0.2) distribution.

Random error was added to the response with $\varepsilon \sim N(0;1)$.

Explanatory Modeling Simulation Details

Each Scenario required 7 DSDs be generated to cover the range of k.

As an example, for Scenario 1, a DSD with $k = 5$ has $n = 13$ trials, a randomly selected model would have 4 main effects of size 6, 1 quadratic effect of size 3, and 1 interaction of size 3.

For each scenario, and number of factors $k = 5, 6, \ldots, 11$ combination 1,000 iterations were performed where the appropriate model was randomly generated for each iteration.

Each simulation iteration was analyzed with various methods and the effects entering the selected models were tallied.

Explanatory Modeling Simulation Details

The following quantities were measured for each scenario, and *k*:

- Power (overall number of times the correct active factors were identified) and the associated Type II error rates.
- Type I error rates (number of times inactive factors were identified as active).
- Average total number of effects identified as active.
- The power to detect the different types of effect, i.e. power to correctly identify the quadratic effects, the interaction effects and the main effects.
- PRESS statistic for the selected models and their responses.

Explanatory Modeling Simulation Details

The following analysis strategies were employed:

- **Forward Selection** using AICc and BIC as a stopping criteria.
- **The Dantzig Selector** using AICc and BIC to choose δ, λ=1.5.
- **All Subsets Regression** using AICc and BIC for the selection criteria with max model size set to 9 effects.

Under the assumption of Effect Sparsity it would be unusual to find more than 9 active effects in a screening design.

Effect Heredity restrictions are not enforced (and cannot be for the DS) in any of the analyses – this is a topic for future research.

The Dantzig Selector is not supported in JMP Pro so the simulation analyses were performed in R.

Model Selection Criteria

AICc is the **Akaike Information Criterion** with a bias correction factor "c" for small data sets.

BIC is the **Bayesian Information Criterion**.

The AICc, assuming the response is normally distributed, has the following formula:

$$AICc = nLn\left(\frac{SSE(p)}{n}\right) + \frac{2p(p+1)}{n-p-2}.$$

Overfitting penalty term

Where *SSE* is sum of squares error, n is the number of observations and p is the number of model terms **including** the intercept and the estimate of σ.

Model Selection Criteria

The BIC is similar in form to the AICc and assuming the response is normally distributed has the form:

$$BIC = nLn\left(\frac{SSE(p)}{n}\right) + pLn(n).$$

Overfitting penalty term

The model with the smallest BIC value is interpreted as the one with the largest **posterior probability given the data**.

The basis for the BIC lies in Bayesian probability theory while the basis for AICc lies in thermodynamics and is related to Boltzmann's entropy.

Generally the two popular criteria for model selection do not agree on a best model due to the different manner in which over fitting is penalized.

There is no consensus as to which may be the preferred criterion.

Model Selection Strategies

The **Dantzig Selector** is a shrinkage method often used when the number of possible effects $p > n$. The solution is found by solving the linear programming problem:

$\hat{\beta}_{DS}$ is the solution to

$$\min_{\hat{\beta} \in B} \left\| \hat{\beta} \right\|_1 \text{ subject to } \left\| X^T \left(Y - X\hat{\beta} \right) \right\|_\infty \leq \delta$$

The tuning parameter δ is determined empirically with AICc or BIC.

A two stage procedure is often used where the initial effects are estimated then any **effect < γ is removed** and the remaining effects estimated by OLS; **γ = 1.5** was used (Marley and Woods, 2010).

Several others Marley and Woods (2010), Draguljić et al. (2014), Weese et al (2015) have shown success in terms of identifying the correct active factors in supersaturated designs using the Dantzig selector.

Simulation Results for DSDs

Below is a summary of the average power, type I error, and type II error by method over all simulation scenarios, effects, and design sizes.

Method	Power	Type I Error	Type II Error
Best AICc	0.657	0.070	0.343
Best BIC	0.660	0.147	0.340
Dantzig BIC	0.705	0.092	0.295
Dantzig AICc	0.607	0.059	0.393
Forward AICc	0.574	0.088	0.426
Forward BIC	0.711	0.285	0.289

Forward Selection with BIC has the highest overall power and a large type I error rate, however for screening this generally would be an acceptable tradeoff.

The DS and All Subsets Regression with BIC had lower power however both had lower type I error rates.

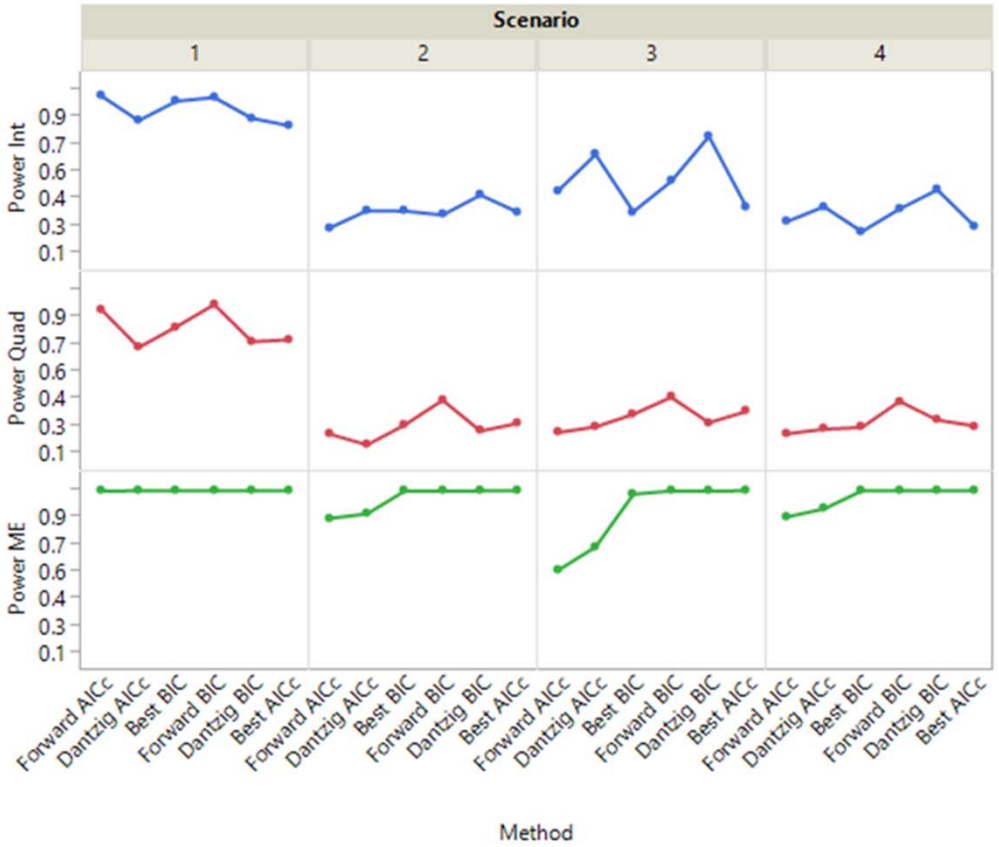
Simulation Results for DSDs

To the right is a graph of results by Scenario and type of effect.

The DS with BIC had the highest power for interactions in 2, 3, and 4.

The power for main effects was high in all 4 Scenarios.

Forward with BIC had the highest power for quadratic effects.



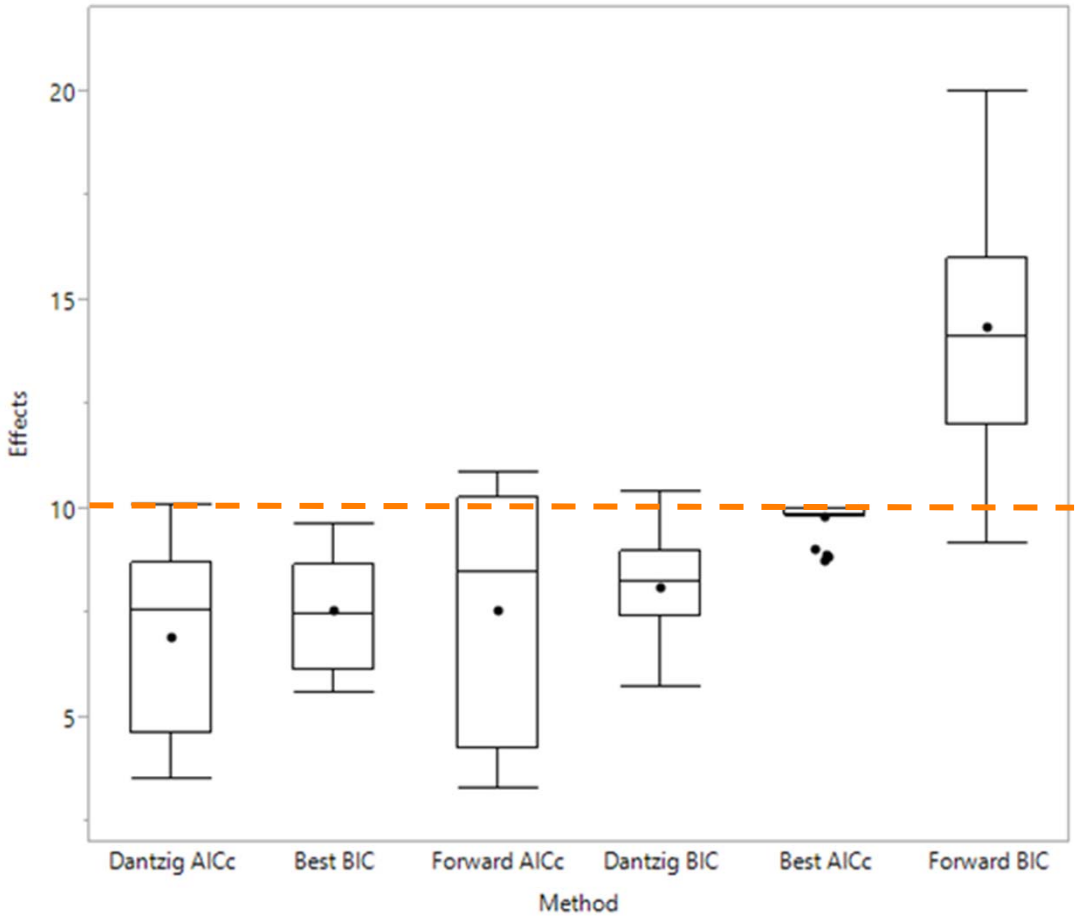
	Main Effects		Quadratic Effects		Interaction Effects		Heredity
Scenario	a	τ	a	τ	a	τ	
1	n/3	6	1	3	1	3	Strong
2	3	6	3	6	3	6	Strong
3	n/4	3	2	9	2	9	Strong
4	3	6	3	6	3	6	Weak

Simulation Results for DSDs

To the right is a graph of the average number of effects found by method over all scenarios.

Forward Selection BIC is noticeably higher and is consistent with a **known tendency of BIC to over fit models when *n* is small relative to *p***.

In fact the largest number of active effects possible in a model across all cases would be **10**.



Simulation Results for DSDs

Prediction error in the simulations can only be explored using "in-sample" measures of prediction error.

The **PRESS** statistic was used as a measure of "in-sample" prediction error and the simulation scenarios and selected models were compared based upon Press.

PRESS mimics the concept of cross-validation ("out-of-sample") prediction error, however a true independent validation data set is the preferred way to measure prediction capability.

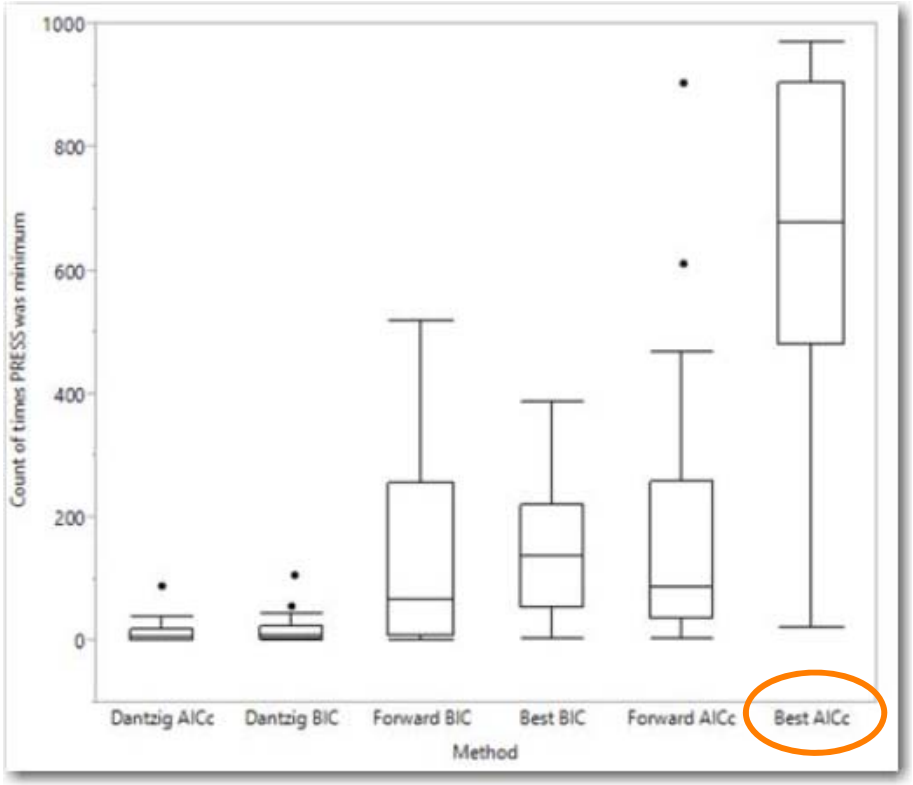
For each model selection strategy the number of times the selected model had the minimum Press among all selected modes was tallied.

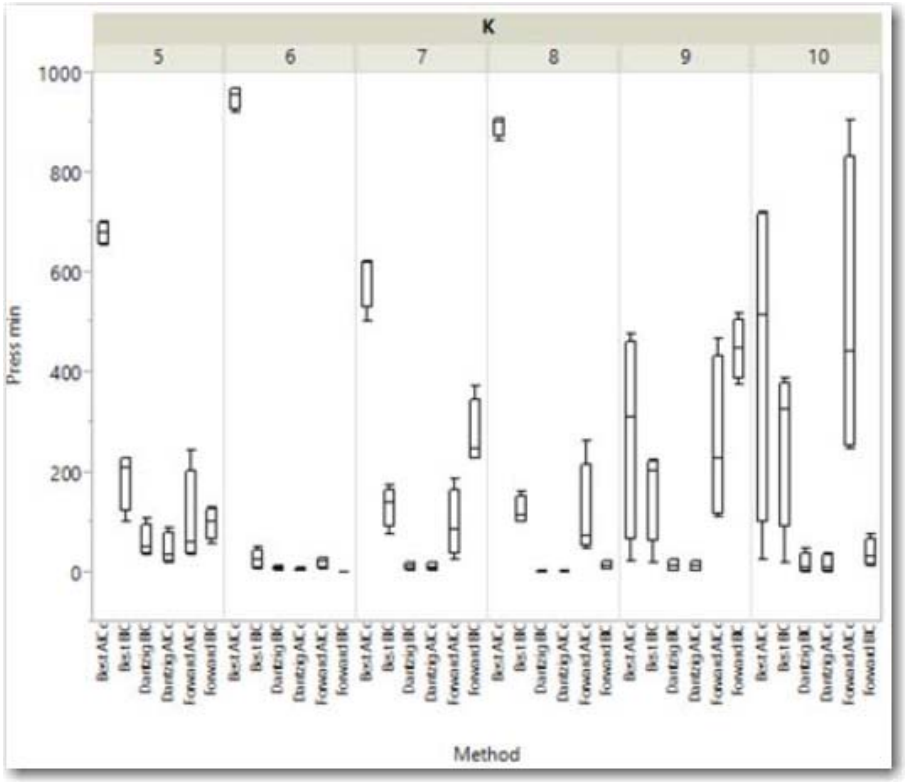
Smaller PRESS indicates better predictive behavior.

Simulation Results for DSDs

Interestingly, the methods with the highest power were not also the methods that had the lowest PRESS.

Best-subsets regression using AICc was dominant for the smaller designs, but as k increased (and n) the effect dissipated.





Simulation Results for DSDs

It is difficult to make a blanket recommendation on the best approach to analyze DSDs as screening designs – explanatory modeling.

The **Dantzig Selector** and **Forward Selection** using BIC did comparatively well in terms of power for interaction and quadratic effects, however the user must decide on the acceptable level of risk for Type I errors.

Given the ease of Forward Selection, even for large models, it may be the simple analysis solution for users tolerant of Type I errors.

The Dantzig Selector and All Subsets Regression most likely would improve in power substantially for all three types of effects if **multiple solutions** were considered instead of a single solution.

Examining different values of γ for the Dantzig Selector most likely would also improve power with an unknown Type I error cost.

Analysis of DSDs for Prediction

In order to best evaluate the best approach to the analysis of DSDs for prediction, an independent validation set is required.

Generally, in design of experiments such validation sets are not available, however we have **available two case studies where a DSD and CCD were run in parallel**.

We will use the CCD as a validation set and the DSD as a training set to determine how well our model selection strategies perform in terms of prediction error on the CCD set.

Prediction error is measured in terms of average squared error (ASE) on the CCD validation set.

$$ASE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

Case Study: Glycoprofiling Proteins

Glycoproteins are the largest group of biologically-derived drugs, however robust analytical methods to profile or characterize the post-transcription glycosylation of proteins are needed.

Yeung and Ramsey (2015) discuss the characterize of an HPAE-PAD analytical method for glycoprofiling employing both a 5 factor DSD and a 5 factor CCD run in parallel.

We are grateful to Dr. Eliza Yeung of Cytovance Biologics, Oklahoma City, OK for making the experimental results available for the talk.

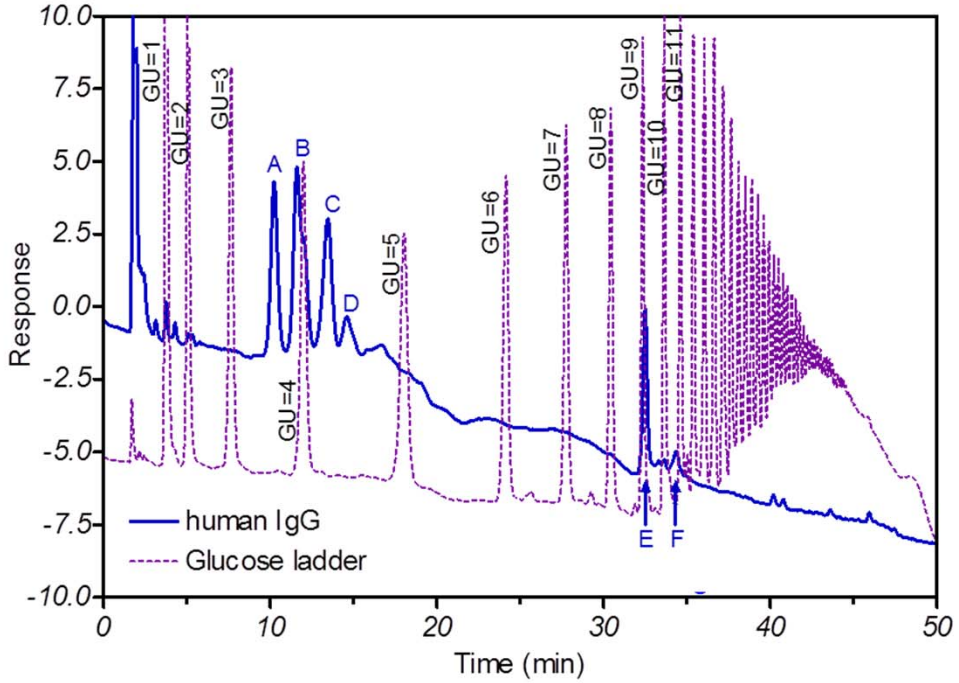
High Performance Anion Exchange Chromatography with pulsed amperometic detection (HPAE-PAD) as a potential and cost effective method for characterization.

Traditionally chromatographic methods for glycoprofiling have been inconsistent – the glycans do not necessarily elute consistently

Case Study: Glycoprofiling Proteins

A new method has been developed by Dr. Yeung whereby a glucose unit ladder (a reference solution) is used to properly identify the different glycoforms from a protein in a human antibody sample.

The approach uses the GU ladder as a reference to identify glycoform peaks from an actual human antibody sample.



Glycoform glycan	G-Unit
A	3.59
B	3.89
C	4.23
D	4.42
E	9.17
F	10.8

Case Study: Glycoprofiling Proteins

Five factors were selected to manipulate in the experiment.

Factor (level)	-1	0	1
Initial %NaOAc (% A)	0	10	20
Initial %NaOH (% B)	30	40	50
Gradient_01 (mM NaOAc /min)	0.415	1.25	2.085
Gradient_02 (mM NaOAc /min)	1.25	2.085	2.915
Gradient_03 (mM NaOAc /min)	4.72	5.555	6.39

* Gradient_01, _02 and _03 are % A (500 mM NaOAc) increases over 12 min, 12 min and 18 min respectively and at constant initial % B (200 mM NaOH,10 mM NaOAc). The values are expressed as mM NaOAc per min.

Case Study: Glycoprofiling Proteins

Seven responses were chosen to optimize in the experiment. However in the present discussion we will focus on RT_G03, which is the time at which Glucose Unit 3 elutes.

RT_03 is the key response in order for the method to work.

Response	Description	Optimization
RT_G03	Retention Time	Target ~ 8.5 min
Resol_G03	Resolution G03-G04	Maximize
Resol_G04	Resolution G04-G05	Maximize
Resol_G05	Resolution G05-G06	Maximize
Resol_G09	Resolution G09-G10	Maximize
Resol_G10	Resolution G10-G11	Maximize
USP Tailing	USP Tailing G04	Monitor (0.8-1.2)

Case Study: Glycoprofiling Proteins

Instead of AICc and BIC we use a form of them with nice theoretical properties for model selection.

It is best to rank models based on AICc differences computed as

$$\Delta_i = AICc_i - AICc_{\min}$$

where $AICc_{min}$ is the smallest AICc among the candidate models.

It is also best to work with BIC differences in ranking models where the differences are computed as

$$B_i = BIC_i - BIC_{\min}$$

Typically models where Δ or B exceed a range of 2 – 4 are excluded from further consideration – this is not a hard and fast rule.

JMP Pro will be used for the analyses except R is used for the DS.

Case Study: Glycoprofiling Proteins

Shown below are JMP Formula Editor formulas to create the Δ_i, B_i and the Akaike weights w_i.

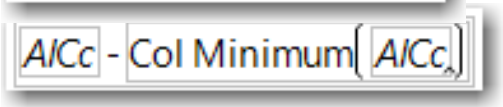
Following the convention of Burnham and Anderson (2002, pg. 75) the weights are

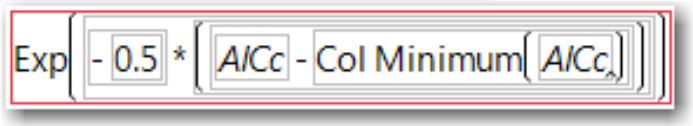
$$w_i = \frac{e^{-0.5\Delta_i}}{e^{-0.5\Delta_{\min}}} = \frac{e^{-0.5\Delta_i}}{1} \quad \text{where } \Delta_{\min} = 0$$

The smaller the weight the less likely that model is the best model in the K-L sense; the weights represent the odds that a given model is the best K-L model.

The JMP formulas for the three quantities are:

$\Delta_i =$ 

$B_i =$ 

$w_i =$ 

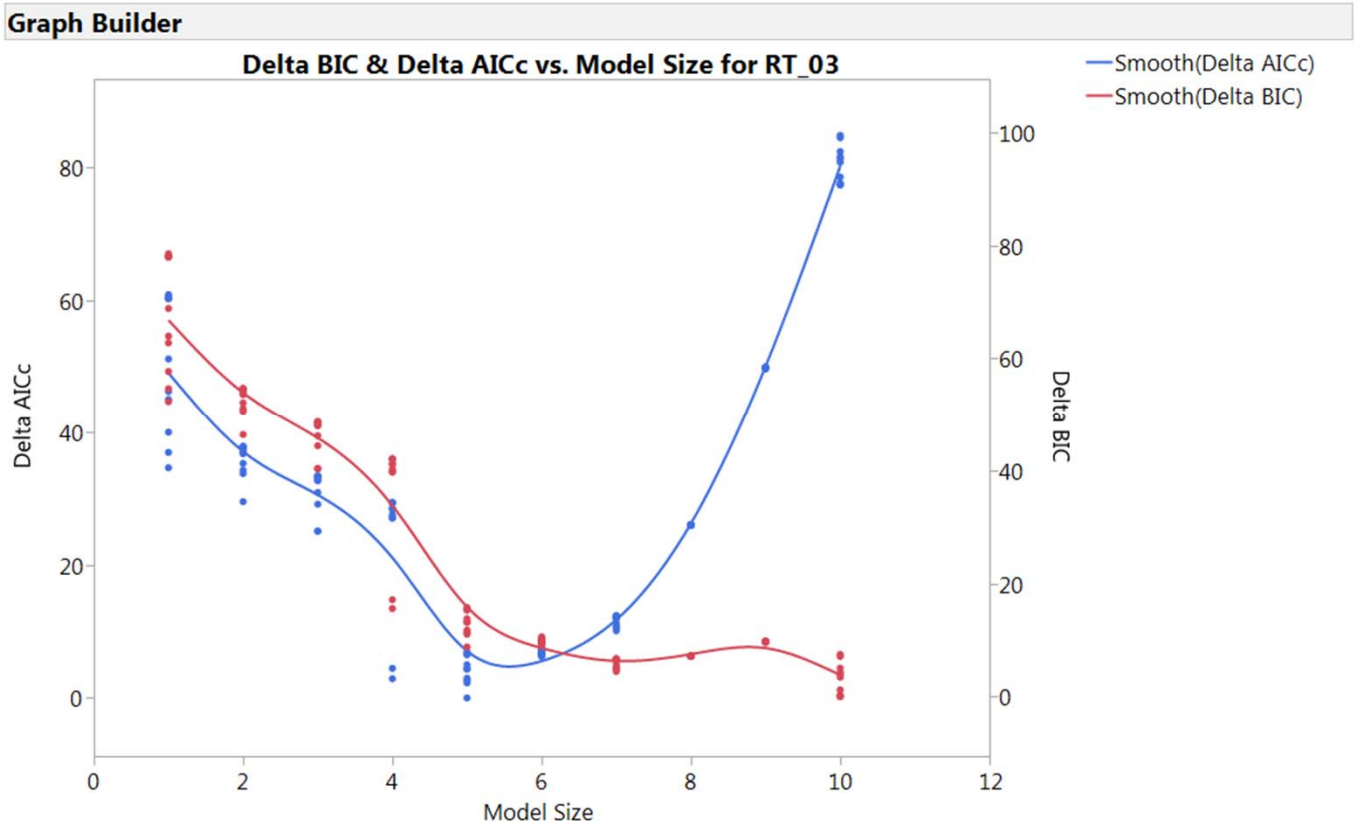
Case Study: Glycoprofiling Proteins

Below is a plot of the delta AICc and BIC results that help us identify potential candidate models; response is **RT_03**.

AICc suggests models with about $p = 5$.

BIC suggesting models with around $p = 7$ *to 10*.

As expected the two criteria do not agree as to best models.



Case Study: Glycoprofiling Proteins

Using Delta AICc only two models are considered potentially best.

	Model	Number	RSquare	RMSE	AICc	BIC	Delta AICc	Delta BIC
1	Initial %Ac(0,20),Initial %NaOH(30,50),Gradient_02(1.25,2.915)...	5	0.9925	0.3504	32.3298	23.7379	0.0000	11.4178
2	Initial %Ac(0,20),Initial %NaOH(30,50),Initial %Ac*Initial %Ac,I...	4	0.9885	0.4150	32.6021	27.9043	0.2724	15.5842

The 4 term model is shown to the right.

Notice it does not have any **Gradient** terms.

This is important given GU_03 elutes before the gradient elutions are applied.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	162.69129	40.6728	236.1447
Error	11	1.89461	0.1722	**Prob > F**
C. Total	15	164.58590		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	5.625	0.169429	33.20	<.0001*
Initial %Ac(0,20)	-3.645	0.131239	-27.77	<.0001*
Initial %NaOH(30,50)	-0.8834	0.131239	-6.73	<.0001*
Initial %Ac*Initial %Ac	1.8866	0.214312	8.80	<.0001*
Initial %Ac*Initial %NaOH	1.041625	0.14673	7.10	<.0001*

Case Study: Glycoprofiling Proteins

Our analysis differs from the simulation in that we consider multiple solutions for All Possible Models analysis done in JMP.

From the All Possible Models solutions 7 models were selected for consideration based upon Δ_i and B_i and a **Fit Group** created.

All of the **Fit Group** models were saved to the JMP data table and were compared in prediction performance on the CCD data using the RASE (square root ASE).

The **Model Comparison** platform in JMP Pro was then used to compare the models.

In addition a prediction average model was created, which uses the average prediction from the 4 of the best models.

Prediction Averaging is used to decrease prediction error.

Case Study: Glycoprofiling Proteins

Multimodel inference has become increasingly important in predictive analytics and has existed for many years in meteorology and econometrics in the form of forecast averaging.

We quote from Burnham:

> "Empirical science in the twenty-first century will increasingly rely on multimodel inference. Models in the life sciences are nearly always oversimplified and it is not reasonable to make inference based on only the one model estimated to have been the best (i.e., 'best' in the sense of K-L information). Rather, *there are a host of advantages in making inference from a weighted combination of results from all the models in the set*." (Burnham, K. et. al., 2011).

Case Study: Glycoprofiling Proteins

Model Averaging is another form of multimodel inference in which one averages the estimated coefficients for each effect across a possibly large set of fitted models.

The model averaged coefficients are typically weighted averages of the coefficients from each of the fitted models where that effect occurred.

Model Averaging is implemented in the *Stepwise* platform (a red triangle option).

AICc weights for each model are used to create the averages of the coefficients for each effect in the full quadratic model.

Please note that not all possible models are estimable due to the fact that linear combinations of two way interactions and quadratic effects can result in near singular models.

Case Study: Glycoprofiling Proteins

Model averaging generally results in a very large over-fit model.

- In this case, the full quadratic model with 20 effects is being fit (in only 13 df).

Over-fitting results in inflated coefficient estimates, which in turn results in inflated prediction error.

The effect of model averaging with weights is to shrink the coefficients for each model effect.

- The coefficient averaging is really a form of regularization whereby the inflated coefficient estimates are shrunken by the averaging, which diminishes over-fitting problem.

See Burnham and Anderson (Chp. 4, 20020 and (2004) for further discussion of model averaging.

Case Study: Glycoprofiling Proteins

Below is the estimated full quadratic model using Model Averaging in the Stepwise platform.

More research is needed on choices of maximum model size, weighting schemes, and the number of models to average.

However, the technique does make it possible to fit a full quadratic model from the supersaturated DSD.

The estimated coefficients are qualitatively quite reasonable.

Note: we find the technique intriguing, but in need of further investigation.

Parameter
Intercept
Initial %Ac(0,20)
Initial %NaOH(30,50)
Gradient_01(0.415,2.085)
Gradient_02(1.25,2.915)
Gradient_03(4.72,6.39)
Initial %Ac(0,20)*Initial %Ac(0,20)
Initial %Ac(0,20)*Initial %NaOH(30,50)
Initial %NaOH(30,50)*Initial %NaOH(30,50)
Initial %Ac(0,20)*Gradient_01(0.415,2.085)
Initial %NaOH(30,50)*Gradient_01(0.415,2.085)
Gradient_01(0.415,2.085)*Gradient_01(0.415,2.085)
Initial %Ac(0,20)*Gradient_02(1.25,2.915)
Initial %NaOH(30,50)*Gradient_02(1.25,2.915)
Gradient_01(0.415,2.085)*Gradient_02(1.25,2.915)
Gradient_02(1.25,2.915)*Gradient_02(1.25,2.915)
Initial %Ac(0,20)*Gradient_03(4.72,6.39)
Initial %NaOH(30,50)*Gradient_03(4.72,6.39)
Gradient_01(0.415,2.085)*Gradient_03(4.72,6.39)
Gradient_02(1.25,2.915)*Gradient_03(4.72,6.39)
Gradient_03(4.72,6.39)*Gradient_03(4.72,6.39)

Case Study: Glycoprofiling Proteins

In the JMP Stepwise platform, Forward Selection using AICc selected the same model as All Possible Models based on minimum AICc.

Forward Selection using BIC fit a near saturated model with 10 effects that was clearly overfit and gave "wild" predictions on the CCD data.

In JMP Pro, Forward Selection can also be performed using the **Generalized Regression** personality in the **Fit Model** platform.

Since the **Dantzig Selector is not supported in JMP Pro** and the nature of the Forward Selection results in JMP Pro, we only considered the All Possible Models results in our Fit Group.

We also consider the Model Averaging prediction model.

Case Study: Glycoprofiling Proteins

Below is a screen shot of the JMP data table with the Fit Group prediction formulas saved to the data table; the p in the column header indicates the number of effects in the model.

Notice that the 8, 9, and 10 effect models selected by BIC are badly overfit yielding unstable predictions on the CCD data.

Design	RT_G03	Pred p=4 RT_G03	Pred p=5 RT_G03	Pred p=6 RT_G03	Pred p=7 RT_G03	Pred p=8 RT_G03	Pred p=9 RT_G03	Pred p=10 RT_G03
CCD	9.100	9.232	9.490	9.240	9.408	-260.964	-221.226	9.159
CCD	13.800	13.082	12.823	13.073	12.905	12.818	239.812	9.634
CCD	3.817	3.708	3.967	3.716	3.548	3.845	7.228	492.755
CCD	5.633	5.625	5.625	5.625	5.457	5.866	5.294	128.076
CCD	5.617	5.625	5.366	5.366	5.366	-129.431	-110.149	-115.215
CCD	5.633	5.625	5.625	5.625	5.625	5.612	5.671	5.671
CCD	3.817	3.708	3.450	3.600	3.432	-267.398	234.608	244.818
CCD	3.817	3.708	3.450	3.200	3.368	273.767	-227.532	-241.038
CCD	11.033	11.157	11.157	11.157	11.157	-124.111	9.177	9.534
CCD	5.767	5.625	5.883	5.883	5.883	-128.824	-109.628	-114.672
CCD	5.783	5.625	5.625	5.625	5.625	275.709	5.890	6.014
CCD	9.367	9.232	8.973	9.123	9.292	9.066	9.327	9.204
CCD	3.917	3.708	3.967	4.117	4.285	3.738	4.587	-477.645

Case Study: Glycoprofiling Proteins

Below is a screen shot of the **Model Comparison** results sorted by RASE.

The best model in terms of RASE has 4 effects and is the second best model in terms of AICc.

Notice that the Model Averaging and Prediction Averaging models are second and third best respectively.

Model Comparison

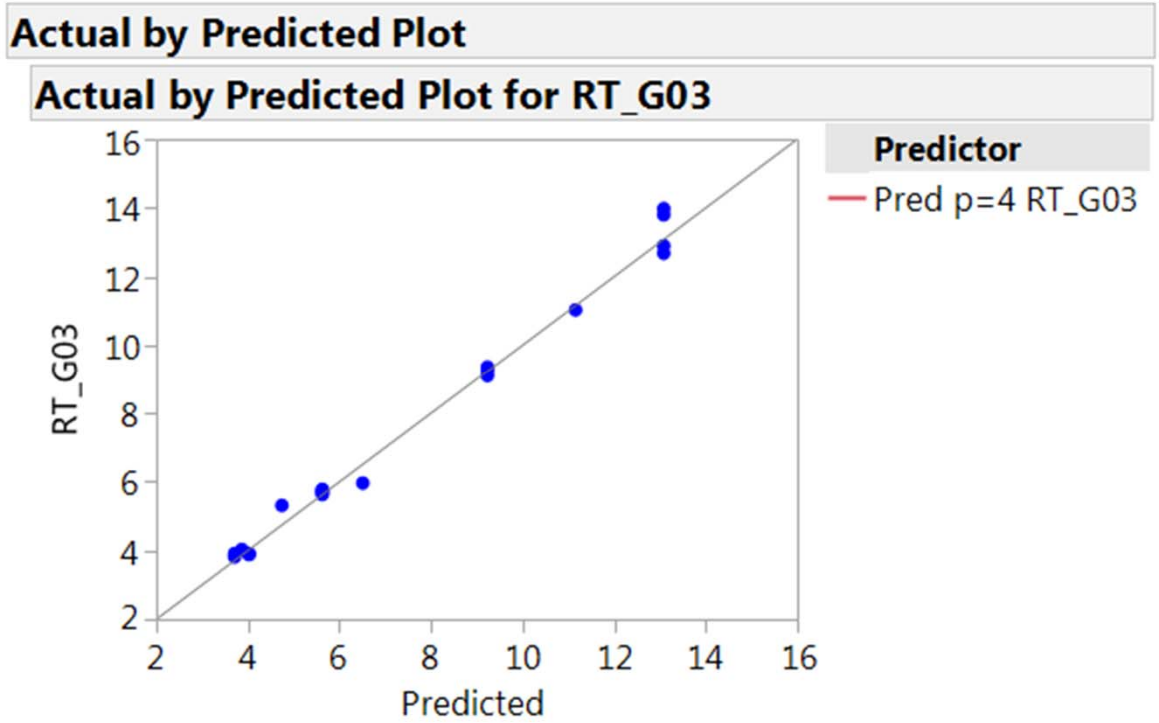
Measures of Fit for RT_G03

Predictor	Creator	.2.4.6.8	RSquare	RASE	AAE	Freq
Pred p=4 RT_G03	Fit Least Squares		0.9919	0.2969	0.2039	28
Model Avg Predicted RT_G03	Fit Stepwise		0.9914	0.3052	0.2078	28
RT_G03 Avg Predictor	Model Comparison Model Averaged		0.9909	0.3136	0.2429	28
Pred p=6 RT_G03	Fit Least Squares		0.9893	0.3411	0.2765	28
Pred p=7 RT_G03	Fit Least Squares		0.9881	0.3594	0.2904	28
Pred p=5 RT_G03	Fit Least Squares		0.9880	0.3603	0.2776	28

Case Study: Glycoprofiling Proteins

One can use an actual by predicted plot to determine how well the DSD model predicts the CCD responses.

Below is a screen shot of the actual by predicted plot for the best model. The appears quite good.

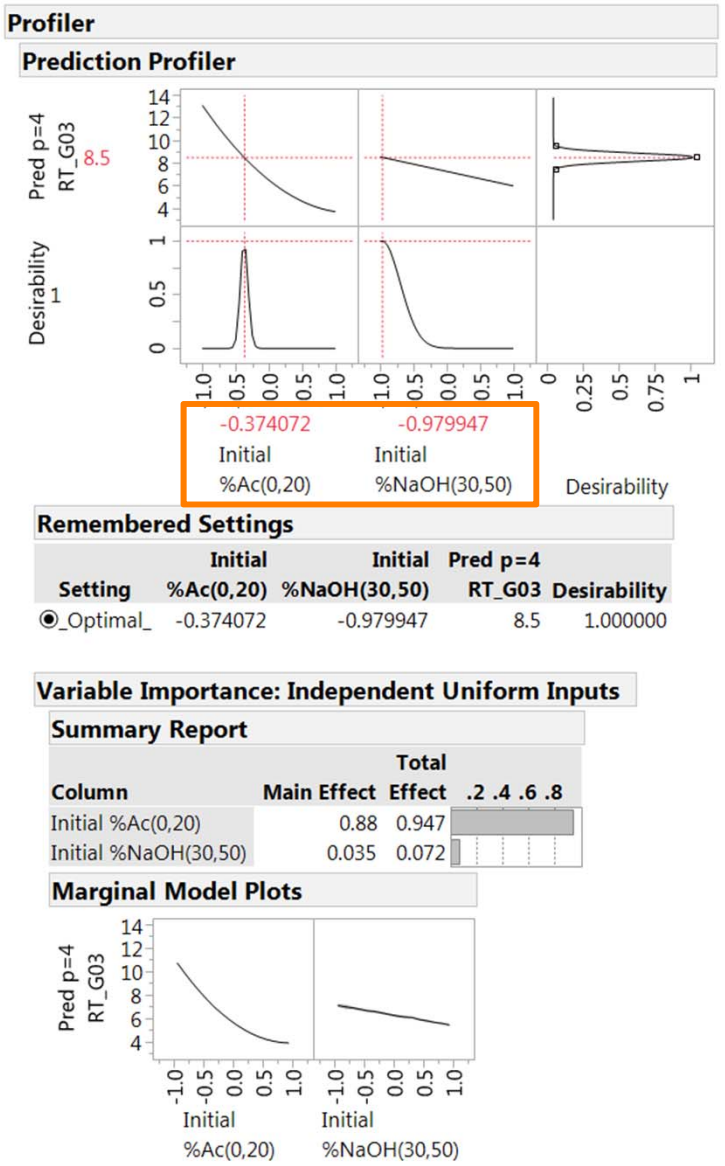


Case Study: Glycoprofiling Proteins

Finally the goal is to determine settings of the input factors that achieve the most desired goal for **RT_03**, which is to hit a target of 8.5 minutes elution time.

Using the **Profiler** in JMP we can find settings of the two inputs to achieve the target time.

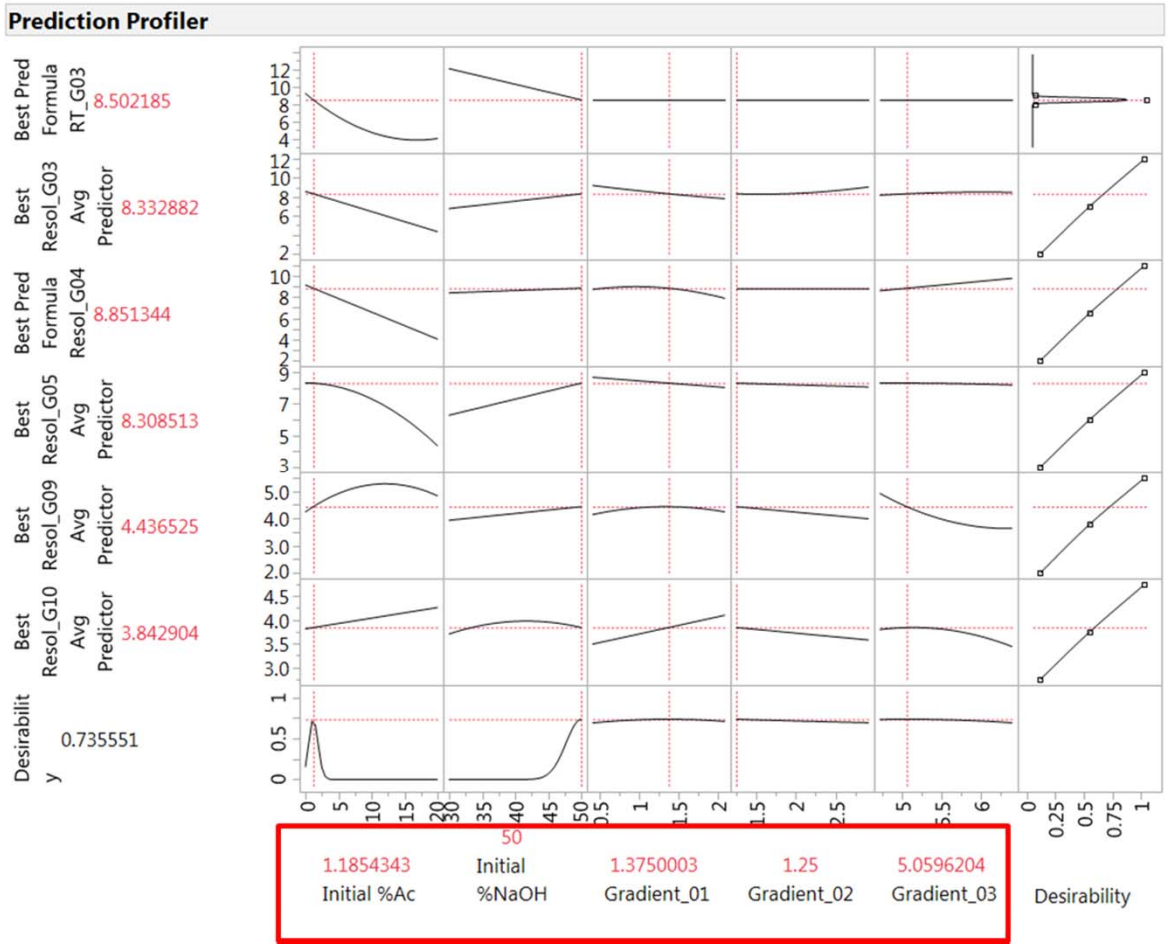
Also, we can assess **variable importance** and see that Initial %AC is by far the most important input to control.



Case Study: Glycoprofiling Proteins

Recall that there are 6 responses, so the modeling exercise was repeated for the other 5 responses and best models selected based upon RASE on the CCD data.

The profiler is then used to simultaneously optimize all 6 responses.



Conclusions on Case Studies

At this point it is difficult to make blanket recommendations from the case studies.

All Possible Models with AICc, using **JMP** did best in terms of selecting a 4 effect model with the lowest RASE on the CCD validation set.

All Possible Models with BIC using JMP tended to overfit models resulting in the best models (largest) generating "wild" predictions on the CCD validation set.

Both Prediction Averaging and Model Averaging resulted in models with relatively small RASE on the CCD validation set and performed similarly to the best AICc model.