DISCOVERY SUMMIT
EXPLORING DATA • INSPIRING INNOVATION
SEPTEMBER 15 - 18, 2014
SAS WORLD HEADQUARTERS - CARY, NC

Modeling and Analysis Strategies for Definitive Screening Designs using JMP Pro 11

Philip J. Ramsey, Ph.D.
North Haven Group &
University of New Hampshire
Durham, NH
pjrstats@aol.com

Mia Stephens, MS
JMP Academic Programs
SAS Institute, JMP Division
Cary, NC
mia.stephens@jmp.com

Outline

- Introduction
- Definitive Screening Designs
- Model Selection
- The Fermentation Experiments
- All Possible Models Analysis
- Prediction Averaging
- Model Averaging
- Relative Effect Importance
- Validation and Comparison
- Characterization and Optimization

Introduction

"The best way to predict the future is to create it." Peter Drucker.

The very efficient **Definitive Screening Designs** are receiving a lot of interest in bio-manufacturing due in part to the FDA and EMA recommendations on Quality by Design (QbD) requirements for pharmaceuticals and biologics.

QbD requires definition and characterization of a process **design space**, which entails modeling the relationships between process inputs and critical to quality attributes (CQA).

Design of experiments is the most efficient way (perhaps only way) to achieve this characterization during process development.

Experimental trials in biologics are costly and time consuming; e.g., a single trial can take one month to complete and cost $$$.

Definitive Screening Designs

A new class of screening designs were developed by Jones and Nachtsheim (2011a, 2011b) referred to as **Definitive Screening Designs** (DSD); the designs have been enhanced by Xiao (2012).

- For K factors DSDs require $2K+1$ runs if K is even and $2K+3$ if K is odd (to ensure main effect orthogonality).
- All factors are run at three levels in a factorial arrangement.
- Main effects are orthogonal and free of aliasing (partial or full) with quadratic effects and two-way interaction effects.
- No quadratic or two-way interaction effect is fully aliased with another quadratic or two-way interaction effect.
- It is possible to estimate every term of a full quadratic model, **but not in a single model**.

Prediction vs. Exploration

In building predictive models we have two competing issues:

- **Under-fitting** the model resulting in biased or inaccurate prediction;
- **Over-fitting** the model resulting in inflated prediction error.

Although the classic approach to the under- and over-fitting problem is to find a single, best compromise model, this is not necessarily an optimal strategy.

In **no way can one consider a single model to be correct**; a correct model only exists in a simulation study.

Modern computing power and statistical algorithms available allow us to look for models, or a combination of models, that **best predict** the behavior of the physical system.

Model Selection

Two widely accepted measures of fit for a model are:

AICc = bias corrected Akaike Information Criterion;

BIC = Bayesian Information Criterion;

Both criteria punish under- and over-fitting, but in a different way. So, they may not agree on the best model(s) – they often do not (see Burnham and Anderson, 2002).

There is not agreement in the statistical community as to whether AICc or BIC criterion is preferred; it almost surely depends on the application.

For both the AICc and BIC **smaller values indicate better predictive models**.

Model Selection

The AIC is derived from information theory by Hirotugu Akaike in 1974 and is related to **Boltzmann entropy** from thermodynamics.

The AICc measures the information in the data lost by fitting a model, therefore models with smaller AICc are preferred.

AICc can subsequently be used to rank a set of fitted models in terms of their relative support given the data.

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.

Δ_i estimates the **Kullback-Leibler distance or divergence** between two models and models with large Δ_i values are not considered further.

Model Selection

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**.

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

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

The B_i can be interpreted as a natural log **Bayes Factor** (Kass and Raftery, 1996).

The Fermentation Experiments

We now focus on an experiment to optimize the biomolecule **Yield** of the fermentation step in a bio-process.

For the experiment $K = 5$ factors were identified:

1. **pH** (6.8, 7.2) = fermentation solution pH;
2. **Dissolved Oxygen** (%DO) (target values 20%, 40%);
3. **Induction Tempe** (39.5 C, 42.5 C) = Temperature at which the biomolecule production is induced in the E. Coli cells.
4. **Induction OD_{600}** (20, 40) = biomass at which the induction is initiated as measured by optical density at 600 nm.
5. **Feed Rate** (1.9, 3.5 mL/hr) = feed rate of a growth media containing 50% glycerol added to the fermentation solution when induction is initiated.

The Fermentation Experiments

The two goals of the experiment were to **characterize the fermentation** step and to **maximize the Yield** of a biomolecule (X) produced by E. Coli cells.

Note: The goal was not necessarily to maximize the mass of the E. Coli community. It is possible to substantially increase the mass of a microbial community without maximizing biomolecule production.

The three responses of interest are:

1. **Yield** = biomolecule titer measured in units of mg/L;
2. **OD600** = measure of biomass by optical density at 600 nm;
3. **WCW** = wet cell weight in units of g/L.

Yield was the primary response and the focus of our analysis.

Analysis of the Fermentation Experiment

We discuss and compare **5 modeling strategies**.

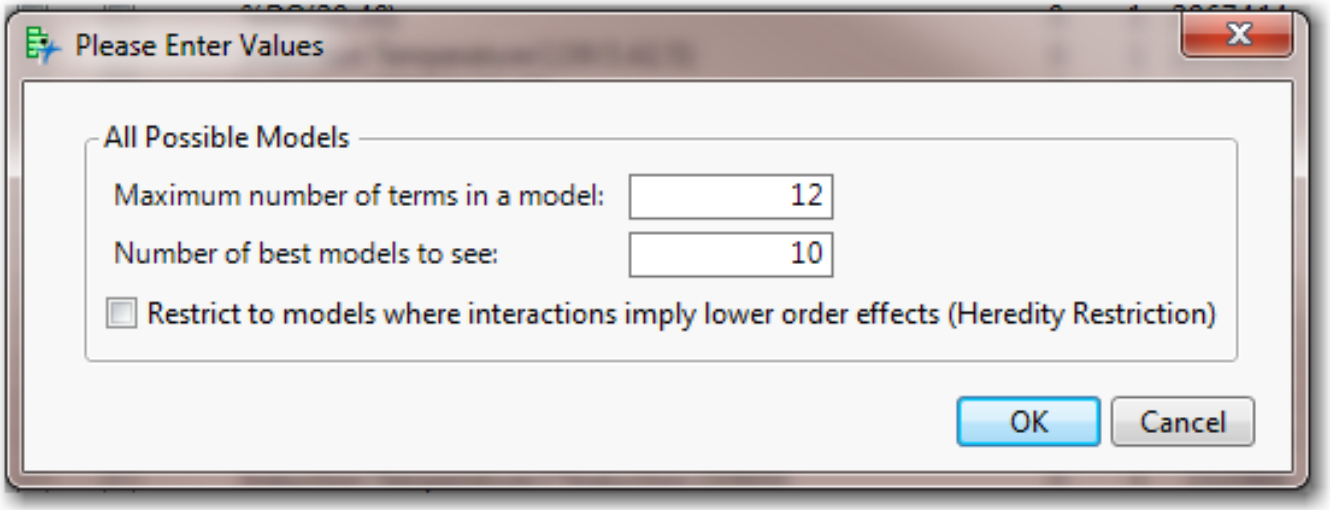
The first four strategies are based upon **All Possible Models** regression and the last method is based on the adaptive LASSO.

1. Find a *single best predictive model* using AICc and BIC.
2. *Prediction Averaging*: find a subset of best models using AICc and BIC and use the average prediction from the models.
3. *Model Averaging*: fit the entire full quadratic model using coefficient averaging across the large set of all possible models.
4. *Relative Effect Importance* based upon exponential AICc weights (see Burnham and Anderson, 2002, page 167).
5. Generalized regression using the *Adaptive LASSO*.

All Possible Models Analysis

To initiate the All Possible Models analysis (**we'll do this in JMP**):

- Use *Analyze* > *Fit Model*
- Use *Macros* > *Response Surface* to fit a full quadratic model
- Change the *Personality* to *Stepwise* and click on *Run*
- Select *All Possible Models* from the top red triangle



All Possible Models Analysis

Once the Fit Group is formed, the 8 models can be compared using residual plots, Actual by Predicted plots, Lack of Fit tests, Press, etc.

Press (Prediction Error sum of Squares) is a measure of how well a model might predict if applied to validation data not used to fit the model – smaller values are preferred.

Caution, Press does not necessarily protect against over fitting but is useful in comparing a set of competing models.

The Fit Group can then be used to Prediction Average or to select a best model.

We now examine both the individual models and prediction averaging based on the 8 selected models.

All Possible Models Analysis

The best Δ_{min} and B_{min} of the 8 candidate models models are highlighted in the table below.

The smallest Press occurs for one of the B_{min} model, which is smaller than even the other two 8 effect models.

Number	MS(Press)	AICc	BIC	Δ_i	w_i	B_i	BIC wts
4	61.43	170.42	156.61	2.76	0.2514	12.67	0.0018
4	62.71	171.51	165.26	3.85	0.1457	13.76	0.0010
5	45.74	167.66	156.61	0.00	1.0000	5.12	0.0775
5	48.35	169.75	158.70	2.09	0.3516	7.21	0.0273
7	41.79	182.75	153.13	15.10	0.0005	1.63	0.4428
8	18.21	199.42	151.50	31.76	0.0000	0.00	1.0000
8	33.43	200.98	153.06	33.33	0.0000	1.57	0.4570
8	31.93	201.12	153.20	33.47	0.0000	1.70	0.4265

Prediction Averaging

Prediction averaging is a form of multimodel inference in which the predictions across the subset of models are averaged.

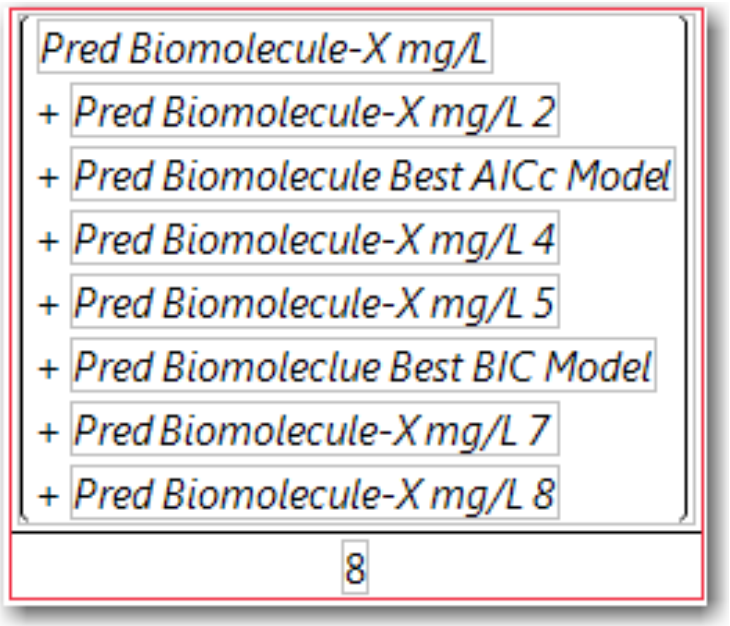
The averages may be arithmetic or weighted by some criterion such as AIC weights or the analogous BIC weights.

- An advantage of averaging predictions over a set of models is that it provides *protection against over and under fitting*.
- The prediction averaging has the effect of *shrinking the prediction error* that occurs if one under or over fits a single model.

Prediction Averaging

In this case, we will use the arithmetic average of the predictions from the 8 models selected as best, so models with 4 through 8 effects are used to generate the average.

A simple formula is used to generate the average using the JMP **Formula Editor**.



Note: In JMP Pro the prediction average can be generated in the *Model Comparison* platform using *Model Averaging*.

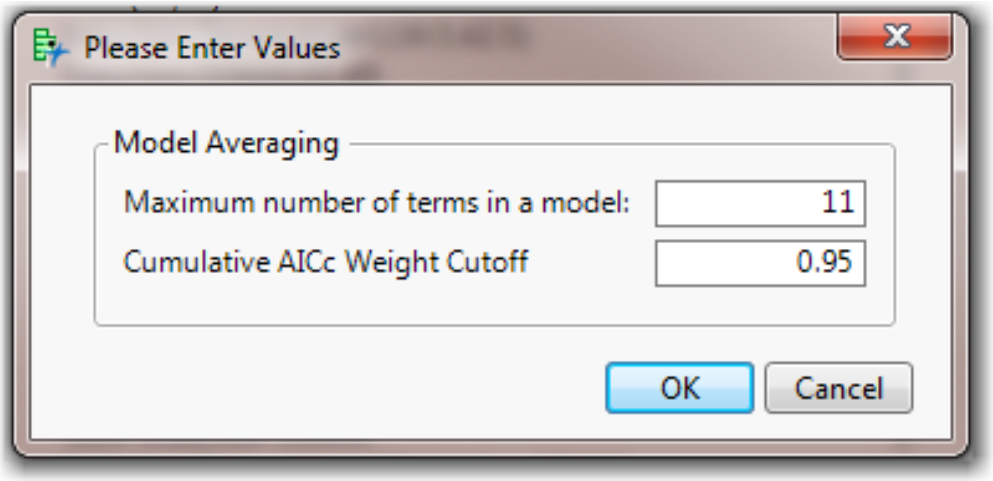
Model Averaging

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.



Model Averaging

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

In this case, the full quadratic model is being fit (in only 11 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, 2002) and (2004) for further discussion of model averaging.

Model Averaging

Here we estimate the full quadratic model using the Model Averaging option in the Stepwise platform.

The technique makes it possible to fit a full quadratic model from the supersaturated DSD.

The estimated coefficients are qualitatively quite reasonable.

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

Model Averaging

Averaging models with 1 to 11 terms, using a cutoff AICc weight quantile of 0.95, which resulted in using 1502 out of 151807 models fit

Parameter	Estimate	Std Error
Intercept	354.421	.
pH(6.8,7.2)	-15.209	8.26125
%DO(20,40)	-28.700	11.01958
Induction Temperature C(39.5,42.5)	-1.801	4.63338
Induction OD600(20,40)	0.600	3.30131
Feed rate(1.9,3.5)	91.112	15.61479
pH*pH	-2.502	6.97003
pH*%DO	16.890	11.45002
%DO*%DO	-14.991	11.94275
pH*Induction Temperature C	-4.197	7.24643
%DO*Induction Temperature C	15.231	10.09181
Induction Temperature C*Induction Temperature C	-6.428	9.40379
pH*Induction OD600	0.273	4.99308
%DO*Induction OD600	0.216	4.99105
Induction Temperature C*Induction OD600	0.774	5.27164
Induction OD600*Induction OD600	1.623	7.09676
pH*Feed rate	-1.476	5.68849
%DO*Feed rate	2.565	6.28416
Induction Temperature C*Feed rate	-23.455	11.03514
Induction OD600*Feed rate	-0.892	5.38469
Feed rate*Feed rate	-37.325	15.35863

Relative Effect Importance Modeling

Relative Effect Importance Modeling is used to determine which model effects have the highest importance in terms of prediction.

The technique is discussed in Burham and Anderson (2002, pg. 167) and involves fitting a set of models; perhaps All Possible Models.

Using their notation let w_{j+} represent the sum of the AICc weights w_i (or possibly other types of weights) across all of the fitted models where the j^{th} effect occurs.

Effects with higher w_{j+} values have higher relative importance compared to the effects with smaller values.

A Pareto type plot can then been used to show visually show the relative effect importance values (**we'll show this in JMP**).

Relative Effect Importance Modeling

Although this plot is informative by itself for scientists, it can also be used to fit models based on the most important effects.

A model comprised of the 5 most important effects had the lowest AICc, BIC, and MS(Press) among this reduced set of possible models.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	360.56286	11.61172	31.05	<.0001*
Feed rate(1.9,3.5)	91.955556	11.44931	8.03	<.0001*
%DO(20,40)	-45.38889	11.44931	-3.96	0.0033*
Feed rate*Feed rate	-82.20377	16.23377	-5.06	0.0007*
pH(6.8,7.2)	-34.03611	11.44931	-2.97	0.0156*
Feed rate*Induction Temperature C	-59.53864	13.09979	-4.55	0.0014*

Note: One has to be cautious in that a model comprised of a set of most important effects **could potentially be singular** for a given combination of quadratic and interaction effects.

Generalized Regression Modeling

An additional approach that can be used is a generalized or **regularized regression modeling**, which is available in JMP Pro.

The Lasso is a shrinkage and selection method for linear regression.

As usual, sum of squared errors is minimized.

However, a bound on the sum of the absolute values of the coefficients is used to help reduce the impact of over fitting.

There are many variations of Lasso (see Tibshirani, 1996).

We restrict our focus to the ***Adaptive Lasso with BIC for validation***, which results in an 11 effect model.

Note: The use of generalized regression for model selection in *supersaturated designs* is discussed by Marley and Woods (2010).

Validation and Model Comparison

An entire **31 run Central Composite Design** was run in parallel with the DSD as a basis for validation and comparison.

This allows a comparison of the best CCD models, using the same modeling strategies, vs. the models estimated from the DSD.

Using All Possible Models with the CCD data results a 6 effect model having $\Delta_{min} = 0$ and $B_i = 1.1$ (well within the cutoff of 2.0).

We will **validate and compare the DSD models by determining how they performed when applied to the CCD data**.

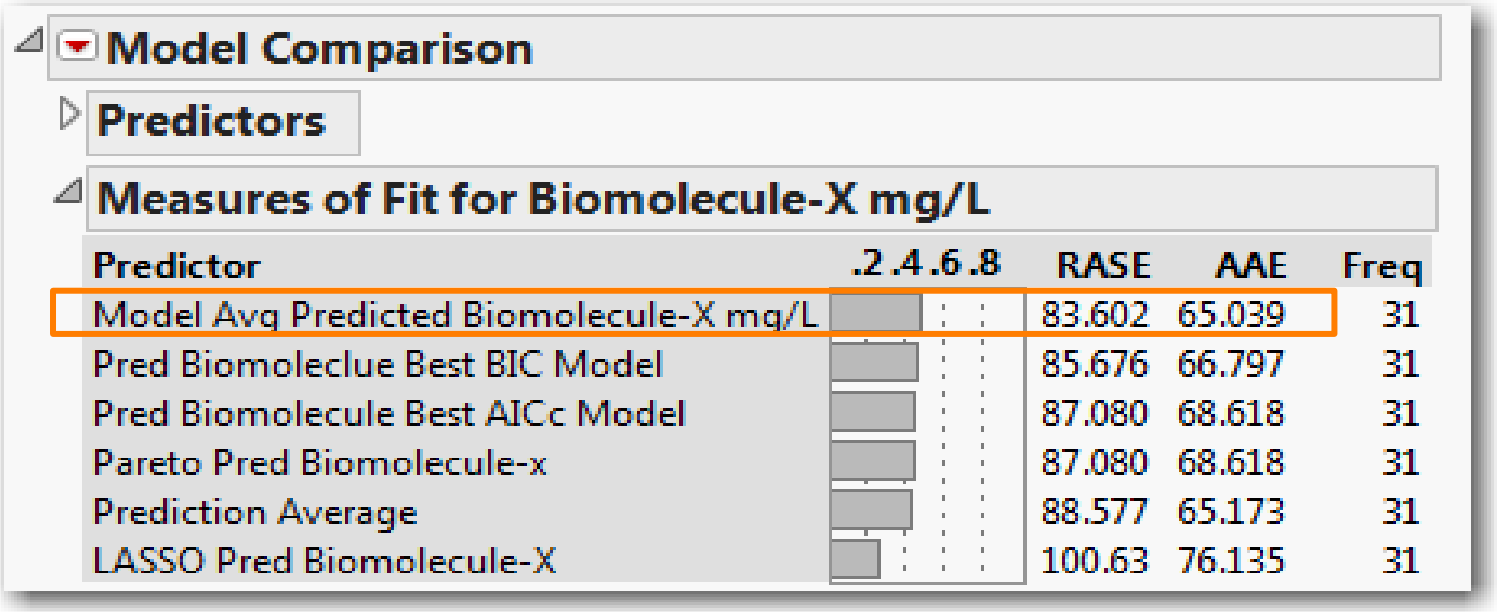
Two statistic used to compare models in validation sets are the *root average squared error (RASE)* and the *average absolute error (AAE)*, which are both measures of the prediction error on the validation trials.

Validation and Model Comparison

We use the **Model Comparison** platform in JMP Pro to compare our models our 6 models based on the 31 CCD validation trials.

The **Model Average** and the **Best BIC** models performed best.

All models actually did well, with the exception of the LASSO.



Model Comparison

Predictors

Measures of Fit for Biomolecule-X mg/L

Predictor	.2 .4 .6 .8	RASE	AAE	Freq
Model Avg Predicted Biomolecule-X mg/L		83.602	65.039	31
Pred Biomoleclue Best BIC Model		85.676	66.797	31
Pred Biomolecule Best AICc Model		87.080	68.618	31
Pareto Pred Biomolecule-x		87.080	68.618	31
Prediction Average		88.577	65.173	31
LASSO Pred Biomolecule-X		100.63	76.135	31

Characterization and Optimization

It is also useful to compare all of the potential models in terms of how they characterize the fermentation process and the suggested factor settings to maximize Yield.

The **Prediction Profiler** has two excellent tools to accomplish these tasks: the *Desirability Functions* and the *Variable Importance* analysis.

Variable Importance analysis uses monte carlo techniques to ascertain the relative importance of the factors in the model both as main effects and as combined effects with interactions, etc. (see Sobol, 1993 and Saltelli, 2002).

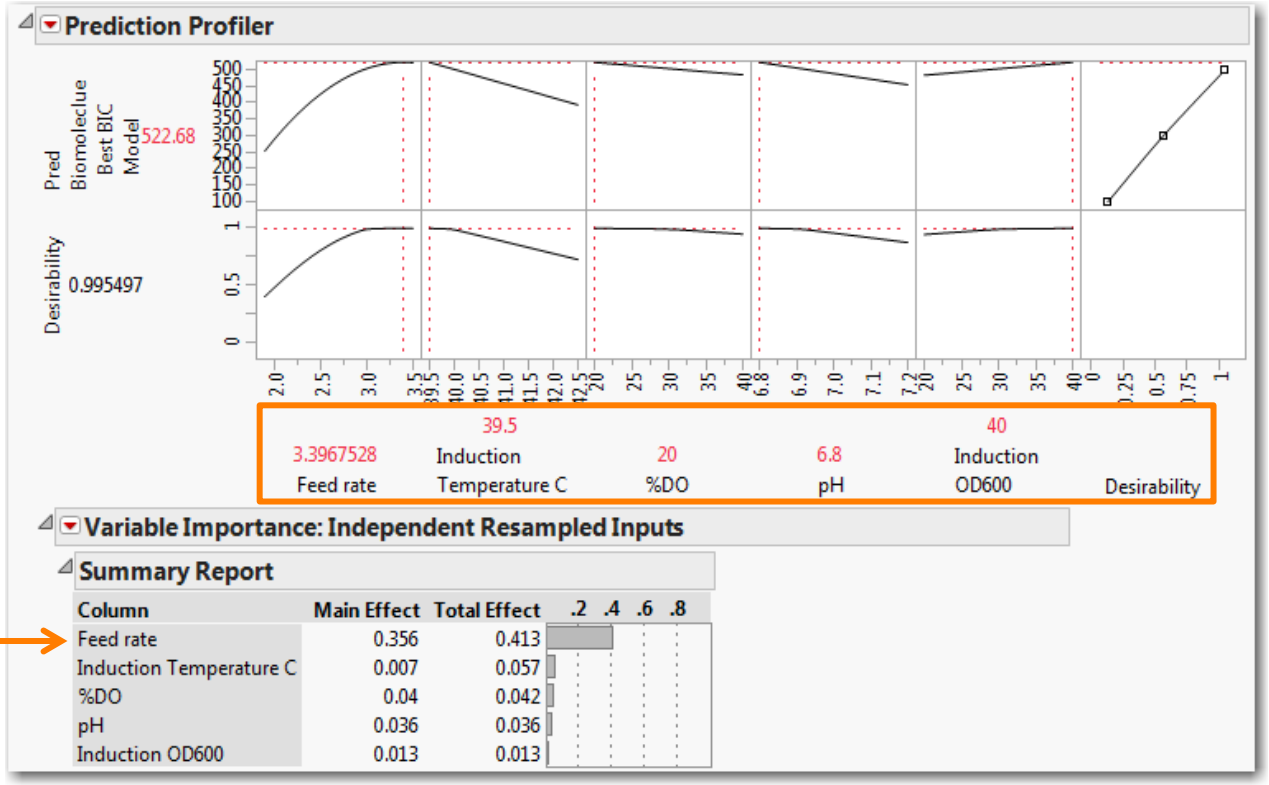
We will examine all of the models in terms of variable importance and suggested optimum settings for the process factors.

Characterization and Optimization

The factor settings to optimize Yield and the variable importance report for the best 8 effect BIC model are shown.

The predicted max average Yield is 523.

Feed rate is the most important input affecting Yield.



Characterization and Optimization

All models lead to the same conclusions concerning factor importance.

BIC Model: Variable Importance: Independent Resampled Inputs

Summary Report

Column	Main Effect	Total Effect
Feed rate	0.356	0.413
Induction Temperature C	0.007	0.057
%DO	0.04	0.042
pH	0.036	0.036
Induction OD600	0.013	0.013

Model Avg: Variable Importance: Independent Resampled Inputs

Summary Report

Column	Main Effect	Total Effect
Feed rate	0.425	0.427
%DO	0.044	0.053
Induction Temperature C	0.005	0.017
pH	0.015	0.015
Induction OD600	0.004	0.004

Pred Avg: Variable Importance: Independent Resampled Inputs

Summary Report

Column	Main Effect	Total Effect
Feed rate	0.356	0.373
%DO	0.081	0.088
Induction Temperature C	0.01	0.036
pH	0.029	0.029
Induction OD600	0.007	0.007

AIC Model: Variable Importance: Independent Resampled Inputs

Summary Report

Column	Main Effect	Total Effect
Feed rate	0.353	0.404
%DO	0.065	0.065
Induction Temperature C	0.007	0.058
pH	0.037	0.037

Pareto Model: Variable Importance: Independent Resampled Inputs

Summary Report

Column	Main Effect	Total Effect
Feed rate	0.355	0.407
%DO	0.066	0.066
Induction Temperature C	0.001	0.054
pH	0.038	0.038

LASSO: Variable Importance: Independent Resampled Inputs

Summary Report

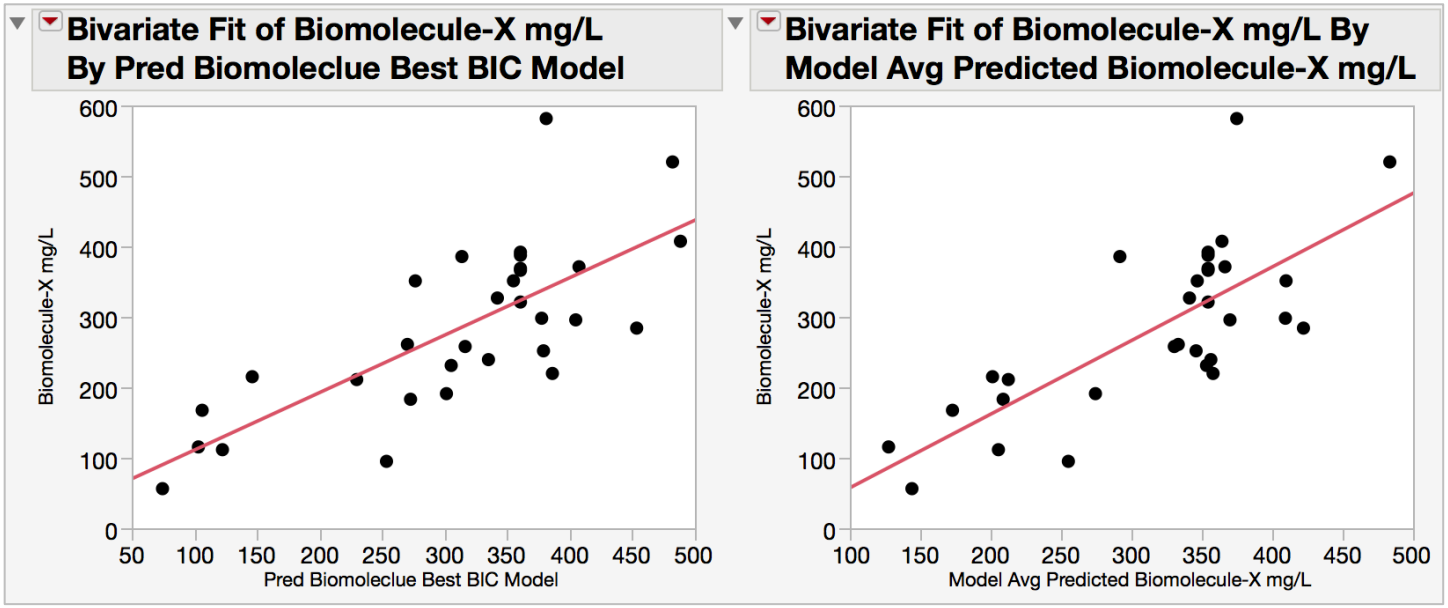
Column	Main Effect	Total Effect
Feed rate	0.332	0.332
%DO	0.09	0.14
Induction Temperature C	0.008	0.05
pH	0.024	0.024
Induction OD600	0.006	0.006

Characterization and Optimization

It is also helpful to look at **Actual by Predicted** plots on the CCD validation trials to confirm that the prediction is unbiased.

Ideally, the fitted line will have an intercept 0 and slope 1, with no obvious lack of fit patterns.

The selected models (below) are predicting well on the CCD trials.



Conclusions and Recommendations

Based on the evaluation of the various modeling strategies for DSDs the **best BIC model** and **Model Averaging** performed best in terms of predicting the results of the 31 CCD trials.

However, the best AICc model and the prediction average models performed nearly as well.

The **Lasso** approach performed poorly and is generally not a good strategy to use with the small sample sizes in DSDs.

For process characterization and optimization in terms of variable importance and desirability (and control strategies for a process) all of the modeling strategies (except Lasso) lead to similar conclusions.

Conclusions and Recommendations

Although we do not show the results, the analysis of the **CCD** experiment, using the same modeling strategies, **leads to similar conclusions** in terms of variable importance and optimization.

This reinforces that the DSDs, combined with proper modeling strategies, are a cost effective, viable alternative to the much larger response surface designs.

We urge those doing research in design of experiments to spend a good bit more time studying analysis strategies.

Burnham and Anderson (2002, 2004) point out that model selection strategies have not been properly studied.

Conclusions and Recommendations

Finally:

Validation is a critical part of modern experimentation and must be emphasized more strongly in Design of Experiments training.

This is particularly true when the objective of the experiment is prediction of future performance.

More work needs to be done on what constitutes sound validation strategies.

For pharmaceuticals and biologics, validation will be increasingly important as QbD methodology becomes mainstream and possibly required in the future by the FDA, EMA, etc.

References

Bickel, P. J., Ritov,Y. and Tsybakov,A. (2008) *Simultaneous Analysis of Lasso and Dantzig selector*. Ann. Statist.

Burnham, K P, and D R Anderson (2002). *Avoiding Pitfalls When Using Information -Theoretic Methods*. Journal Of Wildlife Management 66 (3):912-918.

Burnham, K P, and D R Anderson (2002). *Model Selection and Multimodel Inference, Second Edition*. Springer: New York.

Burnham, K P, and D R Anderson. *Multimodal Inference: Understanding AIC and BIC in model selection*. Sociological Methods and Research 33, Nr. 2 (2004): 261-304.

Burnham, K. P., Anderson, D. R. and Huyvaert, K (2011). *AIC Model Selection and Multimodel Inference in Behavioral Ecology: Some Background, Observations, and Comparisons*. Behav Ecol Sociobiol 65:23-35.

Erler, A., de Mas, N., Ramsey, P., Henderson, G. (2013). *Efficient biological process characterization by definitive-screening designs: the formaldehyde treatment of a therapeutic protein as a case study*. Biotechnology Letters, 35, 3.

FDA, CDER, CBER, and CVM (2011). Process Validation: General Principles and Practices. Guidance for industry.

International Conference On Harmonization (ICH) Guidelines: http://www.ich.org/products/guidelines/quality/article/quality-guidelines.html

Goos, P, and Jones, B. *Optimal Design of Experiments: A Case Study Approach*. John Wiley & Sons, LTD, 2011.

References

Hansen, B.E. (2008). *Least-squares Forecast Averaging*. Journal of Econometrics 146, p. 342-350.

Kass, R.E. and Raftery, A.E. (1995). *Bayes Factors*. Journal of the American Statistical Association 90:430, 773-795.

Jones, B. and Nachtsheim, C. (2011a). *Efficient Designs With Minimal Aliasing*. Technometrics, 53, 1, 62 – 71.

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

Marley, C.J. and Woods, D.C. *A Comparison of Design and Model Selection Methods for Supersaturated Experiments*. Computational Statistics and Data Analysis 54 (2010) 3158 – 3167.

Saltelli, A. (2002), *Making best use of model evaluations to compute sensitivity indices*, Computer Physics Communications, 145, 280–297.

Sobol, I. M. (1993), *Sensitivity estimates for nonlinear mathematical models*, Mathematical Modeling and Computational Experiment, 407–414.

Tibshirani, R. (1996). *Regression shrinkage and selection via the lasso*. J. Royal. Statist. Soc B., Vol. 58, No. 1, pages 267-288).

Xiao, L, Lin, D, and Bai, F. *Constructing Definitive Screening Designs Using Conference Matrices*. Journal of Quality Technology 44, Nr. 1 (2012): 2-8.