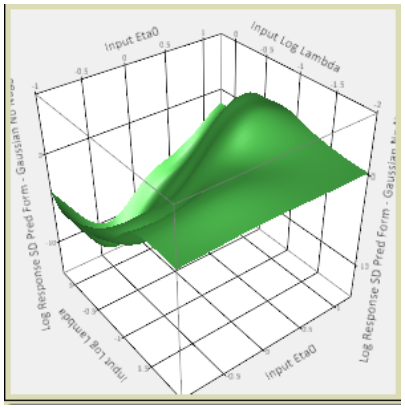
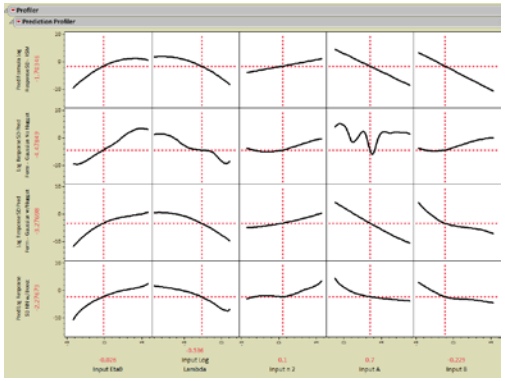




Meta Modeling of Computational Models

Challenges and Opportunities

Bill Worley and Cy Wegman

Procter & Gamble Co.

Objectives

1. Introduce simulation experiment issues.
2. Recognize limitations in analysis techniques.
3. Recognize limitations in traditional DOX approach.
4. Discuss potential solutions to these limitations.

Unique Issues with Computational Experiments

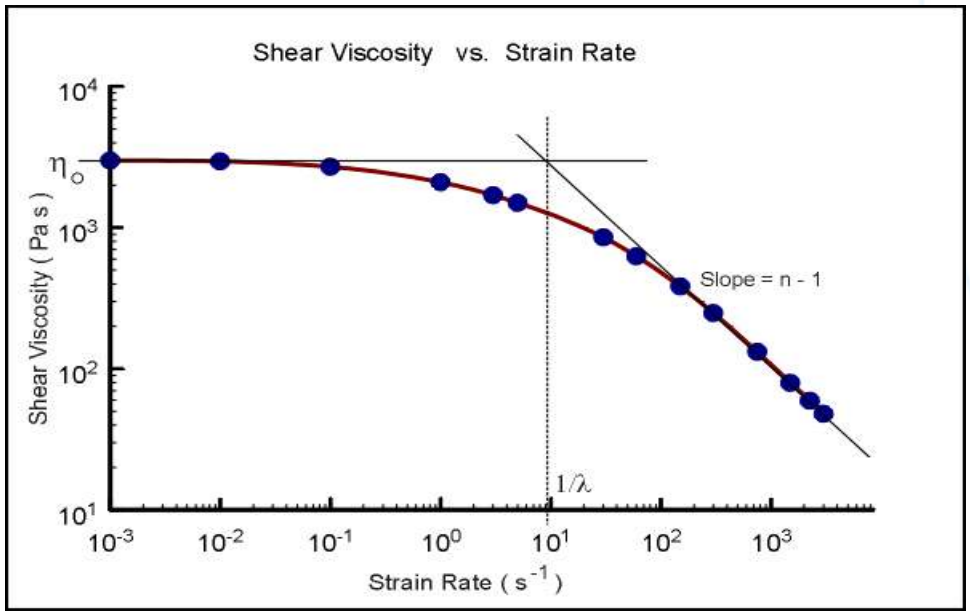
- No replication error
  - Can cause issues with regression analysis
- May have issues with summary characterization
- There may be issues with lack of full convergence
- May be highly non-linear
  - May have issues with bias error
- Use of non-parametric fitting techniques can cause interpretation issues
- Use of non-parametric fitting techniques may cause model over fit
- Computational model may not be representative of reality.

Problem Description

- A typical simulation will have millions of data points of species data. In this case, species is the fraction of two materials being mixed.
- Standard deviation of all the data is used to characterize the performance of a mix tank.
- Standard deviation varies as a function of the material properties.

Experimental Factors

- Rheology Model
  - 3 Rheology parameters may be varied in simulation
    - η_0 - zero shear viscosity
    - λ – time constant calculated from the reciprocal of the strain rate
    - n – power-law component of the flow curve intersect



Problem Definition

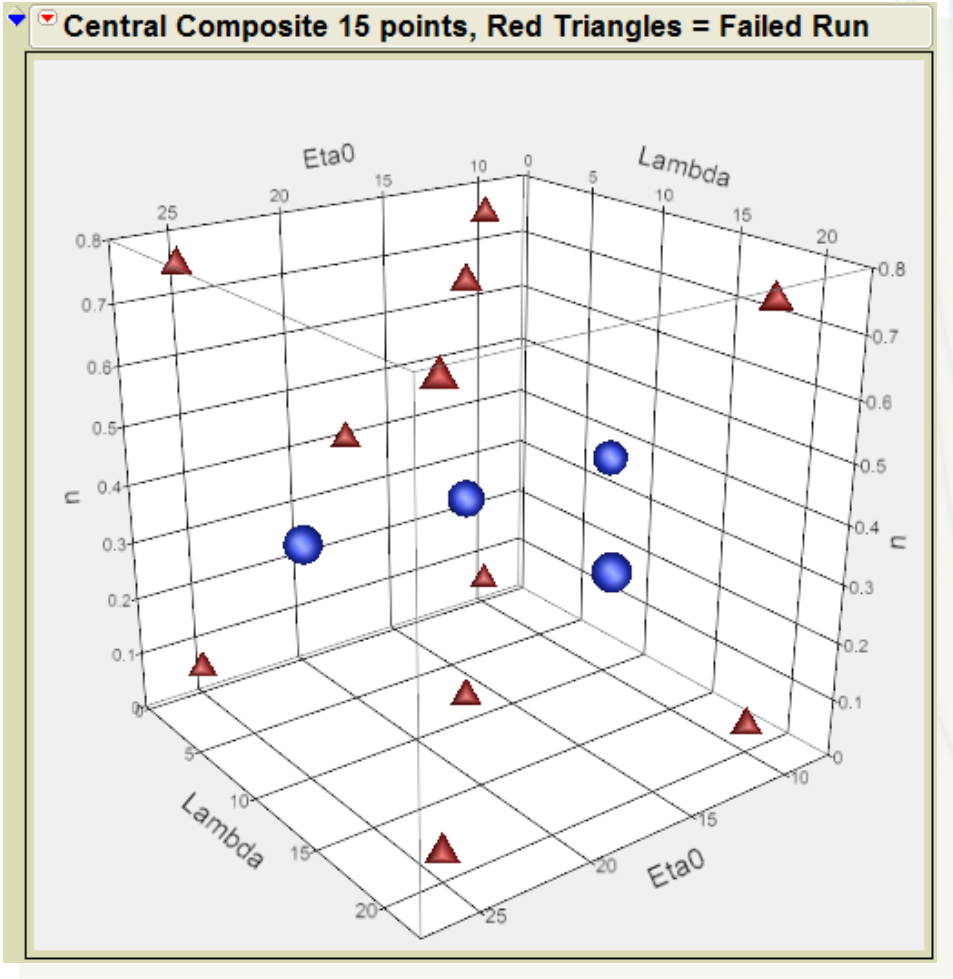
– Define and predict standard deviation at 10 minutes blend time in a mix tank as a function of the three rheology parameters.

Rheology Model Inputs for DOX Design

Factor	Low	High
Eta0	-1	+1
Lambda	-1	+1
n	-1	+1

Response is ‘Standard Deviation’ at 10 minutes

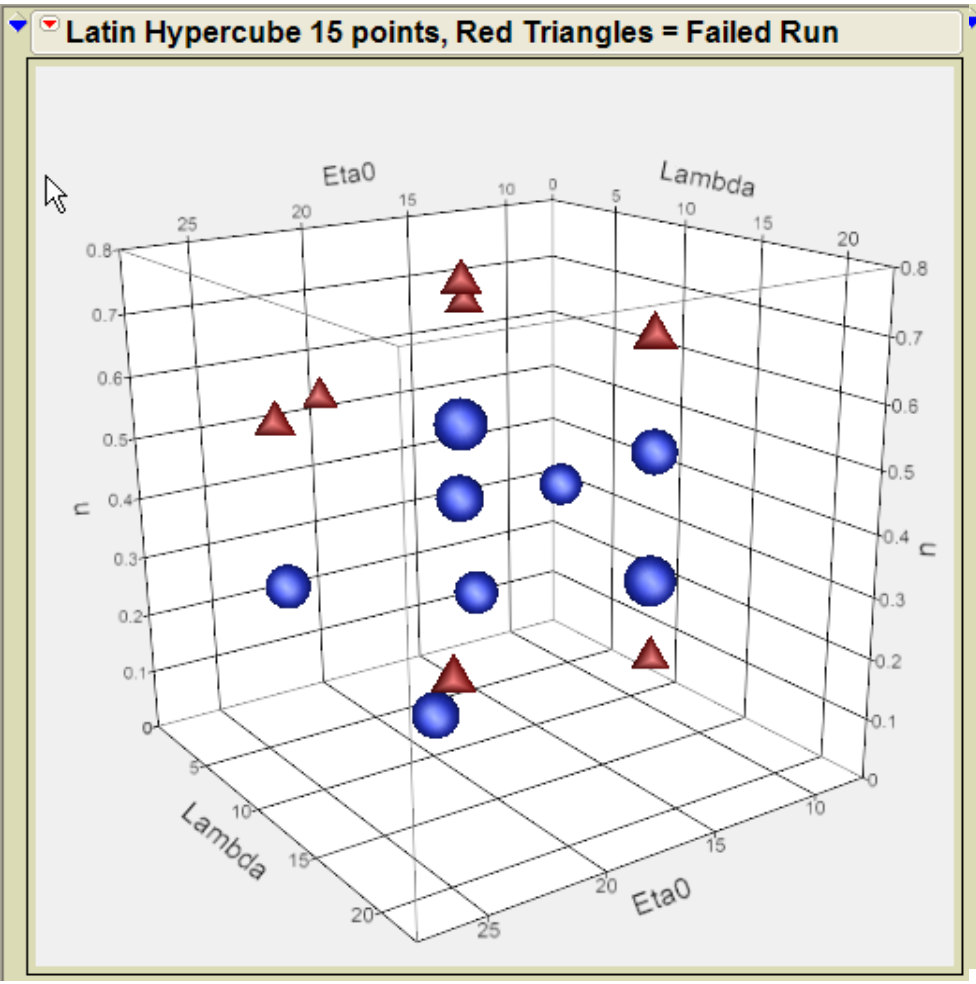
Example Problem Review
Central Composite 3D Scatterplot

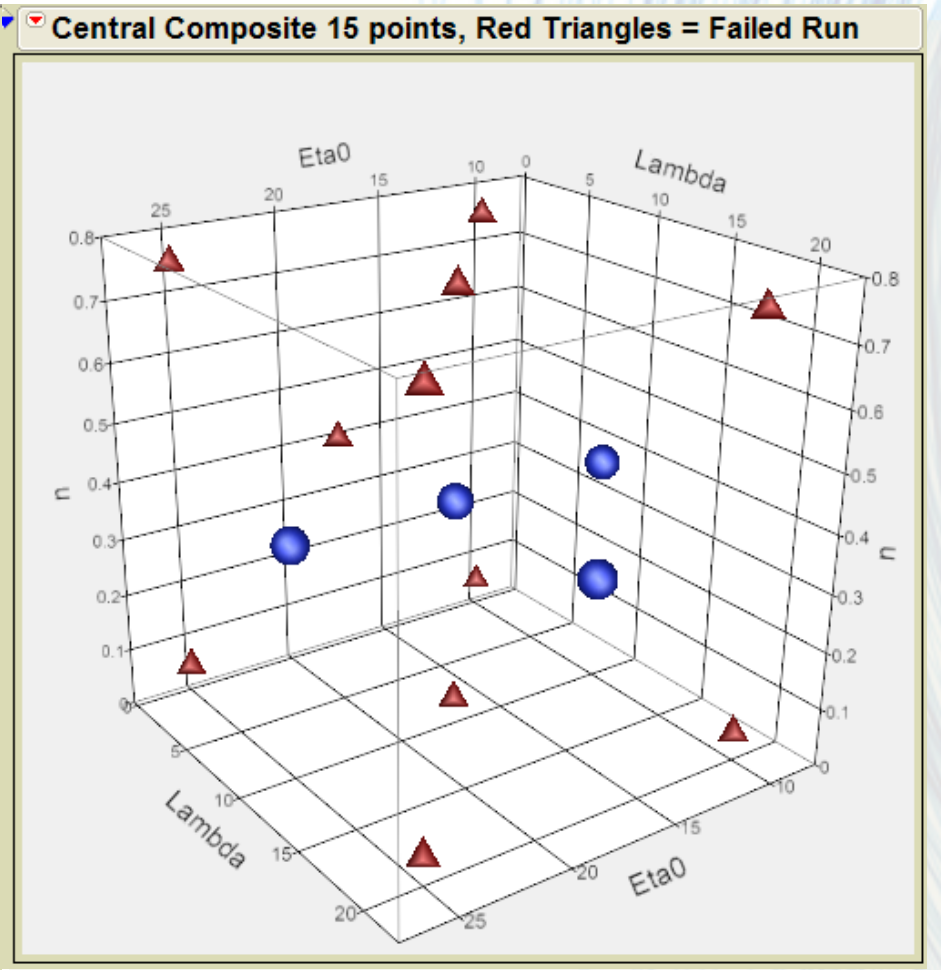


Another Option

- Develop a space filling experimental design based on selected parameter limits.
- Run simulations based on this design with standard deviation being the response.

Comparison of Space Filling and RSM Design Space



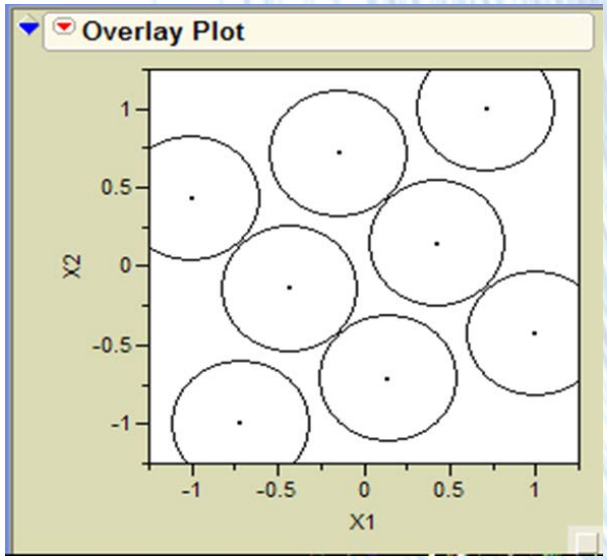


Potential Solutions

- Further constrain the low to high ranges of each factor and augment the design
- Determine if there are any linear equations that can be used to further constrain the design and then use DOE Augment to add additional conditions.
- Add more data points in the region where convergence occurs.

Space Filling - Main Characteristics

- Useful where the form of the model is more important than concern for run-to-run variability.
- For systems with no variability, randomization and blocking are irrelevant. No need for repeating the same run because the results will be the same.
- Primary use in computer simulation.
- Latin Hypercube is well known for space filling design option.



Potential Issues

- Many computational models have highly nonlinear behavior.
- In the interest of keeping the model simple; If the response over the experimental space is smooth and can be approximated well with a polynomial then choose this type of design.
- If the response is highly nonlinear, standard designs will be inadequate.
- Stochastic systems, e.g. Extend through put models, may be best estimated with standard designs.

Initial Number of Runs

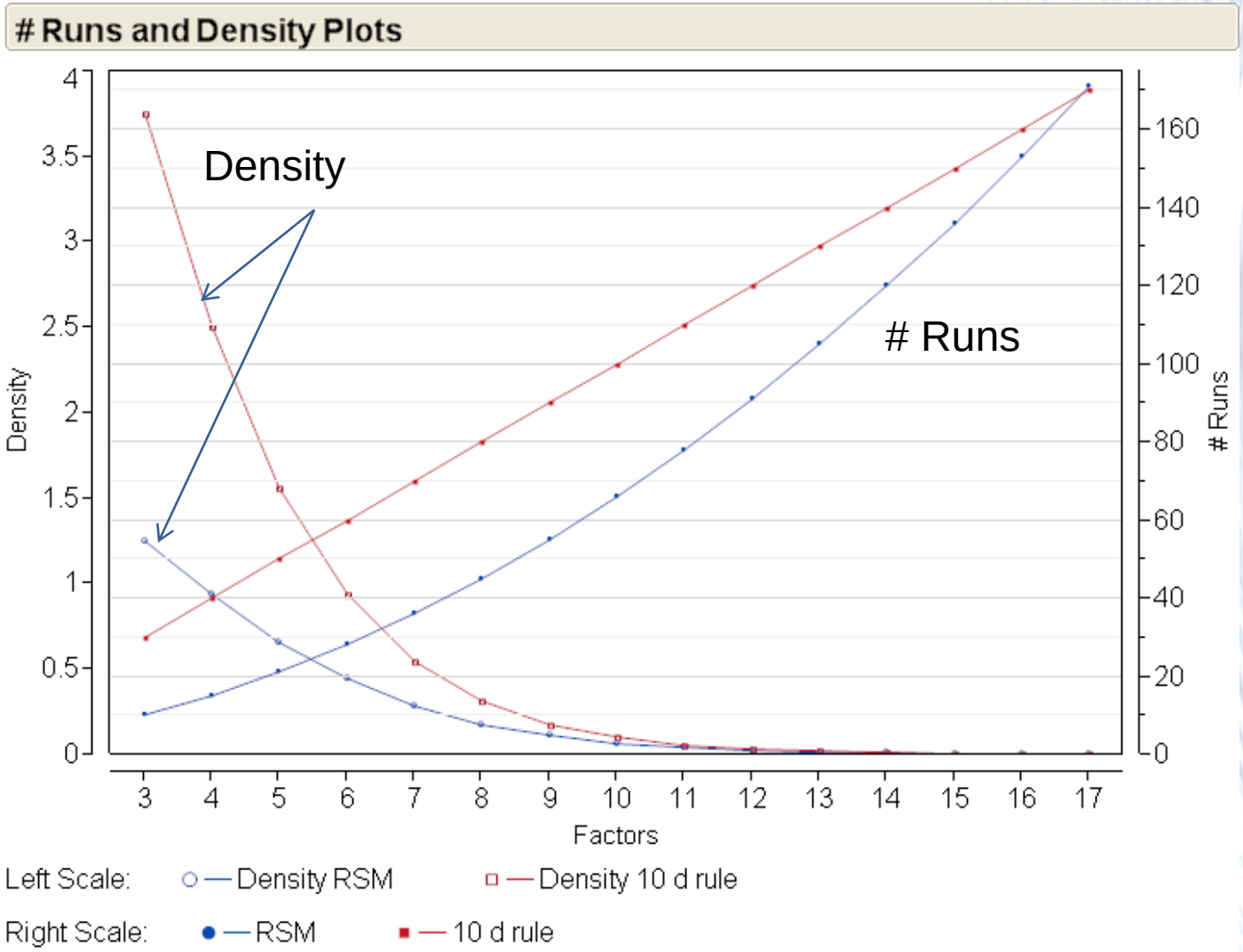
Rules of Thumb on Number of RSM Parameters

- Where N = Number of factors
- Recommended Runs = 10*N

If you can't afford this then:

- Recommended Runs = Intercept (1) + Main Effects (N) + Quadratic Effects (N) + Interactions (N*(N-1) / 2).
  - This is the number of terms in a response surface model

of Runs & Density of Points



OK – I Have Data, Now What?

What Is The Best Fit Model For My Data?

Modeling Comparison

1. Discuss Gaussian and Neural Net Fit Models.
2. Recognize the advantages and disadvantages of these fit models versus a Response Surface model.

Gaussian Models

- Models the response between a continuous response and one or more continuous predictors.
- These models are commonly used for computer simulation experiments and they often perfectly fit the data points.
- Gaussian processes can deal with no-error-term models, in which the same input values always result in the same output value.
- The main purpose of using this platform is to obtain a prediction formula that can be used for further analysis and optimization.

Model Options in Gaussian Fit

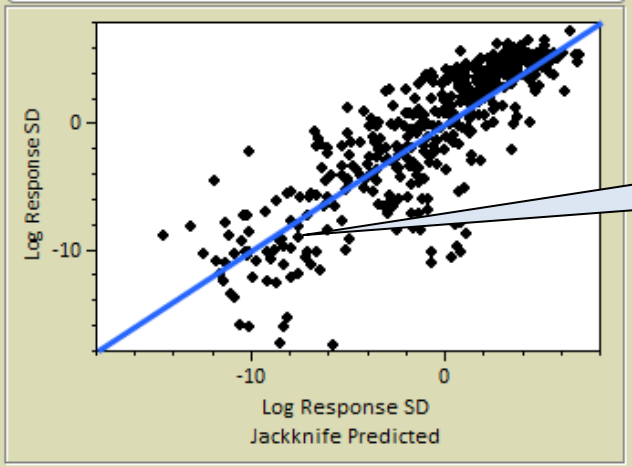
- **Estimate Nugget Parameter** - This is useful if there is noise or randomness in the response, and you would like the prediction model to smooth over the noise instead of perfectly fitting.
- **Correlation Type** - lets you choose the correlation structure used in the model.
  - **Gaussian** allows the correlation between two responses to always be non-zero, no matter the distance between the points.
  - **Cubic** allows the correlation between two responses to be zero for points far enough apart.
  - **Minimum Theta Value** - allows you to set the minimum theta value used in the fitted model.

Gaussian Process for Space Filling Designs

Scaling Meta Model - Gaussian Process of Log Response SD - JMP Pro [2]

Gaussian Process Model of Log Response SD

Actual by Predicted Plot



Ideally, the data points fit tightly on a 45 degree line

Model Report

Column	Theta	Total Sensitivity	Main Effect	Input Eta0 Interaction	Input Log Lambda Interaction	Input n 2 Interaction	Input A Interaction	Input B Interaction
Input Eta0	2.5572947	0.2576159	0.1741162	.	0.0286391	0.0022158	0.0450819	0.0075629
Input Log Lambda	7.8079059	0.3073644	0.1796935	0.0286391	.	0.0256219	0.0696167	0.0037932
Input n 2	0.6215812	0.0624372	0.0156556	0.0022158	0.0256219	.	0.0179156	0.0010283
Input A	42.081389	0.2532285	0.0936381	0.0450819	0.0696167	0.0179156	.	0.0269762
Input B	1.0496321	0.0732001	0.0338395	0.0075629				.

μ	σ^2
0.2065513	24.125521

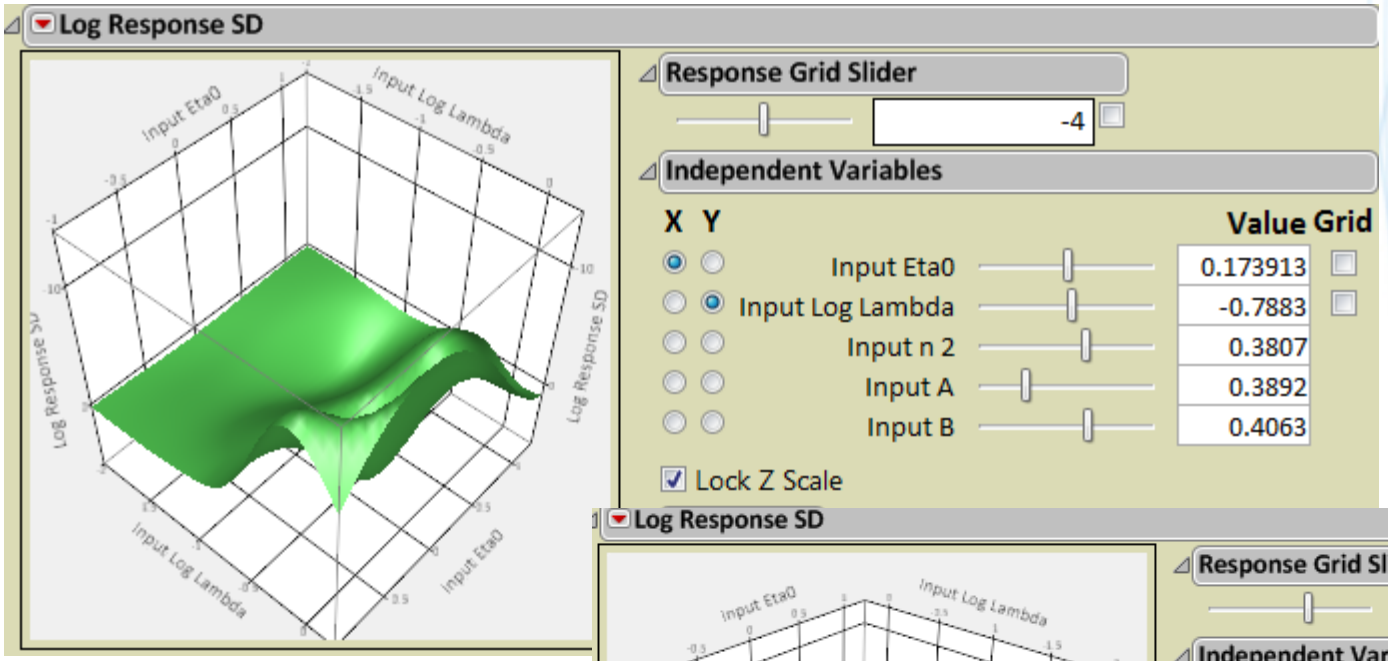
-2*LogLikelihood

2238.9666

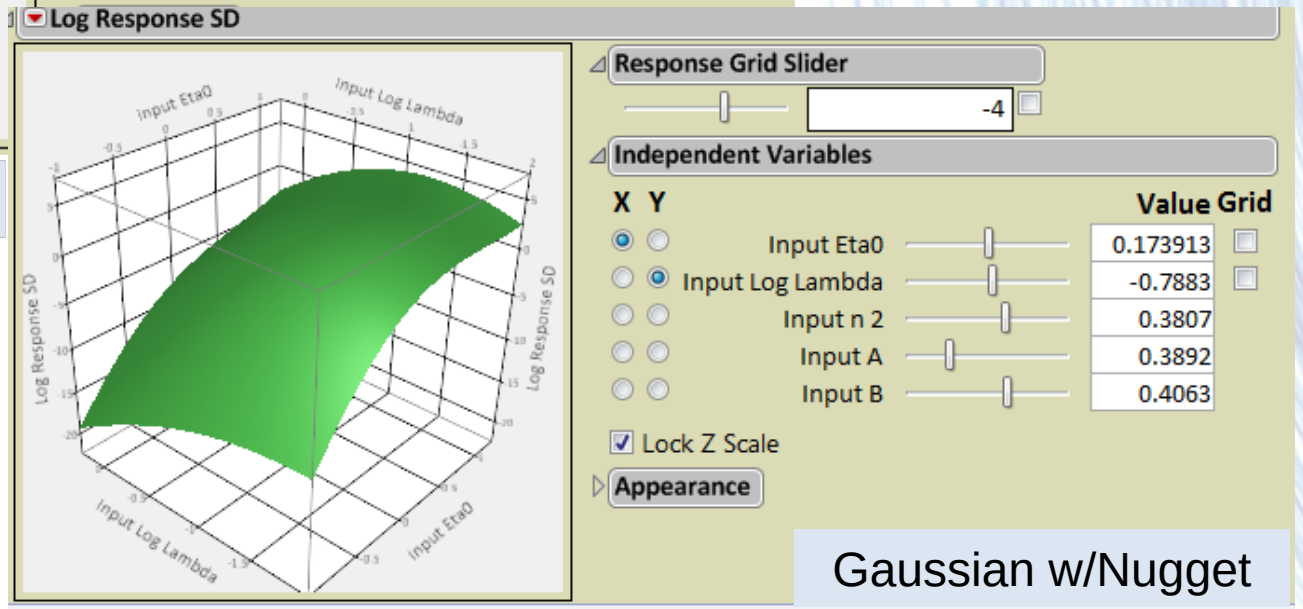
Fit using the Gaussian correlation function.

Estimates for the fraction of the variance each term explains are above

Example of What a Nugget Parameter Will Do



Gaussian – No Nugget

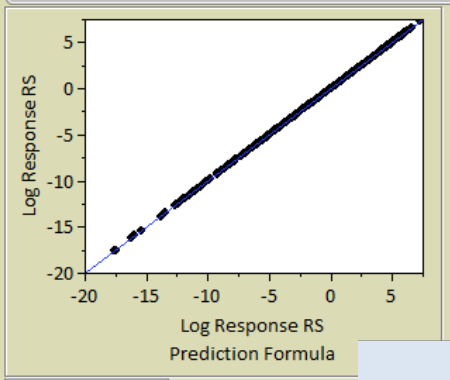


Gaussian w/Nugget

Comparison of Prediction Results

Fit Y by X Group

Bivariate Fit of Log Response RS By Log Response RS Prediction Formula



Linear Fit

No Nugget

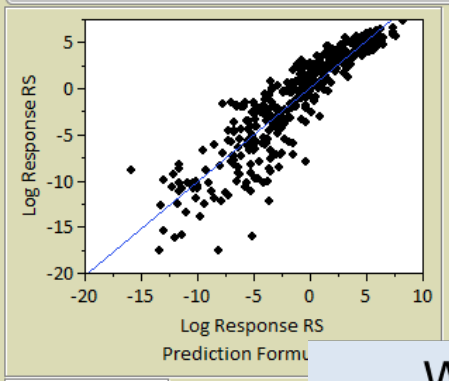
Linear Fit

Log Response RS = 6.14e-14 + 1*Log Response RS Prediction Formula

Summary of Fit

RSquare	1
RSquare Adj	1
Root Mean Square Error	1.327e-7
Mean of Response	-0.74068
Observations (or Sum Wgts)	415

Bivariate Fit of Log Response RS By Log Response RS Prediction Formula 2



Linear Fit

With Nugget

Linear Fit

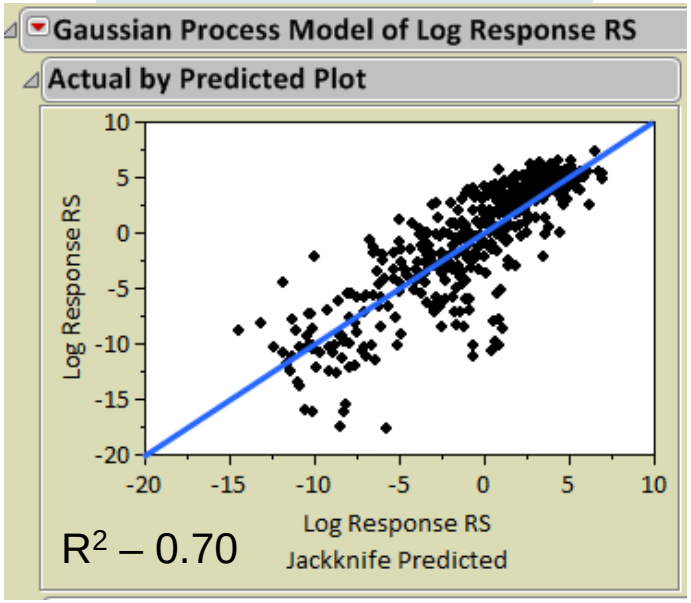
Log Response RS = 0.0137181 + 1.0185209*Log Response RS Prediction Formula 2

Summary of Fit

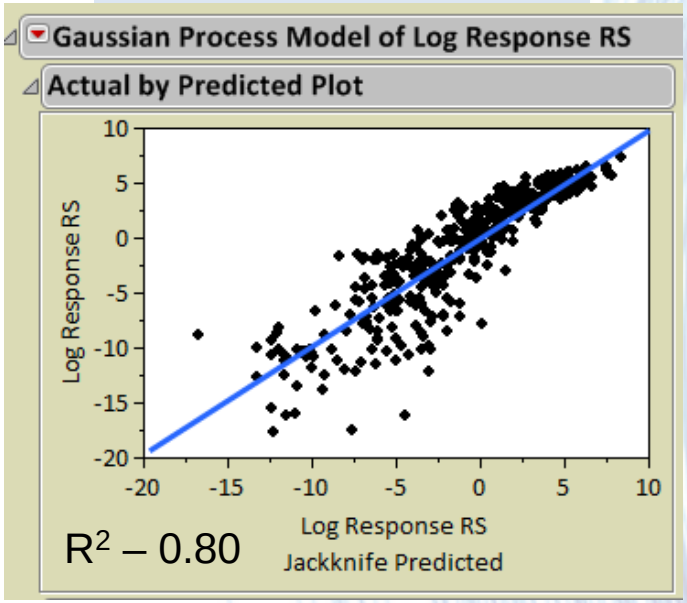
RSquare	0.835869
RSquare Adj	0.835471
Root Mean Square Error	2.222718
Mean of Response	-0.74068
Observations (or Sum Wgts)	415

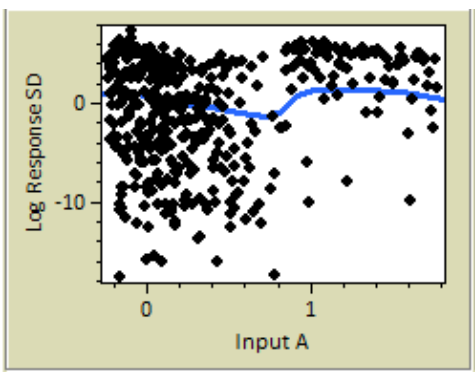
Comparison Showing Smoothing

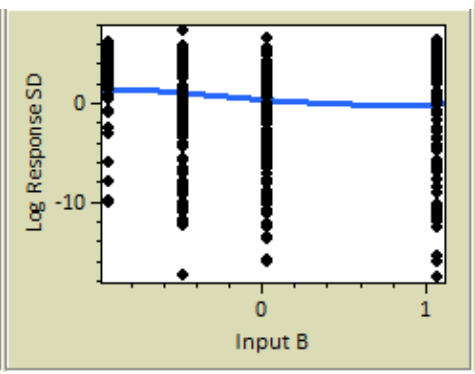
No Nugget

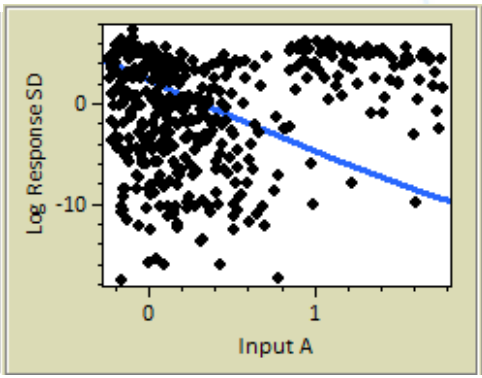


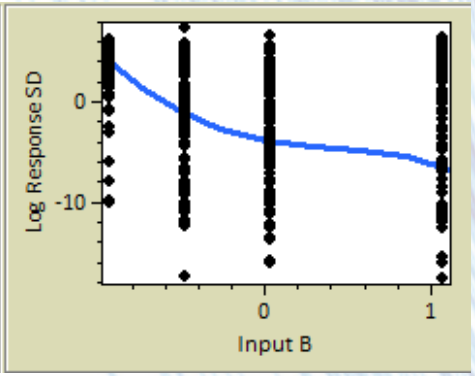
With Nugget









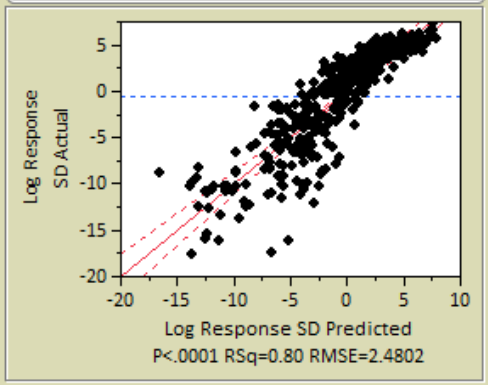


Comparing Gaussian to Response Surface

Response Surface Fit Model Statistics

Response Log Response SD

Actual by Predicted Plot



R^2 and adjusted R^2

RMSE

Summary of Fit

RSquare	0.801087
RSquare Adj	0.79515
Root Mean Square Error	2.480174
Mean of Response	-0.74068
Observations (or Sum Wgts)	415

Analysis of Variance

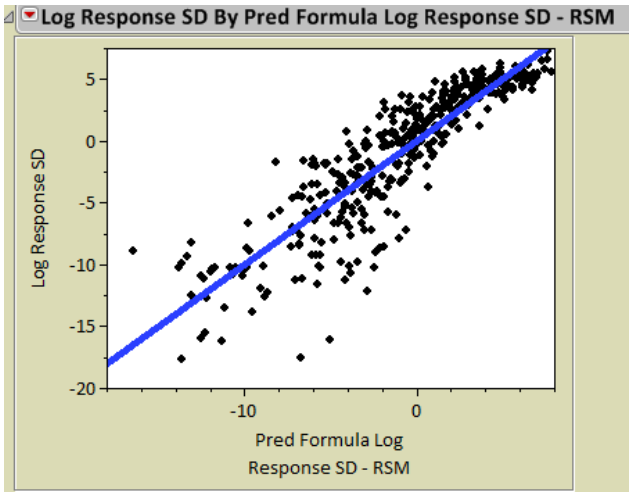
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	12	9958.822	829.902	134.9157
Error	402	2472.808	6.151	Prob > F
C. Total	414	12431.630		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1.598679	0.237895	-6.72	<.0001*
Input Eta0	5.1787656	0.210327	24.62	<.0001*

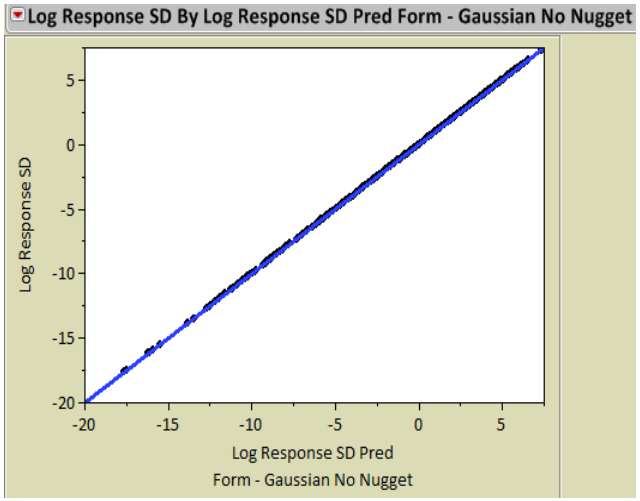
Comparison of GP vs. RSM Prediction Values

Response Surface



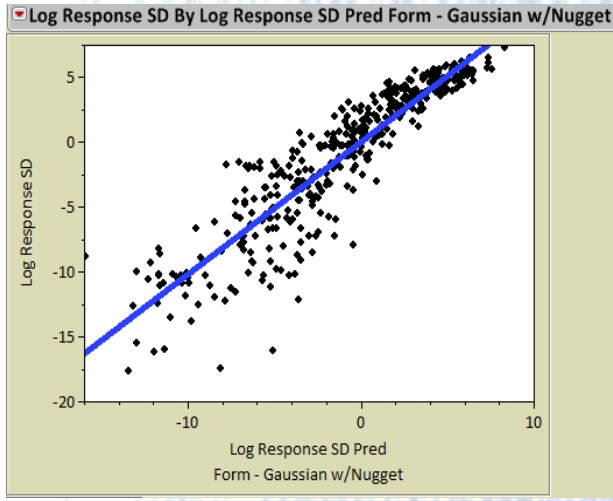
Summary of Fit	
RSquare	0.801087
RSquare Adj	0.800606
Root Mean Square Error	2.446922
Mean of Response	-0.74068
Observations (or Sum Wgts)	415

Gaussian - No Nugget



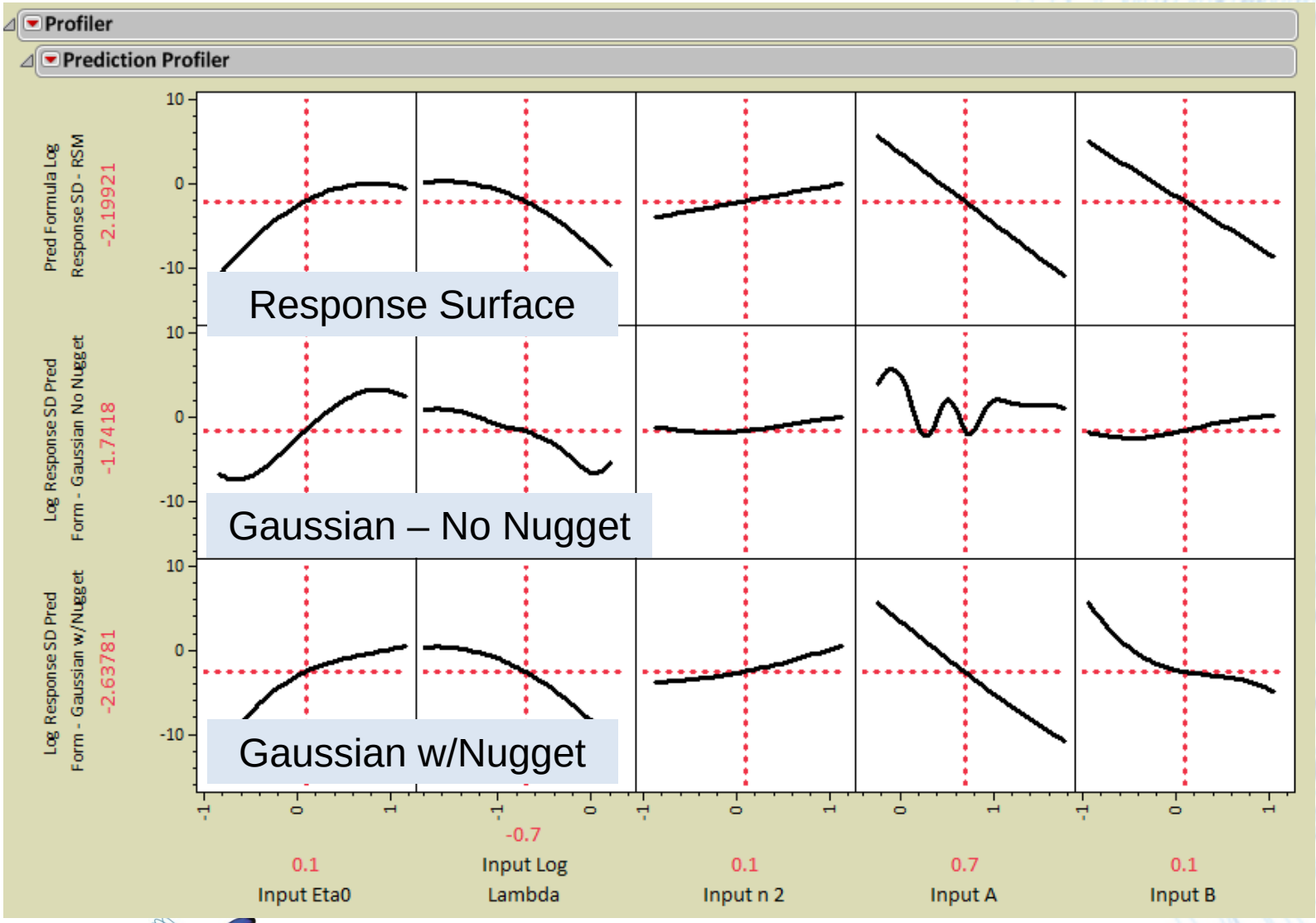
Summary of Fit	
RSquare	1
RSquare Adj	1
Root Mean Square Error	1.484e-7
Mean of Response	-0.74068
Observations (or Sum Wgts)	415

Gaussian w/ Nugget



Summary of Fit	
RSquare	0.835869
RSquare Adj	0.835471
Root Mean Square Error	2.222718
Mean of Response	-0.74068
Observations (or Sum Wgts)	415

Comparison of GP vs. RSM Prediction Values Profiler



Neural Net Models

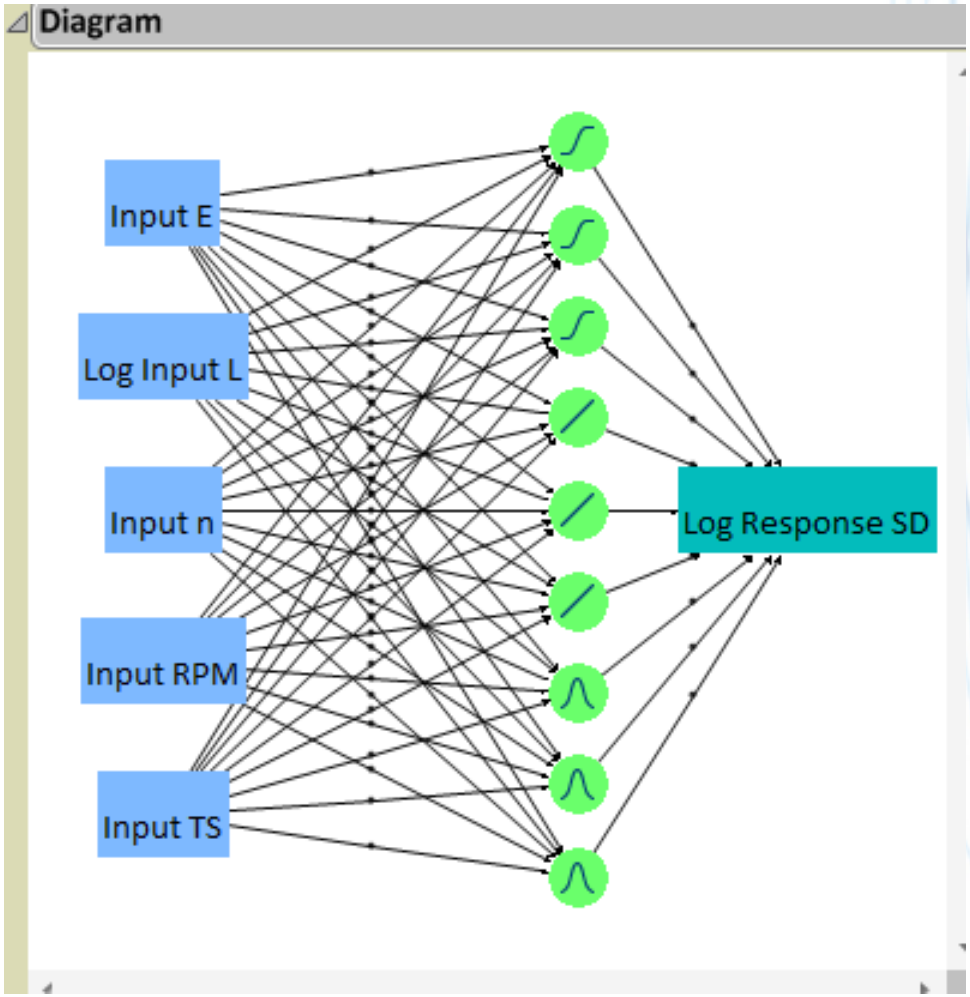
- A Neural Net (NN) is a set of nonlinear equations that predict response values from input variables in a flexible way using layers of linear regressions and S-shaped functions. JMP fits the neural net using standard nonlinear least-squares regression methods.
- The advantage of a neural net is that it can efficiently and flexibly model different response surfaces. Any surface can be approximated to any accuracy given enough hidden nodes.
- NN models have also been found to be effective in modeling deterministic data – i.e. computer simulations.

Neural Net Disadvantages

- It is easy to over fit a set of data

- The fit is not parametrically stable, so that in many fits the objective function converges long before the parameters settle down.
- There are local optima, so that you can get a different answer each time you run it, and that a number of the converged estimates will not be optimal.
- With the new platform it is really flexible and you can get a lot of answers and can take a lot of trials.

Neural Net Diagram



Cross-validation

- Cross-validation - reserves a portion of an existing data set that is used to verify the model generated by the non-held-back, training data.
- There are three cross-validation options in the NN platform
  - Random Holdback is best suited for large data sets where there is enough data to reasonably fit complex models.
  - K-Fold cross-validation is better suited for smaller data sets. The data are partitioned randomly into a number of groups that you specify (K) – 5 groups is the JMP default.
  - 3 Level Random Validation Set – Training, Validation, Test

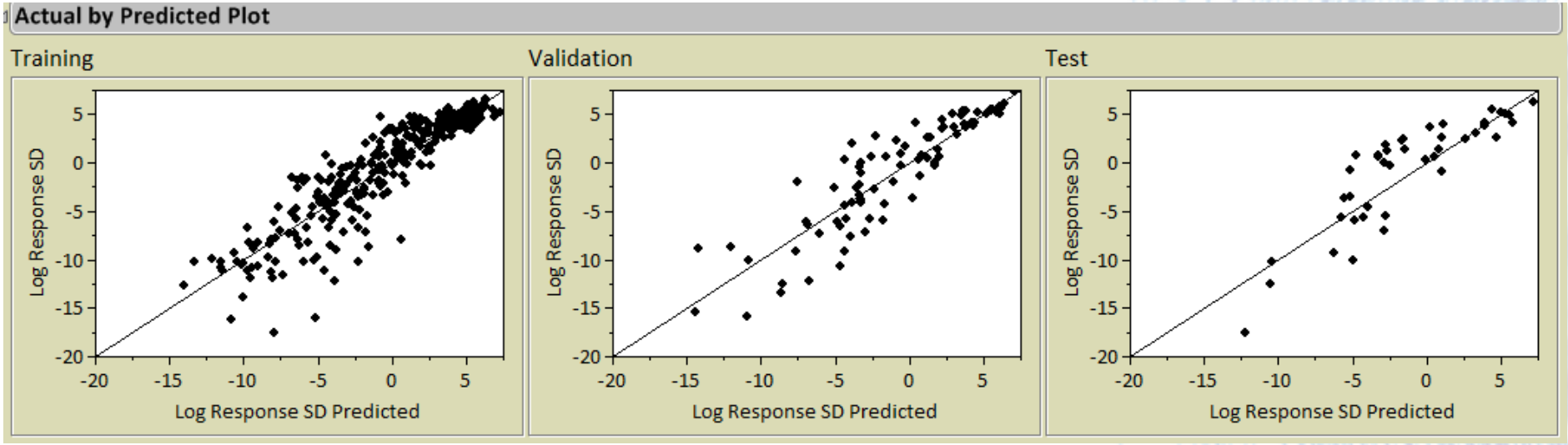
Ways to Minimize Overfitting

- Fit NN models with a 3 level validation set to cross-validate the estimates on data that are not used in the estimation process.
- Use an over fit penalty which puts a penalty on the size of the parameter estimates.
- Run as many tours as you can afford.
- Use K-fold for small data sets to obtain a more reliable fit.
- Use single nodes for each activation function type in the first layer and then use a large number of models for boosting.

Comparing Fit Models

- So which fit model do you choose and why?
  - The one that gives the best fit with the least amount of model terms, with the best model statistics. Keep it simple.
  - **"Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful." George Box**

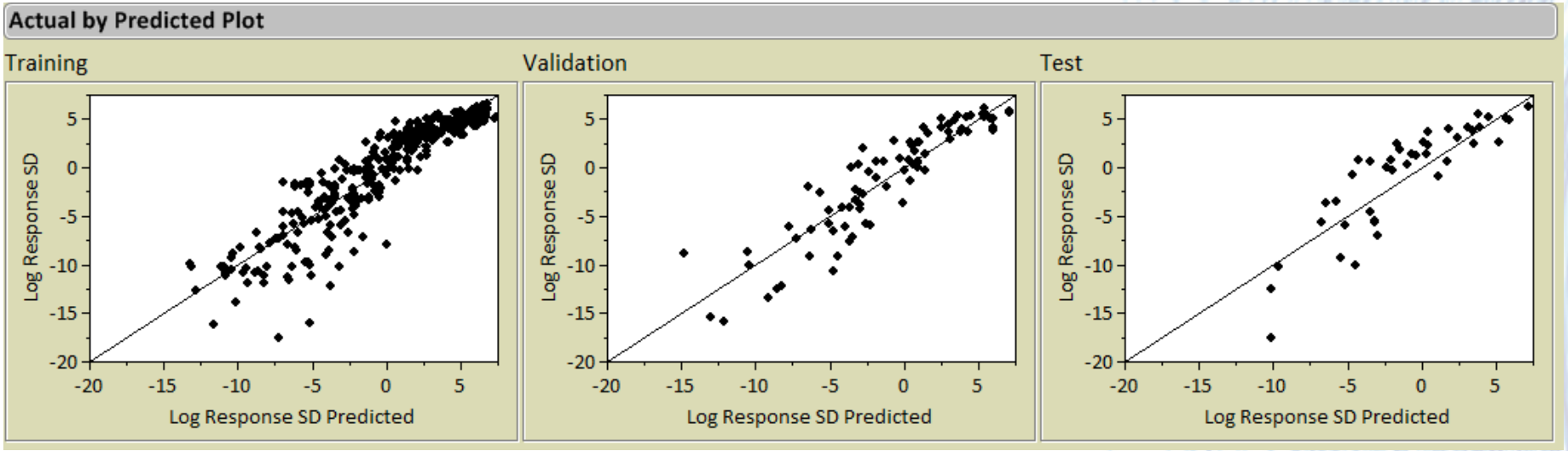
Neural Fit with 3-Level Validation Set no Boosting



Model NTanH(3)NLinear(3)NGaussian(3)

Training		Validation		Test	
Log Response SD	Measures	Log Response SD	Measures	Log Response SD	Measures
RSquare	0.8142496	RSquare	0.803091	RSquare	0.7505163
RMSE	2.3361634	RMSE	2.5145687	RMSE	2.7072024
Mean Abs Dev	1.6467843	Mean Abs Dev	1.8659866	Mean Abs Dev	2.1290544
-LogLikelihood	657.56007	-LogLikelihood	194.30631	-LogLikelihood	101.42388
SSE	1582.7212	SSE	524.81364	SSE	307.81569
Sum Freq	290	Sum Freq	83	Sum Freq	42

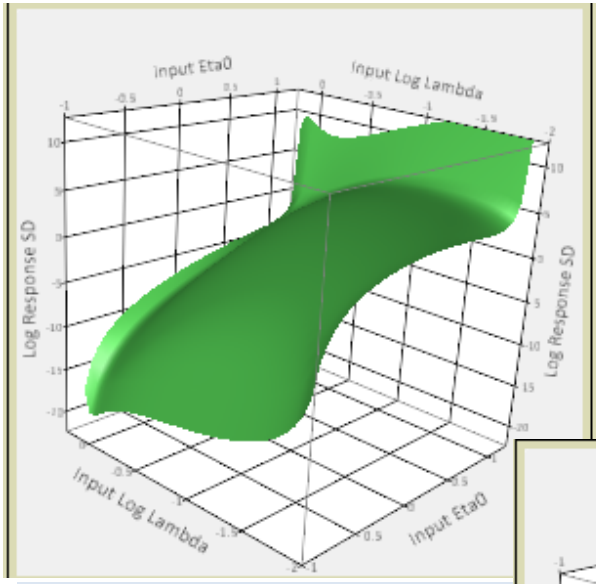
Neural Fit with 3-Level Validation Set and Boosting



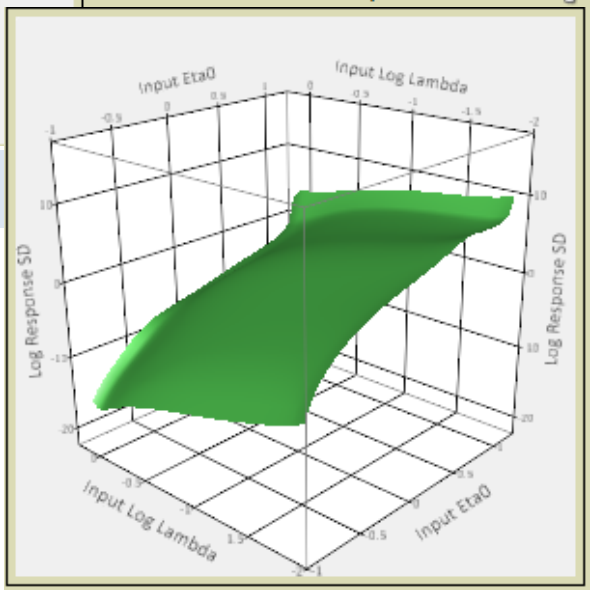
Model NTanH(2)NLinear(2)NGaussian(2)NBoost(30)

Training		Validation		Test	
Log Response SD	Measures	Log Response SD	Measures	Log Response SD	Measures
RSquare	0.8201076	RSquare	0.8377169	RSquare	0.7514989
RMSE	2.299031	RMSE	2.2827975	RMSE	2.7018661
Mean Abs Dev	1.653624	Mean Abs Dev	1.7460766	Mean Abs Dev	2.2207552
-LogLikelihood	652.91361	-LogLikelihood	186.28024	-LogLikelihood	101.34101
SSE	1532.8076	SSE	432.52665	SSE	306.60337
Sum Freq	290	Sum Freq	83	Sum Freq	42

Example of What Boosting Will Do - Neural Fits



Neural No Boosting



Neural w/Boosting

Response Grid Slider

-4

Independent Variables

X	Y		Value	Grid
●	○	Input Eta0	0.173913	
○	●	Input Log Lambda	-0.7883	
○	○	Input n 2	0.3807	
○	○	Input A	0.3892	
○	○	Input B	0.4063	

Response Grid Slider

-4

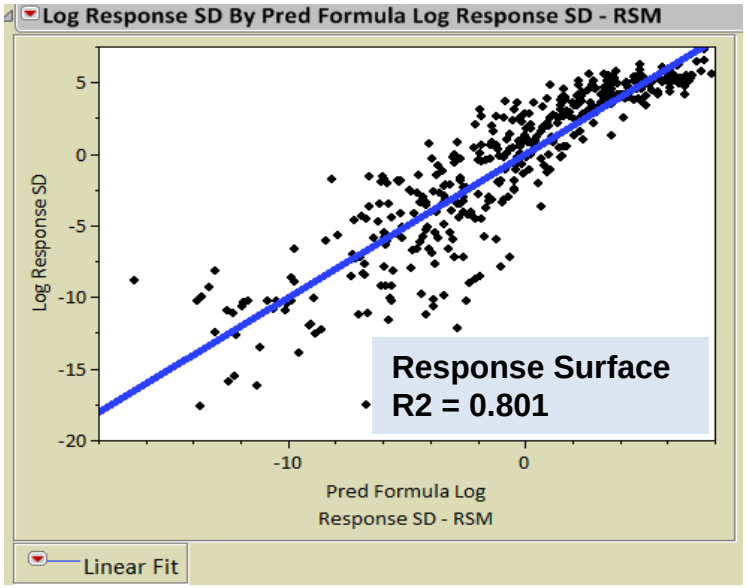
Independent Variables

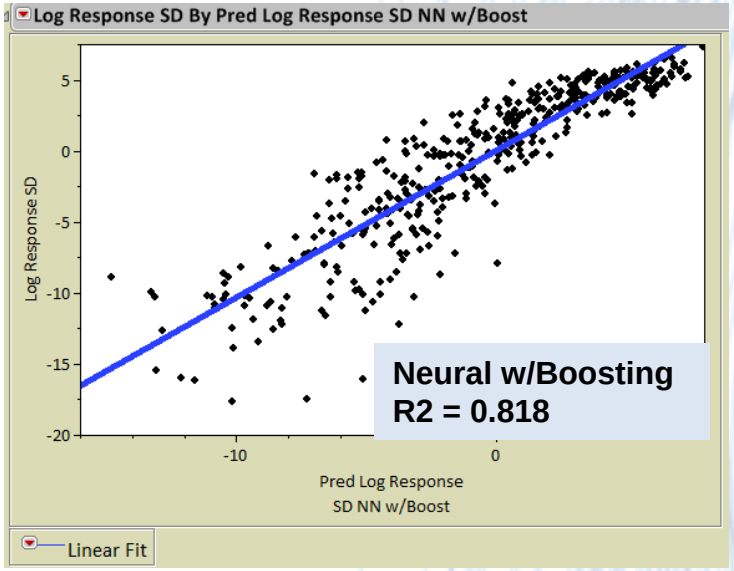
X	Y		Value	Grid
●	○	Input Eta0	0.173913	
○	●	Input Log Lambda	-0.7883	
○	○	Input n 2	0.3807	
○	○	Input A	0.3892	
○	○	Input B	0.4063	

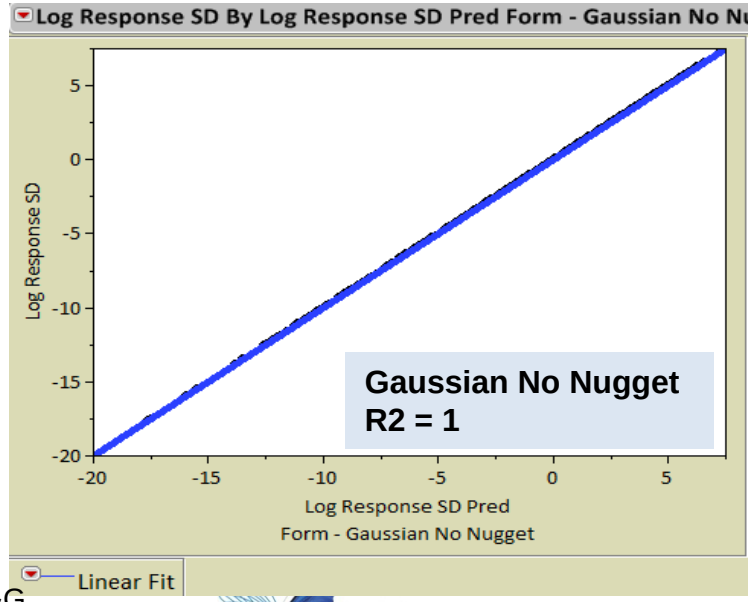
☑ Lock Z Scale

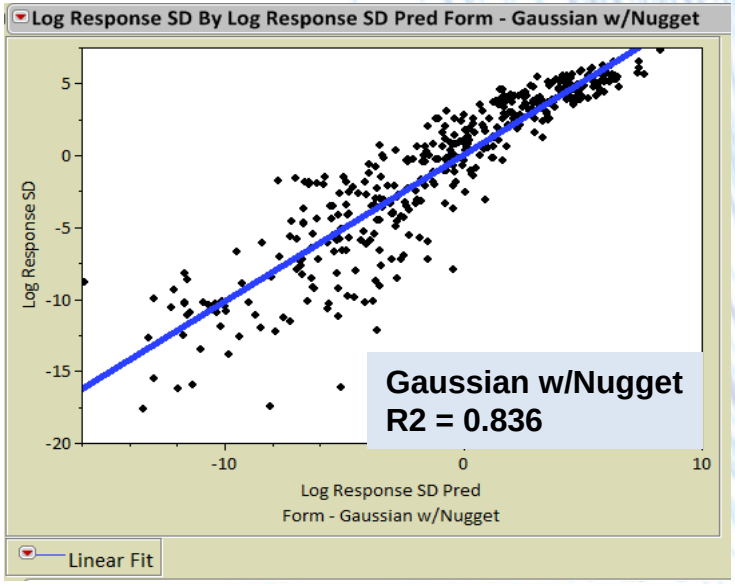
Appearance

Fit Model Comparison

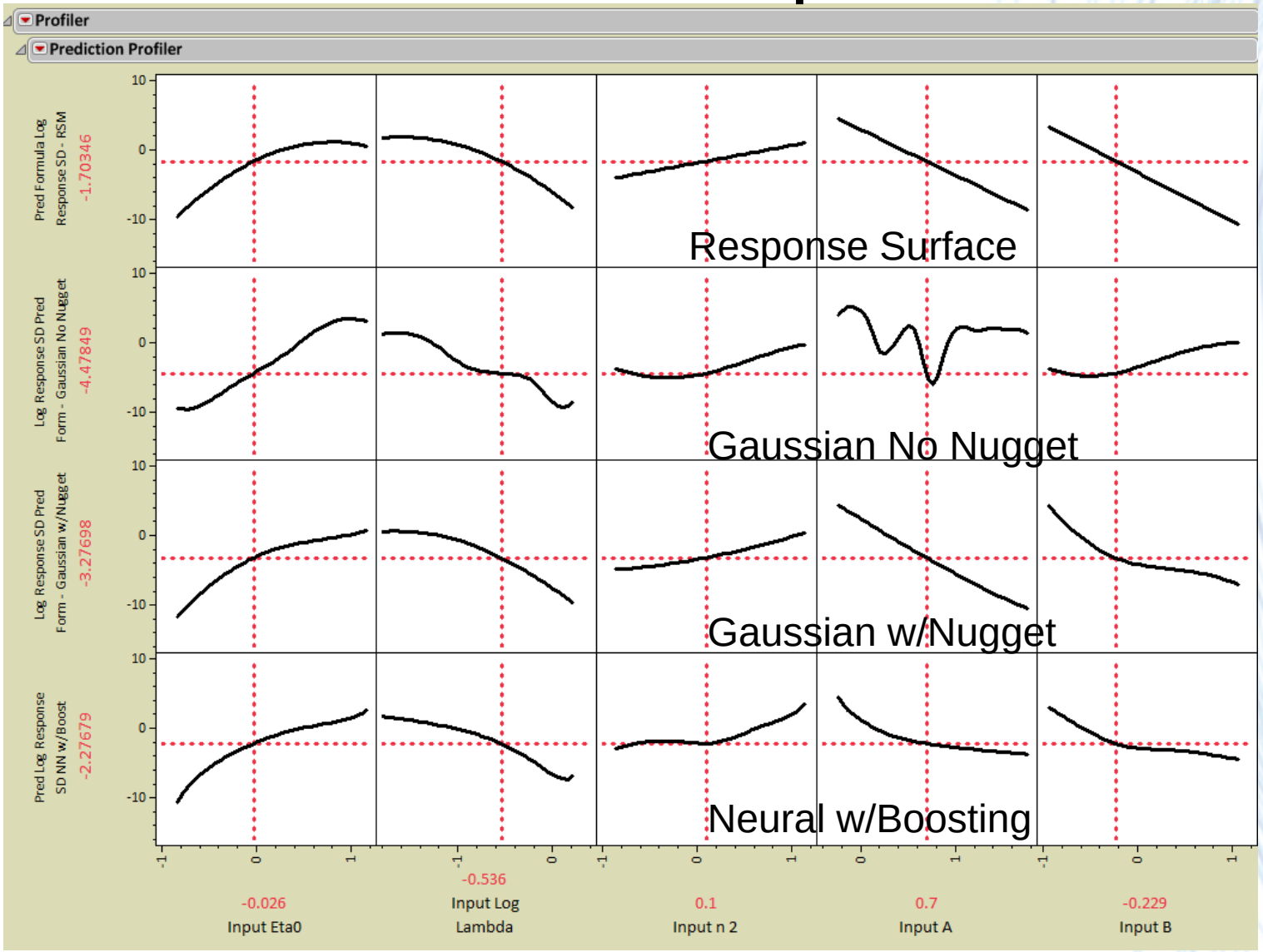




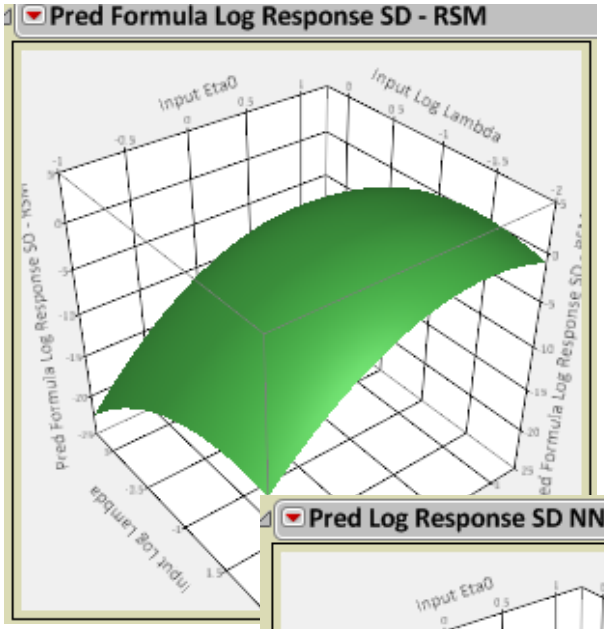


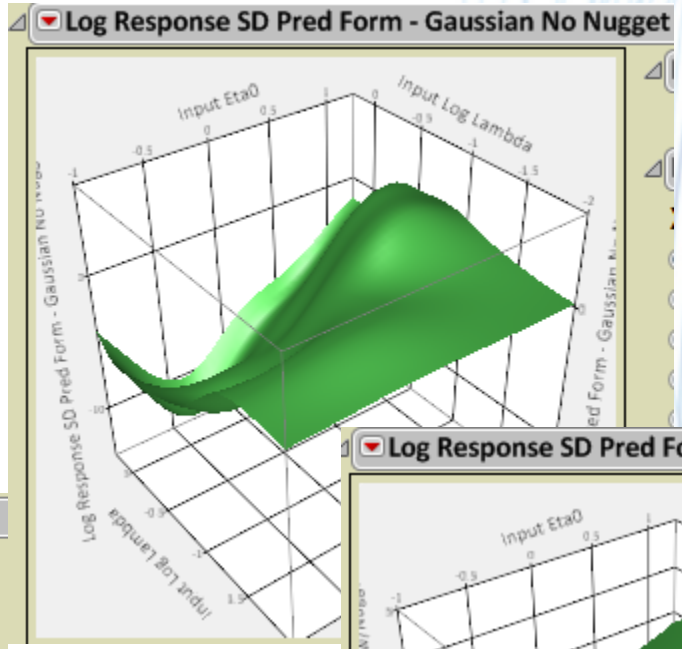


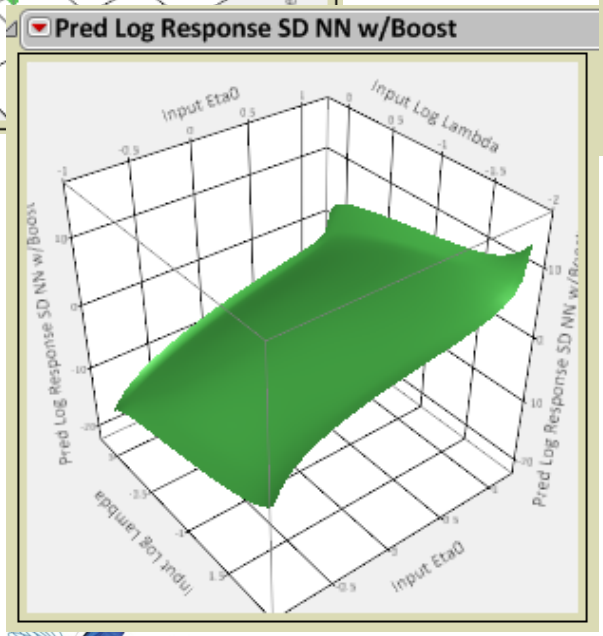
Use Profilers to Help Decide

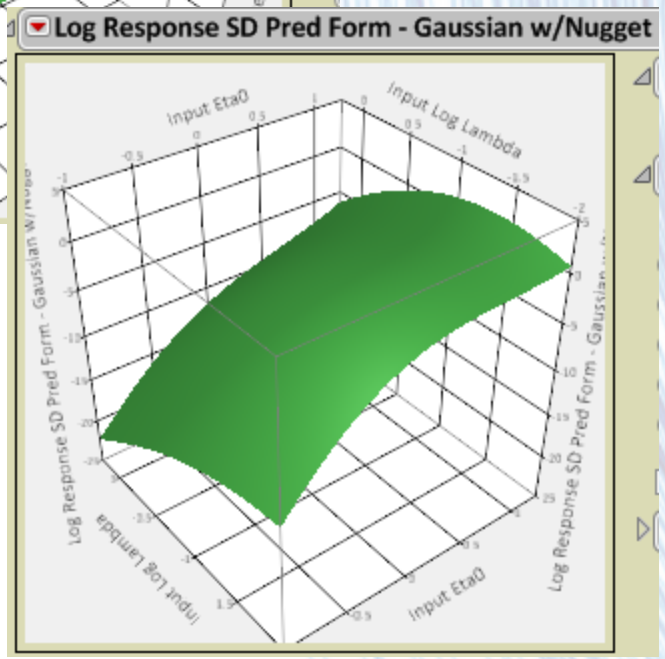


Comparing Surface Profilers

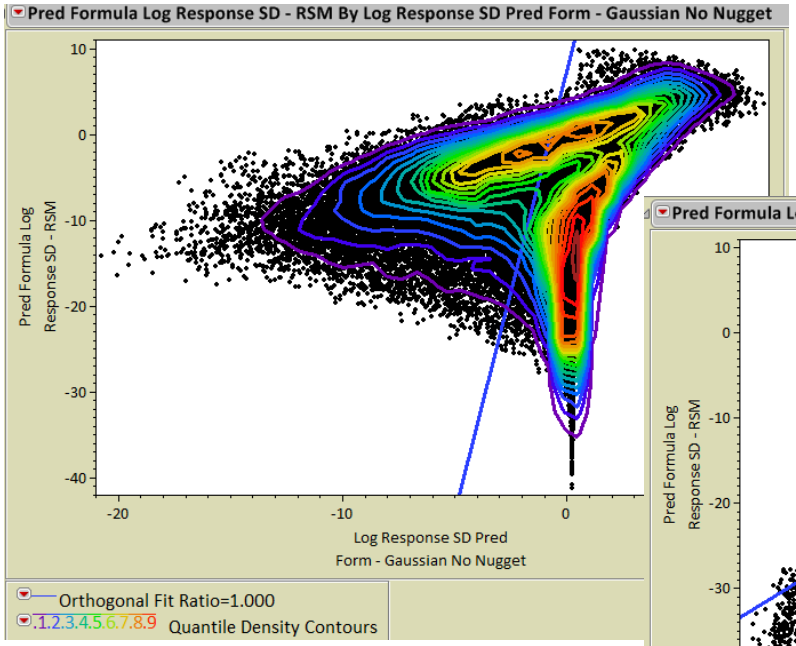


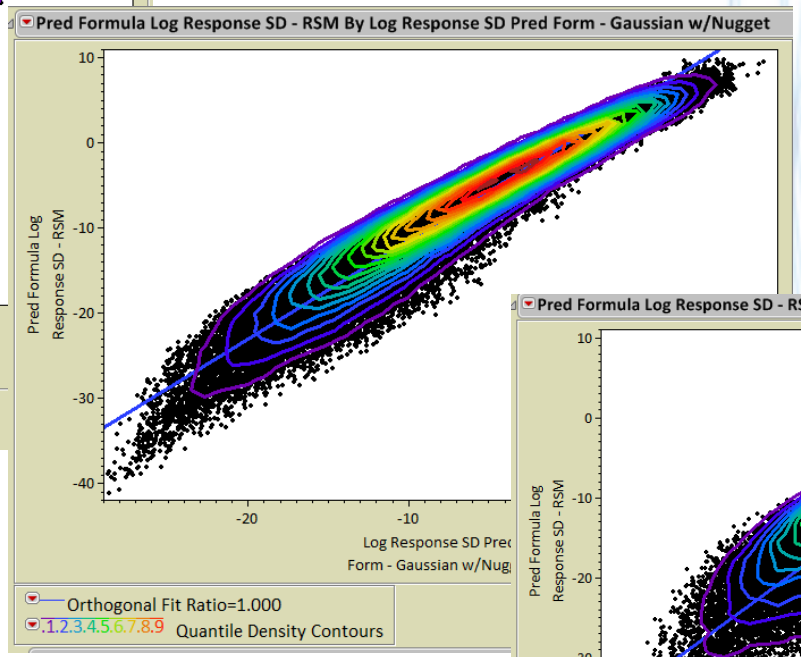


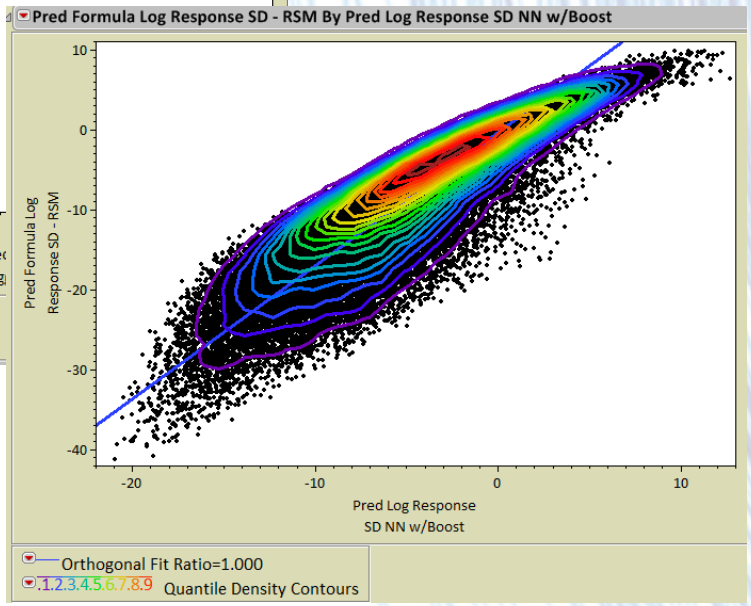




Point Density Comparison







Summary For Surrogate Model Experimental Designs

- Use space filling experimental designs – Latin Hypercube
  - Limit the bias error by maximizing the spread the design points while keeping the spread as uniform as possible.

- Add more data points using JMP DOE Augment and:
  - Add linear constraints
  - Change ranges of inputs

Summary For Fit

- Do Gaussian with and without nugget and check Jackknife fit.
- Neural Net models offer a good alternative to Gaussian models but can be more complicated. These sometimes outperform Gaussian models.
- Use the smoothing function for Neural fits.
- Don’t rely on the R^2 alone when deciding on the best fit model.
- Picking the right fit model is about keeping the model as simple as possible while still getting reasonable prediction capability.