Symbolic Regression in Design of Experiments: A Case Study with Linearizing Transformations

Flor A. Castillo **Ken A. Marshall** **Jim L. Green** **Arthur K. Kordon**

The Dow Chemical Company
Freeport, TX 77541
USA

Abstract

The paper presents the potential of genetic programming (GP)-generated symbolic regression for linearizing the response in statistical design of experiments when significant Lack of Fit is detected and no additional experimental runs are economically or technically feasible because of extreme experimental conditions. An application of this approach is presented with a case study in an industrial setting at The Dow Chemical Company.

1 INTRODUCTION

The complexity of some industrial chemical processes requires that first-principle or mechanistic model be considered in connection with empirical models. At the basis of empirical models is that underlying any system there is a fundamental relationship between the inputs and the outputs that can be locally approximated over a limited range of experimental conditions by a polynomial or a linear regression model.

Suitable statistical techniques such as design of experiments (DOE) are available to assist in this process (Box *et al*, 1978). The capability of the linear model to represent the data can be assessed through a formal Lack of Fit (LOF) test when experimental replicates are available (Montgomery, 1999). Significant LOF in the model indicates a regression function that is not linear; i.e. the polynomial initially considered is not appropriate. A more adequate model may be found by fitting a polynomial of higher order by augmenting the original design with additional experimental runs. Specialized designs such as the Central Composite Design are available for this purpose (Box *et al.*, 1978).

However, there are many practical cases where runs are very expensive or technically unfeasible because of extreme experimental conditions, thus making the fit of a higher order polynomial impractical. This problem can be handled if appropriate input transformations are used, provided that the basic assumption of least-square estimation regarding the probability distributions of errors is not affected. These assumptions require that errors be uncorrelated and normally distributed with mean zero and constant variance.

Some useful transformations are discussed in Box and Draper.(1987). Unfortunately, transformations that linearize the response without affecting the error structure are not always obvious and are often developed based on experience or theoretical insight. Genetic programming (GP)- generated symbolic regression provides a unique opportunity to rapidly develop and test these transformations. Symbolic regression includes the finding of a functional mathematical expression that fits a given set of data (Koza, 1992).

GP-generated symbolic regression is an evolution-based algorithm for automatically generating nonlinear input-output models. Several possible models of the response as a function of the input variables are obtained by combining basic functions, inputs, and numerical constants. This multiplicity of solutions offers a rich set of possible transformations of the inputs. At the same time, the most significant challenge of GP-generated transforms is that most models are not parsimonious and include chunks of inactive code or terms that do not contribute to the overall fitness (Banzhaf *et al*, 1998) and that may prove inefficient in producing a linearizing transformation. This problem can be managed to some degree at the expense of extra-computation time by appropriate algorithms that quickly test the ability of transforms to linearize the response without altering error structure.

The application of GP in DOE and the potential of combining them offer a unique set of opportunities that is beginning to grab the attention of researchers and industry. Experimental design techniques have already been used to evaluate the effects of GP parameters (Spoonger, 2000). An excellent discussion of algorithm-driven regression based on genetic programming for solving supersaturated designs is presented in Cela *et.al* (2001).

In this paper, a novel approach of integrating GP with DOE is presented. This approach has the potential to improve the effectiveness of empirical

model building by saving time and resources in situations where experimental runs are quite expensive or technically unfeasible because of extreme experimental conditions. GP is applied to the development of variable transforms that linearize the response in statistically designed experiments for a chemical process in The Dow Chemical Company.

2 METHODOLOGY

A series of experimental runs were performed in a lab scale reactor in four variables. The response variable was the selectivity of one of the products. These experiments were statistically analyzed and the effect of the variables as well as a prediction of the response within the area of experimentation was well understood. LOF was induced by removing one experimental run to simulate a common situation in which LOF is significant and additional experimental runs are impractical due to the extreme cost of experimentation or because it is technically unfeasible due to extreme experimental conditions. In this system the potential of GP-generated transforms was studied allowing the comparison of results with a well-known system.

The appropriateness of GP-generated transforms to linearize the response without affecting error structure was assessed by performing the transformations presented in the functional form of the GP model. Then a linear regression model was fit in the transformed inputs. This model, referred to as the *transformed linear model*, was examined for Lack of Fit and appropriate error structure. Both models, the transformed linear model and the GP model, were tested considering 9 additional experiments in the region of the design. The validity of the results was determined by comparing model predictions with the previously analyzed experiments and with a fundamental kinetic model (FKM) that was earlier developed. The results indicate that GP-generated transformations have the potential of linearizing the response in those cases where additional experimental runs are not possible.

3 THE EXPERIMENTAL DESIGN

The experiments conducted in lab-scale thermal chlorination reactor system consisted of a complete 2^4 factorial design in the factors x_1, x_2, x_3, x_4, with three center points. A total of 19 experiments were performed. The response variable, S_k, was the yield or selectivity of one of the products. The factors were coded to a value of –1 at the low level, +1 at the high level, and 0 at the center point. The complete design in the coded variables is shown in Table 1

To develop a base case and test for variable transformations, LOF was induced by removing run number 1 of the experimental design. The response S_k, was fit to the following first-order linear regression equation

Table 1: 2^4 factorial design with three center points

RUNS	x_1	x_2	x_3	x_4	S_k
1	1	-1	1	1	1.598
2	0	0	0	0	1.419
3	0	0	0	0	1.433
4	-1	1	1	1	1.281
5	-1	1	-1	1	1.147
6	1	1	-1	1	1.607
7	-1	1	1	-1	1.195
8	1	1	1	-1	2.027
9	-1	-1	-1	1	1.111
10	-1	1	-1	-1	1.159
11	-1	-1	-1	-1	1.186
12	1	-1	-1	1	1.453
13	1	1	-1	-1	1.772
14	-1	-1	1	-1	1.047
15	-1	-1	1	1	1.175
15	1	1	1	1	1.923
17	1	-1	-1	-1	1.595
18	1	-1	1	-1	1.811
19	0	0	0	0	1.412

considering only terms that are significant at the 95% confidence level.

$$S_k = \beta_o + \sum_{i=1}^{k} \beta_i x_i + \sum\sum_{i<j} \beta_{ij} x_i x_j \qquad (1)$$

Table 2 shows the corresponding Analysis of Variance showing evidence of Lack of Fit (p = 0.0476). Therefore, the hypothesis that a first-order model can adequately describe this system is rejected.

Table 2 - Analysis of variance for the linear model

Source	**DF**	**Sum of Squares**	**Mean Square**	**F Ratio**
Model	8	1.5091186	0.188640	107.6350
Error	9	0.0157733	0.001753	**Prob > F**
C. Total	17	1.5248919		<.0001
Lack Of Fit				
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Ratio**
Lack Of Fit	7	0.01555519	0.002222	20.3775
Pure Error	2	0.00021810	0.000109	**Prob > F**
Total Error	9	0.01577329		0.0476
				Max RSq
				0.99

The corresponding residual plot, presented in Figure 1, suggested non-constant variance, which is one of the necessary conditions of the error structure for least-square estimation.

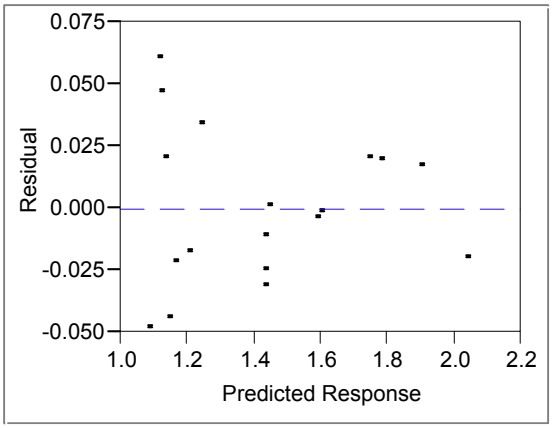

Figure 1 - Residual plot for first-order linear model suggesting non-constant variance

Under these circumstances, a variance-stabilizing power transformation of the response (y) was performed (Box and Cox, 1964). The response was transformed to y^{λ} where the parameter λ varies from –2 to 2 and the choice of λ that results in the minimum residual sum of squares of the transformed model is the maximum likelihood estimation of λ and the best transformation of the response. In the present case, however, the power transformation resulted in a λ value of 1 indicating that no transformation of the response was helpful. Cases like this are quite common in industrial processes. The next alternative to be investigated is the transformation of the input variables by means of GP-generated symbolic regression.

3.1 THE GP-GENERATED TRANSFORMATIONS

The GP approach will be used to search for potential transforms of the input variables. The GP algorithm was applied to the original data set, considering the response variable as the output and the four variables, x_1, x_2, x_3, x_4, in uncoded form as inputs. This resulted in a series of non-linear equations that satisfied the data. The functional form of these equations produced a rich set of possible transforms that were tested for the ability to linearize the response without altering error structure. An advantage of this approach is that experience or physical interpretation may be used to identify promising transforms, which were previously unavailable to the experimenter. An additional advantage is that GP generates a sensitivity analysis ranking all the input variables in order of importance to the fitness of the equations (Kordon and Smits, 2001) allowing to verify significant factors in the linearized models.

The GP algorithm is implemented as a toolbox in MATLAB. The initial functions for GP included: addition, subtraction, multiplication, division, square, change sign, square root, natural logarithm, exponential, and power. Function generation takes 20 runs with 500 population size, 100 number of generations, 4 reproductions per generation, 0.6 probability for function as next node, 0.01 parsimony pressure, and correlation coefficient as optimization criteria. A snapshot of the input/output sensitivity is shown in Figure 2, which shows x_1 as the most important input.

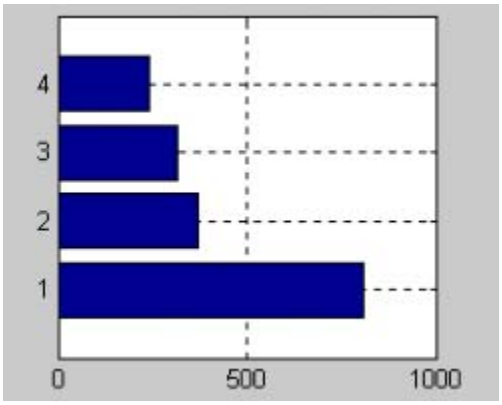

Figure 2 GP-based Input/output sensitivity of the four input variables

The selection of the best candidates is based on a trade-off between the fitness of the function and the ability to linearize the response while producing an acceptable error distribution. From the set of potential non-linear equations the best fit between model prediction and empirical response was found for the following analytical function:

$$S_k = \frac{3.13868\times10^{-17} e^{\sqrt{2x_1}} \ln\left[(x_3)^2\right] x_2}{x_4} + 1.00545 \quad (2)$$

Where x_1, x_2, x_3, x_4 are the input variables and S_k is the output.

The correlation coefficient between the analytical function and the empirical data was 0.95. This nonlinear equation indicates an exponential relationship with x_1, a logarithmic relationship with x_3, a linear relationship with x_2, and an inverse relationship with x_4, as shown in Table 3. To test the capability of these transforms to linearize the response, the following transformations were applied to the input variables as supplied by the GP function (2).

Table 3 - Variable transformations suggested by GP model

Original Variable	Transformed Variable
x_1	$Z_1 = \exp\left(\sqrt{2x_1}\right)$
x_2	$Z_2 = x_2$
x_3	$Z_3 = \ln[(x_3)^2]$
x_4	$Z_4 = x_4^{-1}$

Then a first-order linear regression model (i.e., the transformed linear model) was fit to the transformed variables. Table 4 shows the corresponding parameter estimates. The analysis of variance, presented in Table 5,shows no evidence of LOF indicating that the GP-generated transformations were succesful in linearizing the response.

Table 4 - Parameter estimates for transformed linear model

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	–0.29	0.0464	–6.26	0.0033
Z_1	0	1.3e-15	75.83	<.0001
Z_3	0.165	0.0079	20.80	<.0001
Z_4	0.147	0.0372	3.98	0.0164
Z_2	0.092	0.0070	13.17	0.0002
$(Z_1-7.e12)*(Z_3-3.2)$	0	2.7e-15	18.02	<.0001
$(Z_1-7.e12)*(Z_4-0.8)$	0	1.3e-14	7.58	0.0016
$(Z_3-3.199)*(Z_4-0.8)$	–0.701	0.08058	–8.70	0.0010
$(Z_1-7.e12)*(Z_2-3.9)$	0	2.4e-15	5.38	0.0058
$(Z_3-3.199)*(Z_2-3.9)$	0.050	0.01481	3.40	0.0274
$(Z_4-0.835)*(Z_2-3.9)$	0.180	0.06895	2.62	0.0590
$(Z_1-7.e12)*(Z_3-3.2)*(Z_2-3.9)$	0	5.1e-15	–5.04	0.0073
$(Z_1-7.e12)*(Z_4-0.8*(Z_2-3.9)$	0	2.4e-14	3.43	0.0264
$(Z_3-3.199)*(Z_4-0.8)*(Z_2-3.9)$	0.603	0.14952	4.04	0.0156

Table 5 - Analysis of variance for transformed linear model

Source	**DF**	**Sum of Squares**	**Mean Square**	**F Ratio**
Model	13	1.5241819	0.117245	660.5351
Error	4	0.0007100	0.000177	**Prob > F**
C. Total	17	1.5248919		<.0001
Lack Of Fit				
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Ratio**
Lack Of Fit	2	0.00049190	0.000246	2.2554
Pure Error	2	0.00021810	0.000109	**Prob > F**
Total Error	4	0.00071000		0.3072
				Max RSq
				0.9999

The transformed linear model itself is less parsimonious than the nonlinear GP model including even third order iterations. However, the model is very significant.
The corresponding residual plot for the transformed linear model is presented in Figure 3. This plot indicates no violation regarding basic assumptions for the probability distribution of errors required by least squares, indicating that the GP-generated transformations linearized the response without altering the error structure of the model produced. One observation is that the residual of one center point is larger than the residuals of the other two center points. However, in the original analysis, this data point had also been excluded due to problems with experimental conditions during the run.

The transformed linear model and the nonlinear GP model were used to predict the selectivity to the output at the conditions of experiment 1 (the experiment removed from the original data set in order to induce Lack of Fit).

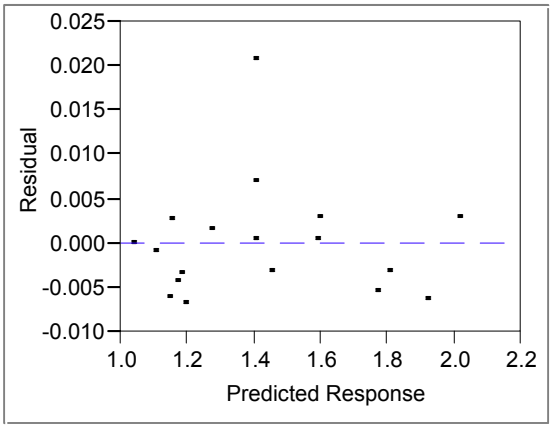

Figure 3 - Residual plot for the transformed linear model

Figure 4 shows the plot of predicted versus actual values for the two models.

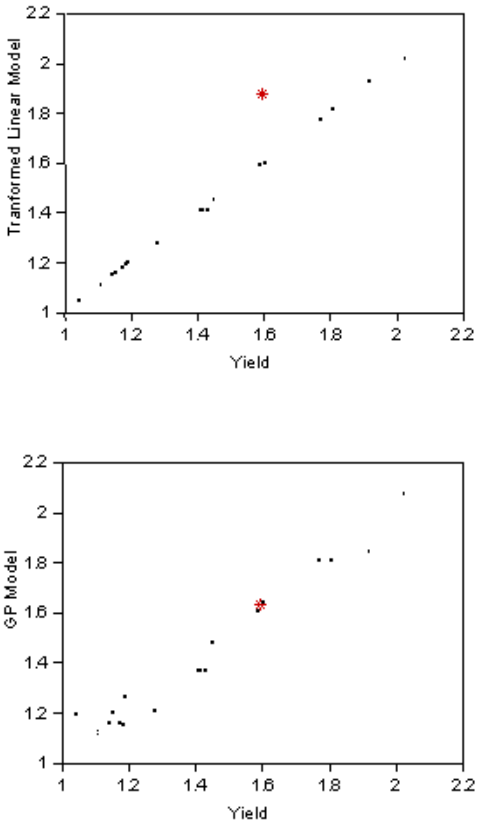

Figure 4 - Predicted versus actual values for the transformed linear and the nonlinear GP model

The point corresponding to experiment 1 is indicated in the figure by an asterisk. The performance of both models was very good. The correlation coefficient was 0.99 for the transformed linear model, and 0.978 for the GP model.
The nonlinear GP model gives a more accurate prediction for the value of the removed point. But both models predict an increase response by operating at conditions of high x_1, x_2, x_3, and low x_4. These results were consistent with the results obtained previously by analyzing the full design and by a fundamental kinetic model.

3.2 THE TESTING DATA SET

The prediction capabilities of the transformed linear model and the GP model were tested with nine additional experimental points within the range of experimentation. This is a relative small data set because of the cost and difficulty of experimentation. Plots of the predicted response for the transformed linear and the GP model versus the actual values, presented in Figure 5, indicate good performance of both models indicating that the models are comparable in terms of prediction with additional data inside the region of the design. The correlation coefficient was 0.99 for the transformed linear model, and 0.98 for the GP model. The selection of one of these models over the other would be driven by the requirements of a particular application. For example, in the case of process control, the more parsimonious model would generally be preferred.

4 CONCLUSIONS

In the course of conducting designed experiments, Lack of Fit is often encountered, indicating that the proposed linear regression model fails to adequately describe the data. One traditional approach to address this problem is to introduce higher order terms to the linear model. This is accomplished by adding experiments to the original design, which can be time-consuming, costly, or may be technically unfeasible because of extreme experimental conditions.
A second approach is to use transformations to avoid additional experimentation. One technique is transformation of the responses, but this is not always effective. In those cases, transformation of the input variables may be the only alternative to remove Lack of Fit and provide an appropriate model. Unfortunately these transformations are not always obvious and are often developed based on experience or theoretical insight. GP provides a way to rapidly develop and test these transformations so that those appropriate linear models are developed.
The genetic programming (GP) algorithm was successfully applied to the results of DOE in a chemical process in Dow Chemical Company. Experimentation in this system is difficult and time-consuming due to the severe conditions of the experiments. Data for the interested output were manipulated to induce Lack of Fit. Genetic programming was used to generate a nonlinear model for the output as a function of four experimental variables. The form of the nonlinear model was used to suggest input variable transformations for a linear model. The resulting transformed linear model showed no evidence of Lack of Fit. No additional experimental data had to be used in the analysis to achieve this result. The success of this industrial application illustrates the great potential of using GP to address Lack of Fit in linear regression problems. This approach can improve the effectiveness of empirical model building by saving time

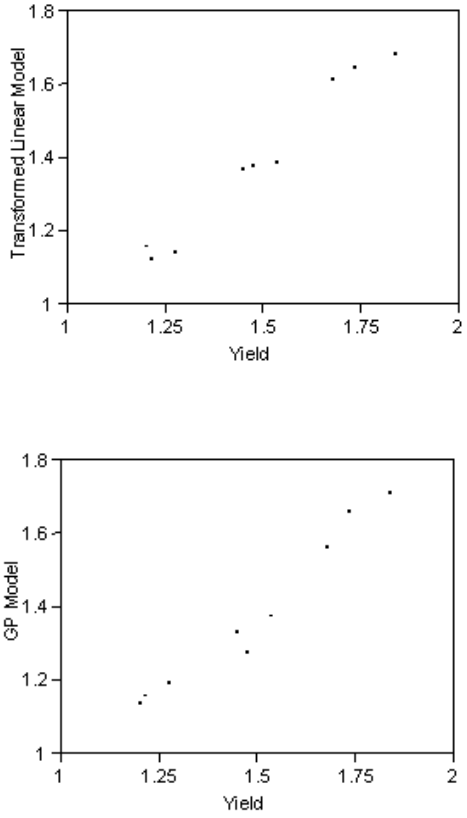

Figure 5 - Predicted versus actual values for additional data

and resources when experiments are expensive or difficult. However, more systematic research in the area of defining a methodology for robust nonlinear response surface generated by GP is recommended..

References

Banzhaf W., P. Nordin, R. Keller, and F. Francone, *Genetic Programming: An Introduction*, Morgan Kaufmann, San Francisco, 1998.

Box, G.E.P, and Draper, N. R., *Empirical Model Building and Response Surfaces*, John Wiley and Sons, New York, 1987.

Box, G.E.P., and Cox, D.R., "An Analysis of Transformations", *J. Roy. Stat. Soc.*, Series B, V. 26, p. 211, 1964.

Box, G.E.P., Hunter, W.G., and Hunter, J.S., *Statistics for Experiments: An Introduction to Design, Data Analysis, and Model Building*, John Wiley and Sons, New York, 1978.

Cela, R., Martinez, E., and Carro, A. M, Supersaturated experimental Design: New Approaches to building and Using it Part II Solving Supersaturated Designs by Genetic Programming Algorithms*, Chemometrics and Intelligent Laboratory Systems*, **57**, pp. 75-92, 2001

Kordon, A. K. and G. F. Smits, Soft Sensor Development Using Genetic Programming, Proceedings of the GECCO'2001, San Francisco, pp. 1346-1351.

Koza, J. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, MIT Press, Cambridge, MA, 1992.

Montgomery, D. C., *Design and Analysis of Experiments*, John Wiley and Sons, New York, 1999.

Spoonger. *Using Factorial Experiments to Evaluate the Effects of Genetic Programming parameters*. In Proceedings of EuroGP'2000, Edinburgh, LNCS Ul802, pp. 2782, 2000.