Symbolic Regression in Multicollinearity Problems

Flor Castillo, Carlos Villa

The Dow Chemical Company

Outline

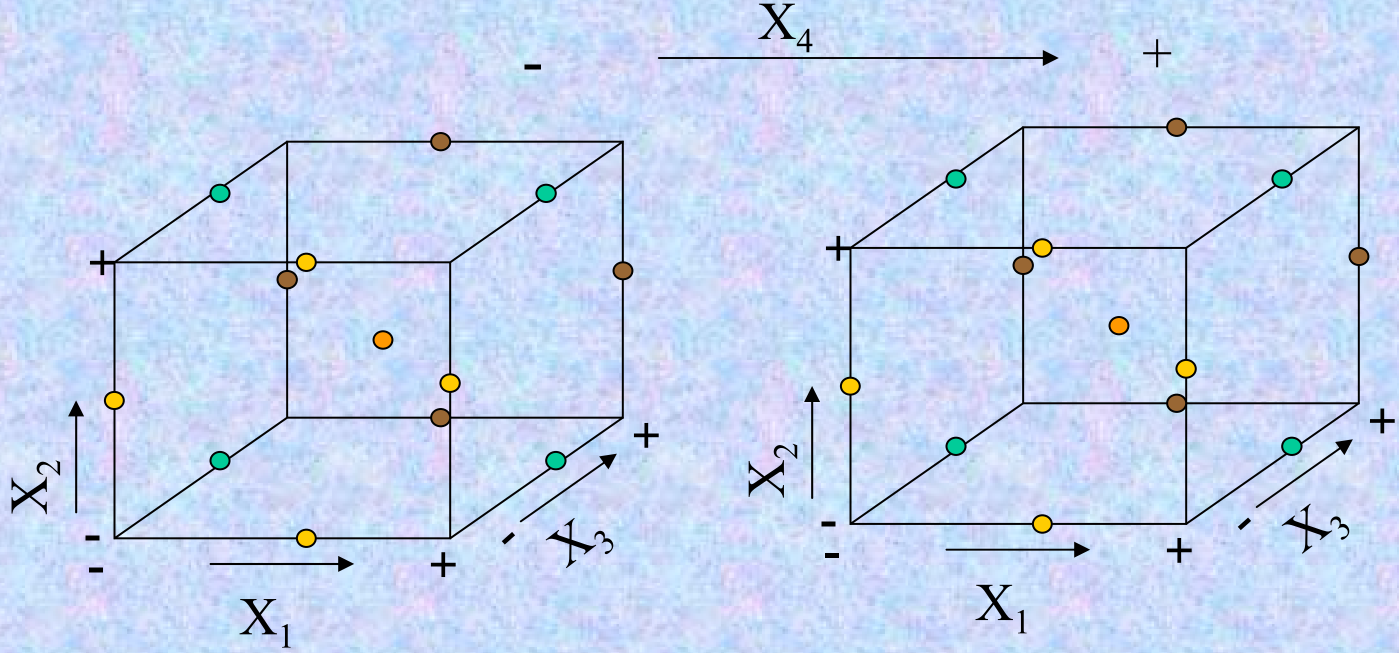
- **Why we need Symbolic Regression in multicollinearity**
- **A case study**
- **The proposed approach using GP**
- **Results**
- **Conclusions**

GP in Multiple Linear Regression (MLR) Models

- GP has been used in two situations
  - design of experiments (DOE) scheme to solve lack of fit situations (LOF)

  - MLR with historical (plant data) to minimize multicollinearity (strong relationship among inputs)

Box-Behnken Experimental Design



Response (Output):

Particle size distribution of a chemical compound

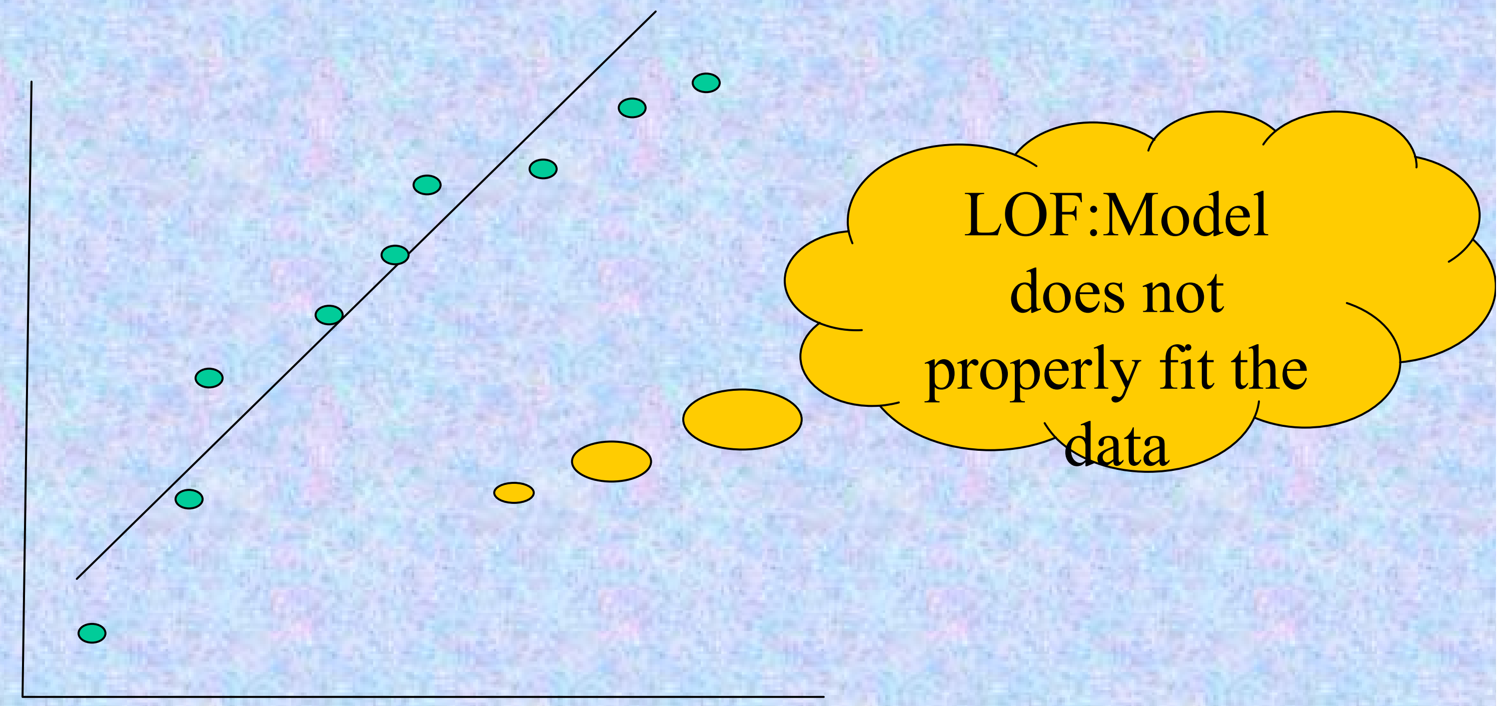
$$S_k = \beta_o + \sum_{i=1}^{k} \beta_i X_i + \sum\sum_{i<j} \beta_{ij} X_i X_j + \sum \beta_{ii} X_i^2$$

What if LOF is statistically significant?

Inputs:

- X_1, X_2, X_3, X_4

LOF



Statistical test can detect LOF

p value for LOF<0.05: Significant LOF

Possible LOF Solutions

- Ignore it
  - Possible limitations on conclusions
- Collect more data
  - Induce correlation
  - Cost of additional sampling, etc.
- Try a different more complex model
  - Current data may not support new model
- Try a different transformed model
  - Transformation to try not obvious (**Genetic Programming (GP) can help**)

Box-Behnken Data Analysis

$$S_k = \beta_o + \sum_{i=1}^{k} \beta_i X_i + \sum\sum_{i<j} \beta_{ij} X_i X_j + \sum \beta_{ii} X_i^2$$

Full Model

$R^2 = 0.88$

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	14	4.711	0.336	7.78
Error	15	0.649	0.043	Prob > F
C. Total	29	5.360		0.0002

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	10	0.609	0.061	7.61
Pure Error	5	0.040	0.008	Prob > F
Total Error	15	0.649		0.0185
				Max RSq
				0.993

Parameter Estimates

Term	Estimate	t Ratio	Prob>|t|
Intercept	83.2	979.64	<.0001
X1(700,2100)&RS	-0.0417	-0.69	0.4984
X2(20,40)&RS	0.30833	5.13	0.0001
X3(3,15)&RS	0.28333	4.72	0.0003
X4(8,16)&RS	0.31667	5.27	<.0001
X1*X1	-0.0375	-0.47	0.6437
X2*X1	0.125	1.20	0.2481
X2*X2	-0.1125	-1.42	0.1772
X3*X1	0.125	1.20	0.2481
X3*X2	-0.25	-2.40	0.0296
X3*X3	-0.225	-2.83	0.0126
X4*X1	0.025	0.24	0.8133
X4*X2	-0.3	-2.88	0.0114
X4*X3	-0.225	-2.16	0.0471
X4*X4	-0.125	-1.57	0.1365

Significant Lack-of-fit in full model

All Terms involving X_1 are not significant

Reduced model (without X_1 terms)

$R^2 = 0.85$

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	9	4.553	0.506	12.53
Error	20	0.807	0.040	Prob > F
C. Total	29	5.360		<.0001

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	9	0.572	0.064	2.98
Pure Error	11	0.235	0.021	Prob > F
Total Error	20	0.807		0.0460
				Max RSq
				0.956

Parameter Estimates

Term	Estimate	t Ratio	Prob>|t|
Intercept	[illegible]	[illegible]	[illegible]
X2(20,40)&RS	[illegible]	[illegible]	[illegible]
X3(3,15)&RS	[illegible]	[illegible]	[illegible]
X4(8,16)&RS	[illegible]	[illegible]	[illegible]
X2*X2	[illegible]	[illegible]	[illegible]
X3*X2	[illegible]	[illegible]	[illegible]
X3*X3	-0.21964	-2.89	0.0090
X4*X2	-0.3	-2.99	0.0073
X4*X3	-0.225	-2.24	0.0366
X4*X4	-0.11964	-1.58	0.1308

Still, significant Lack-of-fit in reduced model

GP Generated transformations

Fit model in transformed variables

$$S_k = \beta_o + \sum_{i=2}^{4}\beta_i Z_i + \sum_{i<j}\sum \beta_{ij} Z_i Z_j + \sum_{i=2}^{4}\beta_{ii} Z_i^2$$

$$y = \frac{|x_2|^{0.54528}}{\sqrt{|\ln(x_3 x_2 + x_3)| * x_2 x_4}}$$

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	9	4.7080	0.5231	16.045
Error	20	0.6520	0.0326	Prob > F
C. Total	29	5.3600		<.0001

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	9	0.4170	0.0463	2.169
Pure Error	11	0.2350	0.0214	Prob > F
Total Error	20	0.6520		0.1131
				Max RSq
				0.956

Parameter Estimates

Term	Estimate	t Ratio	Prob>|t|	VIF
Intercept	82.8704	748.58	<.0001	.
Z2(4.47214,6.32456)&RS	0.4771	7.31	<.0001	1.573
Z3(0.60768,0.95406)&RS	-0.3578	-6.49	<.0001	1.371
Z4(0.0625,0.125)&RS	-0.4379	-7.14	<.0001	1.477
Z2*Z2	-0.0887	-1.29	0.2128	1.034
Z2*Z3	0.28248	3.50	0.0022	1.415
Z3*Z3	-0.0724	-0.60	0.5556	1.254
Z2*Z4	0.23959	2.75	0.0123	1.151
Z3*Z4	-0.2631	-3.37	0.0030	1.525
Z4*Z4	-0.0166	-0.21	0.8362	1.095

Variable transformations suggested by GP model

Original Variable	Transformed Variable
X_2	$Z_2 = X_2^{0.5}$
X_3	$Z_3 = [\mathrm{Log}(X_3)]^{-0.5}$
X_4	$Z_4 = X_4^{-1}$

Summary of Fit

RSquare	0.878
RSquare Adj	0.824
Root Mean Square Error	0.181
Mean of Response	83
Observations (or Sum Wgts)	30

Notice no significant Lack-of-fit p>0.05

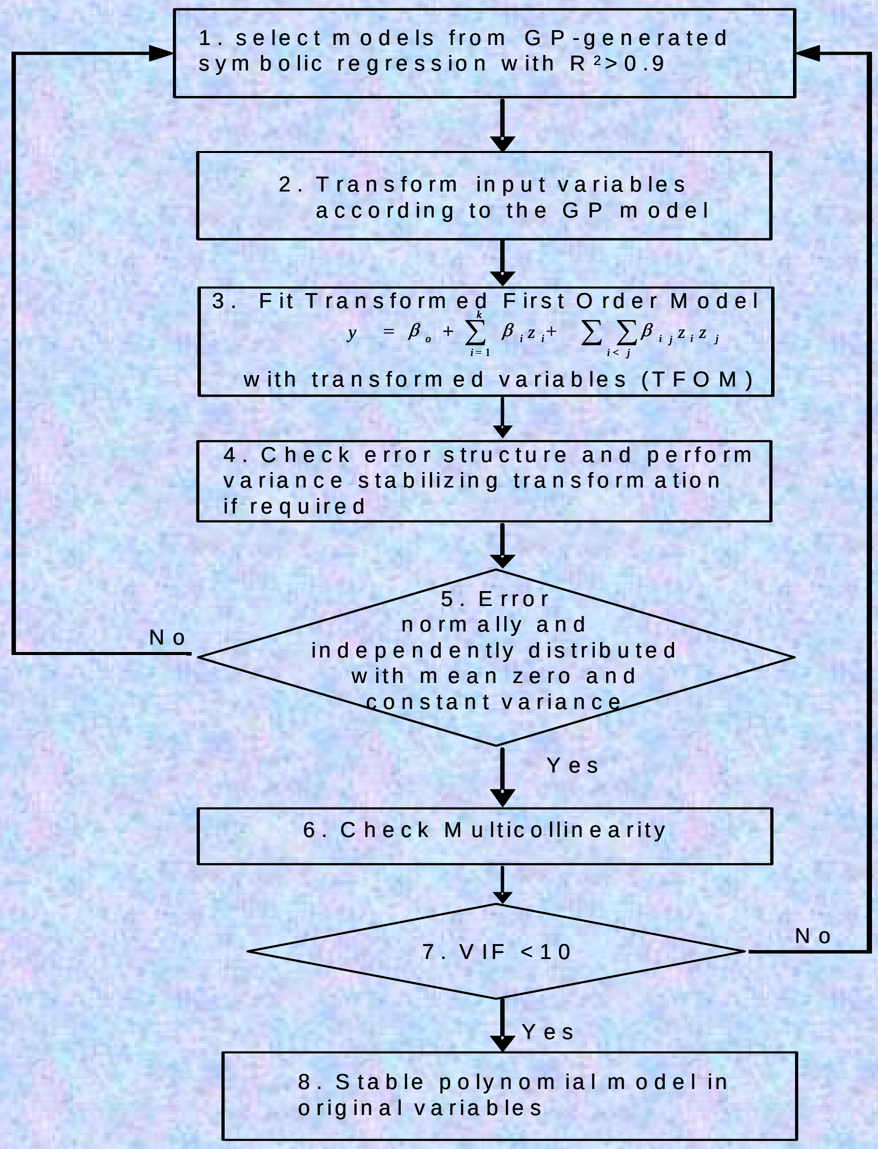
Symbolic Regression in Multicollinearity Problems

- **Plant data often becomes the focus of a modeling exercise.**
- **Initial model consider: multiple regression model (MLR)**
- **Issues with plant data**
  - Data collinearity: relationship between inputs
  - Severe Multicollinearity :
    - Affects the precision of the estimated regression coefficients.
    - Can cause real concerns with the stability, validity, and usefulness of the resulting model

Possible Multicollinearity Solutions

- Use PCA, PLS to create independent meta-variables (linear combinations of inputs)
- Meta-variables are independent of each other however variable interpretation is a challenge (plant people)
- Collect additional data (not always feasible)
- Try a different transformed model
- GP can help minimize multicollinearity in MLR models.

Proposed Approach Using GP to Minimize Multicollinearity



1. Generate several GP models
2. Generate non-linear input transforms according to GP model
3. Fit MLR model in transformed variables
4. Perform statistical analysis and check Multicollinearity (check error structure, residuals, correlations (VIF))
5. Repeat steps 2-4 until a stable MLR model is obtained (multicollinearity is minimized)

Case study with small data set

The data set consisted of three inputs variables (x1-x3) and one response (y) from a chemical process

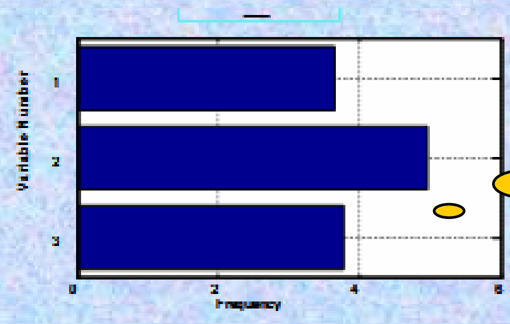
First order polynomial considered by MLR

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3$$

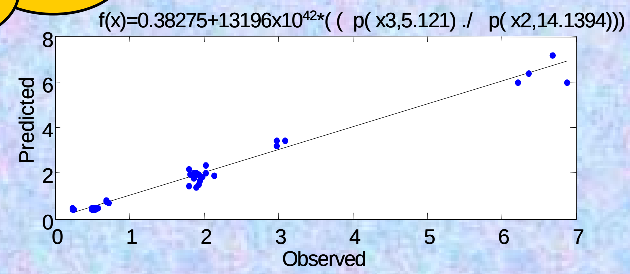
Term	Estimate	t Ratio	Prob>|t|	VIF
Intercept	-0.879	-7.145	<.0001	.
x1	0.265	1.526	0.137	5.46
x2	-4.246	-8.679	<.0001	77.58
x1*x2	0.537	2.701	0.011	3.00
x3	2.549	5.518	<.0001	68.20
x2*x3	0.891	4.318	<.0001	1.69

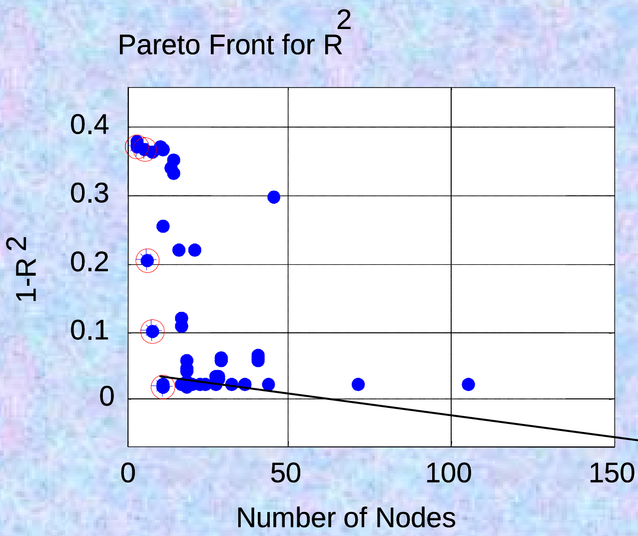
large Multicollinearity observed
VIF>10

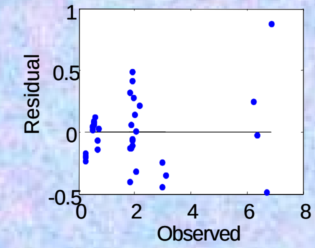
STEP 1. Generate GP models



X_2 was included most often







Error Statistics

Corr. Coeff. : 0.98906
Std. Dev. : 0.26842
Rel. Error : 0.14753
R^2 Statistic : 0.97824
RMSEP : 0.26842
Ratio Nodes : 11

$$y = \frac{0.38275 + 1.3196e42 * x_3^{5.121}}{x_2^{14.1394}}$$

STEP 2. Generate input transforms according to GP models

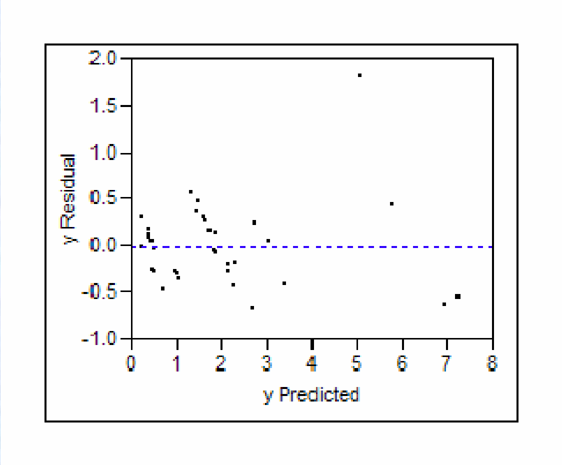
Original Variable	Transformed Variable
x_1	Z_1
x_2	$Z_2 = 1/x_2^{14}$
x_3	$Z_3 = x_3^5$

Relationship revealed by GP model

STEP 3. Fit MLR model in transformed variables

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \beta_{12} z_1 z_2 + \beta_{13} z_1 z_3 + \beta_{23} z_2 z_3$$

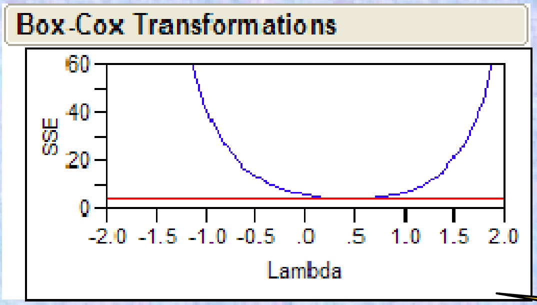
STEP 4. Perform statistical analysis and check (check error structure, residuals, correlations (VIF))



Error structure shows departure from constant variance assumption

Variance stabilizing transformation needed:

- Box and Cox Transformation:



$y = y^{\lambda}$

λ= Value that minimizes the SSE

Box and Cox transformation
$y = y^{0.5}$

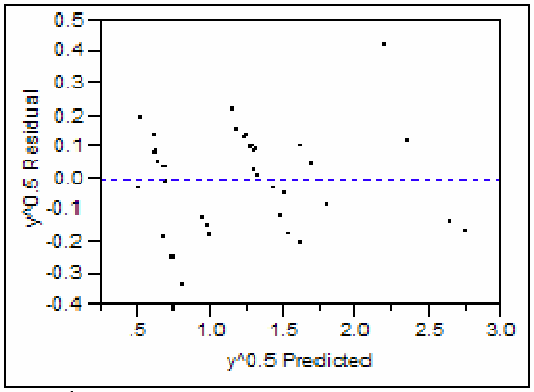
MLR model with variance stabilizing transformation:

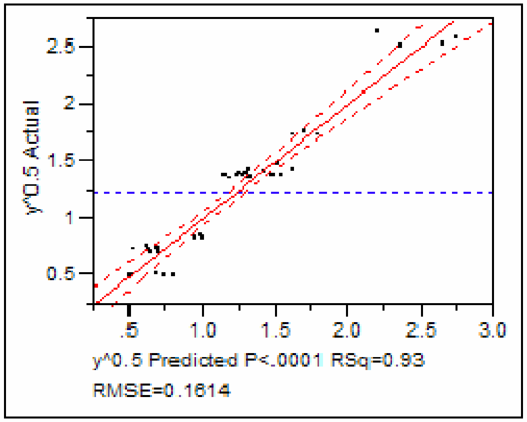
$$y^{\lambda} = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \beta_{12} z_1 z_2 + \beta_{13} z_1 z_3 + \beta_{23} z_2 z_3$$

MLR model with variance stabilizing transformation

Term	Estimate	t Ratio	Prob>|t|	VIF
Intercept	1.370	12.89	<.0001	.
x1	-0.476	-4.92	<.0001	3.60
1/x2^14	0.493	4.52	<.0001	5.35
x1*(1/x2^14)	-0.332	-2.49	0.0180	3.66
x3^5	-0.241	-3.00	0.0051	4.37

Improved model:
Stable polynomial model
No evidence of severe
Multicollinearity
VIF<10





adequate error structure:
Normally and independently distributed errors
with mean zero and constant variance

Case study with larger data set

- In another chemical process, data obtained from 3-month process history was used in empirical modeling effort
- A (detrimental) bi-product concentration was response (output) of interest
- All other variables considered potential inputs
- Can a reasonable empirical model be developed to predict how this bi-product output can be minimized?

Case study with larger data set

The data set consisted of thirteen inputs variables (x1-x13) and one response (*y*) from a chemical process

First order polynomial considered by MLR

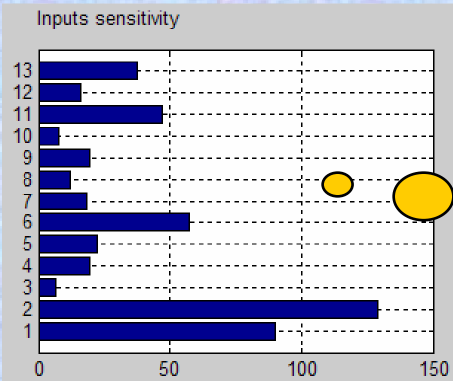
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3$$

Term	β Estimate	t Ratio	Prob>|t|	VIF
Intercept	230.70902	0.33	0.7432	
x1	0.9406677	19.31	<0.0001	3.84056
x2	-2.428614	-22.97	<0.0001	7.05279
x3	0.4005954	2.97	0.0041	9.42801
x4	-10.17105	-0.36	0.7217	861.2503
x5	2.956458	0.20	0.8385	343.7906
x6	10.223555	0.36	0.7164	918.9986
x7	-31.91927	-0.57	0.5686	3431.5002
x8	14.871442	0.35	0.7257	1976.0583
x9	-135.1481	-0.69	0.4919	1000231.8
x10	117.8077	0.68	0.4967	964097.17
x11	16.152238	0.40	0.6930	70850.669
x12	14.186557	0.89	0.3750	77.489476
x13	-19.53814	-0.67	0.5023	19404.123

Undesigned data will often be too unbalanced for standard modeling techniques

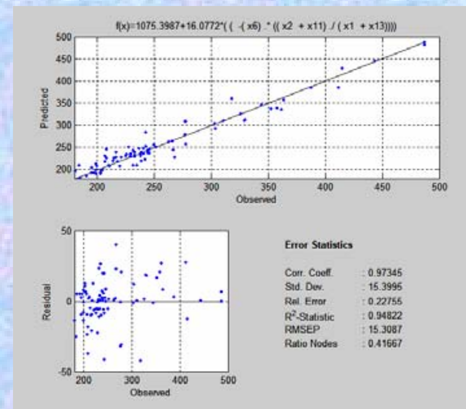
large Multicollinearity observed
VIF>10

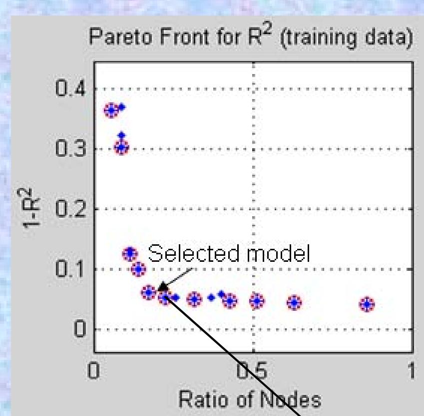
STEP 1. Generate GP models



X_1, X_2, X_6, X_{11}, & X_{13} were included most often

X1, X2 statistically significant inputs p<0.001



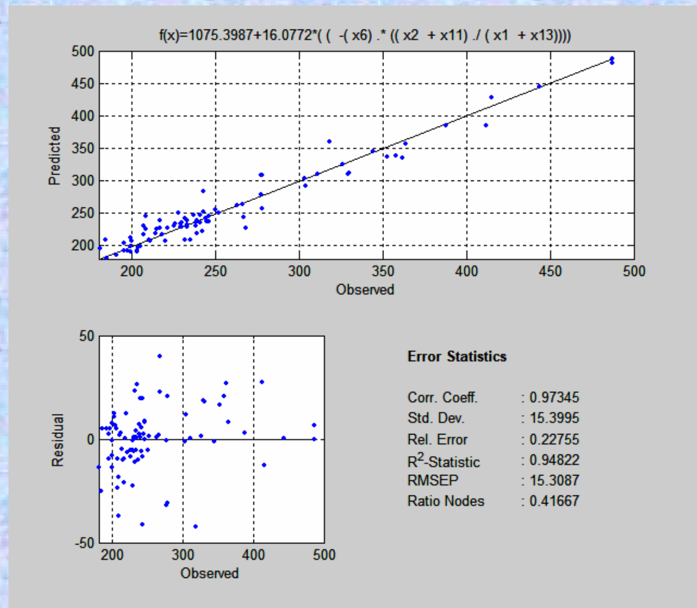


Pareto front optimization used to select model with “best” balance between performance & complexity

$$Y = 10275 - 16078 * \frac{x_6(x_2 + x_{11})}{x_1 + x_{13}}$$

STEP 2. Generate input transforms according to GP models

Original Variable	Transformed Variable
x_{2}, x_{11}	$Z_{1} = (x_2 + x_{11})$
x_{1}, x_{13}	$Z_2 = 1/(x_1 + x_{13})$
x_6	$Z_3 = x_6$



- **STEP 3.** Fit MLR model in transformed variables

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \beta_{12} z_1 z_2 + \beta_{13} z_1 z_3 + \beta_{23} z_2 z_3$$

Term	β Estimate	t Ratio	Prob>|t|	VIF
Intercept	2955.597	16.616	<0.0001	
Z_3 = x6	-7.265	-5.812	<0.0001	1.496
Z_1 = x2+x11	-2.148	-32.646	<0.0001	2.504
Z_2 = 1/x1+x13	-908023.43	-21.148	<0.0001	2.392

No multicollinearity problems

Conclusions

Approach using GP to minimized multicollineariy has been applied successfully in the Dow Chemical Company.

Unique features of the proposed approach

- •Combine linear regression models (designed experiments, undesigned data) with GP generated models
- •Uses the unique potential of GP generated models for suggesting variable transforms that minimized multicollinearity
- •Maximizes the use of available data when model extrapolation is required

Advantages of the approach

- •Produces stable polynomial (MLR model) with adequate error structure
- •provides a simple model which is easily understood by engineers and process people and offers
- •statistical analysis: outlier detection on the input space, influential observations and confidence band of the parameters can be applied offering additional assurance on the capabilities of the obtained model
- •Improves model validation (alternative models)