Symbolic Regression in Multicollinearity Problems

Flor A. Castillo
The Dow Chemical Company
2301 N. Brazozport Blvd., B1217
Freeport, TX, 77541
979-238-7554
facastillo@dow.com

Carlos M. Villa
The Dow Chemical Company
2301 N. Brazozport Blvd., B1217
Freeport, TX, 77541
979-238-5554
cmvilla@dow.com

ABSTRACT

In this paper the potential of GP-generated symbolic regression for alleviating multicollinearity problems in multiple regression is presented with a case study in an industrial setting. The main advantage of this approach is the potential to produce a simple and stable polynomial model in terms of the original variables.

Categories and Subject Descriptors

G.3. [**Mathematics of Computing**]: Probability and statistics– *Correlation and regression analysis.*

General Terms:

Experimentation.

Keywords

Multicollinearity, multiple regression, undesigned data

1. INTRODUCTION

In many industrial applications, observational data is collected and stored, to later become the focus of a modeling exercise. The main objective of this is usually process control. Suitable statistical techniques such as multiple regression are available to assist in this process [4]. However, statistically modeling of this type of data provides many challenges because data multicollinearity or strong relationships between inputs is usually present. Techniques such as Principal Component Regression, and Partial Least Squares [6] are often used. However the major limitation of these techniques especially from the point of view of plant personnel is that variable interpretation of the resulting principal components is often difficult.

GP- generated symbolic regression offers a unique opportunity because it produces several possible models of the response as a function of the input variables [3]. This set of possible models not only suggest additional candidate models but also indicate possible relationships among variables that, when applied to multiple regression modeling, have the potential to alleviate the problem of multicollinearity. This provides a stable regression model in terms of the original variables which is easier to implement and understand especially by plant personnel. This paper shows an application of Genetic Programming for data exploration to alleviate the problem of multicollinearity in a small data set.

GECCO'05, June 25-29, 2005, Washington, DC, USA.
ACM 1-59593-010-8/05/0006.

2. THE CASE STUDY

The data set consisted of three inputs variables (x_1-x_3) and one response (y) from a chemical process. A total of 39 observations were obtained from the plant. Table 1 shows the data set in which the input variables have been codified from -1 to 1.

Table 1 Undesigned Data Set

x1	x2	x3	y
0.41	0.48	0.45	0.23
-0.09	0.72	0.81	0.24
1.00	1.00	0.94	0.24
0.59	0.69	0.71	0.25
0.16	-0.06	-0.01	0.71
0.28	-0.20	-0.14	0.68
0.47	-0.12	-0.03	0.68
0.19	-0.60	-0.57	1.80
-0.16	-0.68	-0.44	2.02
-0.16	-0.68	-0.56	1.82
0.06	-0.70	-0.61	1.92
0.38	-0.62	-0.66	1.90
-0.16	-0.58	-0.50	1.92
-0.06	-0.65	-0.49	1.89
-0.09	-0.61	-0.38	1.90
-0.25	-0.89	-0.74	2.98
-0.09	-0.85	-0.69	2.97
-0.19	-0.87	-0.69	3.09
0.63	0.74	0.80	0.52
0.44	0.69	0.75	0.54
0.78	0.66	0.75	0.49
0.75	0.71	0.80	0.50
0.38	0.60	0.74	0.49
0.50	0.67	0.84	0.53
0.66	0.65	0.83	0.58
0.69	0.79	1.00	0.53
0.38	0.53	0.75	0.55
-1.00	-0.58	-0.36	1.98
-0.81	-0.54	-0.10	1.80
-0.81	-0.95	-0.56	6.21
-0.84	-1.00	-0.60	6.69
-0.72	-0.93	-0.50	6.88
-1.00	-0.97	-0.59	6.36
0.50	-0.80	-0.97	1.94
-0.03	-0.86	-0.99	2.03
0.16	-0.83	-0.99	1.87
0.03	-0.82	-0.93	2.14
0.16	-0.86	-1.00	1.87

The following first order polynomial model was initially considered:

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_{12}x_1x_2+\beta_{13}x_1x_3+\beta_{23}x_2x_3 \quad (1)$$

The β_i coefficients are the model parameters estimated by the method of least squares.

The corresponding parameter estimates showing only significant terms at the 95% confidence level is presented in Table 2.

Table 2. Parameter estimates for first order model

Term	**Estimate**	**t Ratio**	**Prob>|t|** [1]	**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

The model shown in Table 2 had a R^2 of 0.97 which shows that this model accounts for 97% of the variability in the data set.

Multicollinearity (correlation structure among the inputs) was examined using Variance Inflation Factors (VIF) [4] which are listed in the last column of Table 2. In general, high VIF's (some authors suggest VIF's greater than 10 [5]) suggest that severe multicollinearity exists and the model is suspected. From the VIF's listed in Table 2, it was obvious that severe multicollinearity issues existed within the data. This happens frequently with undesigned data from industrial situations. Many of the process variables will often vary together resulting in severely unbalanced data. One alternative often suggested is to remove any redundant inputs that may be included in the model. In our example, it was not possible to reduce any apparently redundant input because physically they have different meaning. Another alternative is planning a design of experiments (DOE) to collect additional data. However, in many industrial situations, collecting more data to help with the modeling is not a viable economic solution. Other option was to consider PCR (Principal Component Regression) or PLS (Partial Least Squares). However, this was not a practical solution because plant personnel in general prefer models in terms of the original variables. Under theses circumstances Symbolic regression was considered

2.1 Symbolic Regression Results

Several models of the response as a function of the input variables x_1, x_2, x_3 were obtained through GP-generated symbolic regression. The following model from the set of selected models presented the highest R^2 of 0.97

$$y=0.38275+\frac{1.3196\times10^{42}|x_3|^{5.121}}{|x_2|^{14.1394}} \quad (2)$$

The relationships revealed by the model are shown in Table 3.

Table 3. Transformation uncovered by GP

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

The linear regression model presented previously in equation (1) was re-fitted taking into account the transformed variables presented in Table 3. The resulting model is shown in Table 4. This model had and R^2 of 0.93 with appropriate error structure required by least squares and no evidence of severe multicollinearity as indicated by VIF.

Table 4 Parameter estimates for STM

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

The use of GP in this particular problem provided a simple polynomial form with no significant multicollinearity, which is easily understood by engineers and process people and offers a simple alternative form for monitoring and control. This polynomial model has the additional advantage that statistical analysis such us outlier detection on the input space [2], influential observations [1] and confidence band of the parameters can be applied offering additional assurance on the capabilities of the obtained model.

3. REFERENCES

[1] Cook, R.D. (1977), Detection of Influential observations in linear regression, *Technometrics*, 19, 15-18.

[2] Hoaglin, D. C., and Welsh, R. E. (1978), The Hat Matrix in Regression and ANOVA, *Amer. Stat.*, 32, 17-22

[3] Koza, J. (1992), Genetic Programming: On the Programming of Computers by Means of Natural Selection, MIT Press, Cambridge, MA.

[4] Montgomery, D and Peck, E. (1992) *Introduction to Linear Regression Analysis*, New York, NY: Wiley

[5] Myer, R. H., and Montgomery, D. (1995), *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, New York, NY: Wiley

[6] Wold, H. (1985), "Partial Least Squares," in Samuel Kotz and Norman L. Johnson, eds., *Encyclopedia of Statistical Sciences,* New York: Wiley , 6, 581-591.

[1] The column "prob>|t|" in Table 2 indicates the significance of the terms. At the 95% confidence, terms with prob>|t| values less than or equal to 0.05 are considered to be statistically significant.