Journal of the Korean Statistical Society (2005), 34: 3, pp 209–218

ALL POSSIBLE HIERARCHICAL QUADRATIC REGRESSIONS FOR RESPONSE SURFACES †

SUNG-SOO KIM[1], SOON-SUN KWON[2] AND SUNG-HYUN PARK[2]

ABSTRACT

In response surfaces analysis, we often proceed by supposing that, over a limited region of factor space, a polynomial of only first or second degree might adequately approximate the true function. To find the best subset model, all possible quadratic regressions for response surfaces can be very valuable to get optimum solutions under some reasonable experimentations. However, there is a very hard computational burden to get all possible quadratic regressions. In practice, it is sufficient to consider only hierarchical models. In this paper, we propose an algorithm to get all possible hierarchical quadratic regressions for fitting response surfaces.

AMS 2000 subject classifications. Primary 62G08; Secondary 62G20.
Keywords. Response surfaces, Triangular decomposition, Hierarchical regressions.

1. INTRODUCTION

In response surface analysis, we consider models that help to ensure successful experimentation. Suppose there are p independent variables and a dependent variable y. The second-order response surface model is

$$y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \sum_{j=1}^{p} \beta_{jj} x_j^2 + \sum_{i<j}^{p} \beta_{ij} x_i x_j + \epsilon \ ,$$

where ϵ is the error term which is distributed with mean zero and variance σ^2.

As in regression modelling, it is required to select best subset models, and the most obvious way to choose an appropriate submodel is to perform all of the

Received March 2005; accepted August 2005.
†This work was supported by a grant R01-2003-000-10220-0 from the Korea Science and Engineering Foundation.
[1]Department of Information Statistics, Korea National Open University, Seoul 110-791, Korea (e-mail: sskim@knou.ac.kr)
[2]Department of Statistics, Seoul National University, Seoul 151-747, Korea

possible $2^{p(p+3)/2} - 1$ regressions in the second-order model with p independent variables. The computational burden grows rapidly to get all possible subset models when independent variables are increased. This restricts our applying for all possible subset regressions in response surfaces. However, as Faraway(2004) says, when selecting variables in polynomial models, lower order terms should not be removed from the model before higher order terms in the same variable because of the scale change problem. Also in second-order response surfaces, we would not normally consider removing interaction terms without simultaneously considering the removal of related quadratic terms for the practical interpretation. These mean that it is sufficient to consider only hierarchical quadratic models for response surfaces. Hierarchical models are the models which include all the lower order interactions and main effects that are marginal to any higher order interaction in the model. Considering only all possible hierarchical quadratic regressions, there still remains hard computational burdens.

Many algorithms for doing all possible subset regressions are proposed, and these algorithms are generally based on sweepings or QR decompositions. Algorithms based on sweepings or QR decompositions depend on the previous results of sequential process to get all possible regressions. Hence applying these algorithms for all possible quadratic regression models is not practical because of serious computational burden and new sequences should be developed to do all possible hierarchical quadratic regressions. Maybe regressions by leaps and bounds proposed by Furnival and Wilson (1974) could be a possible solution in selecting variables in second-order quadratic regressions. Recently new algorithm based on triangular decomposition was proposed by Kim (2000). Unlike above algorithms, the algorithm based on triangular decomposition does not depend on the sequential sweepings or adjacent transpositions of columns. The results of triangular decomposition for each model can be obtained from those of initial full model. Also computational burden is not serious compared with that of sweepings or QR decompositions. In this paper we will provide an algorithm to get all possible hierarchical quadratic regressions based on triangular decomposition.

We will review in Section 2 all possible regressions by triangular decomposition and order of generation to get all possible regressions. In Section 3, we will provide a procedure for producing residual sums of squares for all possible hierarchical quadratic regression models. Also, we will provide an example for response surfaces.

2. REVIEW OF ALL POSSIBLE REGRESSIONS

Recently, all possible subset regressions using the triangular decomposition was proposed by Kim (2000), which is simple but computationally efficient compared with sweepings or QR decomposition methods. We review this algorithm briefly.

[Definition of Triangular Decomposition]

Let S be an $m \times m$ positive definite symmetric matrix. Then there exists a unique unit lower triangular matrix L and a unique diagonal matrix D with positive diagonal elements, such that $S = L^{-1}D(L^{-1})'$. The matrices L^{-1} and D are of the forms :

$$L^{-1} = \begin{bmatrix} 1 \ldots & 0 \\ \ldots\ \ldots\ \ldots & \\ l_{ij} \ \ldots & 1 \end{bmatrix}, D = \begin{bmatrix} d_1 \ \ldots & 0 \\ \ldots\ \ldots\ \ldots & \\ 0 \ \ldots & d_m \end{bmatrix}, d_j > 0,\ i, j = 1, 2, \ldots, m$$

[Property of Triangular Decomposition]

Let $Z = (Y : X)$ where Y is a dependent variable vector, X is an $n \times p$ design matrix, and L and D be the resulting matrices of the triangular decomposition of $Z'Z$. We denote L and D as follows .

$$L = \begin{bmatrix} 1 & 0 & 0 \ldots 0 \\ \lambda_{1y} & 1 & 0 \ldots 0 \\ \lambda_{2y} & \lambda_{21} & 1 \ldots 0 \\ \vdots & \vdots & \vdots \ \vdots \ \vdots \\ \lambda_{py} & \lambda_{p1} & \lambda_{p2} \ \ldots 1 \end{bmatrix}, D = diag(d_y, d_1, d_2, \ldots, d_p)$$

Then the residual sums of squares of the regression of Y on a subset of the other variables, say $(X_1, X_2, \ldots, X_k)$, is

$$SSE_{Y \cdot (X_1, X_2, \ldots, X_k)} = (\tfrac{1}{d_y} + \sum_{i=1}^{k} \tfrac{\lambda_{iy}^2}{d_i})^{-1} \ , \ k \leq p \ .$$

Using this property, we can obtain the residual sums of squares for the fitted model and also for the other models fitted in sequential order. For example, if four independent variables labelled $ABCD$ are fitted in that order, then the residual sums of squares for the models A, AB, ABC and $ABCD$ can be obtained sequentially. To obtain the residual sums of squares of all possible subset models,

detailed solutions were provided by Kim (2000) : (1) how to order the independent variables and (2) how to obtain the triangular decomposition for a simpler model using L and D for the full model without decomposing the design matrix of the simpler model. Here the full model means the model which includes all the independent variables. For the all possible hierarchical subset regressions, the order of generations is important, so we briefly introduce the order of deleting variables in case of p independent variables.

[Order of generation]

When the order of independent variables is defined, we can obtain the residual sums of squares sequentially in that order. For example, if $p = 4$ and the independent variables are labelled $ABCD$, then the following $2^{4-1} = 8$ orders can be implemented:

Order	*Model*
1 2 3 4	$A, AB, ABC, ABCD$
2 3 4	B, BC, BCD
1 3 4	AC, ACD
3 4	C, CD
1 2 4	ABD
2 4	BD
1 4	AD
4	D

Here (1234) means the independent variables ($ABCD$) respectively. This ordering can be obtained by deleting (1),(2),(12),(3),(13),(23),(123) from (1234) sequentially. This deleting sequence can be generalized as follows:

k	*Sequence*(S_k)
1	1
2	1 2 12
3	1 2 12 3 13 23 123
4	1 2 12 3 13 23 123 4 14 24 124 34 134 234 1234

Here $S_k = (S_{k-1}, k, T_{k-1})$, $(k = 2, 3, \cdots, p-1; S_1 = 1)$ where $T_{k-1} = (S_{k-1} : k)$. Using this sequence, we can efficiently construct the orders to obtain the residual sums of squares for all possible regressions.

3. ALL POSSIBLE HIERARCHICAL QUADRATIC REGRESSION MODELS FOR RESPONSE SURFACES

In response surfaces, we consider main effects, two-interaction effects and quadratic effects for practical reasons. Here, we propose the order of generation to obtain the residual sums of squares of all possible hierarchical quadratic regressions. Also we provide an example to select the best subset model. After selecting the best subset model, we can proceed canonical analysis for response surfaces.

3.1. Procedure to get All Possible Hierarchical Quadratic Regressions

In the second order response surface model,

$$y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \sum_{j=1}^{p} \beta_{jj} x_j^2 + \sum_{i<j}^{p} \beta_{ij} x_i x_j + \epsilon \ ,$$

if we use all possible algorithm based on triangular decomposition to select a best subset model for p main independent variables, the required number of triangular decomposition is $2^{p(p+3)/2-1}$. This number is incredibly increasing compared to the number of $2^{(p-1)}$ which is the required number of triangular decomposition considering only main effects. Hence, it is impractical to do all possible subset regressions in quadratic response surface models because of serious computational burdens. However, as Faraway (2004) says, practically in response surface analysis, if higher order terms are included, the lower order terms should be included. Considering this fact, it only suffices to consider the hierarchical regression models. To generate all possible hierarchical quadratic regression models using triangular decomposition, the sequential order of deleting variables is important. We provide the procedures for generating all possible hierarchical quadratic regressions.

In hierarchical regression models, if lower order terms are deleted, higher order terms related to the lower order terms do not exist in the model. Also in second-order response surfaces, we would not normally consider removing interaction terms without simultaneously considering the removal of related quadratic terms. Considering these facts, we can devise new stepwise deleting order of variables from the initial full model. This stepwise deleting order consists of two-steps: deleting order of main effects and deleting order of interacting terms under the condition that main effects exist in the model. This stepwise deleting order is like the followings:

TABLE 3.1 *Deleting sequences for p=3 variables*

order	A	B	C	AB	AC	BC	A^2	B^2	C^2
F	1	2	3	4	5	6	7	8	9
1	.	2	3	.	.	6	.	8	9
2	1	.	3	.	5	.	7	.	9
12	.	.	3	.	.	.	.	.	.
3	1	2	.	4	.	.	7	8	.
4	1	2	3	.	5	6	7	8	9
5	1	2	3	4	.	6	7	8	9
6	1	2	3	.	.	6	.	8	9

(1) Initially, make the full design matrix consisting of main effects, interaction terms and quadratic terms.

(2) If lower order terms are deleted, then related higher order terms are also deleted. For example, if main effect A is deleted, then related higher order terms AB and A^2 are also deleted. Hence, first we generate deleting order of main effects similar to the order used in Kim (2000). For example, for $p = 4$(called 1234) variables, the order of deleting variables is (1)-(2)-(12)-(3)-(13)-(23)-(123)-(4)-(14)-(24)-(34). Here the last variable terms (4)-(14)-(24)-(34) are added since we should consider the interaction terms. After deleting these variables, if there are more than 3 variables, then the deleting order of Section 2 is applied to interaction terms.

(3) The second step is to generate the deleting order of interaction terms under the condition that all main effects are included. This step is like the order of all possible subset regressions, so the same order as the deleting order of generation of Section 2 is only applied to interaction terms.

For easy understanding, we will give two examples for $p = 3$ and $p = 4$ variables. The order of deleting sequences for $p = 3$ and $p = 4$ variables are given in [Table 3.1] and [Table 3.2] respectively. The required number of triangular decomposition for $p = 3, 4, 5$ to get all possible hierarchical quadratic models is only 8, 55, 723 respectively, while the corresponding numbers of triangular decompositions to do all possible subset regressions of second-order response surface model are 256, 8192, 524288 respectively.

TABLE 3.2 *Deleting sequences for p=4 variables*

order	*A*	*B*	*C*	*D*	*AB*	*AC*	*AD*	*BC*	*BD*	*CD*	A^2	B^2	C^2	D^2
F	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	.	2	3	4	.	.	.	8	9	10	.	12	13	14
1-8	.	2	3	4	.	.	.	.	9	10	.	12	13	14
1-9	.	2	3	4	.	.	.	8	.	10	.	12	13	14
1-89	.	2	3	4	.	.	.	.	.	10	.	.	13	14
2	1	.	3	4	.	6	7	.	.	10	11	.	13	14
2-6	1	.	3	4	.	.	7	.	.	10	11	.	13	14
2-7	1	.	3	4	.	6	.	.	.	10	11	.	13	14
2-67	1	.	3	4	.	.	.	.	.	10	.	.	13	14
12	.	.	3	4	.	.	.	.	.	10	.	.	13	14
3	1	2	.	4	5	.	7	.	9	.	11	12	.	14
3-5	1	2	.	4	.	.	7	.	9	.	11	12	.	14
3-7	1	2	.	4	5	.	.	.	9	.	11	12	.	14
3-57	1	2	.	4	.	.	.	.	9	.	.	12	.	14
13	.	2	.	4	.	.	.	.	9	.	.	12	.	14
23	1	.	.	4	.	.	7	.	.	.	11	.	.	14
123	.	.	.	4	.	.	.	.	.	.	.	.	.	.
4	1	2	3	.	5	6	.	8	.	.	11	12	13	.
4-5	1	2	3	.	.	6	.	8	.	.	11	12	13	.
4-6	1	2	3	.	5	.	.	8	.	.	11	12	13	.
4-56	1	2	3	.	.	.	.	8	.	.	.	12	13	.
14	.	2	3	.	.	.	.	8	.	.	.	12	13	.
24	1	.	3	.	.	6	.	.	.	.	11	.	13	.
34	1	2	.	.	5	.	.	.	.	.	11	12	.	.
5	1	2	3	4	.	6	7	8	9	10	11	12	13	14
6	1	2	3	4	5	.	7	8	9	10	11	12	13	14
56	1	2	3	4	.	.	7	8	9	10	11	12	13	14
7	1	2	3	4	5	6	.	8	9	10	11	12	13	14
57	1	2	3	4	.	6	.	8	9	10	11	12	13	14
67	1	2	3	4	5	.	.	8	9	10	11	12	13	14
567	1	2	3	4	.	.	.	8	9	10	11	12	13	14
8	1	2	3	4	5	6	7	.	9	10	11	12	13	14
58	1	2	3	4	.	6	7	.	9	10	11	12	13	14
68	1	2	3	4	5	.	7	.	9	10	11	12	13	14
568	1	2	3	4	.	.	7	.	9	10	11	12	13	14
78	1	2	3	4	5	6	.	.	9	10	11	12	13	14
578	1	2	3	4	.	6	.	.	9	10	11	12	13	14
678	1	2	3	4	5	.	.	.	9	10	11	12	13	14
5678	1	2	3	4	.	.	.	.	9	10	.	12	13	14
9	1	2	3	4	5	6	7	8	.	10	11	12	13	14
59	1	2	3	4	.	6	7	8	.	10	11	12	13	14
69	1	2	3	4	5	.	7	8	.	10	11	12	13	14
569	1	2	3	4	.	.	7	8	.	10	11	12	13	14
79	1	2	3	4	5	6	.	8	.	10	11	12	13	14
579	1	2	3	4	.	6	.	8	.	10	11	12	13	14
679	1	2	3	4	5	.	.	8	.	10	11	12	13	14
5679	1	2	3	4	.	.	.	8	.	10	.	12	13	14
89	1	2	3	4	5	6	7	.	.	10	11	12	13	14
589	1	2	3	4	.	6	7	.	.	10	11	.	13	14
689	1	2	3	4	5	.	7	.	.	10	11	12	13	14
5689	1	2	3	4	.	.	7	.	.	10	11	.	13	14
789	1	2	3	4	5	6	.	.	.	10	11	12	13	14
5789	1	2	3	4	.	6	.	.	.	10	11	.	13	14
6789	1	2	3	4	5	.	.	.	.	10	11	12	13	14
56789	1	2	3	4	.	.	.	.	.	10	.	.	13	14

If we use the QR decomposition procedure proposed by Smith and Bremner (1989) to do all possible quadratic regressions, the minimum numbers of sequences are 502, 16369, 1048555 respectively. While the ordering of subsets of sweeping methods proposed by Garside (1965) are 511, 16383, 1048575 respectively. Hence computational burden of our proposed procedure is considerably diminished, and

TABLE 3.3 *An example in Myers(1976), p223*

x_1	x_2	x_3	y	x_1	x_2	x_3	y
-1	-1	-1	57	0	0	0	63
1	-1	-1	40	-2	0	0	28
-1	1	-1	19	2	0	0	11
1	1	-1	40	0	-2	0	2
-1	-1	1	54	0	2	0	18
1	-1	1	41	0	0	-2	56
-1	1	1	21	0	0	2	46
1	1	1	43				

all possible hierarchical quadratic regressions for response surfaces can be used in real problems.

3.2. Example

We provide example using the p=3 independent variables of [Table 3.3]. First we make full design matrix of X after checking collinearity. In fact, it is easy to check the collinearity since the diagonal elements d_j that are near to zero indicate that the jth variable can nearly be written as a linear combination of its predecessors. After checking the collinearity the design matrix for response surfaces is given as $X = (1, x_1, x_2, x_3, x_1x_2, x_1x_3, x_2x_3, x_1^2, x_2^2, x_3^2)$ and the results of all possible hierarchical quadratic regressions are shown in [Table 3.4]. From this result, we can select a best subset model using the criterion like R^2_{adj} or C_p. If we use criterion R^2_{adj} , the model based on variables $(x_1, x_2, x_1x_2, x_1^2, x_2^2)$ can be selected. After fitting regression model, we can proceed further canonical analysis for response surfaces.

4. CONCLUDING REMARKS

When we are interested in variable selection in response surface analysis, getting all possible subset models is a best solution. However the amount of calculations required to perform all possible subset models is formidable. In second-order response models, it is practical that if higher order terms are included, the lower order terms should be included. Also in second-order response surfaces, we would not normally consider removing interaction terms without simultaneously considering the removal of related quadratic terms. Considering these facts, we can only consider the all possible hierarchical quadratic regression

TABLE 3.4 *All possible hierarchical quadratic regressions for $p = 3$ variables*

N	R^2	R^2_{adj}	Cp	*Variables*
1	0.0058	-0.0707	16.8624	x_1
1	0.0179	-0.0577	16.5225	x_2
1	0.0038	-0.0729	16.9180	x_3
2	0.0237	-0.1391	18.3610	x_1, x_2
2	0.0217	-0.1414	18.4167	x_2, x_3
2	0.0095	-0.1555	18.7565	x_1, x_3
3	0.0274	-0.2378	20.2552	x_1, x_2, x_3
3	0.0229	-0.2435	20.3808	x_2, x_3, x_2x_3
3	0.0102	-0.2598	20.7382	x_1, x_3, x_1x_3
3	0.1629	-0.0654	16.4581	x_1, x_2, x_1x_2
4	0.1667	-0.1666	18.3523	x_1, x_2, x_3, x_1x_2
4	0.3181	0.0454	14.1092	x_2, x_3, x_2x_3, x_2^2
4	0.1087	-0.2479	19.9784	x_1, x_3, x_1x_3, x_1^2
4	0.2614	-0.0340	15.6983	x_1, x_2, x_1x_2, x_1^2
4	0.0281	-0.3607	22.2369	x_1, x_2, x_3, x_1x_3
4	0.0287	-0.3598	22.2193	x_1, x_2, x_3, x_2x_3
5	0.1674	-0.2952	20.3340	x_1, x_2, x_3, x_1x_2, x_1x_3
5	0.3821	0.0388	14.3171	x_2, x_3, x_2x_3, x_2^2, x_3^2
5	0.2323	-0.1941	18.5130	x_1, x_3, x_1x_3, x_1^2, x_3^2
5	0.7893	0.6722	2.9054	x_1, x_2, x_1x_2, x_1^2, x_2^2
5	0.0294	-0.5099	24.2010	x_1, x_2, x_3, x_1x_3, x_2x_3
5	0.1680	-0.2943	20.3164	x_1, x_2, x_3, x_1x_2, x_2x_3
5	0.3239	-0.0518	15.9477	x_1, x_2, x_3, x_2x_3, x_2^2
6	0.1686	-0.4549	22.2981	x_1, x_2, x_3, x_1x_2, x_1x_3, x_2x_3
6	0.1278	-0.5263	23.4412	x_1, x_2, x_3, x_1x_3, x_2x_3, x_1^2
6	0.2665	-0.2837	19.5566	x_1, x_2, x_3, x_1x_2, x_2x_3, x_1^2
6	0.3878	-0.0713	16.1556	x_1, x_2, x_3, x_2x_3, x_2^2, x_3^2
7	0.2671	-0.4658	21.5383	x_1, x_2, x_3, x_1x_2, x_1x_3, x_2x_3, x_1^2
7	0.6557	0.3114	10.6483	x_1, x_2, x_3, x_1x_3, x_2x_3, x_1^2, x_2^2
7	0.7943	0.5887	6.7637	x_1, x_2, x_3, x_1x_2, x_2x_3, x_1^2, x_2^2
8	0.7950	0.5216	8.7454	x_1, x_2, x_3, x_1x_2, x_1x_3, x_2x_3, x_1^2, x_2^2
8	0.6823	0.2587	11.9029	x_1, x_2, x_3, x_1x_3, x_2x_3, x_1^2, x_2^2, x_3^2
8	0.8209	0.5822	8.0183	x_1, x_2, x_3, x_1x_2, x_2x_3, x_1^2, x_2^2 ,x_3^2
9	0.8216	0.5004	10	x_1, x_2, x_3, x_1x_2, x_1x_3, x_2x_3, x_1^2, x_2^2, x_3^2

models. To apply sweeping methods or QR decomposition methods to all possible hierarchical methods, the new procedures of ordering or sequences should be developed and this is to be the subject of future work. To do all possible subset regressions of second-order model, the branch and bound algorithm proposed by Furnival and Wilson (1974) maybe a possible solution.

Here we provide the procedures to do all possible hierarchical quadratic regressions for response surfaces. This is an extension of the procedures to get all possible subset models based on the triangular decomposition procedures proposed by Kim (2000). Like the procedures proposed by Kim (2000), each triangular decomposition results of each new model can be obtained using those of initial full model. This gives the possibility of adjusting the procedures to the environment of parallel distributed computing systems.

The proposed procedures to get all possible hierarchical quadratic regressions can be adjusted to select a best subset model containing dummy independent models. In fact, when dummy independent variables are included in the model, it is general to consider the interaction effects. In this case, we can only consider the hierarchical models without quadratic terms.

For convenience, we provide a program written in *R*. This program is for the model including two-factor interactions and quadratic terms. For reference, the user cpu time running Example 3.2 is 0.03 on the Pentium 4 CPU 3.20GHZ Window XP system.

ACKNOWLEDGEMENTS

We are most grateful to the referees for many helpful comments, which greatly improved the presentation of the paper.

REFERENCES

FARAWAY, J.J(2004).*Linear Models with R*, CRC Press.

FURNIVAL, G.M. AND WILSON, R.W.JR.(1974). "Regressions by leaps and bounds", *Technometrics*, **16**, 499-511.

GARSIDE, M.J.(1965). "The best subset in multiple regression analysis", *Applied Statist*, **14**, 196-200.

KIM,SUNG-SOO(2000). "All possible subset regressions using the triangular decomposition", *Journal of Statistical Computation and Simulation*, **65**, 81-94.

MYERS, R.H.(1976). *Response Surface Methodology*, Vlacksburg : VA.

SMITH, D. M. AND BERMNER, J. M.(1989). "All possible subset regressions using the QR decomposition", *Computational Statistics and Data Analysis*, **7**, 217-235.