Journal of the KSQC
Vol. 6, No. 1 (1978)

공업제품의 질을 관리하기 위한 반응표면 실험의 응용
-통계적 모형 적합과 반응의 예측을 중심으로-

(An Application of Response Surface Experiments to Control the Quality of Industrial Products : Model Fitting and Prediction of Responses)

朴 聖 炫 *

SUMMARY

In response surface experiments, a polynomial regression model is often used to fit the response surface to explore the functional relationship between a response variable and several independent variables, and to determine the optimum operating conditions, which would be desirable to control the quality of industrial products.

The problem considered in this paper is that of selecting subsets of polynomial terms from a given polynomial model so as to achieve "improved" response surfaces in estimation of the response. Such improvement in fitting the response surfaces would be very helpful to determine the optimum operating conditions and to explore the functional relationship with better precision. A criterion is proposed for selection of polynomial terms and illustrated with an industrial example.

要約 (FORWORD)

반응표본 실험에 있어서 반응변수와 여러개의 독립변수와의 함수관계를 규명하기 위하여 다항회귀모형이 많이 사용되고 있으며 또한 이 다항회귀모형은 최적반응조건을 결정하고 제품의 질을 조절하기 위하여서도 쓰여진다.

이 논문에서 연구하는 문제는 다항회귀모형을 구성하고 있는 많은 項 중에서 어떤 항들을 선택하여 주는 것이 精度있게 추정하기 위하여 적절한가 하는 문제이다. 精度가 향상되는 반응표면을 발견한다는 것은 최적반응조건을 결정하고 변수간의 함수관계를 정확하게 구하는데 도움을 준다.

다항회귀모형에서 적절한 項들을 선택하기 위하여 이 논문에서는 하나의 기준을 제시할 것이며, 실제로 공장에서 응용될 수 있는 예제를 들어 설명하고 있다.

序論 (INTRODUCTION)

반응변수(response variable)와 여러 독립변수 사이의 관계를 조사할 때 반응 변수는 보통 다음과 같이 가정된다

$$\eta(x) = \beta_0 + \sum \beta_i x_i + \sum \beta_{ij} x_i x_j + \sum \beta_{ijk} x_i x_j x_k + \cdots \cdots$$

이 식에서 x들은 실험변수들의 변환들인데, x들은 취급하는 영역(즉 R, 이 영역내에서 가정된 다항식이 사용된다)의 중심에 위치한다.

$\eta(x)$로 주어진 다항모형을 사용하므로써 다음과 같은 의문을 가질 수 있다.

반응표면을 나타내는 다항식에서 다항식의 차수(degree)가 주어졌을때, 이 주어진 차수 이하의 모든 항들은 포함시켜야 하는가? 또한, 다항모형에서 몇개의 항을 없애므로써, $\hat{y}(x)$의 精度(precision)의

* 서울大学校自然大学計算統計学科教授

관점에서 볼때, "좀더 나은"반응표면을 얻을 수 있을까? (단, $\hat{y}(x)$는 $\eta(x)$의 최소자승추정량이다)

사실상, 선형회귀에서 변수들의"가장 좋은"부분집합(subset)을 결정하는 문제는 많은 응용통계학자들의 관심을 모아오고 있다. 참고문헌으로 Draper & Smith(3), Mallows(7), Allens(1), Helms(4), Hocking(6), 박(8, 9, 10)등이 있다 문헌들에서 변수들을 골라내기 위해서 제공된 대부분의 기준들은, 단지 실험점들에서만의 $\hat{y}(x)$의 精度에 관계되어 있지만 일반적으로 반응곡면실험에서 적합시킨 방정식 $\hat{y}(x)$는 실험점들 뿐만 아니라 실험자가 취급하는 x의 어떤 영역 내에서도 사용될 수 있다. 만일 판단기준이 반응표면의 精度를 바탕으로 만들어졌다면, 전 영역 R에서 반응표면이 어느 정도로 잘 적합되었는가를 평가하는 것이 타당할 것이다.

박(8)은 취급하는 영역에서 $\hat{y}(x)$의 精度가 주된 관심사인 경우에, 반응표면을 적합시키는 완전한(full) 다항모형으로 부터 멱항(polynomial)들의 부분집합을 추출하는 기준을 제시했다.

이 논문의 주된 목적은 '박'의 기준이 어떻게 제조공업의 최적의 작업조건을결정하는데 적용되는 가를 예증하고, 반응을 예언하고 산업제품의 품질관리에 사용되는 반응표면을 좀더 적합하게 개선하는데 있다. 다음 절에서 '박'의 기준에 대한 간단한 요약이 있고, 다음에 공업문제의 예가 제시될 것이다.

추출기준 (SELECTION CRITERION)

투입된 멱항(polynomial term)들의 t차원 벡터 $x'=(1, x_1, x_2, \cdots, x_1^2, x_2^2, x_1x_2, \cdots\cdots)$에서 $\eta(\geq t)$개의 관측값이 있다고 하면 반응변수는 보통 다음과 같은 벡터 기호로 나타내진다.

$$y=x'\beta+e \qquad (1)$$

단 β는 미지의 회귀계수의 t차원 벡터이고 残差(residual) e는 평균이 0이고 미지의 분산이 σ^2인 독립동일분포이다.

r을 식 (1)에서 제거할 항의 갯수라 하면 $p=t-r$은 마지막 식에 남을 항의 갯수이다. 따라서 식 (1)을 분배된 벡터의 형태로 쓰면

$$y=x_p'\beta_p+x_r'\beta_r+e$$

이 식에서 x_p는 남을 항을 포함하고 x_r은 없어질 항을 포함한다.

$\hat{\beta}_p$와 $\hat{\beta}_r$의 성분을 가진 $\hat{\beta}$를 β의 최소자승법에 의한 추정량이라 하고, $\tilde{\beta}_p$를 완전모형(full model)에서 x_r의 멱항들을 제거한 부분모형(subset model)에서의 β_p의 최소자승추정량이라고 하면 다음과 같이 표시된다.

$$\hat{\beta}=(X'X)^{-1}X'Y, \quad \tilde{\beta}_p=(X_p'X_p)^{-1}X_p'Y$$

단 X와 X_p는 각각, 각 실험점에서 x와 x_p의 항들로 부터 얻어진 값들의 행렬이고 Y는 관찰된 반응들의 n차원 벡터이다. 완전모형을 사용하는 경우에 특정한 x에서의 반응의 추정치는 $\hat{y}(x)=x'\hat{\beta}$이고, x_r이 제거된 부분모형을 사용하면 반응의 추정치를 $\tilde{y}_p(x_p)=x_p'\tilde{\beta}_p$가 된다.

Hocking(5)은 만일 $Var(\hat{\beta}_r)-\beta_r\beta_r'$가 positive semi-definite 이면 $Var(\hat{\beta}_p)-MSE(\tilde{\beta}_p)$가 positive semi-definite 이고 $Var(\hat{y})\geq MSE(\tilde{y}_p)$임을 증명했다. 단 MSE는 평방평균오차(Mean Squared Error)를 의미한다.

이와같은 성질로 부터 유도되어서 제안된 기준은 다음의 식 Q를 최대로 하는 P멱항(polynomial term)들을 골라내는데 있다.

$$Q=\int_R[Var(\hat{y})-MSE(\tilde{y}_p)]dw(x)$$

이 식에서 적분은 취급하는 영역 R전체에서 하고, $W(x)$은 R에서의 확률밀도함수로 취급할 수 있는 가중함수이다. 또한 $W(x)$는 영역이 다른점에서 $\eta(x)$의 추정량의 중요성이 다른 경우도 허용이 되며, 필요하다면 이산집합에도 적용시킬 수 있다.

박(8)은,모수 σ^2과 β_r을 완전모형을 사용한 자료로 부터 계산된 추정치로 대치한 경우에 최대화 될 Q의 값은 다음과 같다는 것을 보였다.

$$\hat{Q}=\hat{\sigma}^2\{Tr[(X'X)^{-1}M]-Tr[(X_p'X_p)^{-1}M_{pp}]\}-\hat{\beta}_r'[A'M_{pp}A-2A'M_{pr}+M_{rr}]\hat{\beta}_r, \qquad (2)$$

단, $M=\int_R xx'\,dw(x)$,

$$M_{ij} = \int_R x_i x_j \, dw(x),$$

$$A = (X_p' X_p)^{-1} X_p' X_r,$$

Tr은 trace를 의미한다.

$\hat{Q}$의 첫 항은 $Var(\hat{y}) - Var(\tilde{y}_p)$의 적분자를 나타내는데 이것은 항상 +이거나 0이다. Q의 끝항은 $\tilde{y}_r$의 편기(biased)의 제곱을 적분한 것이다. 따라서 이 기준은 취급하는 전 영역 R에서 분산의 감소에서 얻어지는 精度의 증가가 편기의 제곱으로 인하여 잃어버리는 精度의 감소에 비하여 큰 부분집합을 찾아내는 것이라 볼 수 있다.

공업적인 예제
(AN INDUSTRIAL EXAMPLE)

이 절에서는, 앞에서 제안한 반응의 통계적 품질관리를 위한 $\hat{\theta}$ 기준에 의해서 어떻게 변항들을 뽑아내는 가를 보여주고, 실제적으로 Mallows의 Cp 통계량이 많이 이용되므로(문헌 (7)참조)이 두 가지 기준을 비교해 볼 것이다.

이 예제의 자료로는, 파라핀 왁스와 폴리에틸렌 첨가물이 첨가된 빵포장지의 봉합강도를 결정하는 문제에서, 최적작업 조건을 결정하는 3가지 요인의 효과를 실험한 Brown (2)의 예제를 사용했다.

이 문제의 종속변수와 독립변수는 다음과 같다.

y=봉합강도, gms/in

x_1=봉합온도

x_2=냉각봉의 온도

x_3=폴리에틸렌 첨가물의 퍼센트

또, 각 변수들의 실제값은 다음과 같이 주어진다.

x_1=(봉합온도−255°)/30

x_2=(냉각온도−55°)/9

x_3=(폴리에틸렌% −1.1 %)/0.6

설계행렬 D와 20개의 관찰값을 나타내는 벡타 Y는 다음과 같다.

$$D = \begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 \end{matrix} \\ \begin{bmatrix} -1 & -1 & -1 \\ 1 & -1 & -1 \\ -1 & 1 & -1 \\ 1 & 1 & -1 \\ -1 & -1 & 1 \\ 1 & -1 & 1 \\ -1 & 1 & 1 \\ 1 & 1 & 1 \\ -1.682 & 0 & 0 \\ 1.682 & 0 & 0 \\ 0 & -1.682 & 0 \\ 0 & 1.682 & 0 \\ 0 & 0 & -1.682 \\ 0 & 0 & 1.682 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \end{array} \quad Y = \begin{bmatrix} 6.6 \\ 6.9 \\ 7.9 \\ 6.1 \\ 9.2 \\ 6.8 \\ 10.4 \\ 7.3 \\ 9.8 \\ 5.0 \\ 6.9 \\ 6.3 \\ 4.0 \\ 8.6 \\ 10.1 \\ 9.9 \\ 12.2 \\ 9.7 \\ 9.7 \\ 9.6 \end{bmatrix}$$

이 실험계획법은 6개의 중심점을 가진 회전중심합성계획임을 유의하라

이 예제에서 취급영역은 단위정육면체, 가중함수 $W(x)$는 평등분포, 영역 R에서 2차 다항모형인 것으로 가정한다.

최소자승법에 의해서 구해진 2차반응 모형은 다음과 같다.

$$\hat{y}(x) = 10.1657 - 1.1038 x_1 + 0.0872 x_2 + 1.0206 x_3 - 0.7602 x_1^2 - 1.0430 x_2^2 - 1.1491 x_3^2 - 0.3500 x_1 x_2 - 0.5000 x_1 x_3 + 0.1500 x_2 x_3,$$

또 분산분석표는 도표1과 같다.

도표1. 분산분석표

변동요인	자유도	평 방 합	평균평방합	F
선형회귀	3	30.9654	10.3218	
2차회귀	6	39.3402	6.5567	
적합결여	5	6.9044	1.3809	1.39
반복오차	5	4.9600	0.9920	
합	19	82.1700		

적합의 결여도가 유의적이 아니기 때문에 2 차다항모형은 이 반응모형에 적합하고, 합동오차는 $\hat{\sigma}^2 = 1.1964$이다.

자세한 계산과정은 생략했으나, 제외되는 멱항들의 갯수가 주어진 각각의 경우에 대해 얻어지는 최량의 부표집합을 도표2에 정리해 두었다.

도표2. $\hat{Q}$와 C_p에 의한 다항들의 추출

제외되는항수	제외되는 항	$\hat{Q}$	C_p
1	x_2	0.4454	8.0875
2	$x_2, x_2 x_3$	0.6810	6.2392
3	$x_2, x_2 x_3, x_1 x_2$	0.7293*	5.0649
4	$x_2, x_2 x_3, x_1 x_2, x_1 x_3$	0.5388	4.7502*
5	$x_2, x_2 x_3, x_1 x_2, x_1 x_3, x_1^2$	-1.6838	9.7548

RSS는 잔차자승합(Residual sum of squares)이고 P를 남은 항의 갯수라고 하면 $C_p = RSS/\hat{\sigma}^2 + 2p - n$이므로 Mallows의 C_p는 최소화 기준이다. 그러나 $\hat{Q}$는 최대화 기준이다.

도표 2. 는 $\hat{Q}$가 3개의 항이 제외되었을때 최대이고 C_p는 4개의 항이 제외되었을때 최소임을 보였다. 따라서 완전이차모형에서 제외되는 항의 최적갯수가 다르다는 것이 $\hat{Q}$와 C_p가 크게 다른점이다.

완전이차모형의 기본반응 분석은 정상점(stationary point, 이 예제에서는 최대 반응점)은 다음과 같다.

$$x_0 = \begin{bmatrix} x_{10} = -1.0098 \\ x_{20} = 0.2602 \\ x_{30} = 0.6808 \end{bmatrix}$$

그리고 이 점에서 추정된 반응은 $\hat{y}_0 = 11.08$이다.

그러나 x_2, $x_2 x_3$, $x_1 x_2$가 제외된 부분모형에서는 최대점이

$$x_0 = \begin{bmatrix} x_{10} = -0.9392 \\ x_{20} = 0.0466 \\ x_{30} = 0.6481 \end{bmatrix}$$

이고 $\tilde{y}_0$ = 11.01이다.

$\hat{\beta}_r' = (\hat{\beta}_{23}, \hat{\beta}_{12})$를 관찰해 보면 다음과 같은 중요한 내용을 얻을 수 있다.

행렬 $Var(\hat{\beta}_r) - \hat{\beta}_r \hat{\beta}_r'$가 positive definite 이기 때문에 정상점 x_0, x_0를 포함한 실험가능영역의 어느 점에서도 $Var(\hat{y}) \geq MSE(\tilde{y}_p)$이다.

결과적으로 완전모형을 사용하는 대신 축소된 부분모형을 사용하면, 영역내의 어디에 존재하든지 정상점은 좀더 나은 精度(이것은 적은 평균평방오차를 의미한다)를 가지고 추정할 수 있다.

참 고 문 헌

1. Allen, D. (1971). Mean square error of prediction as criterion for selecting variables. Technometrics 13, 469-476.
2. Brown, D., Turner, W. & Smith, A. (1958). Sealing strenght of wax-polyethylene blends. Tappi 41, 295-300.
3. Draper, N. & Smith, H (1966). Applied regression analysis. John Wiley & Sons, Inc. New York.
4. Helms, R. (1974). The average estimated variance criterion for the selection of variables problem in general linear models. Technometrics 16, 261- 274.
5. Hocking, R. (1974). Misspecification in regression. The American Statistician 28, 39-40.
6. Hocking, R. (1976). The analysis and selection of variables in linear regression. Biometrics 32, 1- 49.
7. Mallows, C. (1973). Some comments on Cp. Technometrics 15, 661-675.
8. Park, S. (1977). Selection of polynomial terms for response surface experiments. Biometrics 33, 225-229.
9. Park, S. (1977). On screening of variables for response surface experiments with mixtures. The Journal of the Korean statistical society, Vol. 6. No. 2.
10. Park, S. (1978). Selecting contrasts among parameters in Scheffe's mixture models: screening components and model reduction. Technometrics 20 (to be published in August).