비선형모형분석을 위한 탐색적 자료분석

장대홍[1)]

요약

비선형모형분석의 초기 단계에서 초기값(starting value, initial parameter value)를 결정하는 문제는 비선형모형의 모수추정을 위한 반복기법의 수렴속도나 국소값(local minimum)문제에 영향을 주게 된다. 본 논문을 통하여 탐색적 자료분석이 초기값를 결정하는 데 도움을 줄 수 있음을 보이고자 한다.

주요용어: 초기값, 산포도행렬, 평행좌표그림

1. 서론

비선형모형을 적합시키기 위하여는 비선형모형분석의 초기 단계에서 모수의 초기값을 결정하여 주어야 한다. 초기값을 잘못 입력하면 모수추정을 위한 반복기법의 수렴속도가 떨어지거나 오차의 최소제곱합을 찾지 못하고 국소제곱합을 찾는 것으로 끝날 수 있다. 그러므로, 초기값을 결정하는 문제는 비선형모형의 분석시 중요한 초기단계라 하겠다. 초기값을 결정하는 위하여 다음과 같은 방법들을 쓸 수 있다(Bates와 Watts(1988), Seber와 Wild(1989), Draper와 Smith(1998), Myers외 2인(2002)).

1. 모수의 수가 p개라면 반응변수의 오차를 무시하고 실험점을 p개 선택하여 연립방정식을 구하여 이 해를 초기값으로 이용한다.

2. 비선형모형에서 모수는 물리적인 의미를 갖는 경우가 많으므로 모수를 변화시켜 가며 이 비선형모형의 변하는 모습을 참조하여 초기값을 결정한다.

3. 이 비선형모형이 변수변환을 이용하여 선형모형이 된다면 이 선형모형에 대한 모수추정값을 초기값으로 정한다.

4. 격자탐색(grid search)이나 랜덤탐색(random search)을 이용하여 초기값으로 정한다.

이러한 방법들을 쓸 때, 탐색적 자료분석을 이용하면 초기치를 결정하는 데 도움을 줄 수 있다.

2. 초기값 결정을 위한 탐색적 자료분석

비선형모형에서 모수는 물리적인 의미를 갖는 경우가 많으므로 모수를 변화시켜 가며 이 비선형모형의 변하는 모습과 실험점들로 만들어지는 반응표면을 서로 비교하여 초기값을 결정할 때 Maple이나 Mathematica같은 수학패캐지를 사용하면 비선형모형의 변하는 모습을 실시간 확인할 수 있다. 예(예 1)로, 다음 표 1과 같이 설명변수는 기질응축도(substrate concentration, 단위는 ppm)이고, 반응변수는 반응속도(velocity, 단위는 $counts/\ \min^2$)인 puromycin 자료를 생각하여 보자(Myers외 2인(2002)에서 인용하였음.). 이 자료에 적합한 비선형모형으로서 다음과 같은 Michaelis-Menten모형을 사용한다.

$$E(y) = f(x,\ \theta) = \frac{\alpha x}{\beta + x}$$

1) (608-737) 부산광역시 남구 대연3동 599-1 부경대학교 자연과학대학 수리과학부 통계학전공 교수

표 1. puromycin 자료

기질응축도 x	반응속도 y
0.02	47 76
0.06	97 107
0.11	123 139
0.22	152 159
0.56	191 201
1.10	200 207

여기서, θ는 모수벡터이다. 이 모형에서 α는 $x\to\infty$ 일 때의 반응속도, 즉 반응속도의 점근속도이고, β는 절반응축도, 즉, 이 값 β에 도달 시 반응속도는 최대반응속도의 절반에 달하게 된다. Maple 명령어를 이용하여 표 1의 자료를 리스트로 만든 후 실험점들을 그림으로 나타낼 수 있고 'animate'라는 Maple 명령어를 이용하여 α와 β 각각을 변화시켜 가며 움직이는 그림을 그리면 이 비선형모형의 변하는 모습을 실시간 확인할 수 있고, 이 비선형모형의 변하는 모습과 실험점들로 만들어지는 반응곡선을 서로 비교하면 초기값을 결정하는 데 도움을 줄 수 있다. 또한, α와 β의 물리적인 의미를 파악할 수 있다. 이를 시행하기 위한 Maple명령어는 다음과 같다.

```
list1:=[[0.02,47],[0.02,76],[0.06,97],[0.06,107],[0.11,123],[0.11,139],[0.22,152],[0.22,159],[0.56,191],[0.56,
201],[1.1,200],[1.1,207]];
plot(list1,x=0..1.2,y=0..250,labels=[concentration,velocity],style=point,axes=boxed);
animate(alpha*x/(0.06+x),x=0..1.2,alpha=100..300,labels=[concentration,velocity],axes=boxed);
animate(210*x/(beta+x),x=0..1.2,beta=0..0.2,labels=[concentration,velocity],axes=boxed);
```

다른 예(예 2)로서 Khuri와 Cornell(1996)에 나타나는 다음 표2와 같은 예를 보자.

표 2. 암소체중 자료

가축률 (kg/ha) x_1	보충먹이량 (kg/일) x_2	체중증가량 (kg/일) y
242	5.0	0.715
665	1.5	0.591
687	8.5	0.592
1432	0.0	0.248
1530	10.0	0.428
1542	5.0	0.321
2108	1.5	0.124
2469	8.5	0.215
2570	5.0	0.190

이 자료에 비선형모형 $E(y)=f(x,\theta)=\theta_1+\theta_2 e^{\frac{-\theta_3}{x_1}-\theta_4 x_2}$를 적합시키고자 한다. Maple 명령어를 이용하여 표 2의 자료를 리스트로 만든 후 실험점들을 3차원 그림으로 나타낼 수 있고 'animate3d'라는 Maple 명령어를 이용하여 $\theta_1,\theta_2,\theta_3,\theta_4$ 각각을 변화시켜 가며 움직이는 3차

원 그림을 그리면 이 비선형모형의 변하는 모습을 실시간 확인할 수 있고, 이 비선형모형의 변하는 모습과 실험점들로 만들어지는 반응표면을 서로 비교하면 초기값을 결정하는 데 도움을 줄 수 있다. 이를 시행하기 위한 Maple명령어는 다음과 같다.

```
with(plots):
list2:=[[242,5.0,0.715],[665,1.5,0.591],[687,8.5,0.592],[1432,0.0,0.248],[1530,10.0,0.428],[1542,5.0,0.321],[
2108,1.5,0.124],[2469,8.5,0.215],[2570,5.0,0.190]];
pointplot3d(list2,labels=[stock,intake,weight],axes=boxed,color=red);
animate3d(theta1-1.2*exp(-1380/x-0.05*y),x=0..3000,y=0..15,theta1=0..2,labels=[stock,intake,weig
ht],axes=boxed);
animate3d(0.7+theta2*exp(-1380/x-0.05*y),x=0..3000,y=0..15,theta2=-2..2,labels=[stock,intake,wei
ght],axes=boxed);
animate3d(0.7-1.2*exp(-theta3/x-0.05*y),x=0..3000,y=0..15,theta3=1000..2000,labels=[stock,intake,
weight],axes=boxed);
animate3d(0.7-1.2*exp(-1380/x-theta4*y),x=0..3000,y=0..15,theta4=0..0.1,labels=[stock,intake,weig
ht],axes=boxed);
```

랜덤탐색(random search)을 이용하여 초기값으로 정할 때 탐색적 자료분석을 이용하면 효과적이다. 모수공간의 직사각형영역(rectangular region)을 적당히 정하여 난수발생기를 사용하여 난수를 구한 후 산포도행렬(scatterplot matrix)과 평행좌표그림(parallel coordinate plot)을 이용하면 초기값을 결정하는 데 도움을 줄 수 있다. 한 예(예 3)로서 Seber와 Wild(1989)에 나와 있는 다음과 같은 표 3의 자료에 대하여

표 3. 인위적인 자료

설명변수 x	반응변수 y
-2	0
-1	1
1	-0.9
2	0

비선형모형 $E(y)=f(x,\theta)=\alpha e^{-\beta x}$를 적합시킬 때 오차의 제곱합이 두 개의 국소값을 갖게 된다. 초기값을 어떻게 잡느냐에 따라 두 개의 국소값($(\alpha,\beta)=(0.087,0.620)$와 $(-0.063,-0.699)$) 중 하나로 수렴하거나 아예 반복기법이 수렴하지 못하여 모수를 추정하지 못하게 된다. 모수공간의 직사각형영역을 $R=\{(\alpha,\beta)|-0.3\leq\alpha\leq0.3,-1.3\leq\beta\leq1.3\}$로 하고, 난수발생기를 사용하여 모수벡터에 대한 1000개의 난수를 구한 후 대응되는 오차제곱합을 구한다. 이 중 이 오차제곱합의 값이 1.8보다 작은 196개에 대하여 산포도행렬을 그리니 다음 그림 1과 같았다. 그림 1에서 오차의 제곱합이 두 개의 극소점을 갖음을 확인할 수 있고, 초기값으로서 $(\alpha,\beta)=(0.09,0.64)$을 선택할 수 있음을 알 수 있다. 통계패캐지 S-Plus의 brush와 spin기능을 이용하면 더 확실하게 알 수 있다. 이 초기값을 이용하여 반복기법을 쓰면 최소제곱합을 찾게 되어 원하는 모수 추정을 하게 된다. 평행좌표그림을 이용하여서도 산포도행렬과 같은 결론을 얻을 수 있다.

예 2에 대하여서도 예 3에서와 같이 산포도행렬과 평행좌표그림을 이용하여 초기값을 찾으니 $(\theta_1,\theta_2,\theta_3,\theta_4)=(0.76,-1.37,1461,0.05)$이었다. 이 초기값을 이용하여 반복기법을 쓰면 최소제곱합을 찾게 되어 원하는 모수 추정을 하게 된다.

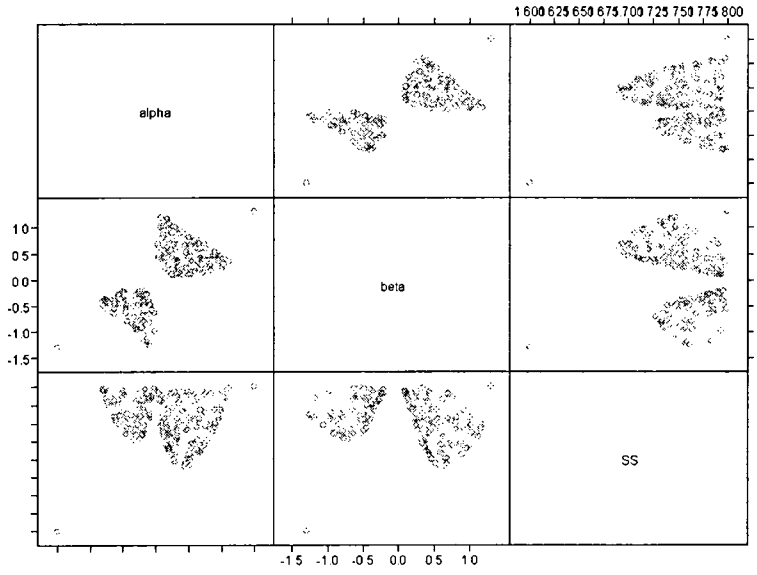

그림 1. 예 3에 대한 산포도행렬

3. 결론

비선형모형분석시 초기값을 결정할 때 탐색적 자료분석이 초기값를 결정하는 데 도움을 줄 수 있다.

참고문헌

[1] Bates, D. M. and Watts, D. G.(1988). *Nonlinear Regression Analysis and Its Application*, John Wiley, New York.
[2] Draper, N. R. and Smith, H.(1998). *Applied Regression Analysis*, 3rd ed., John Wiley, New York.
[3] Khuri, A. I. and Cornell, J. A.(1996). *Response Surfaces*, 2nd ed., Dekker, New York.
[4] Maple 7, Waterloo Maple, Inc., Waterloo., 2001.
[5] Myers, R. H., Montgomery, D. C., and Vining, G. G.(2002). *Generalized Linear Models*, John Wiley, New York.
[6] Seber, G. A. F. and Wild, C. J.(1989). *Nonlinear Regression*, John Wiley, New York.