비선형 회귀모형에서 가중최소제곱법에 의한 수위-유량곡선식 개발

Development of Rating Curve by Weighted Least Squares Method in Nonlinear Regression Model

○이우석*, 이길성**

1. 연구의 필요성 및 목적

하천 유량자료는 홍수기 수방대책, 갈수기 수자원관리 등 물문제를 해결하기 위한 기본자료가 된다. 하천 유량은 주로 실시간으로 상시관측되는 수위자료를 수위-유량관계를 이용하여 얻어지고 있으므로, 정확한 수위-유량관계를 구하는 일은 매우 중요하다. 수위-유량곡선을 개발할 경우 영유량 수위(stage of zero flow) 결정 및 수위에 따른 수위-유량곡선식의 분할 등의 문제가 발생한다. 실무에서는 수위-유량 관측자료를 사용하여 로그변환을 통한 선형회귀분석을 통해 수위-유량곡선식을 구하고 있는데 이러한 개발방법의 특징 및 문제점을 파악하는 연구도 필요하다(Clarke, 1999).

본 연구에서는 비선형회귀모형에서 유량 관측치의 오차에 따라 가중치를 달리하는 가중최소제곱법(WLS)을 제시하였다. 로그변환 선형회귀모형과 비선형회귀모형에 대해 곡선식의 오차 및 신뢰구간을 구하여 각 기법의 장단점을 파악하고자 하였으며, 곡선식 분할에 따른 영향도 검토하였다. 최적화기법으로는 전역최적화기법인 Simulated Annealing 방법을 사용하였는데 이는 곡선식 분할 등에 따라 매개변수 수가 증가하므로 지역최소값에 도달할 가능성을 배제하기 위함이다.

2. 연구배경

수위-유량곡선식은 지수형과 포물선형 두 가지 형태가 사용되고 있으며, 다음 식 (1)과 같은 지수형이 수리학적 원리에 더 적합하므로 많이 사용되고 있다.

$$Q = a(h+b)^c \tag{1}$$

여기서 Q는 유량, h는 수위, $-b$는 유량이 '零'인 수위(stage of effective zero flow), a와 c는 상수이다.

식(1)의 상수를 구하기 위해 실무에서 사용되는 방법은 주로 곡선식의 양변에 로그를 취하여 선형회귀식을 구성한 후 최소제곱법(OLS, Ordinary Least Squares)을 적용하는 것이다. 영유량 수위 $-b$를 결정하는 방법으로 Johnson 방법(Rantz et al., 1982)이 주로 쓰이고 있다. 또한 영유량수위를 가정한 다음 회귀분석을 수행하여 가정 적은 오차를 보여주는 b 값을 최적으로 결정할 수 있다(Mosley and McKerchar, 1993).

하천에서의 흐름은 저수위에서는 단면통제(section control)를 받을 수 있고 고수위에서는 하도통제(channel control)를 받게되어 수위-유량곡선식이 변화하게 된다. 이러한 통제를 고려하여 수위에 따라 곡선식을 분할할 수 있다. 이길성 등(1996)은 저수위, 평수위, 홍수위에 대한 수위-유량곡선을 별도로 작성한 바 있다. 그러나, 곡선식 분할 위치를 구하는 방법은 대상하천에 대한 기술

* 서울대학교 지구환경시스템공학부 박사과정 수료 · 한국수자원공사

** 서울대학교 지구환경시스템공학부 교수

자의 판단과 눈대중(eyefitting) 방법에 의존하고 있어 이를 체계적으로 구하는 노력이 필요하다.

따라서, 수위-유량곡선식의 매개변수를 최적으로 추정하는 방법과 로그변환을 통하지 않고 비선형회귀모형을 이용하여 곡선식을 추정하는 연구가 필요하다.

3. 연구 방법

3.1 선형 회귀분석

식 (1)에서 각 변에 자연대수 e를 밑으로 하는 로그를 취하면 수위-유량곡선식은 다음 식 (2)와 같이 변환된다.

$$\ln Q = \ln a + c \ln (h+b) \tag{2}$$

따라서 $\ln (h+b)$를 설명변수 x로 하고 $\ln Q$를 반응변수 y로 하여 단순선형 회귀모형을 구성할 수 있다. 각 오차의 제곱의 합을 최소로 하는 최소제곱법(LSM, Least Squares Method)으로 식(2)의 매개변수를 추정할 수 있다.

최적의 b 값은 회귀분석 오차가 가장 적을 때의 값이며 이는 하한값에서 상한값까지 0.01 값을 증가시켜 회귀분석을 실시하여 가장 큰 결정계수를 얻는 격자탐색법(grid search)으로 결정할 수 있다.

회귀모형의 적합도를 나타내는 척도로 결정계수(coefficient of determination)가 주로 사용된다. 선형회귀모형 결과의 결정계수는 로그변환 자료의 회귀분석 결과이므로 원자료로 환산하여 결정계수를 재산정해주어야 한다.

또한, 오차 분산의 불편추정량 $\hat{\sigma}^2$은 다음 식(3)과 같이 산정할 수 있다.

$$\hat{\sigma}^2 = \frac{SS_E}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n-2} \tag{3}$$

여기서 SS_E는 오차제곱합이며 y_i 및 $\hat{y_i}$는 각각 $x=x_i$ 에서 실측치와 예측치이다.

선형모형에 의해 추정된 수위-유량곡선식의 신뢰구간은 $x=x_0$ 에서 반응변수 참값의 평균 ($E(Y \mid x_0) = \mu_{Y \mid x_0}$)의 신뢰구간으로 나타낼 수 있다. $\hat{\mu}_{Y \mid x_0}$의 표준편차 SE(Standard Error)는 회귀상수들의 분산을 고려하면 식 (4)와 같이 나타낼 수 있다.

$$\mathrm{SE}(\hat{\mu}_{Y \mid x_0}) = \hat{\sigma}\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]^{\frac{1}{2}} \tag{4}$$

따라서, $x=x_0$ 에서 평균 반응변수의 $100(1-\alpha)\%$ 신뢰구간은 식 (5)와 같이 주어진다.

$$\hat{\mu}_{Y \mid x_0} - t_{\alpha/2,(n-2)}\mathrm{SE} < \mu_{Y \mid x_0} < \hat{\mu}_{Y \mid x_0} + t_{\alpha/2,(n-2)}\mathrm{SE} \tag{5}$$

여기서, $t_{\alpha/2,(n-2)}$는 자유도가 $n-2$인 t분포의 상위 $100(\alpha/2)$ % 백분위수를 나타낸다.

만약 수위-유량곡선식이 분할될 경우 최적의 곡선식 분할위치는 분할된 곡선식의 오차 합이 가장 적을 때이며 이는 영유량 수위를 구하는 방법과 마찬가지로 모든 관측수위에 대해 하한값에서 상한값까지 탐색하여 결정할 수 있다. 곡선식의 신뢰구간은 분할구간내에서 각각 산정되어야 한다.

3.2 비선형 회귀분석

수위-유량곡선식은 반응변수 Y를 유량 Q로, 설명변수 x를 수위 h로 하는 비선형 회귀모형으로 나타낼 수 있다.

$$\begin{aligned} y_i &= f(x_i, \theta) + \varepsilon_i \\ &= a(x_i + b)^c + \varepsilon_i \quad i = 1,2,\ldots,n \end{aligned} \tag{6}$$

여기서, θ는 미지의 p차원 매개변수 벡터로 곡선식을 분할하지 않을 경우 3차원이 된다. ε_i는 오차항으로 평균이 0이고 분산이 σ_i^2인 정규분포를 갖고 서로 독립인 확률변수라 가정한다.

매개변수 θ의 추정방법으로 주로 최소제곱법을 사용한다. 또한, 오차항의 분산이 일정하다면 OLS(Ordinary Least Squares)를 사용할 수 있다. 즉, $\hat{\theta}$은 오차의 제곱합을 최소화하는 θ 값으로 주어지므로 매개변수를 구하는 문제는 다음 식 (7)의 최적화 문제로 바뀐다.

$$\min_{\theta} J(\theta) = \sum_{i=1}^{n} [e_i]^2 = \sum_{i=1}^{n} (y_i - \hat{y_i})^2 \tag{7}$$

여기서 y_i는 i번째 관측유량이며 $\hat{y_i}$는 추정된 매개변수에 의한 i번째 예측유량이며 n은 관측자료의 개수이다. 매개변수는 물리적인 면을 고려하여 보통 상한치와 하한치를 갖는다.

또한, 비선형 회귀모형에서도 결정계수, 식 (3)에 의한 추정치의 표준편차와 식(5)에 의한 곡선식의 신뢰도를 구할 수 있으며, 비선형 회귀모형을 선형화시켜 해석한다고 가정하여 $\hat{\mu}_{Y|x_0}$의 표준오차를 구할 수 있다(Seber and Wild, 1989).

가중최소제곱법 (WLS)

최소제곱법중 식 (7)의 OLS는 오차의 분산이 서로 같다는 등분산성(homoscedasticity) 가정하에서 적용할 수 있다. 그러나 오차의 분산이 서로 다르면 OLS 대신에 오차에 가중치를 부여하는 가중최소제곱법(WLSM, Weighted Least Squares Method)을 사용해야 한다. 만약, $E(\varepsilon_i) = 0$이 성립하면 반응변수 평균값에 대한 식 (8a)가 성립한다. 또한, 오차에 가중치를 부여한다는 것은 반응변수에 가중치를 부여하는 것과 동일하므로 ε_i의 분산 σ_i^2은 반응변수 y_i의 분산과 같으며 다음 식 (8b)와 같이 정의할 수 있다.

$$E(y_i) = \mu_i(\theta) = f(x_i, \theta) \tag{8a}$$

$$Var(y_i) = Var(\varepsilon_i) = \sigma_i^2 = \sigma_W^2 g^2(\mu_i(\theta), \gamma) = \sigma_W^2 / w_i \tag{8b}$$

여기서, w_i는 i번째 가중치이며 μ_i는 x_i에서 반응변수 참값의 평균이며, g는 분산함수(variance function)이고 γ는 분산함수의 매개변수이며 σ_W^2는 일정 상수로 나타내질수 있다.

식 (7)의 목적함수는 오차의 가중치를 고려하여 다음 식 (9)와 같이 변환된다.

$$\min_{\theta} J(\theta) = \sum_{i=1}^{n} w_i e_i^2 \tag{9}$$

WLS를 적용하기 위해서는 오차의 가중치를 추정해야 한다. 본 연구에서는 회귀모형의 잔차를 이용하여 의사최우도법(pseudo-likelihood method)에 의해 최적의 가중치를 추정하였다. 먼저, 반응변수인 관측유량의 분산이 반응변수 평균의 γ승에 비례한 것으로 가정하여 분산함수를 다음 식 (10)과 같이 설정하였는데 이는 오차가 유량이 커질수록 증가한다는 가정에 따른 것이다.

$$g^2(\mu_i(\theta), \gamma) = \mu_i(\theta)^\gamma \tag{10a}$$

$$V(\varepsilon_i) = \sigma_W^2 \mu_i(\theta)^\gamma \tag{10b}$$

따라서, 가중치는 식(11)에 의해 구할 수 있다.

$$\widehat{w_i} = \frac{1}{g^2(\mu_i(\widehat{\theta}), \gamma)} = \frac{1}{\mu_i(\widehat{\theta})^\gamma} \tag{11}$$

최적가중치를 구하기 위한 절차는 다음과 같다. 다양한 γ값에 대해 먼저 매개변수를 가정하여 식 (11)에 의해 가중치를 추정한 후 다시 WLS를 적용하여 매개변수를 재산정하는 과정을 가중치의 변동이 허용한도 보다 작을 때까지 반복시켜 최종 매개변수와 가중치를 구하게 된다. 이후 의사최우도법에 의해 γ 값을 최적 추정한다. 의사최우도법은 매개변수 θ를 기지의 값으로 가정하여 γ의 최우도 추정치를 구하는 방법이다. γ 는 보통 2.0 ~ 3.0의 값을 가지며 1.6 이하이면 식 (8)의 분산함수가 적당하지 않거나 이상치가 존재하는 경우이다(Carroll and Rupport, 1988).

로그선형 회귀분석에서는 곡선식 분할위치 및 영유량수위를 먼저 격자탐색법에 의해 결정한 후 나머지 매개변수를 추정하였지만, 비선형 회귀분석에서는 곡선식 분할위치를 포함한 모든 매개변수를 동시에 추정하는 알고리즘을 채택하였다.

Simulated Annealing 기법

전통적인 최적화기법은 지역최소값에 도달할 가능성이 있어 전역최적값(global optima)을 찾기 위한 다양한 최적화기법들이 개발되었다. 그중 Simulated Annealing(SA), Tabu Search 및 유전자 알고리즘(GA, Genetic Algorithm) 등이 뛰어난 결과를 도출하여 많이 쓰이고 있다 (Rayward-Smith et al., 1996).

SA에서는 최소화문제에서 새로운 해 $(\theta, \mathbf{hs})_j$의 목적함수의 값이 현재 해 $(\theta, \mathbf{hs})_i$ 의 목적함수보다 더 작으면 새로운 해를 무조건 받아들인다. 또한, 새로운 해가 현재 해보다 더 열등해도 유한한 확률에 의해 열등값으로의 이동을 이용하는데 이 확률은 온도와 목적함수의 차로 나타낼 수 있다. 이러한 열등해로의 상승이동은 다음 식 (12)를 만족하면 허용하게 된다.

$$\exp(-\triangle c/T) > R \tag{12}$$

여기서, $\triangle c$는 목적함수값의 차이 $(c_j - c_i)$ 이며 T는 상승이동을 받아들이는 확률을 통제하는 조절계수(control parameter)이고 R은 (0, 1) 구간의 균등난수(uniform random number)이다.

4. 적 용

4.1 가상 수위-유량 관측자료

가상의 수위-유량곡선식에서 유량 크기에 비례하는 분산을 갖는 잡음을 발생시켜 이를 더해 관측값을 만든후 로그변환에 의한 선형 회귀모형, 비선형 회귀모형에서의 OLS 및 WLS기법을 적용 비교하였다.

다음 식 (13)과 같은 가상의 수위-유량곡선에 의한 계산된 유량값이 참값 $\widetilde{Q_i}$(m^3/s)이라고 가정한다.

$$\widetilde{Q} = 277.19 \times (h - 0.42)^{1.3985} \qquad 0.42 \le h \le 8.80 \tag{13}$$

잡음 n_i가 정규분포 $(0, \sigma_{n_i}^2)$에 따른다고 가정하였으며 잡음의 표준오차 σ_{n_i}를 유량 $\widetilde{Q_i}$의 4 %로 상정하였다. 참값에 잡음를 더하여 45개의 관측값 Q_i(m^3/s)를 가진 6개의 자료군을 생성하

여 모형 보정과 검증에 사용하였다.

비선형 WLS를 적용하기 위해 식 (10)의 분산함수의 γ 값을 구한 결과 자료군 1, 2, 3에서 각각 2.1, 2.0, 1.9로 최적추정되었다. 이러한 결과는 잡음의 분산 결과와 일치하므로 의사최우도법에 의한 가중치 추정이 정확하다고 말할 수 있다.

로그선형 회귀모형, 비선형 OLS 및 WLS기법을 적용한 결과를 표 1과 그림 1, 2에 나타내었다. 표 1에서 RMSE(Root Mean Square Errors)는 절대오차의 평균제곱근이고, RMRSE(Root Mean Relative Square Errors)는 유량크기에 따른 상대오차의 평균제곱근이다.

표 1. 가상지점에서 수위-유량곡선식 개발 결과 비교

자료군		$\hat{a}$	$\hat{b}$	$\hat{c}$	보정			검증	
					결정계수	RMSE	RMRSE	RMSE	RMRSE
	True value	277.19	-0.42	1.3985	-	-	-	-	-
1	로그선형	278.63	-0.42	1.3962	0.9954 0.9997[1)]	97.16	0.0430	81.15	0.0425
	비선형 OLS	300.16	-0.43	1.3496	0.9961	89.66	0.1834	92.00	0.1713
	비선형 WLS	276.85	-0.42	1.4029	0.9959	101.88	0.0415	80.60	0.0438
2	로그선형	280.92	-0.42	1.3983	0.9986 0.9998[1)]	56.30	0.0358	81.75	0.0430
	비선형 OLS	274.22	-0.41	1.4125	0.9986	55.30	0.1952	83.59	0.1924
	비선형 WLS	282.20	-0.42	1.3920	0.9984	60.30	0.0349	80.60	0.0446
3	로그선형	278.74	-0.42	1.3953	0.9980 0.9998[1)]	65.04	0.0360	81.47	0.0431
	비선형 OLS	291.14	-0.43	1.3688	0.9982	61.70	0.1813	86.09	0.1772
	비선형 WLS	277.97	-0.42	1.3983	0.9978	66.77	0.0358	80.82	0.0436
평균	로그선형	279.43	-0.42	1.3966	0.9973	72.83	0.0383	81.46	0.0429
	비선형 OLS	288.51	-0.42	1.3770	0.9976	68.89	0.1866	87.22	0.1803
	비선형 WLS	279.01	-0.42	1.3974	0.9974	76.32	0.0374	80.67	0.0440

주 1) 로그변환된 자료의 선형회귀분석 결과임

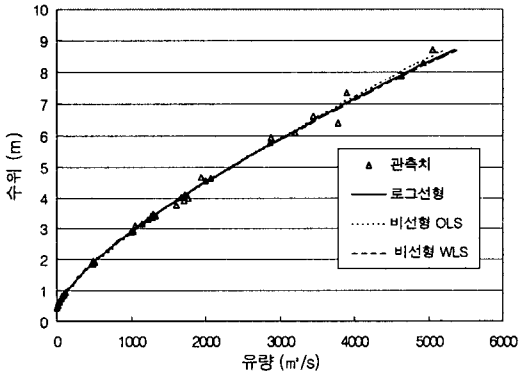

그림 1. 수위-유량곡선식 개발결과 (자료군 1)

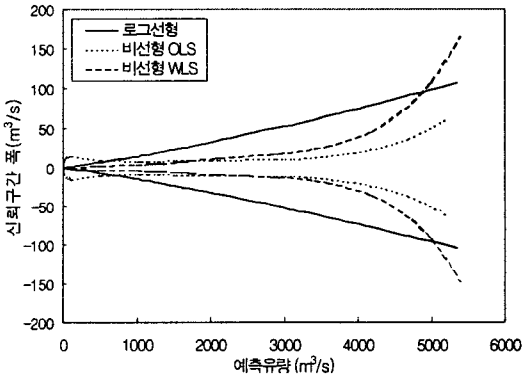

그림 2. 수위-유량곡선식 신뢰구간 (자료군 1)

RMSE 및 결정계수 측면에서 보정단계에서는 비선형 OLS모형, 로그선형 회귀모형, 비선형 WLS모형의 순으로 우수하지만, 검증단계에서는 비선형 WLS가 로그선형보다 약간 우수하였으며, 비선형 OLS가 가장 저조하였다. 또한, 로그선형 회귀모형의 결정계수는 원자료로 환산하여 산정하면 그 값이 낮아짐을 알 수 있다.

상대오차 측면에서는 비선형 WLS 및 로그선형 회귀모형이 우수한 결과를 나타냈으며 비선형 OLS는 저수량에서 오차가 큰 것으로 나타났다. 따라서, 로그선형 모형은 저수량에 큰 가중치를

부여하는 것과 같은 효과를 가지고 있음을 알 수 있다. 그림 2의 수위-유량곡선식의 95 % 신뢰구간를 살펴보면, 비선형 OLS의 경우 곡선식의 저수량에 대한 신뢰도가 다른 모형에 비해 떨어지나 홍수량에 대한 신뢰도는 높음을 알 수 있다.

4.2 실제 유량측정지점에의 적용

대청댐 상류 유량측정지점의 98년도 자료(한국수자원공사, 1998)를 이용하여 비선형 WLS 모형을 적용하여 구한 분산함수의 γ 값은 표 2에 나타내었다.

관측자료의 양상에 따라 다양한 γ 값이 추정되었는데, γ 값이 2에 가까우면 관측유량의 오차가 관측값에 정비례하며, γ 값이 작아지면 저수량(low flow)의 상대오차가 홍수량의 상대오차보다 큼을 의미한다. 비선형 OLS는 곡선식의 홍수량을 주로 적합시키므로 저수위 곡선식의 신뢰도가 떨어진다. 반면에 로그선형 회귀모형은 유량자료를 로그변환함으로 저수량에 대해 큰 가중치를 부여한 것과 동일한 효과를 얻을 수 있는데, 비선형 WLS 모형에서 γ = 2.0 일 경우와 비슷한 결과를 도출하였다. 따라서, 비선형 WLS 모형은 유량측정자료의 오차양상에 따라 의사최우도법에 의해 가중치의 최적추정이 가능하므로 가장 우수한 결과를 도출할 수 있다.

표 2 대청댐 상류 유량측정지점 γ 값

측정지점	자료개수	γ 값	
		곡선식 1개	곡선식 2개
수통	27	1.6	1.5
호탄	27	2.5	1.8
송천	32	1.4	1.6
옥천	45	1.1	1.3
청성	32	1.0	1.0
구룡	30	2.0	2.0

5. 결 론

본 연구에서는 수위-유량관측자료에서 비선형 회귀모형에서 WLS 기법을 적용하여 수위-유량곡선식을 유도하는 방법을 개발하였다. 비선형 WLS 기법에서 매개변수 추정 및 곡선식 최적분할을 위해 전역최적화기법인 Simulated Annealing 기법을 적용하였으며 의사최우도법에 의해 최적가중치를 추정하였다.

잡음이 포함된 가상 수위-유량 관측자료에 적용한 결과, 비선형 WLS 기법은 로그선형 회귀모형 보다 우수한 결과를 도출하였으며 비선형 OLS 기법은 곡선식 저수량의 오차가 크게 발생하였다. 또한, 본 연구에서 개발된 비선형 WLS 모형은 유량측정자료의 오차양상에 따라 의사최우도법에 의해 가중치를 최적추정하므로 최적의 수위-유량곡선식을 추정할 수 있었다.

6. 참고문헌

이길성 등 (1996). **낙동강 수계 실시간 최적 저수관리 시스템 개발 (분석모델 부문) 보고서**. 한국수자원공사.

한국수자원공사 (1998). **1998 수문자료집**.

Carroll, R.J., and Ruppert. D. (1988). *Transformation and Weighting in Regression*. Chapman & Hall.

Clarke, R.T. (1999) "Uncertainty in the Estimation of Mean Annual Flood due to Rating-Curve Indefinition." *Journal of Hydrology*, Vol. 222, pp. 185-190.

Mosley, M.P., and McKerchar, A.I. (1993). *Streamflow, Handbook of Hydrology. Chap. 8,* McGraw-Hill.

Rantz, S.E., and Others (1982). *Measurement and Computation of Streamflow.* USGS.

Rayward-Smith, V.J., Osman, I.H., Reeves, C.R., and Smith, G.D. (1996). *Modern Heuristic Search Methods.* John Wiley & Sons Ltd.

Seber, G.A.F., and Wild, C.J. (1989). *Nonlinear Regression.* John Wiley & Sons, Inc.