Ridge 회귀분석을 사용한 확률홍수량 산정에 관한 연구

A STUDY ON AN ESTIMATION OF PROBABLE FLOOD FLOW USING RIDGE REGRESSION

연세대학교 공과대학 토목공학과 교 수 이 원 환
연세대학교 공과대학 토목공학과 교 수 조 원 철
Colorado State University 박사과정 이 재 응

1. 서 론

수문학자들은 수공구조물의 설계홍수량에 대한 자료가 거의 없거나 사용할 수 없는 지점에서의 설계홍수량을 추정할 필요를 느껴 왔다. 그러한 추정을 하는데 많이 이용되는 방법중의 하나가 다중선형회귀분석을 통하여 확률홍수량의 특징과 유역의 특징을 상관시키는 것으로서, 회귀분석 계수의 표본 추정치는 그 유역의 관측소에서 수집한 년최대유량과 유역특성 자료를 사용하여 구할 수 있다.

유역특성간의 상관성이 거의 없을 경우에는 일반적인 최소자승회귀분석법을 사용하면 합리적인 결과를 얻을 수 있다. 그러나 독립변수인 유역특성간의 상관성이 매우 높을 경우에는 결과가 만족스럽지 못한 것으로 알려져 있다. 회귀분석에 있어서 독립변수들 간의 상관성이 매우 높을 때 다중공선형성이 존재한다고 한다.

다중공선형성이 존재할 때의 문제점은 다음과 같다.

1) 분산과 공분산이 증가한다.

2) 상관계수가 과도하게 증가한다.

3) 계수의 유의값과 신뢰구간의 결정이 어렵게 된다.

Ridge회귀분석을 사용하면 이러한 다중공선형성의 문제점들을 해결할 수 있다.

Ridge회귀분석은 일반적인 최소자승 추정법을 실시하기 전에 독립변수의 단순상관행렬의 대각선 행렬에 상수 k[0≤k≤1] 를 더하여 기본식이 성립되는 것으로, Ridge추정량은 편향되어 있지만 회귀계수추정값이 보다 더 정확하고 예측오차가 작다는 장점을 가지고 있다.

본 논문의 목적은 유역특성과 확률홍수량간의 선형회귀분석 모형에서 회귀계수의 Ridge 추정량과 일반적인 최소자승 추정량을 비교분석하는데 있다. 한강 유역의 확률홍수량 산정모형을 검토하기 위하여 Monte Carlo법을 이용, 모의유량을 발생시켰다.

2. 기본이론

2.1 Ridge회귀분석의 정의

식 (1)과 같이 표준화된 다중선형회귀분석 모형을 생각한다.

$$Y = X\beta + e \tag{1}$$

여기서 X는 독립변수로 (nxp)벡터, Y는 종속변수로 (nx1)벡터, β는 회귀계수로 (px1)벡터이고, e는 부작위교란으로 $E(e)=E(e'e)=\sigma^2 I$인 (nx1)벡터이다.

β의 최소자승추정량은 다음과 같다.

$$\tilde{\beta} = (X'X)^{-1}X'Y \tag{2}$$

$L^2 = (\tilde{\beta}-\beta)'(\tilde{\beta}-\beta)$라고 정의하면 다음 식이 성립된다.

$$E(L^2) = \sigma^2 \, Trace(X'X)^{-1} = \sigma^2 \sum_{i=1}^{p} \lambda_i^{-1}$$

$$E(\tilde{\beta}'\tilde{\beta}) = \beta'\beta + \sigma^2 \sum_{i=1}^{p} \lambda_i^{-1} > \beta'\beta + \sigma^2 \lambda_p^{-1} \tag{3}$$

여기서 $E(L^2)$은 일반적인 최소자승 분석법에 의한 회귀계수추정량의 평균 제곱오차이고 λi는 $X'X$의 고유치로서 독립변수 벡터들이 다중공선형성을 나타낼 때 $\tilde{\beta}$는 β의 만족할 만한 추정량이 못된다. 이러한 문제점을 해결하기 위하여 Ridge추정량이 다음 식과 같이 제안되었다.

$$\tilde{\beta}(k) = (X'X + kI)^{-1} X'Y, \quad (0 \leq k \leq 1) \tag{4}$$

여기서 I는 (p X p)인 단위행렬이다.

2.2 Ridge회귀분석의 특성

회귀계수의 Ridge추정치는 식(1)과 같은 표준화된 회귀분석 모형을 약간 변형시킴으로서 구할 수 있다.

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots\cdots + \beta_p X_p + e \tag{5}$$

Ridge회귀계수에 대한 추정식은 식 (6)과 같다.

$$\begin{aligned}
&(1 + k)\tilde{\beta}_1 + r_{12}\tilde{\beta}_2 + \cdots\cdots + r_{1p}\tilde{\beta}_p = r_{1y}, \\
&r_{12}\tilde{\beta}_1 + (1 + k)\tilde{\beta}_2 + \cdots\cdots + r_{2p}\tilde{\beta}_p = r_{2y}, \\
&r_{1p}\tilde{\beta}_1 + r_{2p}\tilde{\beta}_2 + \cdots\cdots + (1 + k)\tilde{\beta}_p = r_{py},
\end{aligned} \tag{6}$$

여기서 r_{ij}는 i번째 독립변수와 j번째 독립변수간의 단순상관계수이고 r_{iy}는 i번째 독립변수와 종속변수 y사이의 단순상관계수이다.

Ridge추정량과 일반적인 최소자승 추정량은 다음과 같은 관계를 가지고 있다.

$$\begin{aligned}
\tilde{\beta}(k) &= [\, I + k(X'X)^{-1} \,]^{-1} \tilde{\beta} \\
&= z\,\tilde{\beta}
\end{aligned} \tag{7}$$

여기서 $z = [\, I + k(X'X)^{-1} \,]^{-1}$이다.

Ridge추정량의 평균제곱오차는 다음과 같다.

$$\begin{aligned}
E[L^2(k)] &= E[(\tilde{\beta}(k) - \beta)'(\tilde{\beta}(k) - \beta)] \\
&= E[(\tilde{\beta} - \beta)' z' z (\tilde{\beta} - \beta)] \\
&\quad + (z\beta - \beta)'(z\beta - \beta)
\end{aligned}$$

$$= \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \beta'(X'X + kI)^{-2} \beta$$

$$= r_1(k) + r_2(k) \qquad (8)$$

식 (8) 의 우변의 첫번째 항은 Ridge추정량의 분산을 나타내고 두번째 항은 Ridge추정량의 편향성의 제곱을 나타낸다.

그림 1은 분산과 편향성의 제곱, 그리고 Ridge상수 k사이의 관계를 보여준다.

또, Hoerl과 Kennard는 $E[L^2(k)] < E(L^2)$인 k가 항상 존재한다는 중요한 이론을 제안하였다. 즉 Ridge추정량의 평균제곱오차가 일반적인 최소자승 추정량의 평균제곱오차보다 작은 k값이 적어도 하나 존재한다.

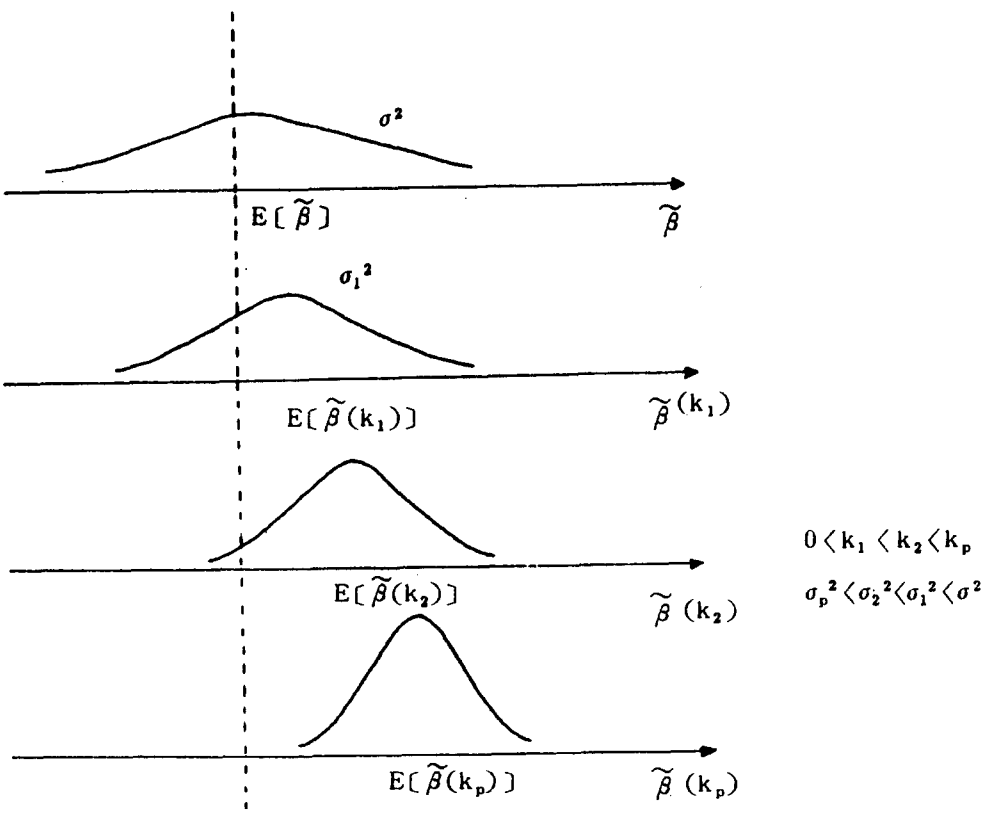

그림 1 분산, 편향성의 제곱, Ridge상수간의 관계

2.3 Ridge상수 k 의 선택

1. Hoerl, Kennard, Baldwin에 의해 제안된 k값

다음과 같은 직교행렬 P(pxp)를 생각한다.

$$P'XX'P = \Lambda \tag{9}$$

여기서 $\Lambda = diag(\lambda_1, \ldots, \lambda_p)$이고 $W = XP$, $\alpha = P'\beta$라고 하면 식 (1)은 다음과 같이 쓸 수 있다.

$$\begin{aligned} Y &= X\beta + e \\ &= XP'P\beta + e \\ &= W\alpha + e \end{aligned} \tag{10}$$

α의 Ridge추정량과 최소자승추정량은 다음과 같다.

$$\begin{aligned} \tilde{\alpha}(k) &= (W'W + kI)^{-1} W'Y \\ &= (\Lambda + kI)^{-1} W'Y \end{aligned} \tag{11}$$

$$\begin{aligned} \tilde{\alpha} &= (W'W)^{-1} W'Y \\ &= \Lambda^{-1} W'Y \\ &= P'\tilde{\beta} \end{aligned} \tag{12}$$

여기서 $\tilde{\alpha}(k)$는 Ridge추정량이고 $\tilde{\alpha}$는 최소자승 추정량이다. 식 (11)과 식 (12) 로부터 $\tilde{\alpha}_i(k)$는 다음과 같이 표시할 수 있다.

$$\tilde{\alpha}_i(k) = \frac{\lambda_i}{\lambda_i + k}\,\tilde{\alpha}_i \tag{13}$$

만일 $X'X = I$ 라면 $k = \dfrac{p\,\sigma^2}{\beta'\beta}$ 일때 평균제곱오차가 최소로 되며 일반적으로 $k_i = \sigma^2 / \alpha_i^2$ 일 때 평균제곱오차가 최소로 된다.

산술평균을 사용하면 편향성을 고려할 수 없기 때문에 단일 k 값을 구하기 위해서는 조화평균을 사용해야 한다. k 의 조화평균을 k_h라 하면 k_h는 식 (15)를 사용하여 구할 수 있다.

$$\frac{1}{k_h} = \frac{1}{P}\sum_{i=1}^{p}\left(\frac{1}{k_i}\right) = \frac{1}{P}\sum_{i=1}^{p}\frac{\alpha_i^2}{\sigma^2}$$

$$= \frac{1}{P\sigma^2}\sum_{i=1}^{p}\alpha_i^2 = \frac{\alpha'\alpha}{P\sigma^2}$$

$$= \frac{\beta'\beta}{P\sigma^2} \qquad (14)$$

따라서 $k_h = \dfrac{P\sigma^2}{\beta'\beta}$ (15)

즉, 위의 결과로부터 k의 최적값은 $\dfrac{P\sigma^2}{\beta'\beta}$ 의 추정치라는 것을 알 수 있다.

$$k = \frac{P\tilde{\sigma}^2}{\tilde{\beta}'\tilde{\beta}} \qquad (16)$$

여기서 P는 β_0 를 제외한 회귀계수의 수이고 $\tilde{\sigma}^2$ 은 잔차의 평균을 제곱한 것이고 $\tilde{\beta}$는 β의 최소자승 추정치이다.

2. Ridge궤적

Ridge궤적이란 k 에 대한 $\tilde{\beta}(k)$의 궤적으로 k가 변함에 따라 회귀계수가 변하는 것을 한 눈에 볼 수 있다는 점이 좋은 점이다. Ridge궤적을 사용함으로써 최소자승법에서의 평균제곱오차보다 작은 평균제곱오차를 갖는 k 값을 찾을 수 있다.

Ridge궤적을 사용하여 k값을 선택하는 방법은 다음과 같다.

1) 어떠한 k값에 대해서 회귀계수값이 안정되어 직교상태의 특성을 갖게 된다.
2) 어떠한 k값에 대해서 k=0에서 부적당한 부호를 갖는 회귀계수들이 올바른 부호로 변화한다.
3) 어떠한 k값에 대해서 잔차의 합의 제곱이 부적정한 값을 갖지 않는다.
4) 어떠한 값에 대해서 회귀계수들의 절대값들이 불합리한 값을 갖지 않는다.

Ridge궤적을 사용하여 선택한 k값은 정확한 값이 아니라 근사값이고, k값을 선택하는 사람의 주관에 따라 변화한다는 점에서 Ridge 궤적을 사용하는데 어려움이 있다.

3. 기본자료 및 해석

3.1 재현기간별 확률홍수량의 산정

본 연구에서는 유역특성인자중 유역면적과 하도경사를 고려하여 한강 유역의 확률홍수량 모형을 다음과 같은 형태로 가정하였다.

$$Q = \alpha A^{\beta 1} S^{\beta 2} \tag{17}$$

여기서 Q는 확률홍수량, A는 유역면적, S는 하도경사이고 α, β_1, β_2는 회귀상수 및 회귀계수이다.

식 (17)을 선형화시키기 위하여 양변에 대수를 취하면 다음과 같은 선형회귀분석모형이 된다.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{18}$$

여기서 $Y = \log Q$, $\beta_0 = \log \alpha$, $X_1 = \log A$, $X_2 = \log S$ 이다.

유역특성인자인 X_1과 X_2 사이의 상관계수를 구해본 결과 -0.9936이 나와 독립변수들간에 다중공선형성이 존재하였다.

3.2 최소자승법에 의한 추정량과 Ridge추정량

50년, 100년, 200년 재현기간을 갖는 확률홍수량에 대한 회귀계수를 k값을 변화시켜 가면서 식(6)을 사용하여 구하였다. 그림 2, 3, 4는 각각 50년, 100년, 200년 재현기간을 갖는 확률홍수량 자료에 대한 Ridge 궤적이다.

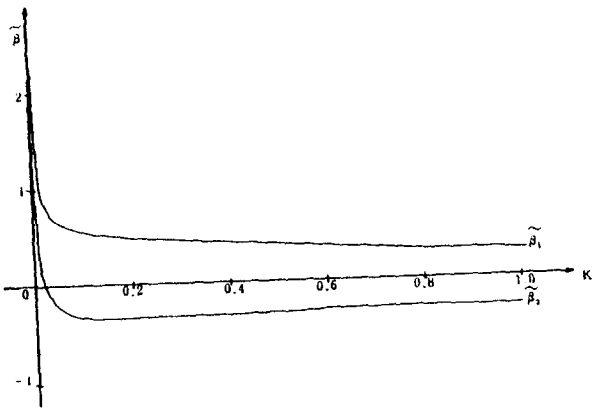

그림 2 Ridge 궤적 : $Q_{50} = \alpha A^{\beta_1} S^{\beta_2}$

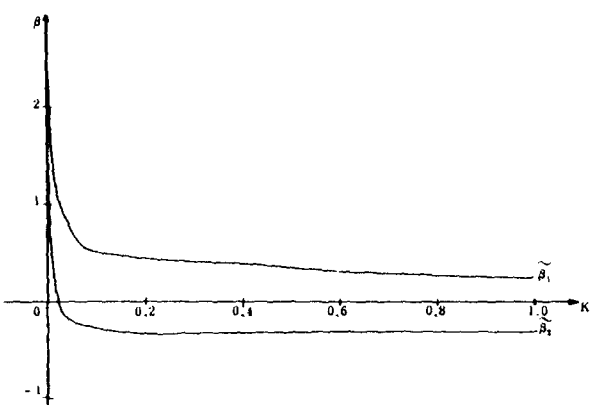

그림 3 Ridge 궤적 : $Q_{100} = \alpha A^{\beta_1} S^{\beta_2}$

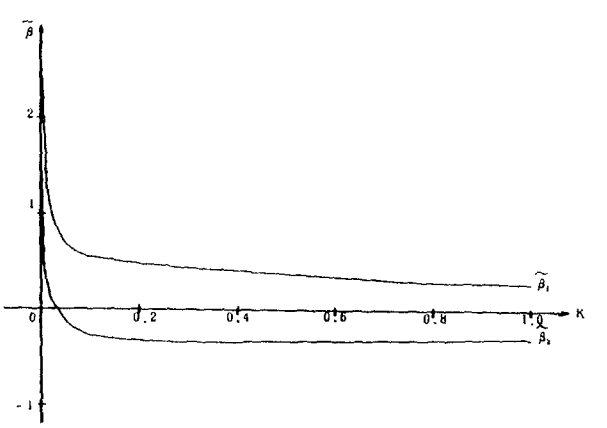

그림 4 Ridge 궤적 : $Q_{200} = \alpha A^{\beta_1} S^{\beta_2}$

식 (16)에 의하여 다음과 같이 k를 구하였다.

$$k_{50} = \frac{P\,\tilde{\sigma}^2}{\tilde{\beta}'\tilde{\beta}} = \frac{2 \times 0.2381}{2.3327^2 \times 1.4141^2} = 0.0640$$

$$k_{100} = \frac{P\,\tilde{\sigma}^2}{\tilde{\beta}'\tilde{\beta}} = \frac{2 \times 0.2391}{2.4605^2 \times 1.5469^2} = 0.0566$$

$$k_{200} = \frac{P\,\tilde{\sigma}^2}{\tilde{\beta}'\tilde{\beta}} = \frac{2 \times 0.2367}{2.5980^2 \times 1.6880^2} = 0.0493$$

위에서 구한 k를 사용하여 Ridge 회귀분석법에 의한 회귀계수와 k가 0일 때의 최소자승 회귀분석법에 의한 회귀계수를 각각 식(2)와 식(6)을 사용하여 구했고 이를 표 1 에 요약하였다.

표 1 재현기간별 확률홍수량에 대한 Ridge 추정량과 최소자승 추정량

	회귀상수	X_1 (log A)	X_2 (log S)	상수	평균제곱오차
OLS_{50}	$\tilde{b}$	2.3327	1.4141	0	
($k=0$)	b	1.9353	1.8687	2.9731	0.4329
$Ridge_{50}$	$\tilde{b}$	0.6174	-0.2694	0	
($k=0.0640$)	b	0.5122	-0.3560	0.7995	0.4108
OLS_{100}	$\tilde{b}$	2.4605	1.5469	0	
($k=0$)	b	2.0455	2.0484	3.2236	0.4331
$Ridge_{100}$	$\tilde{b}$	0.6481	-0.2401	0	
($k=0.0566$)	b	0.5388	-0.3179	0.9138	0.4121
OLS_{200}	$\tilde{b}$	2.598	1.688	0	
($k=0$)	b	2.1492	2.2243	3.4539	0.4311
$Ridge_{200}$	$\tilde{b}$	0.6934	-0.1968	0	
($k=0.0493$)	b	0.5736	-0.2593	0.9849	0.4102

표 2 재현기간별 확률홍수량의 회귀모형 결과식

OLS	Y_{50} = 2.9731 + 1.9353 log A + 1.8687 log S
	Y_{100} = 3.2236 + 2.0455 log A + 2.0484 log S
	Y_{200} = 3.4539 + 2.1492 log A + 2.2243 log S
RIDGE	Y_{50} = 0.7995 + 0.5122 log A - 0.3560 log S
	Y_{100} = 0.9138 + 0.5388 log A - 0.3179 log S
	Y_{200} = 0.9849 + 0.5736 log A - 0.2593 log S

표 1에 의하여 50년, 100년, 200년 재현기간을 갖는 확률홍수량의 최소자승 회귀분석에 의한 회귀모형과 Ridge회귀분석에 의한 회귀모형의 결과식들을 표 2에 표시하였다.

4. 모의발생

4.1 모의발생 모형의 설정

일반적인 최소자승법에 의한 회귀분석과 Ridge회귀분석을 비교하기 위하여 모의발생방법으로 Monte Carlo 방법을 사용하였다. 모형은 식(19)와 같은 형태로 가정한다.

$$Q_{50} = \alpha\ A^{\beta 1}\ S^{\beta 2} \tag{19}$$

식 (19)를 선형화시키기 위하여 양변에 대수를 취하면 다음과 같은 선형회귀분석 모형이 된다.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{20}$$

여기서 $Y = \log_{10} Q_{50}$, $\beta_0 = \log_{10} \alpha$, $X_1 = \log_{10} A$, $X_2 = \log_{10} S$ 이다.

4.2 모의발생 결과의 분석방법

$\tilde{\beta}_1$과 $\tilde{\beta}_2$ (B1O 과 B2O) 가 β_1과 β_2 의 최소자승 추정치를 나타내고 $\tilde{\beta}_1(k)$와 $\tilde{\beta}_2(k)$ (B1R과 B2R)가 Ridge 추정치를 나타낸다고 하면 β_1과 β_2 의 근평균제곱오차는 다음과 같다.

$$\text{RTMSE-B1O} = \left[\frac{1}{NR} \sum_{r=1}^{NR} (\tilde{\beta}_1 - \beta_1)^2_r \right]^{1/2}$$

$$\text{RTMSE-B2O} = \left[\frac{1}{NR} \sum_{r=1}^{NR} (\tilde{\beta}_2 - \beta_2)^2_r \right]^{1/2} \qquad (21)$$

$$\text{RTMSE-B1R} = \left[\frac{1}{NR} \sum_{r=1}^{NR} (\tilde{\beta}_1(k) - \beta_1)^2_r \right]^{1/2}$$

$$\text{RTMSE-B2R} = \left[\frac{1}{NR} \sum_{r=1}^{NR} (\tilde{\beta}_2(k) - \beta_2)^2_r \right]^{1/2}$$

여기서 NR은 반복회수이다.

4.3 모의발생 결과분석

한강유역의 유역면적과 하도경사를 기본자료로 하고 반복회수 NR=70으로 하여 모의발생시킨 결과를 표 3에 요약하였다. 표 3 에서 B1O, B2O는 일반적인 최소자승 회귀분석을 통해 얻은 회귀계수이고, B1R, B2R은 Ridge 회귀분석을 통해 얻은 회귀계수이다. 또, R-RTMSE는 근평균제곱오차의 비를 나타낸다.

표 3 모의발생 결과

회귀상수	회귀상수값	편향상수 (k)		
		평 균	최대값	최소값
B10	1.833			
B20	-7.439			
B1R	0.498	0.2370	1.0000	0.0025
B2R	-5.126			

또, 모의발생 결과 70회의 반복회수 중에 B1의 Ridge 추정치가 최소자승 추정치보다 우월한 횟수는 60회이고 B2의 Ridge 추정치가 최소자승 추정치보다 우월한 횟수는 61회였다. 또 근평균제곱오차의 비는 다음과 같다.

$$R\text{-}RTMSE\text{-}B1 = \frac{RTMSE\text{-}B1R}{RTMSE\text{-}B10} = \frac{3.052}{4.839} = 0.631$$

$$R\text{-}RTMSE\text{-}B2 = \frac{RTMSE\text{-}B2R}{RTMSE\text{-}B20} = \frac{6.918}{10.276} = 0.673$$

일반적인 최소자승 추정량과 Ridge 추정량을 비교하기 위해 식(22)와 같은 평균제곱오차의 비를 계산하였다.

$$RMSE = \frac{MSE - RIDGE}{MSE - OLS} \qquad (22)$$

여기서 RMSE는 평균제곱오차의 비이고 MSE-RIDGE는 Ridge추정량의 평균제곱오차 MSE-OLS는 일반적인 최소자승 추정량의 평균제곱오차이다.

$$RMSE\text{-}Q_{50} = \frac{0.4108}{0.4329} = 0.9489$$

$$RMSE\text{-}Q_{100} = \frac{0.4121}{0.4331} = 0.9515$$

$$RMSE\text{-}Q_{200} = \frac{0.4102}{0.4311} = 0.9515$$

단, $RMSE\text{-}Q_{50}$, $RMSE\text{-}Q_{100}$, $RMSE\text{-}Q_{200}$은 각각 50년, 100년, 200년 재현기간을 갖는 확률홍수량 자료에 대한 평균제곱오차의 비이다.

RMSE가 1보다 작다는 것은 Ridge추정량이 일반적인 최소자승 추정량보다 우월하다는 것을 의미한다.

5. 결 론

본 연구는 Ridge회귀분석을 사용하여 한강유역의 확률홍수량을 산정하기 위한 것으로서 얻어진 결과는 다음과 같다.

1) 한강유역의 유역특성인자인 유역면적과 하도경사 사이의 상관계수가 0.9이상이 되어 다중공선형성이 존재하였다.
2) 한강유역의 50년, 100년, 200년 재현기간을 갖는 확률홍수량의 회귀모형을 일반적인 최소자승 회귀분석과 Ridge회귀분석을 통해 산정하여 평균제곱오차의 비로 비교한 바 Ridge 회귀분석을 사용한 확률홍수량의 회귀모형이 일반적인 최소자승회귀분석을 사용한 확률홍수량의 회귀모형보다 더 양호하였다.
3) 모의발생을 통하여 50 년 재현기간을 갖는 확률 홍수량을 산정한 바 Ridge 회귀분석을 사용한 회귀모형이 일반적인 최소자승 회귀분석을 사용한 확률홍수량의 회귀모형보다 더 양호하였다.