Journal of the Korean Data & Information Science Society
2023, **34**(6), 917–926

http://dx.doi.org/10.7465/jkdi.2023.34.6.917
한국데이터정보과학회지

벌점 분위수 회귀를 이용한 서울시 극단 강수량 예측

임형섭[1] · 이우선[2] · 나종화[3]

[1]한국팜비오 · [2]식품의약품안전처 · [3]충북대학교 정보통계학과

접수 2023년 8월 24일, 수정 2023년 9월 15일, 게재확정 2023년 9월 29일

요 약

지구 온난화로 인한 기상 이변의 시대에 극단 강수량은 안전이나 경제 등에 직접적인 영향을 미치므로 이에 대한 예측은 매우 중요하다. 본 연구에서는 분위수 회귀 모형과 벌점 분위수 회귀 모형을 사용하여 서울시의 월 극단 강수량에 대한 예측을 수행하였다. 분위수 회귀모형의 목적함수에 추가된 벌점항으로는 Lasso, SCAD, MCP 및 Elastic net이 고려되었다. 적합된 모형의 성능 비교는 설명력, BIC 및 예측력의 관점에서 수행하였다. 적합된 분위수 및 벌점 분위수 회귀 모형의 설명력은 전반적으로 우수하게 나타났으며, 이 가운데 SCAD와 MCP 분위수 회귀 모형의 경우 모형을 크게 단순화함을 보였다. 다만 최근의 기상 이변으로 인한 특이 현상을 지금까지의 경험적 자료로 설명하는 데는 한계가 있음을 확인하였으며, 따라서 극단 강수량의 예측을 위해서는 추가적인 정보를 활용하거나 보다 정교한 수치모형의 개발이 요구된다.

주요용어: 극단 강수량, 벌점 분위수 회귀, 분위수 회귀, Lasso, MCP, SCAD.

1. 서론

최근 기후 변화로 인해 발생되는 극단 강수량 (예: 가뭄, 홍수)은 우리나라의 농업이나 경제 등 여러 분야에 부정적인 영향을 미칠 수 있기 때문에 이에 대한 정확한 예측은 매우 중요하다. 이와 같이 실생활에서 우리는 강수량의 평균값보다는 특정 분위수에 해당하는 극단 강수량의 추정에 관심이 있다. 본 논문에서는 조건부 평균을 추정하는 OLS 회귀 모형 대신 조건부 분위수를 추정하는 분위수 회귀 모형과 이 모형의 목적함수에 벌점항을 추가한 벌점 분위수 회귀 모형을 고려하였다.

분위수 회귀 모형과 벌점 분위수 회귀 모형은 다양한 분야에서 유용하게 활용된다. 강수량 예측에 적용한 연구로 Wigena 등 (2014)과 Mondinana 등 (2021)은 분위수 회귀 모형을 이용하여 인도네시아 특정 지역의 월 강수량을 예측하였고, Santri 등 (2016)과 Cahyani 등 (2016)은 Lasso 및 Elastic net 분위수 회귀 모형을 이용하여 월 강수량을 예측하였다. 강수량 이외의 자료에 적용한 연구로 Harbi 등 (2017)은 분위수 회귀 모형을 이라크의 세금 자료에 적용하였고, Lauret 등 (2017)은 일조량 자료에 적용하였다. 또한 Boualegue (2017)은 Lasso 분위수 회귀 모형을 이용하여 태양 총 복사 자료에 적용하였고, Barger (2018)은 Ridge 분위수 회귀 모형을 이용하여 어린이의 빈혈 자료에 적용한 바 있다. 벌점 회귀 및 기계학습 기법 등을 이용한 기상 관련 예측을 수행한 국내 연구로는 Won과 Na (2020), An과 Lim (2020) 및 Kim과 Jeong (2022)이 있다.

[1] (06775) 서울특별시 서초구 논현로 83, 한국팜비오, 회사원.
[2] (28159) 충청북도 청주시 흥덕구 오송읍 오송생명2로 187, 식품의약품안전처, 보건연구사.
[3] 교신저자: (28644) 충북 청주시 서원구 충대로 1, 충북대학교 정보통계학과, 교수.
E-mail: cherin@cbnu.ac.kr

본 논문의 구성은 다음과 같다. 2장에서는 분위수 회귀 모형과 벌점 분위수 회귀 모형을 소개한다. 벌점 분위수 회귀 모형으로는 Lasso, SCAD, MCP 및 Elastic net 분위수 회귀 모형을 소개한다. 3장에서는 서울시 월 강수량 자료를 이용하여 각 모형을 적합하고 특정 분위수에 대한 예측 및 모형 평가를 수행한다. 4장에서는 연구의 결론에 대해 기술한다.

2. 벌점 분위수 회귀 모형

2.1. 분위수 회귀 모형

분위수 회귀 모형은 Koenker와 Bassett (1978)이 처음 제안한 방법으로, 반응 변수의 조건부 평균을 추정하는 OLS 회귀 모형과는 다르게 조건부 분위수를 추정하는 통계적 모형이다. 이는 자료의 극단값과 같이 자료 분포의 특정 부분에 대해 관심이 있는 경우 매우 유용하다.

설명 변수 X가 주어질 때 반응변수 Y의 100τ 번째 조건부 분위수는 다음과 같이 정의된다.

$$Q_\tau(Y|X=x) = inf\{y : F(y|x) \geq \tau\}, \qquad 0 < \tau < 1,$$

여기서 $F(y|x) = P(Y < y|x)$ 는 Y의 조건부 분포함수이다.

분위수 회귀 모형의 회귀계수는 다음의 목적함수를 최소로 하는 β를 추정한다.

$$\hat{\beta}(\tau) = \underset{\beta_0,\beta}{arg\,min} \left[\sum_{y_i \geq \beta_0 + x_i^T\beta} \tau|y_i - \beta_0 - x_i^T\beta| + \sum_{y_i < \beta_0 + x_i^T\beta} (1-\tau)|y_i - \beta_0 - x_i^T\beta| \right].$$

위 식은 간단하게 다음과 같이 나타낼 수 있다.

$$\hat{\beta}(\tau) = \underset{\beta_0,\beta}{arg\,min} \left[\sum_{i=1}^{n} \rho_\tau(y_i - \beta_0 - x_i^T\beta) \right].$$

여기서 ρ_τ 함수는 τ 에 의존하는 분위수 회귀 모형의 손실함수로 다음과 같이 정의된다.

$$\rho_\tau(u) = \begin{cases} \tau u, & u \geq 0 \\ (\tau - 1)u, & u < 0. \end{cases}$$

2.2. 벌점 분위수 회귀 모형

2.2.1. Lasso 분위수 회귀 모형

Lasso (Least absolute shrinkage and selection operator)는 Tibshirani (1996)가 제안한 방법으로, 회귀계수의 절댓값의 합에 벌점을 주는 방법이다. Lasso 분위수 회귀 모형은 Lasso 방법을 분위수 회귀 모형과 결합하여 다음의 목적함수를 최소로 하는 $\beta(\tau)$ 를 추정한다.

$$\hat{\beta}(\tau) = \underset{\beta_0,\beta}{arg\,min} \left[\sum_{i=1}^{n} \rho_\tau(y_i - \beta_0 - x_i^T\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \right].$$

위 식에서 첫 번째 항은 분위수 회귀 모형의 목적함수와 같고, 두 번째 항은 Lasso 벌점함수이다. 여기서 λ는 벌점을 조율하는 모수로, λ의 값이 0에 가까울수록 위의 목적함수는 분위수 회귀 모형의 경우와 같아지게 된다. 반대로 λ의 값이 커질수록 회귀계수의 값은 0으로 축소한다. 이는 일부 회귀계수를 0으로 축소하여 변수선택 기능을 가지므로 모형을 단순화하는 장점이 있지만 추정된 회귀계수가 편의를 가질 수 있다는 단점이 있다.

이를 보완한 방법으로 Lasso 추정량의 편의를 줄이는 방법인 SCAD와 MCP를 소개하고, Lasso 추정량의 편의가 늘어나는 비용으로 분산을 줄이는 방법인 Elastic net을 소개한다.

2.2.2. SCAD 분위수 회귀 모형

SCAD (Smoothly clipped absolute deviation)는 Fan과 Li (2001)가 제안한 방법이다. Lasso의 벌점항이 볼록 (convex)인 특징을 가지는 반면, SCAD와 MCP는 오목 (concave)인 벌점 함수를 이용하여 회귀계수 추정량의 편의를 줄이는 방법이다. SCAD 분위수 회귀 모형은 다음의 목적함수를 최소로 하는 $\beta(\tau)$를 추정한다.

$$\hat{\beta}(\tau) = \underset{\beta_0,\beta}{arg\ min} \left[\sum_{i=1}^{n} \rho_\tau(y_i - \beta_0 - x_i^T\beta) + \lambda \sum_{j=1}^{p} P(\beta_j|\lambda,\gamma) \right],$$

여기서 $P(\beta_j|\lambda,\gamma)$ 는 SCAD의 벌점함수로 다음과 같다.

$$P(\beta_j|\lambda,\gamma) = \begin{cases} \lambda|\beta|, & |\beta| \le \lambda \\ \dfrac{2\gamma\lambda|\beta| - \beta^2 - \lambda^2}{2(\gamma-1)}, & \lambda \le |\beta| \le \gamma\lambda \\ \lambda^2, & |\beta| \ge \gamma\lambda. \end{cases}$$

위 식에서 γ는 벌점항의 오목성을 조절하는 모수로 벌점 함수의 기울기를 얼마나 빠르게 감소시킬지를 조절하는 모수이며 $\gamma > 2$의 값을 가진다.

2.2.3. MCP 분위수 회귀 모형

MCP(Minimax Concave Penalty)는 Zhang (2010)이 제안한 방법으로 SCAD와 같이 오목한 벌점함수를 사용하여 회귀계수 추정량의 편의를 줄이는 방법이다. MCP 분위수 회귀 모형은 다음의 목적함수를 최소로 하는 $\beta(\tau)$를 추정한다.

$$\hat{\beta}(\tau) = \underset{\beta_0,\beta}{arg\ min} \left[\sum_{i=1}^{n} \rho_\tau(y_i - \beta_0 - x_i^T\beta) + \lambda \sum_{j=1}^{p} P(\beta_j|\lambda,\gamma) \right],$$

여기서 $P(\beta_j|\lambda,\gamma)$는 MCP의 벌점함수로 다음과 같다.

$$P(\beta_j|\lambda,\gamma) = \begin{cases} \lambda|\beta| - \dfrac{\beta^2}{2\gamma}, & |\beta| \le \gamma\lambda \\ \frac{1}{2}\gamma\lambda^2, & |\beta| \le \lambda. \end{cases}$$

위 식에서 γ는 벌점항의 오목성을 조절하는 모수로 $\gamma > 1$의 값을 가진다.

Figure 2.1은 Lasso, SCAD, MCP의 벌점함수를 나타낸다. Lasso는 회귀계수의 절댓값이 증가할수록 벌점의 크기도 일정하게 증가하는 반면, SCAD와 MCP는 벌점함수의 기울기가 점점 감소하여 을 기준으로 일정한 크기의 벌점을 유지한다. 이를 통해 SCAD와 MCP는 절댓값이 큰 회귀계수에 대해 Lasso만큼 과한 벌점을 가하지 않음으로 Lasso 추정량에 비해 편의를 줄일 수 있다.

2.2.4. Elastic net 분위수 회귀 모형

Elastic net은 Zou와 Hastie (2005)가 제안한 방법으로, Ridge와 Lasso 벌점함수를 결합한 방법이다. Lasso의 변수선택 기능과 함께 Ridge의 해의 유일성 특징을 가진다. Elastic net 분위수 회귀 모형은 다음의 목적함수를 최소로 하는 $\beta(\tau)$를 추정한다.

$$\hat{\beta}(\tau) = \underset{\beta_0,\beta}{arg\ min} \left[\sum_{i=1}^{n} \rho_\tau(y_i - \beta_0 - x_i^T\beta) + \lambda \sum_{j=1}^{p} [\alpha|\beta_j| + (1-\alpha)\beta_j^2] \right],$$

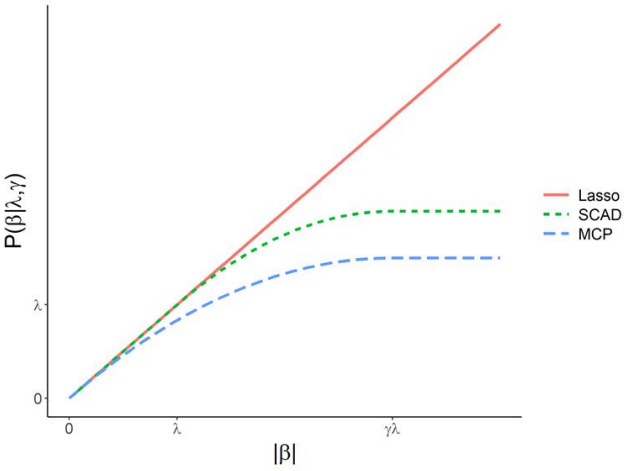

Figure 2.1 Penalty function of Lasso, SCAD and MCP

여기서 λ 는 벌점의 크기를 조절하는 조율모수이고, α 는 Ridge와 Lasso 벌점의 비율을 조절하는 값으로 $[0, 1]$의 값을 가진다. 만약 $\alpha = 1$ 인 경우 위의 벌점항은 Lasso와 같고, $\alpha = 0$ 인 경우는 Ridge와 같다.

3. 분석 결과

3.1. 자료 소개

서울시의 월 강수량의 예측을 위해 기상자료 개방 포털 (https://data.kma.go.kr)에서 제공하는 자료를 이용하였다. 분석에 사용된 자료는 2001년 1월부터 2020년 12월까지의 20년간 월별 자료이다. 본 연구에서는 2001년부터 2018년까지의 자료로 모형을 적합하고 2019년과 2020년의 서울시 월 강수량을 예측하는 것으로 진행하였다.

반응 변수로는 서울시 월 강수량, 설명 변수로는 19개의 기상변수 (평균 기온, 평균 최고기온, 평균 최저기온, 평균 현지 기압, 평균 해면기압, 최고 해면기압, 최저 해면기압, 평균 수증기압, 최고 수증기압, 최저 수증기압, 평균 이슬점온도, 평균 상대습도, 최소 상대습도, 일 최다 강수량, 평균풍속, 최대풍속, 평균 운량, 평균 중하층운량, 합계 일사량)를 이용하였다.

모형 적합에 앞서 서울시의 월 강수량을 상자그림으로 요약하면 Figure 3.1과 같다. 그림에서 장마기간이 포함된 여름철에 강수량이 많은 특징이 잘 나타난다.

Figure 3.2는 주요 변수들 간의 상관분석을 실시한 결과이다. 월 강수량은 일 최다 강수량과 매우 강한 양의 상관관계를 가지며, 평균기온, 평균최고기온, 평균최저기온, 평균수증기압, 최고수증기압, 최저수증기압, 평균이슬점온도, 평균상대습도, 최소상대습도, 평균운량, 평균중하층운량과는 중간 수준 이상의 양의 상관관계를 가진다. 반면, 평균현지기압, 평균해면기압, 최고해면기압, 최저해면기압들과는 중간 수준의 음의 상관관계를 가지며, 평균풍속, 최대풍속, 합계일사량과는 거의 상관성이 없음을 알 수 있다.

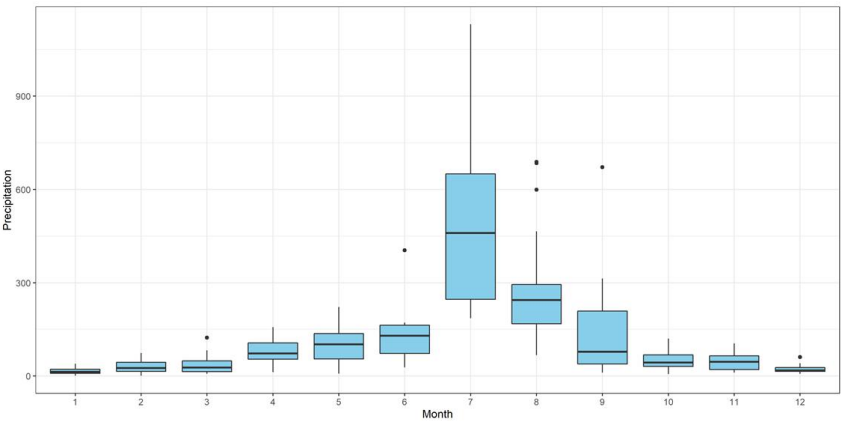

Figure 3.1 Box plot for monthly precipitation in Seoul (2001-2018)

3.2. 모형적합 결과

모형적합에 앞서 모든 설명 변수에 대해 표준화를 수행하였다. 통계패키지 R을 이용하여 분석을 수행하였으며, 구체적인 절차는 다음과 같다. 먼저 분위수 회귀 모형은 {quantreg} 패키지를, Lasso, SCAD, MCP 분위수 회귀 모형은 {rqPen} 패키지를 이용하였으며 동시에 BIC를 최소로 하는 벌점 조율모수 λ를 선택하였다. 아울러 Elastic net 분위수 회귀 모형은 {hqreg} 패키지를 이용하여 손실함수가 최소가 되도록 하는 α 와 λ 를 선택하였으나, 3가지 분위수의 추정 모형에서 모두 $\alpha = 1$ 이 선택되어 Lasso 분위수 회귀 모형과 동일한 결과를 얻게 되어 그 결과를 제시하지 않았다.

Table 3.1은 각 모형을 적합한 결과를 요약한 것이다. 이 표에서 1열은 추정 분위수, 2열은 적합모형, 3열은 선택된 설명변수의 수이고, 4열과 5열은 각각 모형의 설명력을 나타내는 의사결정계수 ($pseudo\ R^2$) 및 선택된 λ 에서의 베이지안 정보기준 (BIC)으로 각각 다음과 같이 정의된다.

$$pseudo\ R^2 = 1 - \frac{\sum_{y_i \geq \hat{y}_i} \tau |y_i - \hat{y}_i| + \sum_{y_i \leq \hat{y}_i} (1-\tau)|y_i - \hat{y}_i|}{\sum_{y_i \geq y_\tau} \tau |y_i - y_\tau| + \sum_{y_i \leq y_\tau} (1-\tau)|y_i - y_\tau|},$$

$$BIC \ = \ log(\sum_{i=1}^{n} \rho_\tau(y_i - \hat{y_i})) + \frac{log(n)}{2n} p, \quad \rho_\tau(u) = \begin{cases} \tau u, & u \geq 0 \\ (\tau - 1)u, & u < 0, \end{cases}$$

여기서 y_i는 관측된 월 강수량, $\hat{y}_i$은 추정값, y_τ는 절편만 포함하는 모형에서의 적합값, p는 설명변수의 수, n은 자료의 수를 의미한다. 의사결정계수는 0∼1의 값을 가지며, 1에 가까울수록 모형 적합이 잘 되었음을 의미한다. BIC는 설명변수의 수를 고려한 모형평가 측도이며, 값이 작을수록 모형 적합이 잘 되었음을 의미한다. 두 측도의 출처는 Koenker 등 (1999)과 Lee 등 (2014)이다.

Table 3.1의 결과를 요약하면 다음과 같다. 고려된 모든 분위수 (τ=0.1, 0.9, 0.95)에서 3종류의 벌점 (Lasso, SCAD, MCP) 분위수 회귀모형의 적합 결과 모든 설명변수 (19개)를 포함하는 분위수 회귀모형에 비해 설명력 ($pseudo\ R^2$)은 다소 떨어지나 모형을 크게 단순화하여 적합도 (BIC)가 우수함을 알 수 있다. 아울러 적합된 벌점 분위수 모형 간에는 설명력에 큰 차이는 없으나, SCAD와 MCP 분위수 모형이 Lasso 분위수 모형보다 좀 더 단순한 추정 결과를 제공한다.

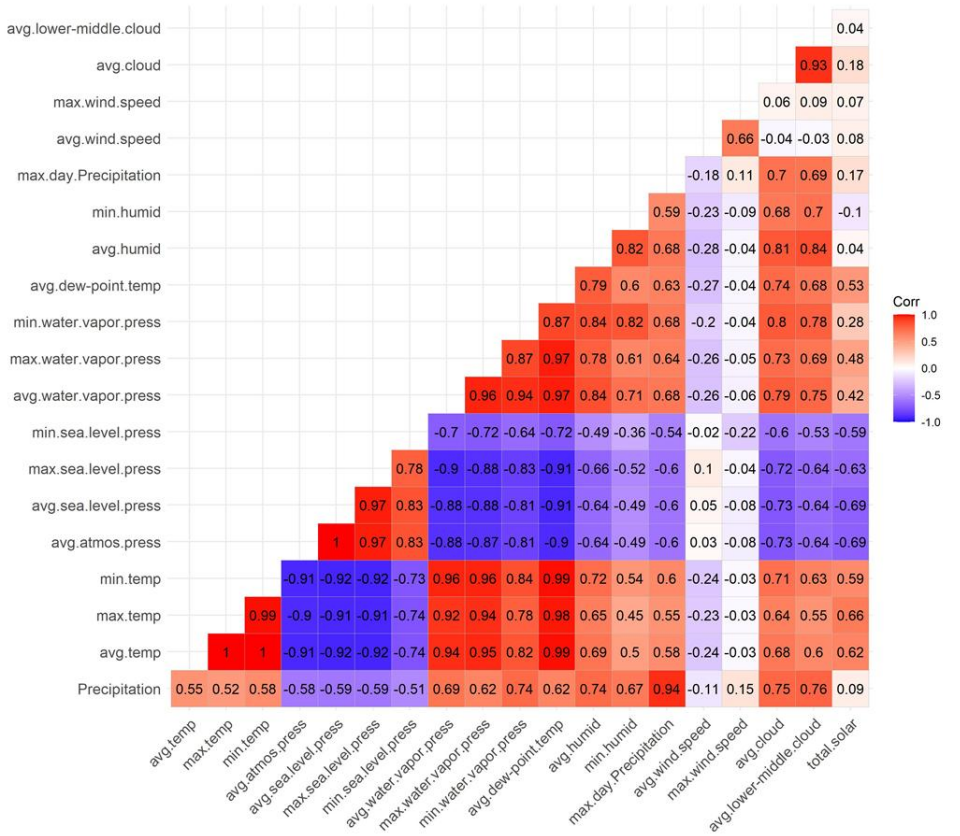

Figure 3.2 Correlation plot between monthly precipitation and meteorological variables

3.3. 예측 성능 비교

이 절에서는 2001년~2018년 자료를 이용한 4가지 적합모형을 이용하여 향후 2년간 (2019년~2020년)의 월 강수량을 예측하고, 그 성능을 비교한다. Table 3.2는 적합된 각 모형의 예측 성능을 비교한 결과이다. Table 3.2에서 3열은 관측값과 추정값 간의 피어슨 상관계수를, 4열은 다음과 같이 정의되는 예측오차를 나타낸다.

$$Prediction\ error = \frac{1}{n}\left[\sum_{y_i \geq \hat{y}_i} \tau|y_i - \hat{y}_i| + \sum_{y_i < \hat{y}_i} (1-\tau)|y_i - \hat{y}_i|\right],$$

여기서 τ는 분위수, y_i 는 관측된 월 강수량, $\hat{y}_i$은 예측값을 의미한다.

Table 3.2에서 관측값과 예측값 간의 상관계수는 모두 0.95이상으로 매우 높게 나타났다. 예측오차는 $\tau = 0.1$ 과 $\tau = 0.9$ 에서는 모든 설명변수를 포함하는 분위수 회귀모형이 타 모형에 비해 다소 작은 결

Table 3.1 Number of variables, pseudo R^2 and BIC for each model

Quantile	Model	Number of variables	Pseudo R^2	BIC
$\tau = 0.1$	Quantile regression	19	0.5499	7.2360
	Quantile Lasso	6	0.5139	7.1512
	Quantile SCAD	6	0.5162	7.1465
	Quantile MCP	6	0.5162	7.1465
$\tau = 0.9$	Quantile regression	19	0.8638	7.4043
	Quantile Lasso	9	0.8611	7.2996
	Quantile SCAD	7	0.8606	7.2785
	Quantile MCP	7	0.8605	7.2789
$\tau = 0.95$	Quantile regression	19	0.8876	6.8556
	Quantile Lasso	9	0.8834	6.7673
	Quantile SCAD	6	0.8813	6.7478
	Quantile MCP	6	0.8813	6.7478

Table 3.2 Correlation and prediction error for each model

Quantile	Model	Correlation	Prediction Error
$\tau = 0.1$	Quantile regression	0.9570	4.2399
	Quantile Lasso	0.9541	5.1145
	Quantile SCAD	0.9551	5.0322
	Quantile MCP	0.9551	5.0322
$\tau = 0.9$	Quantile regression	0.9554	10.4644
	Quantile Lasso	0.9533	11.1299
	Quantile SCAD	0.9526	11.3459
	Quantile MCP	0.9527	11.2707
$\tau = 0.95$	Quantile regression	0.9531	8.6459
	Quantile Lasso	0.9529	8.6602
	Quantile SCAD	0.9515	8.2846
	Quantile MCP	0.9515	8.2846

과를 보이나, $\tau = 0.95$ 에서는 SCAD와 MCP 분위수 회귀모형이 가장 우수한 것으로 나타났다.

다음으로 적합모형의 결과를 SCAD 분위수 회귀모형의 결과를 대표로 제시한다 (나머지 벌점 분위수 회귀모형의 적합 결과가 시각적으로는 모두 유사하므로). Figure 3.3은 2001년~2020년 기간의 월 강수량과 및 추정 (또는 예측) 결과를 나타낸다. 그림에서 실선은 관측된 월 강수량이며 3가지 점선은 적합모형으로부터 각각 10%, 90% 및 95% 분위수를 추정한 값이다. 그 결과 예측 모형으로부터의 추정 또는 값이 실제 관측값의 패턴을 잘 반영하는 것을 알 수 있다.

또한, Figure 3.3에서 우측 음영 부분은 적합 모형으로부터 2년간의 예측 결과로 Figure 3.4는 이를 보다 자세히 나타낸 것이다. 여기서 특이한 점은 2020년 8월의 경우 실제 월 강수량이 상위 5% 분위수 예측값보다 더 큰 것을 확인할 수 있다. 이는 기상이변으로 인해 작년 6월 말부터 시작된 장마가 54일간 이례적으로 지속되어 유래없이 많은 강수량이 관측되었기 때문이다. 최근의 기상 이변은 극단 강수량을 비롯한 각종 기상 관련 예측을 더욱 어렵게 만드는 원인이 되고 있다.

마지막으로, 기상 이변으로 인한 2020년 8월의 월 강수량이 예측값 기준으로 상위 몇 분위수에 해당하는 값인지를 살펴보자. Table 3.3은 3가지 벌점 분위수 회귀 모형을 적합하여 상위 3%와 상위 1% 분위수를 예측한 결과이다. 그 결과 해당 월에 관측된 실 강수량이 상위 3% 수준의 분위수 예측값에 근접함을 확인할 수 있다.

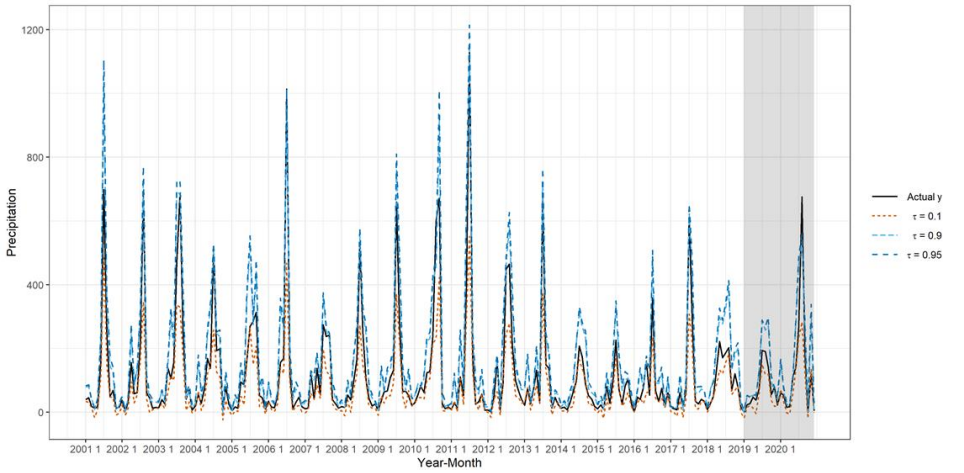

Figure 3.3 Observed and fitted quantile precipitation using quantile SCAD

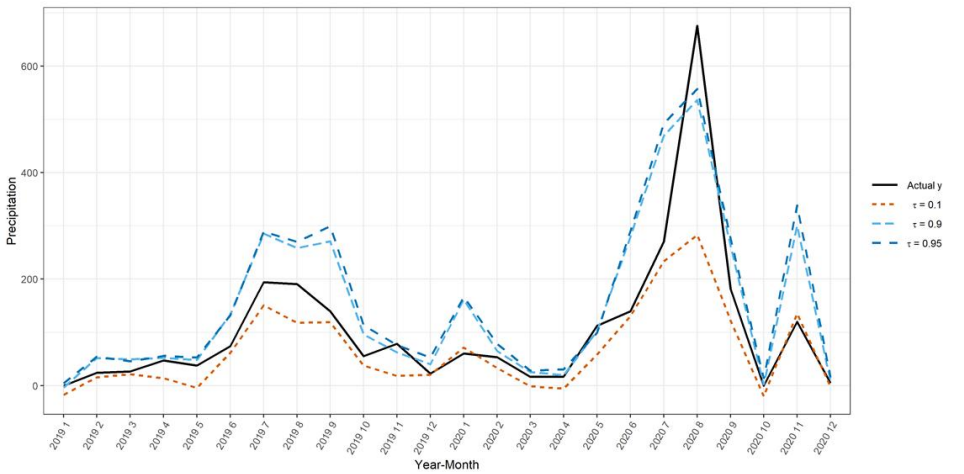

Figure 3.4 Observed and predicted quantile precipitation using quantile SCAD

Table 3.3 Observed and predicted quantile precipitation using penalized quantile regression in august 2020

	$\tau = 0.97$			$\tau = 0.99$		
Actual Precipitation	Quantile Lasso	Quantile SCAD	Quantile MCP	Quantile Lasso	Quantile SCAD	Quantile MCP
675.700	626.672	745.997	656.205	762.667	774.196	769.855

4. 결론

극단 강수량의 예측은 실생활에서 매우 중요한 반면, 최근의 기상 이변으로 인해 극단 강수량의 예측

이 더욱 어려워지고 있다. 본 논문에서는 극단 강수량의 예측을 위해 분위수 회귀 및 벌점 분위수 회귀 모형을 고려하였다. 분석에는 최근 20년간의 서울시 월 강수량 자료를 이용하였으며, 분석 결과를 요약하면 다음과 같다. 먼저 적합모형의 설명력 측면에서 본 논문에서 고려된 분위수 및 여러 가지 벌점 분위수 회귀 모형 모두 우수한 성능을 보였다. 모형의 단순성 측면에서는 SCAD와 MCP 벌점 분위수 모형이 타 모형에 비해 장점을 가지는 것으로 나타났다. 2020년 8월의 이상 폭우는 과거 18년간 자료 기준으로 상위 3 수준의 월 강수량 예측값에 근접함을 확인하였으며, 이러한 현상은 해가 거듭될수록 더욱 심해질 것으로 예상된다. 따라서 향후의 극단 강수량 등의 예측에는 추가적인 정보 또는 새로운 기상정보(예: 위성 관측 자료) 등을 활용하는 보다 정교한 수치 예측모형의 개발이 절실하다.

References

An, S. and Lim, Y. (2020). Forecasting daily PM10 concentration in Seoul Jong-no district by using various statistical technique. *Journal of the Korean Data and Information Science Society*, **31**, 187-198.

Barger, A. S. (2018). Ridge parameter in quantile regression models: An application in biostatistics. *International Journal of Statistics and Applications*, **8**, 72-78.

Statistical postprocessing of ensemble global radiation forecasts with penalized quantile regression. *Meteorologische Zeitschrift*, **26**, 253-264.

Wigena, A. H. and Djuraidah, A. (2016). Quantile regression with elastic-net in statistical downscaling to predict extreme rainfall. *Global Journal of Pure and Applied Mathematics*, **12**, 3517-3524.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association*, **96**, 1348-1360.

Harbi, A. S. M. B. M. and Mohammed, O. B. K. (2017). Using approach quantile regression to determine the factors affecting measuring capacity in Iraq. *American Review of Mathematics and Statistics*, **5**, 35-44.

Kim, M. and Jeong, H. (2022). Development of machine learning based prediction of particulate matter concentration in Seoul. *Journal of the Korean Data and Information Science Society*, **33**, 1095-1111.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica*, **46**, 33-50.

Koenker, R. and Machado, J. A. F. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, **94**, 1296-1310.

Lauret, P., David, M. and Pedro, H. T. C. (2017). Probabilistic solar forecasting using quantile regression models. *Energies*, **10**, 1591.

Lee, E. R., Noh, H. and Park, B. U. (2014). Model selection via bayesian information criterion for quantile regression models. *Journal of the American Statistical Association*, **109**, 216-229.

Mondiana, Y. Q., Zairina, A. and Sari, R. K. (2021). Quantile regression modeling to predict extreme precipitation. *Journal of Physics: Conference Series*, **1918**, 042031.

Santri, D., Wigena, A. H. and Djuraidah, A. (2016). Statistical downscaling modeling with quantile regression using lasso to estimate extreme rainfall. *AIP Conference Proceedings*, **1707**, 080005.

Tibshirani, R. (1996). Regression shirinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267-288.

Wigena, A. H. and Djuraidah, A. (2014). Quantile regression in statistical downscaling to estimate extreme monthly rainfall. *Science Journal of Applied Mathematics and Statistics*, **2**, 66-70.

Won, J. and Na, J. (2020). Prediction pf PM10 in Seoul using penalized regression. *Journal of the Korean Data and Information Science Society*, **32**, 631-640.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, **38**, 894-942.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, **67**, 301-320.

Journal of the Korean Data & Information Science Society
2023, **34**(6), 917–926

http://dx.doi.org/10.7465/jkdi.2023.34.6.917
한국데이터정보과학회지

Prediction of extreme precipitation in Seoul using penalized quantile regression

Hyeongseop Lim[1] · Woosun Lee[2] · Jonghwa Na [3]

[1]Pharmbio Korea Inc.
[2]Pharmaceutical Policy Division, Ministry of Food and Drug Safety
[3]Department of Information and Statistics, Chungbuk National University

Received 24 August 2023, revised 15 September 2023, accepted 29 September 2023

Abstract

In an era of extreme weather events caused by global warming, extreme precipitation has a direct impact on safety and the economy, so predicting it is very important. In this study, prediction of monthly extreme precipitation in Seoul was performed using the quantile regression model and the penalized quantile regression model. Lasso, SCAD, MCP, and Elastic net were considered as penalty terms added to the objective function of the quantile regression model. Performance comparison of the fitted models was performed in terms of explanatory power, BIC, and predictive power. The explanatory power of the fitted quantile and penalty quantile regression models was overall excellent, and among them, the SCAD and MCP quantile regression models showed that the models were greatly simplified. However, it has been confirmed that there are limitations in explaining unusual phenomena caused by recent abnormal weather events with empirical data so far, and therefore, the use of additional information or the development of a more sophisticated numerical model is required to predict extreme precipitation.

Keywords: Extreme precipitation, Lasso, MCP, penalized quantile regression, quantile regression, SCAD.

[1] Employee, Pharmbio Korea Inc., Nonhyeon-ro 83, Seocho-gu, Seoul 06775, Korea.
[2] Assistant director, Pharmaceutical Policy Division, Osong Health Technology Administration Complex, Chungbuk 28159, Korea.
[3] Corresponding author: Professor, Department of Information & Statistics, Chungbuk National University, Chungbuk 28644, Korea. E-mail: cherin@cbnu.ac.kr