강릉 월별 강수량 분석
ARIMA에서 Gamma GLM까지

Taenyoung Lee

Department of Statistics, HUFS

December 22, 2025

발표 흐름

1. 데이터 소개 및 탐색
2. ARIMA 시도와 한계
3. 계절 더미 + SARMA로 시간상관 진단
4. Gamma GLM 채택 및 결과
5. Outlier 분석 및 2011년 이후 변화
6. 강수 변화와 강원도 산불 데이터
7. 결론(의의)

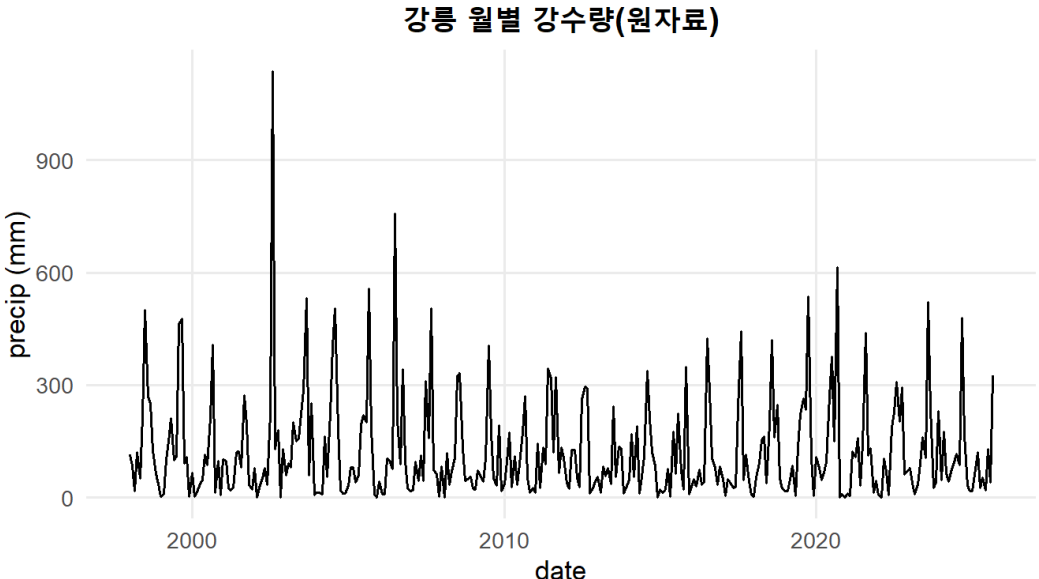
분석 목표

- 강릉 월별 강수량 시계열의 **계절성/시간의존성/분포 특성**을 진단
- 초기 접근: **(Gaussian) ARIMA** 적합 ⇒ 진단상 한계 확인
- 계절 더미 + SARMA 로 자기상관 확인
- 관측 분포(양의 연속형, 우측 긴 꼬리) 반영하여 **Gamma GLM** 선택
- Gamma GLM 기반 **(표준화 deviance residual) outlier** 탐지
- 2011년 기점 이후 **작은 강수 outlier 증가** 관찰 ⇒ 산불 데이터와 함께 확인

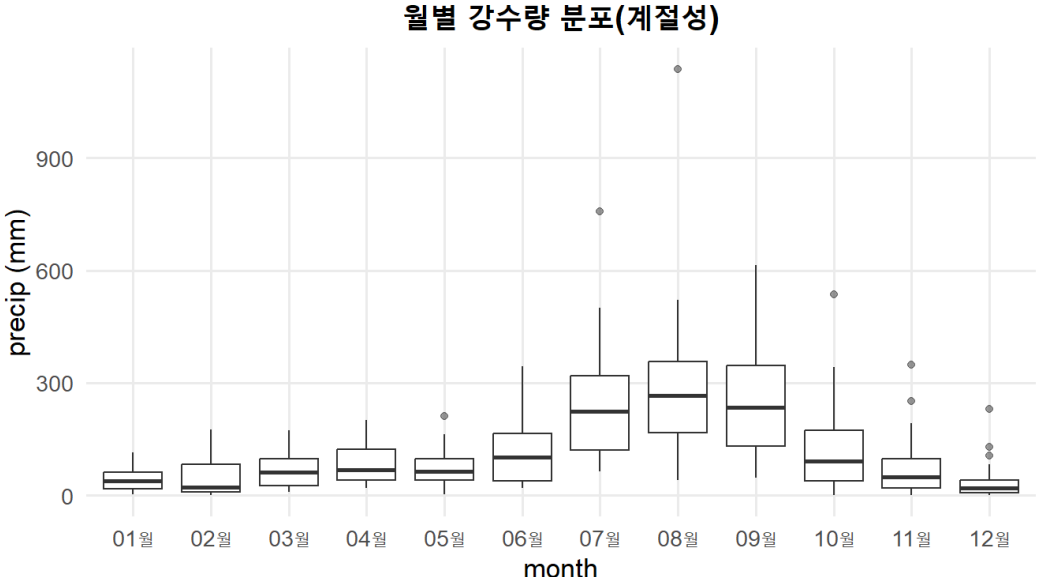
데이터 개요

- 원자료: 강릉 강수량(월별로 집계)
- 데이터 범위: 1998-01 2025-09, 관측치: 331
- 전처리 요약
  - 날짜 문자열(예: “YYYY년 M월”) → `Date` 변환 후 월 시작일로 정렬
  - 월별 강수량 합계(mm)로 집계
  - 파생변수: `month_fac` (01월∼12월), 시점 인덱스 `t`

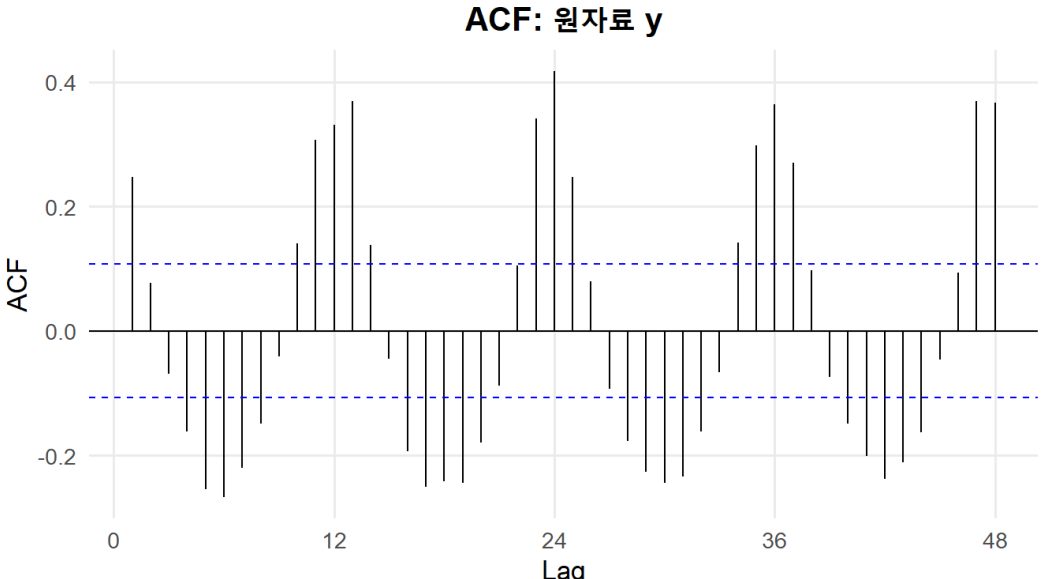
원자료 탐색 1: 시계열 플롯



원자료 탐색 2: 계절성(월별 분포)



원자료 탐색 3: 원자료 ACF



데이터에서 예상되는 어려움

- 강수량은 일반적으로
  - **0 또는 매우 작은 값**이 존재하고,
  - **우측 긴 꼬리(right-skew)**를 가짐.
- 따라서 **정규 오차**를 기본 가정하는 시계열 모형(ARIMA)의 잔차 진단에서 정규성 위반이 자주 나타남.

ARIMA 접근

- 월별 강수량은 계절성이 강하고(연 주기), 시간의존성이 있을 수 있음
- 목표: (1) 계절성 제거, (2) 자기상관 제거, (3) 잔차를 "백색잡음"으로 만들기

AIC 그리드 탐색

	p <int>	q <int>	P <int>	Q <int>	AIC <dbl>
1	2	2	0	1	3923.620
2	2	2	1	1	3924.703
3	2	2	0	2	3924.817
4	2	3	0	1	3924.902
5	3	2	0	1	3924.977

선택 모형

$$\text{ARIMA}(2,0,2)\,(0,1,1)_{12}, \qquad \texttt{include.mean = FALSE}$$

선택된 ARIMA 모형

ARIMA 적합 결과

Table: ARIMA(2,0,2)(0,1,1)$_{12}$ Model Estimation Results

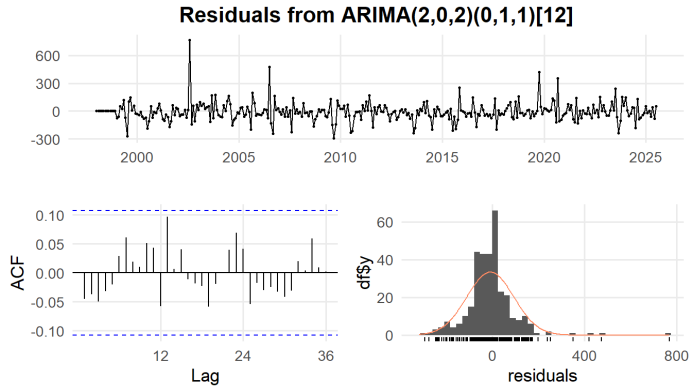
Parameter	Estimate	S.E.	t-value	p-value
AR1 (ϕ_1)	0.6359	0.0546	11.646	0.001
AR2 (ϕ_2)	-0.8962	0.0371	-24.156	0.001
MA1 (θ_1)	-0.7236	0.0387	-18.698	0.001
MA2 (θ_2)	0.9589	0.0313	30.636	0.001
SMA1 (Θ_1)	-0.9996	0.1223	-8.173	0.001

Ljung-Box test

- Q* = 15.224, df = 19, p-value = 0.7083
- Model df: 5. Total lags used: 24

ARIMA 진단: “잔차 정규성” 문제

- ARIMA 잔차에 대해 정규성 검정 결과
  - Shapiro–Wilk: $p \approx 4.6 \times 10^{-16}$
  - Jarque–Bera: $p \approx 0$
  - Anderson–Darling: $p \approx 1.5 \times 10^{-18}$
- 결론: **정규성 가정이 강하게 깨짐** $\Rightarrow$ Gaussian ARIMA로는 분포 특성 반영이 어려움

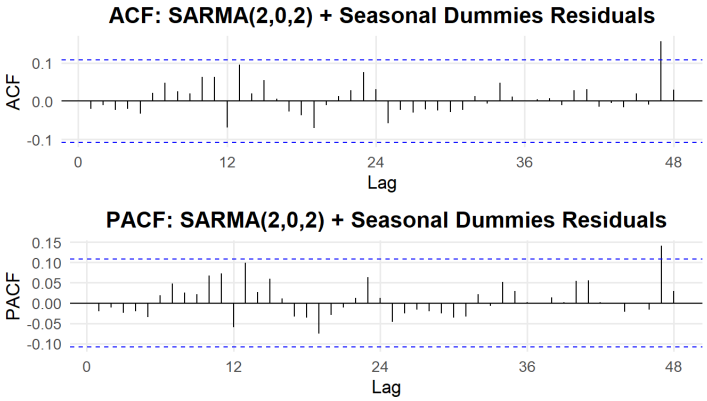


ARIMA 요약: 한계

- 계절/자기상관 측면의 진단은 일정 수준 통과할 수 있어도,
- 핵심 문제는 **강수량 자료의 분포(양의 값, 우측 꼬리)**와 **정규 오차 가정**의 불일치
- 따라서 **분포를 바꾸는 방향**의 모형화가 필요

SARMA 적합 및 진단: ACF/Ljung–Box 결과

- 잔차 진단(ACF)
- Ljung–Box 결과: X-squared = 16.142, df = 24, $p \approx 0.8$ 수준
  $\Rightarrow$ **잔차 자기상관이 유의하지 않음**



해석: “시간상관”보다 “분포”가 핵심

- 월 더미를 넣고 나면,
  - 잔차에서 **자기상관이 크게 남지 않음** (ACF, Ljung–Box 통과)
- 따라서 이후 접근은
  - **강수량 분포를 더 잘 설명하는 모형**을 채택하는 것이 합리적

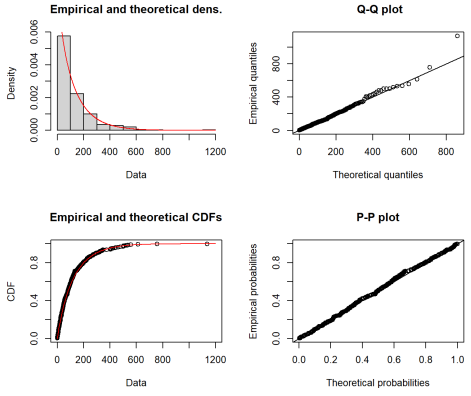
왜 Gamma GLM인가?

- 강수량은 대표적인 **양의 연속형** + **우측 꼬리** 데이터
- Gamma 분포는
  - 지지집합이 $(0, \infty)$이고
  - (형상에 따라) 다양한 수준의 우측 비대칭을 표현 가능
- 로그 링크를 쓰면 평균이 항상 양수:

$$\mathbb{E}[Y_t \mid \text{month} = m] = \exp(\eta_m)$$

- Thom(1958), McKee et al.(1993), Husak et al.(2007), Martinez-Villalobos & Neelin(2019) 등 gamma 분포를 강수량 분포로 사용

Gamma GLM 진단: 분포진단



검정결과

KS statistic: 0.04296419 KS test 결과: 1-mle-gamma "not rejected"
Chi-square stat: 10.98144 df: 14 p: 0.6874933

Gamma GLM 설정

모형

$$Y_t \mid \text{month} = m \sim \text{Gamma}(\mu_m, \phi), \qquad \log(\mu_m) = \beta_0 + \beta_m$$

- 구현: `glm(precip ~month_fac, family = Gamma(link="log"))`
- 주의: Gamma는 0을 허용하지 않으므로 `precip > 0`만 사용

Gamma GLM 결과: 계수 요약

추정 계수

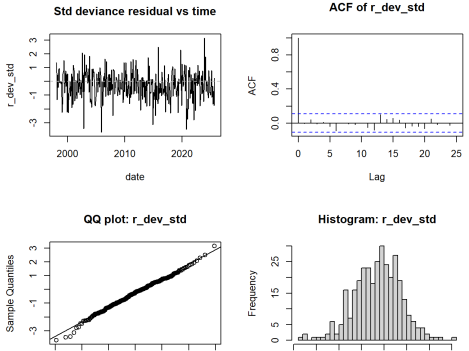
항	Estimate	Std. Error	t value	p-value
(Intercept)	3.7613	0.1628	23.10	$< 2 \times 10^{-16}$
02월	0.1936	0.2371	0.82	0.415
03월	0.4129	0.2303	1.79	0.0739
04월	0.6578	0.2303	2.86	0.00457
05월	0.5304	0.2303	2.30	0.0219
06월	0.9654	0.2303	4.19	3.59×10^{-5}
07월	1.7521	0.2303	7.61	3.21×10^{-13}
08월	1.9158	0.2303	8.32	2.67×10^{-15}
09월	1.8154	0.2303	7.88	5.20×10^{-14}
10월	1.0597	0.2324	4.56	7.34×10^{-6}
11월	0.5305	0.2324	2.28	0.0231
12월	-0.1862	0.2347	-0.79	0.428

Gamma GLM 진단: deviance residual

- 표준화 deviance residual:

$$r_t^{(std)} = \frac{r_t^{(dev)}}{\sqrt{\hat{\phi}(1-h_t)}}$$

- 정규성 검정: Shapiro $p \approx 0.31$, JB $p \approx 0.11$, AD $p \approx 0.33$
- Ljung–Box(24 lag) : $p \approx 0.80 \Rightarrow$ **자기상관 없음**



Outlier 정의(양쪽 꼬리)

- 유의수준 $\alpha = 0.05$ (양쪽 합)
- 임계값: $z_{1-\alpha/2} \approx 1.96$
- Outlier 판정:

$$|r_t^{(std)}| > 1.96 \Rightarrow \text{Outlier}$$

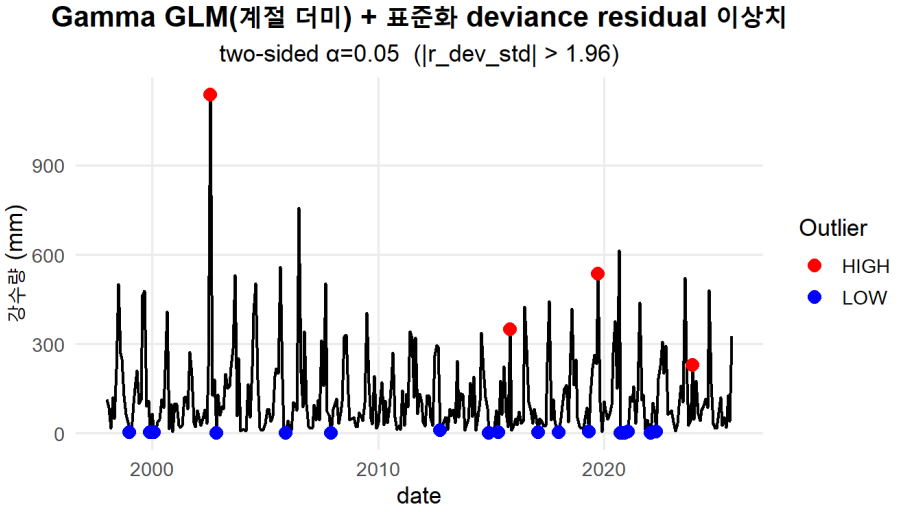
- 부호 해석
  - $r_t^{(std)} > 0$: 관측 강수량이 모델 기대보다 큼(HIGH)
  - $r_t^{(std)} < 0$: 관측 강수량이 모델 기대보다 작음(LOW)

대표 Outlier 예시

$|r^{(std)}|$ 큰 순 정렬 결과 일부

date	precip	fitted_glm	r_dev	r_dev_std	type
2005-12	0.1	35.7	-3.12	-3.70	LOW
2020-10	0.6	124.1	-2.95	-3.48	LOW
2002-11	0.4	73.1	-2.90	-3.43	LOW
2023-12	229.0	35.7	2.67	3.16	HIGH
2014-12	0.4	35.7	-2.65	-3.13	LOW

Outlier 시계열 표시 + 2011년 기점



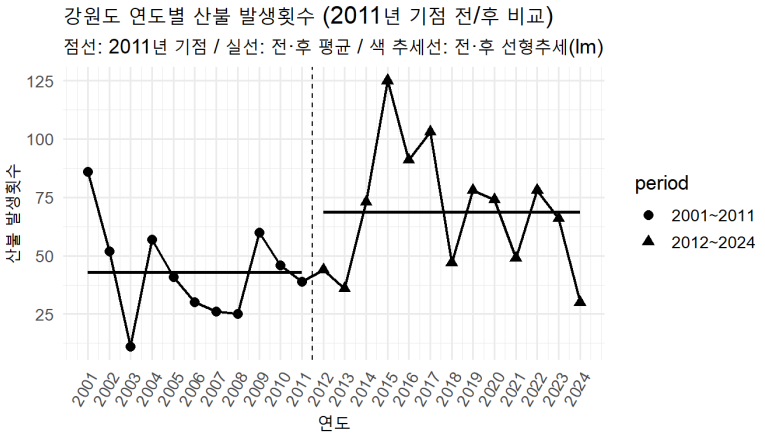
- 중간일(2011년 11월) 기준 outlier 개수: **이전 6개, 이후 11개**

해석: “적은 강수 outlier” 증가

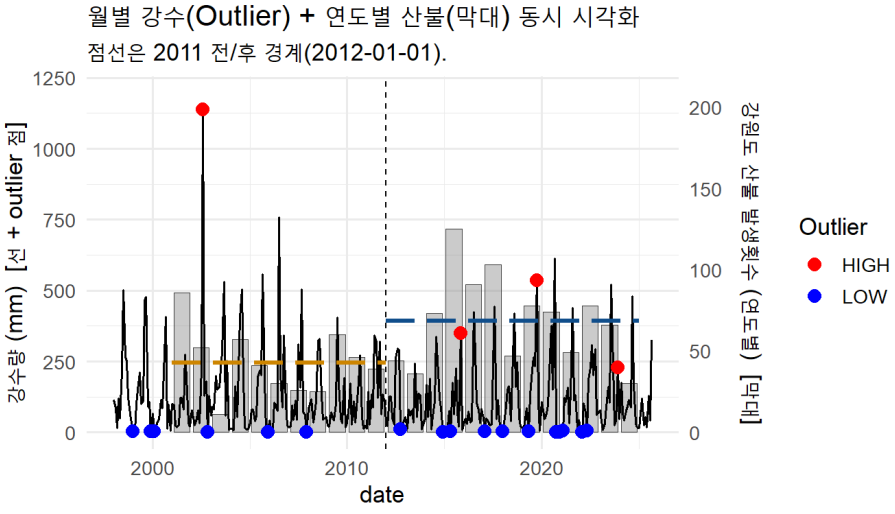
- 관측 사실: 2011년 전/후로 “예상보다 적게 내린 달”이 늘어남
- 가능한 해석(가설)
  - 지역 기후 패턴의 변화(가뭄/건조화)
- 다음 단계: 이런 “건조 시그널”이 실제 위험과 연동되는지 확인

산불 데이터 소개(강원도 연도별 발생 횟수)

- 2001–2024년, 강원도 연도별 산불 발생 횟수(count)
- 동일 기준선: 2011년 전/후(2001–2011 vs 2012–2024)
- 요약
  - 2001–2011 평균: 43.0, 2012–2024 평균: 68.77



강수 outlier + 산불(막대) 동시 시각화



- 시각적으로 “작은 강수 outlier 증가” 시기와 “산불 증가” 시기를 함께 확인

결론 및 의의

- “시계열 분석”을 시작점으로 하되, 진단을 통해 **분포 기반 모형**으로 자연스럽게 전환
- Gamma GLM은 강수량의 특성(양의 값, 우측 꼬리)을 반영하며 진단도 양호
- 표준화 deviance residual로 outlier를 정의하여 기존 정규분포 기반 outlier 탐지 방법으로 탐지 불가능했던 **건조한 달**을 탐지 가능
- 2011년 이후 LOW outlier 증가 관찰 $\Rightarrow$ 산불 증가와의 연관 가능성을 제시

감사합니다

Q & A