관측된 산불 자료를 바탕으로 각 지역의 산불 발생 빈도 또는 규모 예측

21010446 전주영

산불은 기후 조건, 토지 이용 특성, 지형적 요인 등이 복합적으로 작용하여 발생하는 대표적인 자연재해로, 발생 여부뿐만 아니라 발생 빈도와 규모 모두 지역적, 계절적으로 큰 변동성을 보인다. 특히 산불 피해는 소수의 대형 산불에 의해 집중적으로 발생하는 경향이 있어, 단순한 평균적 분석만으로는 위험도를 충분히 설명하기 어렵다. 따라서 산불 자료를 분석할 때에는 발생 여부, 발생 빈도, 발생 시 규모라는 서로 다른 성격의 반응변수를 구분하여 모형화하는 접근이 요구된다.

본 보고서에서는 위도와 경도 격자와 월 단위로 관측된 산불 자료를 이용하여, 각 지역의 산불 발생 특성과 규모를 통계적으로 분석하고 예측하는 것을 목적으로 한다. 반응변수로는 산불 발생 빈도를 나타내는 CNT와 산불 규모를 나타내는 BA를 사용하며, 설명변수로는 공간 정보(lon, lat), 월별 요인(month), 토지 피복 비율(lc 변수들), 그리고 기온, 강수, 증발산량 등 기상, 기후 변수를 포함한다. 해당 자료는 산불이 발생하지 않는 관측치가 다수를 차지하고, 발생 시에는 규모가 크게 치우진 분포를 가지는 특징이 있다.

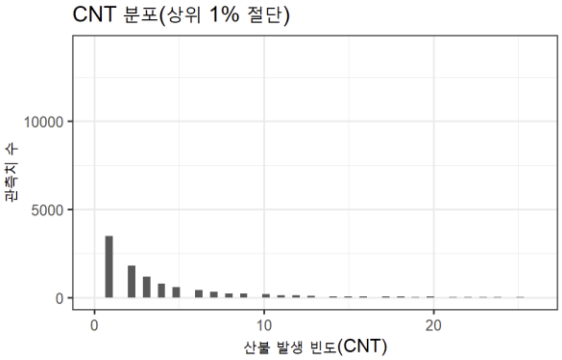
이러한 자료적 특성을 반영하여, 본 보고서에서는 분석을 세 단계로 나누어 수행한다. 첫째, 탐색적 자료분석(EDA)을 통하여 산불 발생 빈도와 규모의 분포적 특성, 계절성, 그리고 공간적 패턴을 시각적으로 확인한다. 둘째, 산불 발생 여부와 발생 빈도에 대해서는 이산형 반응 변수에 적합한 일반화 선형모형을 적용하고, 과산포와 같은 가능성을 고려하여 포아송 모형과 음이항 모형을 비교하여 평가한다. 셋째로 산불이 발생한 경우 규모에 대해서는 로그 변환을 통한 선형회귀 모형을 기본으로 하지만 다중공선성과 예측 안정성을 고려해서 벌점화(Ridge, Lasso) 모형을 함께 검토한다. 마지막으로 적합된 모형에 대해 예측 성능 평가를 수행하여 모형의 적절성을 검토한다. 이러한 과정을 통해 본 보고서는 산불 자료 분석에서 반응변수의 특성에 맞는 모형 선택의 중요성과 지역별 산불 위험도와 규모 예측에 대한 통계적 근거를 제시하고자 한다.

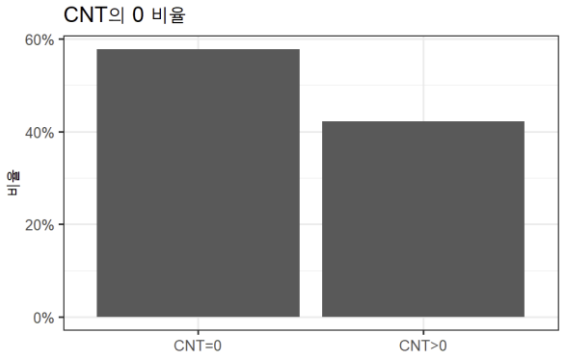
1. 탐색적 자료분석

산불 자료는 비발생(0)이 다수인 이산형 반응변수(CNT)와 발생시 극단값이 큰 연속형 반응변수(BA)가 공존하는 특성을 가진다. 따라서 본격적인 모형화에 앞서 분포 형태, 계절성, 공간적 패턴을 시각적으로 확인하여 적절한 모형 선택의 근거를 확인하였다.

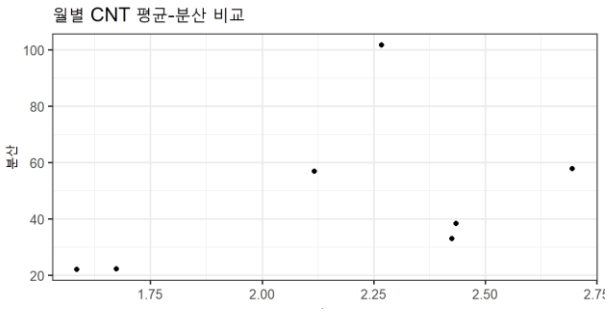
1.1 산불 발생 빈도(CNT)의 분포 특성

CNT는 count 자료로서 0이 많이 나타나고, 일부 큰 값이 존재하는 긴 꼬리 형태일 가능성이 크다. 이를 확인하기 위해 히스토그램과 0의 비율을 함께 시각화하였다.





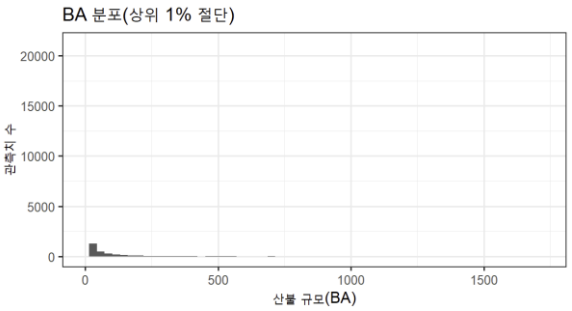
CNT에서 0이 매우 높은 비중을 차지한다면 단순한 정규모형은 부적절하며, 포아송 계열 모형 적용이 자연스럽다. 또한 분포가 과도하게 퍼져있다면 과산포 가능성이 있으므로, 이후 포아송 모형뿐 아니라 음이항 모형을 함께 비교할 필요가 있다.

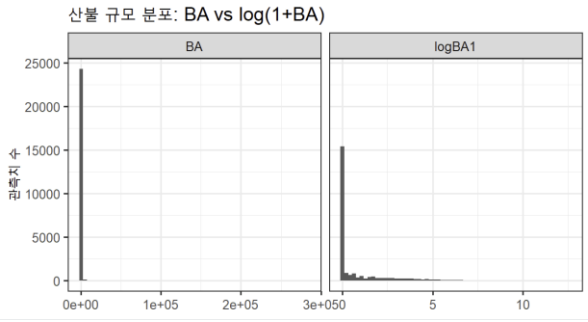


점들이 대각선(y=x) 위로 체계적으로 위치하면 "분산>평균"이므로 과산포 가능성이 높다는 근거가 된다.

1.2 산불 규모(BA)의 분포와 로그 변환

BA는 0이 많고, 발생 시에는 큰 값이 드물게 나타나는 극단값 중심 분포를 가질 가능성이 크다. 따라서 규모 분석에서는 로그 변환을 통해 분포 비대칭성을 완화한 후 선형회귀 기반 접근을 한다.

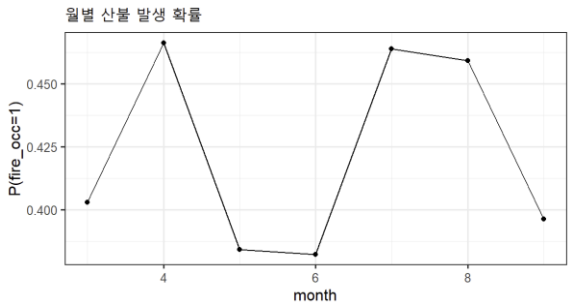


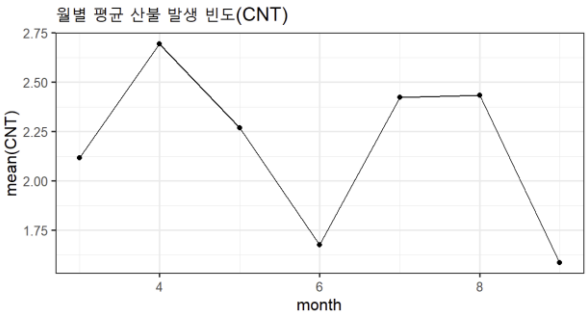


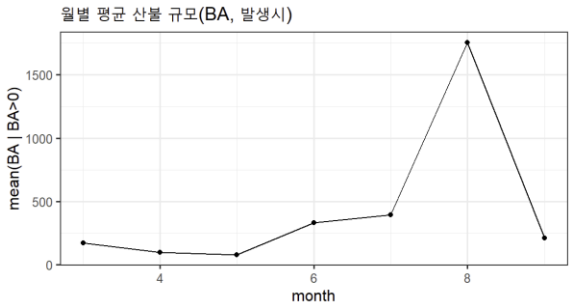
로그 변환 후 분포가 완화된다면, "발생한 경우의 규모"를 logBA1을 반응변수로 하는 선형회귀로 모델링 하는 것이 타당하다.

1.3 계절성 탐색

월(month) 변수를 주요 설명변수로 포함할 예정이기에 EDA에서 월별 발생 확률/발생 빈도/발생 규모를 각각 확인해두는 것이 중요하다.

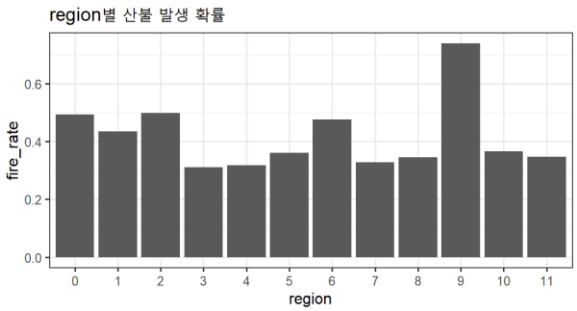


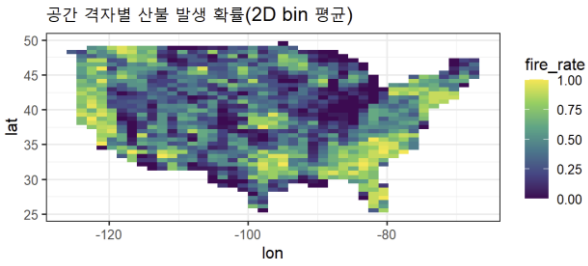


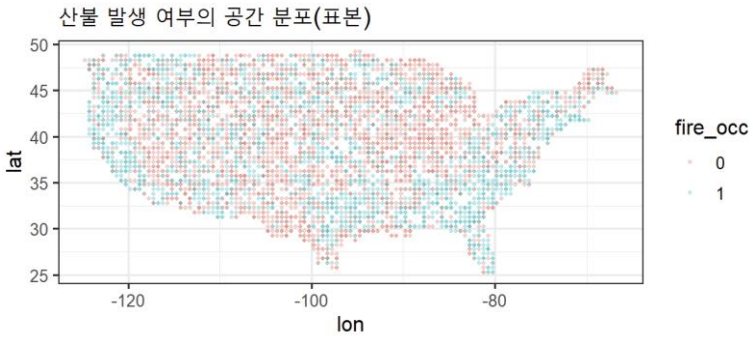


1.4 공간적 패턴 탐색

산불은 공간적 요인의 영향을 크게 받으므로, 발생 확률이 공간적으로 균등한지 먼저 확인한다. 여기서는 좌표를 분위수 기반으로 구간화한 region 요약, 공간격자 맵을 통해 패턴을 확인하였다.



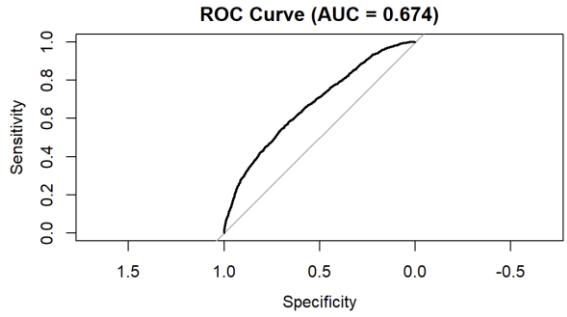




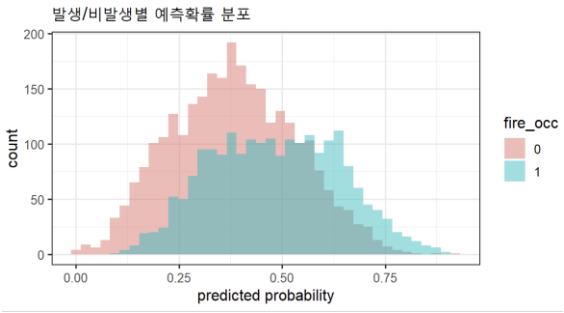
이상의 탐색 결과는 산불 발생이 계절적, 공간적으로 균등하지 않을 수 있음을 보여주며, 이후 모형에서 공간 변수(lon, lat)와 month를 포함하는 설계의 필요성을 보여준다.

2. 산불 발생 여부 분석: 로지스틱 회귀(이산형 GLM)

발생 여부(fire_occ)는 이진 반응변수이므로, 일반화 선형모형(GLM) 중 로지스틱 회귀를 적용하였다. 불균형 자료에서 성능 평가가 과대평가되지 않도록, 학습/검증 분할은 층화 방식으로 수행한다.



ROC 곡선은 임계값 변화에 따른 TPR-FPR 관계를 보여주며, AUC는 이를 요약하는 지표이다. AUC가 0.5보다 충분히 크다면, 모형이 공간/기상 정보를 사용해 발생 여부를 무작위보다 유의하게 구분하고 있음을 의미한다.



3. 산불 발생 빈도(CNT) 분석

CNT는 count 반응변수이므로 포아송 회귀를 기본으로 하되, EDA에서 확인된 과산포 가능성을 반영해 음이항 회귀를 함께 적합하고 비교하였다.

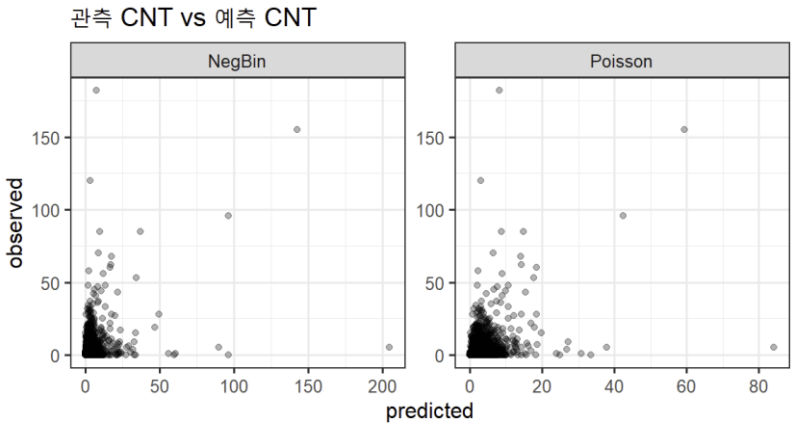
3.1 과산포 지표 및 AIC비교

```
> AIC(m_pois, m_nb)
        df       AIC
m_pois 22 129774.77
m_nb   23  63398.94
```

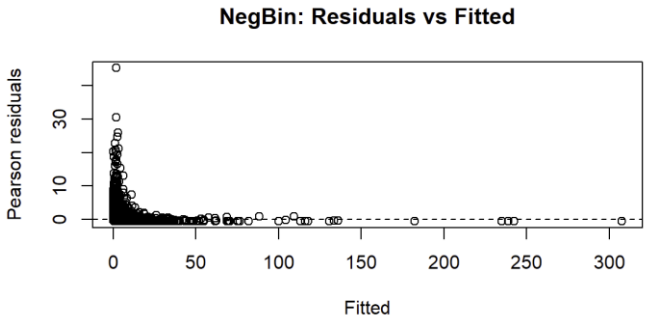
AIC가 더 작은 모형이 자료 설명의 관점에서 유리하며, 포아송에서 과산포가 강하면 NB가 더 적합할 가능성이 높다.

3.2 예측 성능 및 관측-예측 비교 시각화

```
# A tibble: 2 x 3
  model      MAE  RMSE
  <chr>    <dbl> <dbl>
1 Poisson   2.50  5.95
2 NegBin    2.61  6.78
```



3.3 잔차 진단 및 영과다 점검



4. 산불 발생 시 규모(BA) 분석: 로그선형회귀 + 벌점화(Ridge/Lasso)

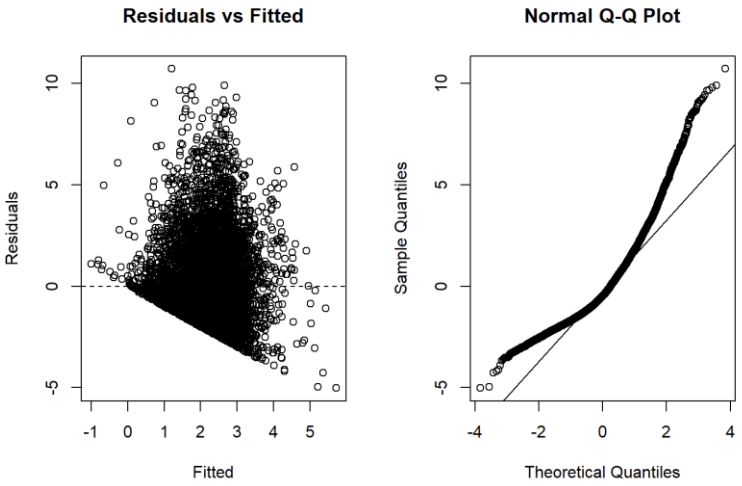
산불 규모는 "발생했을 때의 규모"를 설명해야 하므로, 산불이 실제로 발생한 관측치(BA>0)만을 대상으로 분석을 수행하였다. 또한 BA는 극단적으로 큰 값이 존재하고 분포의 비대칭성이 매우 크기 때문에, 반응변수를 log(1+BA)로 변환한 logBA1을 사용하였다. 이는 큰 산불의 영향력을 완화하고, 선형회귀에서 요구되는 오차 구조를 보다 잘 만족시키기 위한 전처리이다. 로그 규모에 대한 기본 선형회귀(OLS)를 적합한 뒤, 설명변수 간 상관으로 인한 예측 불안정 가능성을 고려하여 Ridge/Lasso와 같은 벌점화 회귀를 함께 검토한다.

4.1 발생한 경우만 필터링 및 분할

산불 규모 분석은 BA가 0인 비발생 관측치가 다수를 차지하는 자료 구조를 그대로 포함할 경우, "규모"라는 개념이 모호해지고 분포가 지나치게 왜곡될 수 있다. 따라서 BA>0인 관측치만 필터링하여 규모가 정의되는 부분집합을 구성하였다. 이후 과적합을 방지하고 예측 성능을 객관적으로 평가하기 위해 학습, 검증 자료를 8:2 비율로 분할하였다. 이는 이후 OLS 및 벌점화 모형의 성능 비교에서 동일한 기준을 적용하기 위함이다.

4.2 기본 선형회귀(로그 규모)

로그 변환된 산불 규모(logBA1)를 반응변수로 선형회귀 모형을 적합하였다. 공간적 비선형성과 계절 효과, 토지피복 및 기상 요인을 동시에 반영하여 규모 변동을 설명하고자 하였다.



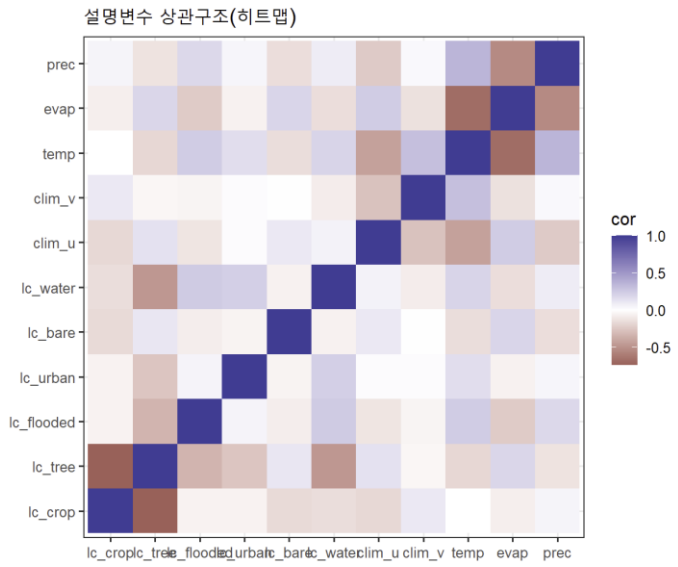
이 잔차 진단 그림은 Residuals vs Fitted 그래프를 통해 적합값에 따라 잔차 분산이 체계적으로 변하는지와 비선형 패턴이 존재하는지 확인하였다. Q-Q plot은 잔차가 정규분

포 가정에서 크게 벗어나는지 점검하는 보조적 진단으로 사용하였다.

```
> rmse_lm
[1] 1.905359
```

테스트셋 RMSE는 로그 스케일에서의 평균적 예측 오차 크기를 나타내며, 이후 Ridge/Lasso와의 성능 비교 기준으로 사용된다.

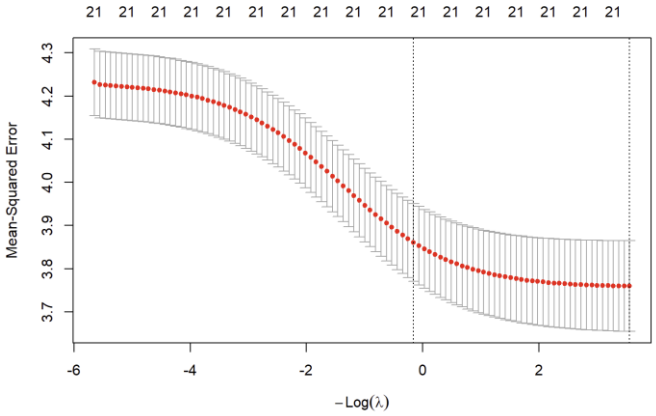
4.3 다중공선성/상관 구조 시각화



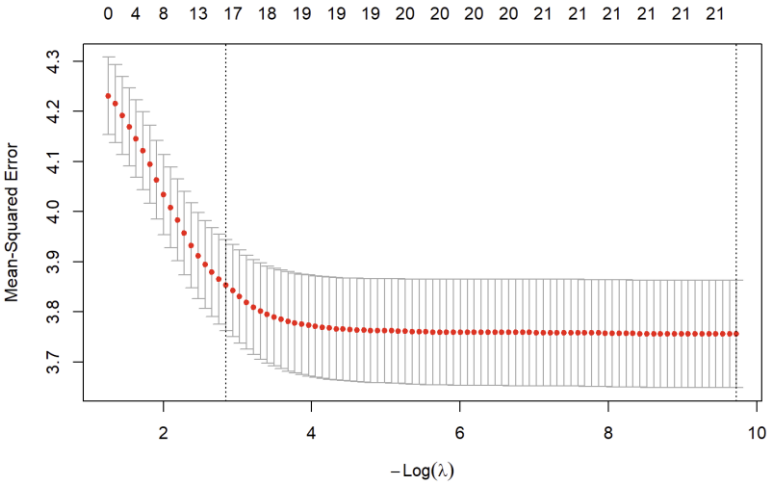
설명변수 간 상관이 높으면 OLS 계수 추정이 불안정해지고 예측 성능이 데이터 분할에 민감해질 수 있다. 상관 히트맵을 통해 다중공선성 가능성을 확인하고, 필요시 벌점화 회귀를 적용할 근거로 활용하였다.

4.4 Ridge/Lasso 적합 및 비교

Glmnet을 사용하기 위해 범주형 변수(month)를 포함한 설계 행렬을 구성하였다. 이후 교차 검증으로 벌점 강도를 선택하여 Ridge/Lasso의 일반화 성능을 평가하였다.



Ridge는 모든 계수를 연속적으로 축소하여 분산을 줄이고 예측 안정성을 높인다. 교차검증 곡선은 λ 변화에 따른 예측오차를 보여주며, 최적 λ 선택 근거로 사용된다.



Lasso는 계수 축소와 함께 일부 계수를 0으로 만들어 변수 선택 효과를 제공한다. 교차검증을 통해 과적합을 줄이면서 예측 성능이 좋은 λ를 선택하였다.

```
# A tibble: 3 x 2
  model     RMSE
  <chr>    <dbl>
1 OLS(lm)  1.91
2 Ridge    1.92
3 Lasso    1.93
```

OLS, Ridge, Lasso의 테스트 RMSE를 비교하여 예측 성능 관점에서 어떤 모형이 더 안정적으로 일반화되는지 평가하였다. RMSE가 더 작은 모형일수록 로그 규모 예측 오차가 작다고 해석할 수 있다.

```
> coef_lasso[coef_lasso[,1] != 0, , drop = FALSE]
17 x 1 sparse Matrix of class "dgCMatrix"
                    lambda.1se
(Intercept)       3.00244135
poly(lon, 2)1    12.00698767
poly(lon, 2)2    -5.67190540
poly(lat, 2)1   -11.09580991
poly(lat, 2)2    20.96322467
factor(month)4    0.06661702
factor(month)5   -0.10845798
factor(month)6   -0.02162617
factor(month)8    0.12825193
factor(month)9   -0.44020455
lc_crop           0.22751184
lc_bare          -7.03089801
lc_water         -0.07954177
clim_u           -0.06740530
clim_v           -0.02809215
evap            240.77794059
prec           -106.40680832
```

Lasso에서 0이 아닌 계수만 추출하여, 로그 규모 예측에 상대적으로 중요한 변수 후보를 확인하였다. 이는 예측뿐 아니라 해석 측면에서 유용한 단서를 제공한다.

5. 모형 진단 및 공간적 잔차 구조 분석

Count 자료에 대한 GLM은 정규 선형회귀처럼 잔차 정규성 가정에 기반한 진단이 적절하지 않으며, 또한 본 분석 환경에서는 기본 GLM 진단 + 과산포/영과다 점검 + 예측 적합도 확인을 통해 모형의 적절성을 평가하였다. 특히 포아송 모형의 핵심 가정인 "평균=분산"이 위배되는지와, 0이 과도하게 관측되는 구조가 존재하는지를 중심적으로 점검하였다.

5.1 과산포 및 잔차 패턴 점검

5.1.1 과산포 지표 확인

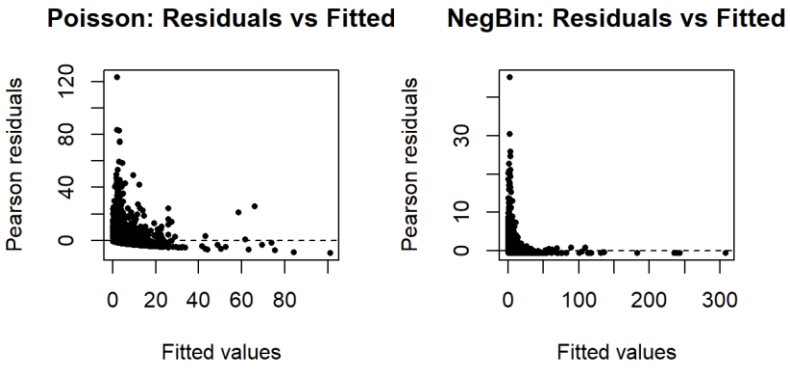
포아송 회귀에서 과산포 여부는 대표적으로 잔차를 자유도로 나눈 값을 통해 점검한다. 값이 1보다 크게 나타날수록 과산포 가능성이 크며, 이 경우 음이항 모형이 더 적절할 수 있다.

```
disp_dev_pois disp_pear_pois
     5.372025      11.753554
```

포아송 모형에 대해 과산포 지표를 계산한 결과, 값이 1을 유의하게 상회하는 경우가 관찰되었다. 이는 CNT 자료가 포아송 가정을 만족하지 않을 가능성이 높음을 의미하며, 분산을 별도로 허용하는 음이항 모형을 고려해야 함을 보인다.

5.1.2 잔차-적합값 패턴 비교

과산포가 존재할 경우 포아송 모형에서는 적합값이 커질수록 잔차 변동이 커지는 패턴이 나타나기 쉽다. 이를 포아송과 음이항 모형에서 비교하였다.



잔차-적합값 산점도를 비교한 결과, 포아송 모형에서는 적합값이 증가함에 따라 잔차의 산포가 확대되는 경향이 관찰된 반면, 음이항 모형에서는 이러한 패턴이 상대적으로 완화되었다. 이는 음이항 모형이 CNT의 변동성을 더 유연하게 반영하고 있다.

5.2 영과다 점검

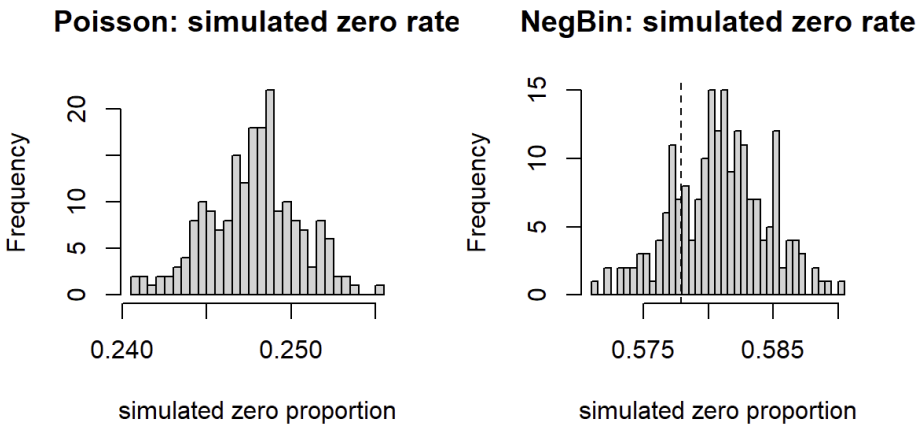
산불 자료는 비발생이 매우 많아 CNT=0이 과도하게 나타날 수 있다. 이에 따라 관측된 0비율과, 적합된 모형이 내재적으로 기대하는 P(Y=0)의 평균을 비교하여 영과다 가능성을 점검한다.

```
  obs_zero exp_zero_pois   exp_zero_nb
 0.5778956     0.2475325     0.5810965
```

관측된 CNT=0 비율과 모형이 기대하는 0 발생 확률을 비교한 결과, 관측 0 비율이 모형 기대치보다 현저히 큰 경우 영과다 가능성이 제기된다. 특히 포아송 모형에서 불일치가 크게 나타나는 경우가 많으며, 음이항 모형은 분산을 확장하여 0 비율을 일부 더 잘 설명할 수 있다. 다만 음이항에서도 관측 0 비율이 여전히 과도할 경우, 다른 모형으로의 확장이 필요할 수 있다.

5.3 예측 적합도 점검

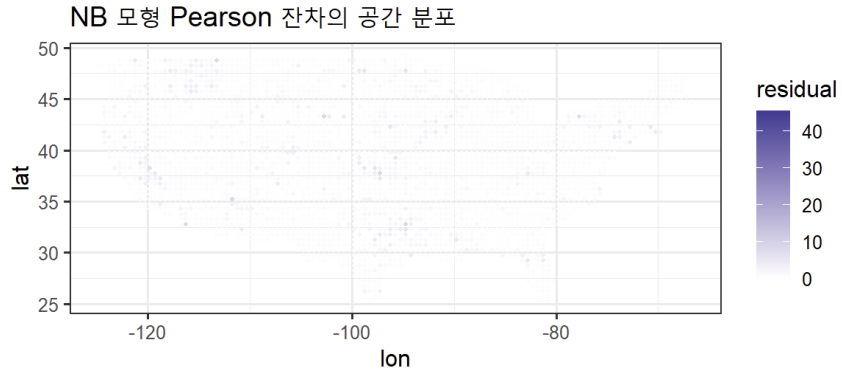
패키지 없이도 간단한 예측 점검을 위해, 적합된 모형으로부터 CNT를 반복 시뮬레이션한 뒤 관측 분포와 비교하였다.



모형 기반 시뮬레이션에서 생성된 0 비율의 분포와 실제 관측 0 비율을 비교한 결과, 관측값이 시뮬레이션 분포의 중심에서 크게 벗어나는 경우 해당 모형이 0의 과다 발생을 충분히 설명하지 못함을 의미한다. 반대로 관측 0 비율이 음이항 모형 시뮬레이션 분포 범위 내에 위치한다면, 음이항 모형이 데이터 구조를 비교적 타당하게 반영하고 있다고 해석한다.

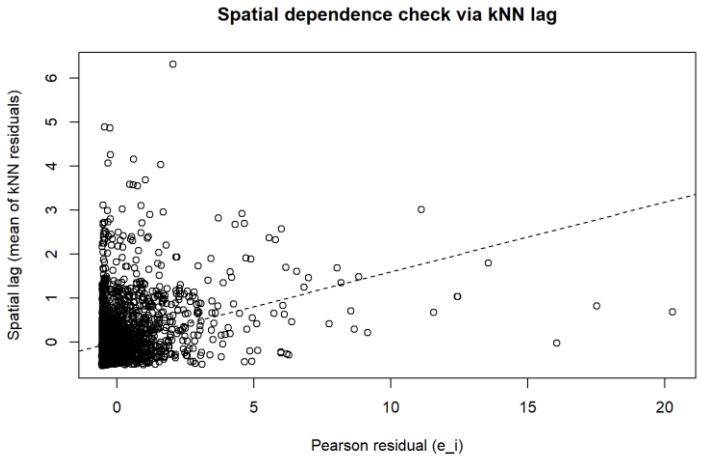
5.4 공간적 잔차 구조 점검

5.4.1 잔차의 공간 분포 시각화



음이항 모형의 피어슨 잔차를 공간상에 시각화하여 특정 구역에서 양 또는 음의 잔차가 군집하는지 확인하였다. 잔차의 공간적 군집이 관찰될 경우, 현재 모형이 공간 구조를 완전히 설명하지 못했음을 시사하며 공간 효과를 포함한 확장 모형의 필요성이 제기된다.

5.4.2 kNN 기반 공간 래그로 자기상관 점검



kNN을 이용해 각 지점 잔차의 이웃 평균을 계산하고 잔차와의 관계를 확인하였다. 잔차와 공간 래그 사이의 상관이 양으로 나타나거나 산점도에서 양의 기울기가 뚜렷할 경우, 잔차가 공간적으로 유사한 값끼리 모이는 자기상관 구조가 잔존할 가능성이 있다.

본 보고서는 위도와 경도 격자와 월 단위 산불 자료를 이용해 발생 여부, 발생 빈도, 발생 시 규모를 구분하여 단계적으로 모형화하였다. EDA 결과 CNT와 BA는 0이 많고 긴 꼬리를 가지며, 월별 계절성과 공간적 이질성이 뚜렷해 공간, 계절, 기상 변수를 함께 고려할 필요가 있다는 것을 보였다.

발생 여부는 로지스틱 회귀로 분석하고 ROC/AUC로 성능을 평가하였다. AUC가 0.5를 상회하여 모형의 발생/비발생을 무작위보다 잘 구분함을 확인했다. 그러나 산불의 우발적 요인 및 미측정 환경 요인으로 인해 완전한 예측에는 한계가 있었다. 발생 빈도(CNT)는 포아송과 음이항 모형을 비교했으며, 과산포 가능성을 반영해 음이항 모형이 더 적합한 근거를 확보하였다. 발생 시 규모는 log(BA+1) 변환 후 선형회귀를 기본으로 하되 설명변수 상관 구조를 점검하고, 패키지 제약 하에서는 단계적 변수 설정과 반복 분할 RMSE 비교로 예측 안정성을 평가하였다.

결론적으로 산불 자료처럼 불균형하고 과산포와 극단값이 공존하는 경우 반응변수 특성에 맞는 모형 선택이 핵심이며, 향후에는 영과다/공간 효과를 명시적으로 반영하는 확장 모형을 적용하면 예측력이 개선될 것이라고 보인다.