Proceedings of Joint Conference of Korean Data And Information Science Society
and The Korean Data Analysis Society,
April 28 - 29, 2006, pp. 345-352

마코프 로지스틱 회귀모형을 이용한 강수 확률예측

박정수[1]

요 약

현 기상의 시점에서 강수 확률 예측을 위해 가장 적절한 모형은 공간적 종속성과 시간적 종속성을 고려한 모형이 선택되어져야 한다. 보통 마크프 연쇄 모형과 예보인자를 이용하는 회귀모형이 모두 고려된 모형을 사용한다. 본 논문에서는 강수 형태를 세 개의 상태로 나눈 경우, 즉 맑은 경우, 흐린 경우, 비온 경우로 나누어 마코프 로지스틱 회귀모형을 세우고 강수확률을 예측할 수 있도록 하였다. 또한 서울 지역의 강수 자료를 이용하여 기존의 마코프 회귀모형과 마코프 로지스틱 회귀모형을 서로 비교하여 실제적 적용 문제를 다루었다.

1. 서론

강수량 자료와 같이 매일 측정 가능한 자료에 대한 정확한 예측을 위해서, 강수량에 영향을 주는 여러 가지 요인을 고려한 통계적 방법들이 사용되고 있다. 그 중에 강수 형태를 비가 온 경우와 오지 않은 경우로 분류하고 시간적 종속성을 고려한 마코프 연쇄 모형이 있다. Kirk and Fraedrich(1998)와 Fraedrich and Leslies(1988)의 논문들은 두 상태의 마코프 회귀모형을 이용하여 강수확률 예측에 적용하였다. 한국 기상청에서는 마코프 성질을 고려하지 않은 단순회귀모형을 사용하였다(조주영, 최준태, 1995). 본 논문에서는 강수 형태를 세 개의 상태로 나눈 경우, 즉 맑은 경우, 흐린 경우, 비온 경우로 나누어 마코프 로지스틱 회귀 모형을 세우고 강수확률을 예측할 수 있도록 하였다. 또한 서울 지역의 강수 자료를 이용하여 기존의 마코프 회귀모형과 마코프 로지스틱 회귀모형을 서로 비교하였다.

2. 로지스틱 회귀모형

X는 설명변수, Y는 1과 0의 값을 갖는 이진 반응변수라고 할 때, $P(Y=1 \mid X)$는 다음 로지스틱 회귀모형으로 표현된다고 가정한다.

$$P(Y=1 \mid X)=\frac{\exp(\beta_0+\beta x')}{1+\exp(\beta_0+\beta x')}$$

[1]500-757 광주광역시 북구 용봉동 300번지, 전남대학교 자연대 통계학과 교수.
전화: 062-530-3445, E-mail: jspark@chonnam.ac.kr

이 식의 모수추정은 최대우도추정법을 사용한다. 범주의 수를 g개로 확장한 모형을 고려하기 위해서 다음과 같이 확장된 로지스틱 모형이 필요하다(Albert and Anderson, 1984). $x' = (x_0, \cdots, x_p)$ 이고, H 는 $(H_1, \cdots, H_g)$의 집단이고, H_i는 $y = i (i = 0, \cdots, n)$ 인 집단이라고 표시하자. 그리고 편의상 $x_0 \equiv 1$이라고 놓는다.

$$P(H_s \mid x) = \exp(\beta_s' x) P(H_g \mid x) \quad (s = 1, \cdots, g-1),$$

$$P(H_g \mid x) = \frac{1}{1 + \exp(\beta_1' x) + \cdots + \exp(\beta_{g-1}' x)}$$

여기에서, $\beta_s' = (\beta_{s0}, \cdots, \beta_{sp})$ $(s = 1, \cdots, g-1)$ 이고 $\beta_g' = 0$ 이다.

이제, 범주가 세 개인 경우을 생각하여, $P(H_0 \mid x) = P(y = 0 \mid x)$ 일 때, $Y = 0$ 일 확률은

$$P(H_0 \mid x) = \frac{1}{1 + \exp(\beta_1 x) + \exp(\beta_2 x)}$$

이다. g=3인 경우, 즉 1, 2, 3 일 때, $\beta_3 = 0$ 이라 놓으면, 로그 우도함수는 다음과 같다.

$$\begin{aligned} &\log L(y_1, y_2, \cdots, y_n \mid x, \beta) \\ &= \sum_{i \in E_1} \log\left(\frac{\exp(\beta_1' x)}{1 + \exp(\beta_1' x) + \exp(\beta_2' x)}\right) + \sum_{i \in E_2} \log\left(\frac{\exp(\beta_2' x)}{1 + \exp(\beta_1' x) + \exp(\beta_2' x)}\right) \\ &+ \sum_{i \in E_3} \log\left(\frac{1}{1 + \exp(\beta_1' x) + \exp(\beta_2' x)}\right) \end{aligned}$$

여기서 β_1 는 그룹1과 그룹3에서, β_2 는 그룹2와 그룹3에서 추정한 모수이다.

반응값이 3개 이상인 경우, 우도함수의 최대화 계산이 복잡해지기 때문에, Begg and Gray(1984)은 대안으로써 단순(반응변수가 0, 1)로지스틱을 두 번 하는 것을 제안했다. 그들은 자신들의 방법이 원 우도함수를 최대화해서 구한 회귀모형과 크게 다르지 않음을 이론적 및 실증적으로 보였다. 본 논문에서도 반응변수가 3가지의 범주일 경우에는 두 번의 로지스틱 회귀분석을 이용했다. 최우추정법이 수학적으로 직접 해를 구하기 어려우므로 Newton-Raphson방법과 같은 수치적방법을 사용한다. 실제로 SAS의 Proc Logistic을 이용했다(허명회, 1989).

3. 마코프 로지스틱 회귀모형

3.1 마코프 회귀모형

1 또는 0의 값을 갖는 이진변수를 $Y(i, t)$라고 할 때, 마코프 회귀모형은 아래와 같이 표현된다.

전 단계가 Y=0일 때,

$$Y(i,t)=\alpha_0+\alpha_1 x_1(i,t)+\cdots+\alpha_p x_p(i,t)$$

이고, 여기에서 $\alpha'=(\alpha_0,\cdots,\alpha_p)$ 는 전 단계가 $Y=0$ 일 때의 회귀모형의 모수이다. 전 단계가 Y=1일 때,

$$Y(i,t)=\beta_0+\beta_1 x_1(i,t)+\cdots+\beta_p x_p(i,t)$$

이다. 여기서, $\beta'=(\beta_0,\cdots,\beta_p)$ 는 전 단계가 $Y=1$ 일때의 회귀모형의 모수이다. 그리고, 전 단계가 Y=0일 때 현 단계의 Y의 추정치를 보면

$$\widehat{Y}(i,t+1)=\widehat{\alpha_0}+\widehat{\alpha_1} x_1(i,t+1)+\cdots+\widehat{\alpha_p} x_p(i,t+1)$$

이다. Fraedrich and Leslie (1987, 1988)은 예측치가 0보다 작을 때는 $\widehat{Y}$ 를 0으로 놓았고, $\widehat{Y}$ 가 1보다 클 경우에는 1로 간주 하였다. 마코프 회귀분석 방법이 계산은 간편하지만, $\widehat{Y}$=4 일 경우에도 1, $\widehat{Y}$=1.1 일 경우에도 1로 놓기 때문에 문제가 있다.

3.2 2-상태 마코프 로지스틱 회귀모형

1 또는 0의 값을 갖는 이진 반응변수 $Y(i,t)$, $i=1,\cdots,N, t=0,1,\cdots,T$ 를 고려하자. 여기서 N은 표본의 크기를 말하고, t는 시점을 말한다. $X(i,t)=\{X_1(i,t),\cdots,X_p(i,t)\}$ 는 p-차의 공변량(covariates)벡터이다. 고정된 i에 대해서 $Y(i,t)$, $t=0,1,\cdots,T$ 는 시점 t에 따라서 변화하는 2-상태 마코프 연쇄라고 가정하고 추이확률은 다음과 같이 표시한다.

$$P\{Y(i,t)=1|Y(i,t-1)=0\}=p_{it},\ P\{Y(i,t)=0|Y(i,t-1)=1\}=q_{it}.$$

p_{it} 는 X값에 따라서 달라진다. 공변량들은 이진반응 변수와 로지스틱 함수 관계가 있다고 가정하면, 마코프 로지스틱 회귀모형은 다음과 같다(Zhang and Medoff-Cooper, 1996).

$$\log\left\{\frac{p_{it}}{1-p_{it}}\right\}=X(i,t)\,\alpha=\alpha_o+X_1(i,t)\alpha_1+\cdots+X_p(i,t)\alpha_p\ ,$$

$$\log\left\{\frac{q_{it}}{1-q_{it}}\right\}=X(i,t)\,\beta=\beta_o+X_1(i,t)\beta_1+\cdots+X_p(i,t)\beta_p\ .$$

초기의 확률을 $\pi_{i0}(k)=P\{Y(i,0)=k\}$ 으로 표기하는데, $\pi_{i0}(Y(i,0))$ 은 확률변수 $Y(i,0)$ 이 갖는 값에 대한 초기확률이다. 추이확률을 $\pi(k,l)=P\{Y(i,t)=l|Y(i,t-1)=k\}$ 로 표기하자. 또

한 $\pi\,(Y(i,t-1),Y(i,t))$ 은 확률변수 $Y(i,t-1)$ 가 갖는 값에서 확률변수 $Y(i,t)$가 갖는 값으로의 추이확률을 표현한다. 그러면 고정된 i 에서 $\{Y(i,t),t=0,1,\cdots,N\}$ 에 대한 로그 우도함수는 다음과 같다.

$$L_i(Y(i,t),t=0,1,\cdots,T)=\log\pi_{i0}(Y(i,0))+\sum_{t=1}^{T}\log\pi\,(Y(i,t-1),Y(i,t)) \tag{3.1}$$

우변의 두 번째 항을 정리하면,

$$\begin{aligned}\sum_{t=1}^{T}\log\pi_{it}(Y(i,t-1),Y(i,t)) &= \sum_{t=1}^{T}\log\pi_{it}(0,Y(i,t))+\sum_{t=1}^{T}\log\pi_{it}(Y(i,t-1),0)\\ &=L_i(\alpha)+L_i(\beta)\end{aligned} \tag{3.2}$$

이므로 식(3.1)의 로그 우도함수는 $L_i(Y(i,t),t=0,1,\cdots,T)=\log\pi_{i0}(Y(i,0))+L_i^r(\alpha)+L_i^s(\beta)$ 이 된다. 식(3.2)에서 $Y(i,t-1)$는 0 또는 1의 값을 갖는데, $Y(i,t-1)=0$ 의 조건에서 위의 식은 다음과 같다.

$$\log\pi_{it}(0,Y(i,t))=a_{01}\log\frac{p}{1-p}+(a_{00}+a_{01})\log(1-p) \tag{3.3}$$

여기서 $a_{kl}(i,t)$ 은

$$a_{kl}(i,t)=\begin{cases}1, & Y(i,t-1)=k \quad \text{and} \quad Y(i,t)=l,\\ 0, & \text{그 외 경우}\end{cases}$$

이다. 식(3.3)을 식(3.2)의 $L_i(\alpha),L_i(\beta)$ 에 각각 대입하여 주면 아래와 같다. $Y_{(i,t-1)}=0$ 일 때,

$$L_i(\alpha)=\sum_t\left[a_{01}(i,t)\log\left\{\frac{p_{it}}{1-p_{it}}\right\}+\{a_{00}(i,t)+a_{01}(i,t)\}\log\{1-p_{it}\}\right]$$

$Y_{(i,t-1)}=1$ 일 때,

$$L_i(\beta)=\sum_t\left[a_{10}(i,t)\log\left\{\frac{q_{it}}{1-q_{it}}\right\}+\{a_{11}(i,t)+a_{10}(i,t)\}\log\{1-q_{it}\}\right]$$

윗 식들에 $\log\frac{p_{it}}{1-p_{it}}=X(i,t)\,\boldsymbol{\alpha}$ 과 $\log\frac{p_{it}}{1-p_{it}}=X(i,t)\,\boldsymbol{\beta}$ 를 각각 대입하면,

$$L_i(\alpha)=\sum_t[a_{01}(i,t)X(i,t)\alpha-\{a_{00}(i,t)+a_{01}(i,t)\}\log\{1+e^{X(i,t)\alpha}\}]$$

$$L_i(\beta)=\sum_t[a_{10}(i,t)X(i,t)\beta-\{a_{11}(i,t)+a_{10}(i,t)\}\log\{1+e^{X(i,t)\beta}\}]$$

로 유도되어, 결국 상수항을 제외한 로그 우도함수는 다음과 같다.

$$l(\alpha,\beta)=\sum_{i=1}^{N}L_i(\alpha)+\sum_{i=1}^{N}L_i(\beta)=l(\alpha)+l(\beta).$$

앞의 유도 과정으로부터 모수 α와 β는 독립적으로 취급되는 $l(\alpha)$과 $l(\beta)$ 로부터 각각 따로 추정할 수 있다. 모수 α와 β를 추정하기 위하여 최우추정법을 이용한다.

3.3 3-상태 마코프 로지스틱 회귀모형

3.2절 내용을 0, 1, 2 즉, 3-상태로 확장하여 생각해보자. 고정된 i에 대해서 $Y(i,t)$, $t=0,1,\cdots,T$ 는 시점 t에 따라서 변화하는 3-상태 마코프 연쇄라고 가정하고, 추이 확률을 다음과 같이 표시하자.

$$P\{Y(i,t)=0|Y(i,t-1)=1\}=p_{it},\quad P\{Y(i,t)=0|Y(i,t-1)=2\}=q_{it},$$
$$P\{Y(i,t)=1|Y(i,t-1)=0\}=r_{it},\quad P\{Y(i,t)=1|Y(i,t-1)=2\}=s_{it},$$
$$P\{Y(i,t)=2|Y(i,t-1)=0\}=t_{it},\quad P\{Y(i,t)=2|Y(i,t-1)=1\}=u_{it}$$

그러면 마코프 로지스틱 회귀모형은 아래와 같이 6개가 된다(일부 생략).

$$\log\left\{\frac{p_{it}}{1-p_{it}}\right\}=X(i,t)\,\boldsymbol{\alpha}=\alpha_o+X_1(i,t)\alpha_1+\cdots+X_p(i,t)\alpha_p$$

$$\log\left\{\frac{q_{it}}{1-q_{it}}\right\}=X(i,t)\,\boldsymbol{\beta}=\beta_o+X_1(i,t)\beta_1+\cdots+X_p(i,t)\beta_p$$

$$\log\left\{\frac{r_{it}}{1-r_{it}}\right\}=X(i,t)\,\boldsymbol{\gamma}=\gamma_o+X_1(i,t)\gamma_1+\cdots+X_p(i,t)\gamma_p$$

2-상태 경우와 같은 방법으로 로그우도함수를 유도하면

$$L_i(Y(i,t),t=0,1,\cdots,T)=\log\pi_{i0}(Y(i,0))+\sum_{t=1}^{T}\log\pi_{it}(Y(i,t-1),Y(i,t))$$
$$=\log\pi_{i0}(Y(i,0))+L_i(\alpha)+L_i(\beta)+L_i(\gamma)+L_i(\varepsilon)+L_i(\delta)+L_i(\eta)\qquad(3.6)$$

이다. $L_i(\alpha),L_i(\beta),L_i(\gamma),L_i(\varepsilon),L_i(\delta),L_i(\eta)$ 은 3.2절에서와 같은 과정을 거쳐 다시 아래와 같이 유도된다 (일부 생략).

$$L_i(\gamma)=\sum_t\left[a_{01}(i,t)X(i,t)\,\boldsymbol{\gamma}-\{a_{00}(i,t)+a_{01}(i,t)+a_{02}(i,t)\}\log\{1+e^{X(i,t)\gamma}\}\right]$$

$$L_i(\delta)=\sum_t\left[a_{02}(i,t)X(i,t)\,\boldsymbol{\delta}-\{a_{00}(i,t)+a_{02}(i,t)+a_{01}(i,t)\}\log\{1+e^{X(i,t)\delta}\}\right]$$

$$L_i(\alpha)=\sum_t[a_{10}(i,t)X(i,t)\alpha-\{a_{11}(i,t)+a_{10}(i,t)+a_{12}(i,t)\}\log\{1+e^{X(i,t)\alpha}\}]$$

결국, 상수항을 제외한 전체의 로그 우도함수는 다음과 같다.

$$\begin{aligned} l(\alpha,\beta,\gamma,\varepsilon,\delta,\eta) &= \sum_{i=1}^{N}L_i(\alpha)+\sum_{i=1}^{N}L_i(\beta)+\sum_{i=1}^{N}L_i(\gamma)+\sum_{i=1}^{N}L_i(\varepsilon)+\sum_{i=1}^{N}L_i(\delta)+\sum_{i=1}^{N}L_i(\eta) \\ &= l_i(\alpha)+l_i(\beta)+l_i(\gamma)+l_i(\varepsilon)+l_i(\delta)+l_i(\eta) \end{aligned}$$

모수 $\alpha,\beta,\gamma,\varepsilon,\delta,\eta$ 는 $l_i(\alpha), l_i(\beta), l_i(\gamma), l_i(\varepsilon), l_i(\delta), l_i(\eta)$ 로부터 각각 독립적으로 취급되어 따로 추정할 수 있다. 실제로 최우추정치는 계산상의 복잡함 때문에 Begg & Gray의 방법을 택한다. 전 단계가 0일 경우, $Y=0$ 일 확률은

$$P(Y_t=0 \mid x, Y_{t-1}=0)=\frac{1}{1+\exp(\gamma x)+\exp(\delta x)}$$

이고, $Y=1$ 일 확률은

$$P(Y_t=1 \mid x, Y_{t-1}=0)=\frac{\exp(\gamma x)}{1+\exp(\gamma x)+\exp(\delta x)}$$

이다. 전 단계가 1일 경우와 2일 경우의 확률도 같은 방법으로 구할 수 있다.

4장. 실제자료분석

서울지역 강수자료에 대해 기존의 마코프 회귀모형과 마코프 로지스틱 회귀모형을 비교하여 실제적 적용 문제를 다루었다. 자료의 기간은 1993년 3월 1일부터 2004년 6월 30일(아침 9시, 저녁 9시)까지이고, 반응변수는 RAIN-맑음(0), 흐림(1), 비(2)이다. 독립변수(10개 변수)는 SP(표면기압), VR850(상대와도)-850hPa, VV850(풍속)-850hPa, S850(바람의 남북성분)-850hPa, DW850(Dewpoint Depression)-850hPa, SHA700(비습이류)-700hPa, cloud1, cloud2, cloud3, cloud4(운량)-맑음(0-2), 구름조금(3-5), 구름많음(6-7), 흐림(8-10)이다. 자료를 4계절로 나누었고, 1993년~2001년 자료를 이용하여 모수를 추정했고, 2002년~2004년 자료에 대해 예측값과 실제값을 비교했다.

마코프 회귀분석을 할 때 $\widehat{Y}$ 즉, 예측값이 1을 초과하거나 음수를 가지는 경우가 발생하는데, 실제 자료 분석을 할 때에는 1을 초과할 경우 1로, 음수이면 0으로 처리하였다. 예측치와 실제값이 일치하는지 알아보기 위해 EMSEP (Empirical MSE of Prediction) 즉, $\sum_{i=1}^{n}(\widehat{y_i}-y_i)^2/n$ 을 계산하였다. 마코프 로지스틱 회귀분석에서는 $\frac{1}{n}\sum_{i=1}^{n}[I(y=i)-\hat{p}(Y_i=y_i)]^2$ 를 계산한다. SAS를 이용하여 자료분석을 한 결과, 아래와 같이 EMSEP를 얻었다. 표 <4-1>에서 나타나는 봄인 경우에 결과를 보면 전 단계가 0인 경우, 마코프 로지스틱 회귀분석이 더 좋은 결과를 준다. 여름, 가을, 겨울

인 경우 모두에서 마코프 로지스틱 회귀분석이 더 좋았다. 실제값과 판별값을 비교하여 표<4-2>와 같은 결과를 얻었다. 본 논문에서는 전단계가 0인 경우만 제시하였다.

<표 4-1. 서울지역의 봄인 경우의 EMSEP>

전단계 / 회귀모형	0(맑음)	1(흐림)	2(비)
마코프 로지스틱 회귀모형	0.10576	0.11639	0.20666
마코프 회귀모형	0.13911	0.23366	0.51382

5. 결론

현 기상의 시점에서 강수 확률 예측을 위해 가장 적절한 모형은 공간적 종속성과 시간적 종속성을 고려한 모형이 선택되어져야 한다. 이를 위해 강수형태를 두 개의 상태로 나눈 경우의 마코프 로지스틱 회귀모형을 제안하였다. 나아가 3-상태(0은 맑음 상태, 1은 흐린 상태, 2는 비가 온 상태)로 확장하여 모형을 세우고, 확률 예측 식을 유도하였다. 서울 지역의 강수 자료를 이용하여 기존의 마코프 회귀모형과 마코프 로지스틱 회귀모형을 비교하여, 다음과 같은 결론을 얻었다. 첫째, 마코프 회귀모형에서 $\widehat{Y}$ 의 값이 1을 초과하거나 음수를 가지는 경우가 생기는데, 실제로는 1을 초과하면 1로, 음수이면 0으로 처리하여 비올 확률을 예보하고 있는데, 이러한 어색한 점을 개선하기 위하여, 또 이론적으로 더 바람직하므로, 3단계 마코프 로지스틱 회귀모형을 사용하여 강수 확률을 예보하는 것이 더 좋다. 둘째, 서울지역을 4계절로 구분하여 분석을 하였는데, 마코프 회귀모형과의 비교 결과 마코프 로지스틱 회귀모형이 예측 면에서 더 좋은 결과를 주었다.

<표 4-2. 서울지역의 봄인 경우의 예측값과 판별값의 비교>

(1) 전단계가 0인 경우 (맑은 경우)			
판별값 / 실제값	0(맑음) (%)	1(흐림) (%)	2(비) (%)
0(맑음)	114 (93.44)	8 (6.56)	0 (0.00)
1(흐림)	6 (8.00)	57 (76.00)	12 (16.00)
2(비)	0 (0.00)	1 (50.00)	1 (50.00)

참고문헌

[1] 조주영, 최준태 (1995). *통계적 방법에 의한 강수 확률 예보*, 연구보고서 95-4, 기상청 예보국 수치예보과.

[2] 허명회 (1989). *SAS 범주형 자료분석*, 자유아카데미, 서울.

[3] Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models, *Biometrika*, 71, 1, 1-10.

[4] Begg, C. B. and Gray, R. (1984). Calculation of polychotomous logisitc regression parameters using individualized regressions, *Biometrika,* 71, 11-18.

[5] Fraedrich, K. and Kirk, E. (1998). Probability of Precipitation: Short-Term Forecasting and Verification, *Contribution to Atmospheric Physics,* 71, 2, 263-270

[6] Fraedrich, K. and Leslie, L. M. (1988). Real-time short-term forecasting of precipitation at an Australian tropical station, *Weather and Forecasting,* 3, 104-114.

[7] Zhang P. and Medoff-Cooper, B. (1996). A Markov Regression Model for Nutritive Sucking Data, *Biometrics,* 52, 112-124.