The Korean Journal of Applied Statistics (2013) DOI: http://dx.doi.org/10.5351/KJAS.2013.26.6.923
26(6), 923–932

Prediction of the Number of Food Poisoning Occurrences by Microbes

In-Kwon Yeo[a,1]

[a]Department of Statistics, Sookmyung Women's University

(Received August 19, 2013; Revised October 15, 2013; Accepted October 28, 2013)

Abstract

This paper proposes a method to predict the number of foodborne disease outbreaks by microbes. The weekly data of food poisoning occurrences by microbes in Korea contain many zero-valued observations and have dependency between outbreaks. In order to model both phenomena, the number of food poisonings is predicted by an autoregressive model and the probabilities of food poisoning occurrences by microbes (given the total of food poisonings) are estimated by the baseline category logit model. The predicted number of foodborne disease outbreaks by a microbe is obtained by multiplying the predicted number of foodborne disease outbreaks and the estimated probability of the food poisoning by the corresponding microbe. The mean squared error and the mean absolute value error are evaluated to compare the performances of the proposed method and the zero-inflated model.

Keywords: Autoregressive model, baseline category logit model, zero-inflated negative binomial regression.

1. 서론

식품의약품안전처 조사에 의하면 우리나라에서 발생하는 식중독 중 60% 정도는 세균이나 바이러스로 인한 것으로 알려져 있으며 주요 원인균으로는 노로바이러스, 병원성 대장균, 살모넬라, 장염비브리오균, 황색포도상구균 등이 있다. 이들 균의 활성도는 온도나 습도와 같은 기후변수에 영향을 받는 것으로 알려져 있는데 이와 관련된 연구가 Bentham과 Langfod (1995), Patrick 등 (2004), Magny 등 (2008)과 Fleuly 등 (2006)에 의해 수행되었으며 우리나라의 경우 Choi 등 (2008), Jung 등 (2012), Yeo (2012)에 의해 연구되었다.

Jung 등 (2012)의 분석에 의하면 원인균별 식중독 발생건수는 각기 다른 기후변수에 영향을 받고 있으며 대부분의 식중독균은 기온이 올라갈수록 활성도가 높아지는데 반하여 노로바이러스에 의한 식중독은 기온이 떨어지면 오히려 증가하는 추세를 가지는 것으로 나타났다. 이것은 원인균별로 기후변수의 영향력이 서로 상쇄되어 전체 식중독 건수를 예측하는데 기후변수의 설명력을 떨어뜨리는 요인으로 작용할 수 있다는 것을 의미한다. Yeo (2012)의 비교분석에 의하면 기후변수를 설명변수로 설정한 포아송회귀모형으로 예측했을 때보다 이전 식중독 발생건수를 설명변수로 가지는 자기회귀모형에 의한 예측력이

This research was supported by the research fund of Sookmyung Women's University (2012).
[1]Professor, Department of Statistics, Sookmyung Women's University, Chongpa-dong 2ga, Yongsan-gu, Seoul 140-742, Korea. E-mail: inkwon@sm.ac.kr

훨씬 우수한 것으로 나타났다. 실제로 우리나라의 경우 단체급식이나 식당에서의 식중독 발생비율이 상대적으로 높다. 이는 위생 및 보건시설의 인프라 구축이나 행정적 조치 등에 영향을 받을 수 있으며 식중독이 발생했을 때 확산을 막을 조치를 취하지 않으면 전염성이 높은 식중독 경우 지속적으로 증가하는 경향이 있다.

식중독 예방 및 처치는 원인균에 따라 다르게 적용해야 하기 때문에 원인균별로 식중독 발생여부나 건수에 대한 예측은 매우 중요하다. 원인균별 식중독 발생건수에 대한 Jung 등 (2012)의 연구에서는 각 원인균별로 주요 변수를 파악하는데 도움이 되었지만 모형에 시계열적 요소를 포함하지 않고 있어 식중독 발생건수를 예측 하는데 한계가 있었다.

이 논문에서는 기후변수를 설명변수로 설정한 다범주 로짓모형으로 원인균별로 식중독 발생 비율을 추정하고 자기회귀모형을 이용하여 예측된 전체 식중독 발생건수를 곱해 원인균별로 식중독 발생 건수를 예측한다. 이에 대한 타당성을 확인해 보기 위해 Jung 등 (2012)에서 사용했던 영과잉 음이항회귀모형과 비교한다. 이를 위해 2003년 5월부터 2010년 4월까지 전국에서 신고 된 주별 전체 식중독 발생건수와 원인균별 식중독 발생건수, 해당 주의 전국 60개 기상관측소에서 관측된 평균기온, 평균습도, 평균일조량 자료를 이용하여 모형을 추정하고 예측값의 평균제곱오차와 평균절대값오차를 통해 비교하고자 한다.

2. 원인균별 식중독 예측 방법론

먼저 기상자료가 식중독발생에 어떻게 영향을 주는지 알아보기 위해 기상자료의 추이와 식중독발생건수에 대한 시계열 그림을 그려보았다. Figure 2.1에서 평균기온, 습도, 일조량은 부분적으로 차이는 있지만 계절주기를 가지는 것을 볼 수 있다. 이에 반해 식중독 발생건수는 기상자료와는 상당히 다른 패턴을 보이고 있다. 이것은 일반화선형모형과 같은 모형으로 기상자료와 식중독 발생건수의 관계를 유도하면 좋은 결과를 얻을 수 없다는 것을 보여준다. Yeo (2012)는 교차상관관계를 통해 기후변수가 식중독 발생건수 간에 시차를 두고 영향을 주는 것을 확인했으나 기후변수와 이전 식중독 발생건수가 설명변수인 포아송회귀모형은 분석모형으로 적절하지 않은 것을 실증분석을 통해 보였다.

식중독 발생은 기후요인에 영향을 받지만 전염성이 있고 식재료관리와 조리법에 영향을 많이 받기 때문에 어떻게 대처하는가가 발생건수에 더 중요한 요인이 될 수 있다. 또한 Jung 등 (2012)에서 언급한 것처럼 원인균별로 기후변수에 대한 활성도가 다르고 어떤 경우에는 역으로 반응하는 경우가 있어 전체 식중독 발생건수를 예측하는데 어려움이 발생한다. 원인균별로 따로 분석할 경우 0의 값을 가지는 관측값이 상대적으로 많아져 영과잉모형을 사용해야 한다. 영과잉모형을 사용했던 Jung 등 (2012)의 분석결과를 보면 원인균별 식중독 발생건수의 예측에서 여전히 만족할 만한 결과를 보이고 있지 않다.

Yeo (2012)의 비교분석에서는 포아송회귀모형보다 자기회귀모형이 전체 식중독발생건수를 예측하는데 더 좋은 예측력을 가지는 것으로 나타났다. 이 논문에서는 자기회귀모형으로 전체 식중독발생건수를 예측하고 다범주로짓모형을 통해 해당시점에서 각 원인균별로 식중독이 발생할 확률을 추정하여 예측된 전체 식중독발생건수에 각각의 확률을 곱해 원인균별 식중독 발생건수를 계산하는 방법을 제안한다. 이론적 내용과 자세한 예측 과정은 다음과 같다.

식중독 원인을 p개로 분류할 수 있고 시점 t에서 j-번째 원인에 의해 발생한 식중독 건수를 Y_{jt}라고 하자. 이 시점에서 발생한 전체 식중독 건수를 N_t라고 하고 각각의 식중독 발생은 하나의 원인에 의해서만 발생한다고 가정하면 $N_t = \sum_{i=1}^{p} Y_{it}$가 된다. 각 원인균에 의한 식중독 발생건수는 서로 독립이고 평균이 μ_{jt}인 포아송 분포를 따른다고 가정한다. 그러면 N_t는 평균이 $\mu_t = \sum_{i=1}^{p} \mu_{it}$인 포아송 분포를 따르게 된다.

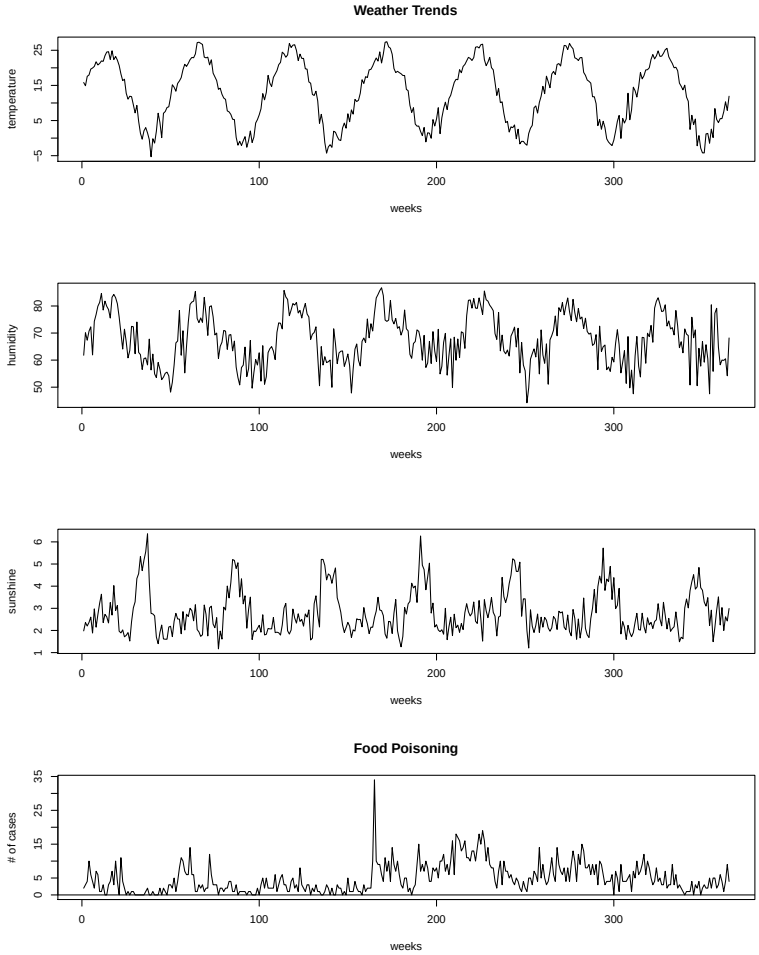

Figure 2.1. Time series plots of temperature, humidity, sunshine and the number of food poisoning

시점 t에서 전체 식중독 건수가 n_t로 주어졌을 때, 원인균별 식중독 발생건수 $\boldsymbol{Y}_t = (Y_{1t}, \ldots, Y_{pt})^T$는 확률이 $\boldsymbol{\pi}_t = (\pi_{1t}, \ldots, \pi_{pt})^T$인 다항분포를 따른다. 여기서 $\pi_{jt} = \mu_{jt}/\mu_t$의 관계를 가지며 $\boldsymbol{\pi}_t$는 k개의 기후자료 $\boldsymbol{x}_t = (x_{1t}, \ldots, x_{kt})^T$에 영향을 받는다고 가정한다. 만약 p-번째 원인균의 확률을 기준으로 정하면 기준범주 로짓(baseline category logit)모형에 의한 관계식은 다음과 같이 정의된다.

$$\log\left(\frac{\pi_{jt}}{\pi_{pt}}\right) = \beta_{j0} + \beta_{j1}x_{1t} + \cdots + \beta_{jk}x_{kt} = \boldsymbol{\beta}_j^T\boldsymbol{x}_t, \quad j = 1, \ldots, p-1.$$

이 기준범주 로짓의 관계식을 이용하면 t 시점에서 j-번째 원인균에 의한 식중독 발생 확률은 다음과 같이 유도할 수 있다.

$$\pi_{jt} = \frac{e^{\boldsymbol{\beta}_j^T\boldsymbol{x}_t}}{1 + \sum_{i=1}^{p-1} e^{\boldsymbol{\beta}_i^T\boldsymbol{x}_t}}, \quad j = 1, \ldots, p, \tag{2.1}$$

여기서 p-번째 원인에 의한 발생확률을 계산할 때에는 $\boldsymbol{\beta}_p = \mathbf{0}$으로 대체한다.

모수 $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_{p-1}^T)$는 최대가능도법을 이용하여 추정할 수 있다. 관측값 $\mathbf{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)^T$와 $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$가 주어지고 각각의 $\boldsymbol{y}_t$들이 독립이라는 가정 하에서 $\boldsymbol{\beta}$의 로그가능도함수는 다음과 같다.

$$l(\boldsymbol{\beta}; \mathbf{Y}, \mathbf{X}) = \sum_{t=1}^{n} \sum_{j=1}^{p} y_{jt} \left\{ \boldsymbol{\beta}_j^T \boldsymbol{x}_t - \log \left(\sum_{k=1}^{p} e^{\boldsymbol{\beta}_k^T \boldsymbol{x}_t} \right) \right\}.$$

이를 최고로 만드는 $\boldsymbol{\beta}$를 수치해석학적으로 구할 수 있다. $\hat{\boldsymbol{\beta}}$를 최대가능도추정값이라고 하면 이를 식 (2.1)에 대입하여 구한 추정확률을 $\hat{\pi}_{jt}$라고 표시한다.

다음 단계로 자기회귀이동평균모형을 적용하여 전체 식중독 발생건수를 추정한다. 앞에서 식중독 발생건수는 포아송 분포를 가정했으며 포아송 분포의 특징 중 하나가 평균과 분산이 같다는 것이다. 이에 반해 자기회귀이동평균모형에서의 추론은 동일한 분산을 가지는 정규분포를 가정 하에서 이루어진다. 포아송분포와 자기회귀이동평균 간의 갭을 해결하기 위해 포아송 분포에서 분산상수화 변환인 제곱근 변환을 반응변수에 적용하여 $Z_t = \sqrt{Y_t}$로 모형을 적합하는 방법도 고려하였다.

$$Y_t = \delta + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{y-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q},$$
$$Z_t = \delta + \phi_1 Z_{t-1} + \cdots + \phi_p Z_{y-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}.$$

위 모형에서의 모수 추정은 조건부 최소제곱법, 비조건부 최소제곱법, 최대가능도법 등 다양한 방법이 있으며 어떤 방법을 사용해도 무방하다. 통상적인 방법으로 추정된 모형으로부터 예측값 $\hat{Y}_t$ 또는 $\hat{Z}_t$를 구하는데 예측값이 음수가 나올 수 있기 때문에 이런 경우를 대비해 $\hat{Y}_t = \max(0, \hat{Y}_t)$와 $\hat{Z}_t = \max(0, \hat{Z}_t)$로 계산하는 단계를 추가한다. Z_t는 Y_t의 제곱근으로 식중독 발생건수를 예측하기 위해서는 $\hat{Z}_t^2$을 이용해야 한다. 그러나 비선형 함수 g에 대해 $E(Y) \neq g^{-1}(E(g(Y)))$이므로 $\hat{Z}_t^2$로 식중독 발생건수를 예측하면 편향이 생길 수 있다. 이러한 문제는 Taylor 전개, 잭나이프, 붓스트랩 등의 방법을 이용하여 해결할 수 있다. 이 논문에서는 편의상 Taylor 전개를 통해 편향을 추정하고 추정된 편향을 제거하는 방법을 사용한다. 임의의 함수 h가 μ에서 2차 미분가능하고 연속이면,

$$E(h(Z)) \approx h(\mu) + \frac{\mathrm{Var}(Z) h''(\mu)}{2}$$

가 성립한다. 제안 방법의 경우 $h(\mu) = \mu^2$이므로 변환한 자료분석 결과를 이용한 원래 척도의 예측값은 다음과 같이 구한다.

$$\hat{Y}_t = \hat{Z}_t^2 + \hat{\sigma}_t^2,$$

여기서 $\hat{\sigma}_t^2$는 Z_t의 분산 추정값을 의미한다. 결론적으로 위의 두 단계에서 구한 $\hat{Y}_t$와 $\hat{\pi}_{jt}$를 곱하면 j-번째 원인균에 의한 t시점에서의 식중독발생건수에 대한 예측값을 구할 수 있다.

이에 반해 Lambert (1992)의 영과잉모형은 개별 원인균별로 식중독 발생건수를 예측하는 것으로 Y_{jt}의 확률질량함수는 다음과 같이 쓸 수 있다.

$$P(Y_{jt} = y) = f(y; \phi_{jt}, \mu_{jt}) = \begin{cases} \phi_{jt} + (1 - \phi_{jt}) f^*(y; \mu_{jt}), & y = 0, \\ (1 - \phi_{jt}) f^*(y; \mu_{jt}), & y = 1, 2, \ldots, \end{cases}$$

여기서 $f^*(y; \mu)$는 평균이 μ인 음이항분포의 확률질량함수를 의미하고 ϕ_{jt}는 이 분포에서 초과된 0 관측값의 발생확률을 설명한다. $f^*(y; \mu)$로 포아송분포를 가정할 수 있으나 음이항분포가 보다 과산

Table 3.1. Estimates and Standard Errors in the Baseline Category Logit Model

원인균	절편		기온		습도	
	추정값	SE	추정값	SE	추정값	SE
노로바이러스	−0.804	0.650	−0.086	0.011	0.012	0.011
대장균	−2.025	0.702	0.067	0.015	−0.006	0.012
살모넬라균	−0.607	0.726	0.074	0.017	−0.032	0.013
비브리오균	−7.221	1.081	0.164	0.027	0.031	0.017
포도상구균	−2.033	0.821	0.011	0.016	0.002	0.014

포를 잘 설명할 수 있으며 추가 모수의 극한으로 포아송분포를 포함하고 있기에 이 논문에서는 음이항분포를 이용한다. 음이항분포의 경우, 임의의 $\gamma > 0$에 대해, 일반화선형모형에서 재모수화된 (re-parameterized) 확률질량함수는 다음과 같이 주어진다.

$$f^*(y;\mu) = \frac{\Gamma(y+1/\gamma)}{y!\Gamma(1/\gamma)}\left(\frac{1}{1+\gamma y}\right)^{\frac{1}{\gamma}}\left(\frac{\gamma\mu}{1+\gamma\mu}\right)^{y}.$$

모수 ϕ_{jt}와 μ_{jt}는 설명변수 $\boldsymbol{x}_{jt}$에 영향을 받으며 임의의 연결함수 h_1과 h_2에 대해 아래의 관계를 가진다고 가정한다.

$$h_1(\phi_{jt}) = \alpha_{j0} + \alpha_{j1}x_{1t} + \cdots + \alpha_k x_{kt} = \boldsymbol{\alpha}^T\boldsymbol{x}_t,$$
$$h_2(\mu_{jt}) = \beta_{j0} + \beta_{j1}x_{1t} + \cdots + \beta_k x_{kt} = \boldsymbol{\beta}^T\boldsymbol{x}_t.$$

일반적으로 연결함수 $h_1(\cdot)$는 로짓함수를, $h_2(\cdot)$로 로그연결함수를 사용하며 과산포 해결을 위한 추가적인 모수 γ를 추정해 예측값을 구할 수 있다. 추정방법의 자세한 내용은 Miller (2007) 또는 SAS (2008)를 참조하면 된다.

3. 식중독 예측비교

이 논문에서는 원인균을 노로바이러스, 병원성 대장균, 살모넬라균, 장염비브리오균, 황색포도상구균, 기타로 구분하였으며 2003년 5월부터 2010년 4월까지 총 365주 동안 전국에서 신고 된 원인균별 식중독발생 건수와 해당 주의 전국 평균기온과 평균습도를 이용하였다. 초기분석에는 일사량도 포함시켰으나 대부분의 경우 유의하지 않는 것으로 나와 분석에서는 제외하였다. 위 기간동안 보고된 식중독 발생건수는 총 1784건이며 원인균별로 보면 노로바이러스 287, 대장균 218, 살모넬라균 167, 비브리오균 148, 포도상구균 144, 기타 820건으로 살모넬라균, 비브리오균, 포도상구균에 의한 식중독이 상대적으로 적은 것을 볼 수 있다. 또한 365주 중 한 주 동안 식중독이 발생하지 않은 주는 39주였고 원인균별로 노로바이러스 230, 대장균 245, 살모넬라균 247, 비브리오균 287, 포도상구균 256, 기타 96주가 0 관측값의 빈도를 가지는 것으로 조사되었다.

분석에서는 기준범주 로짓모형에서의 기준으로 기타에 의한 식중독을 설정하였다. 자기상관함수와 부분자기상관함수를 이용하여 자기회귀이동평균모형의 차수를 유도하였는데 두 모형 모두 자기회귀의 차수만 4까지 유효한 것으로 나타났으며 유의한 계절성을 발견하지 못했다. 또한 AR(4)에 대한 모형 진단에서 심각한 문제가 없는 것으로 나타났다.

먼저 다범주 로짓모형에 대한 모수 추정값을 Table 3.1에 정리하였다. 이 표에 의하면 노로바이러스의 경우 기온에 해당되는 모수가 음수로 기타를 기준으로 했을 때와 비교해 기온이 높아지면 발생확률이 감소하는 경향을 보이고 있으며 다른 균의 경우 반대로 발생 가능성은 높아지는 경향을 가지고 있다. 살모넬라균에 의한 발생확률도 기타에 비해 습도가 높아지면 감소하는 경향을 보이고 있다.

Table 3.2. Estimates and Standard Errors in AR(4) Models

모수	Y_t		Z_t	
	추정값	SE	추정값	SE
δ	0.923	0.827	0.328	0.216
ϕ_1	0.353	0.052	0.323	0.052
ϕ_2	0.193	0.055	0.214	0.055
ϕ_3	0.151	0.055	0.166	0.054
ϕ_4	0.111	0.053	0.127	0.053

Table 3.3. Estimates and Standard Errors in Zero-Inflated Negative Binomial Regressions

모수		노로바이러스		대장균		살모넬라균	
		추정값	SE	추정값	SE	추정값	SE
$\boldsymbol{\beta}$	절편	−1.800	0.925	−0.782	0.985	0.625	0.856
	기온	−0.026	0.015	0.066	0.028	0.032	0.027
	습도	0.031	0.015	−0.007	0.017	−0.022	0.015
$\boldsymbol{\alpha}$	절편	−16.934	6.755	1.222	4.228	−0.049	3.647
	기온	0.278	0.158	−0.107	0.078	−0.315	0.110
	습도	0.140	0.079	−0.014	0.076	0.040	0.062
γ		1.386	0.290	0.635	0.664	0.059	0.178

모수		비브리오균		포도상구균		기타	
		추정값	SE	추정값	SE	추정값	SE
$\boldsymbol{\beta}$	절편	−6.254	1.310	0.789	0.468	0.789	0.468
	기온	0.117	0.036	0.031	0.009	0.031	0.009
	습도	0.046	0.022	−0.004	0.008	−0.004	0.008
$\boldsymbol{\alpha}$	절편	91.786	0.116	−0.909	2.304	−0.909	2.304
	기온	−138.573	1.168	−0.031	0.044	−0.031	0.044
	습도	19.512	7.742	−0.013	0.039	−0.013	0.039
γ		0.997	0.302	0.340	0.101	0.340	0.101

Table 3.2에 Y_t와 Z_t에 대한 AR(4) 모형의 모수 추정값과 표준오차가 제시되었다. 식중독 발생건수는 바로 전주의 식중독 건수에 영향을 가장 많이 받고 시차가 멀어질수록 영향력은 약해지는 것을 볼 수 있다. $\hat{Z}_t$를 제곱하여 식중독 발생건수를 예측하는 과정에서 편향을 줄여주기 위한 Z_t의 분산 추정값은 $\hat{\sigma}^2 = 0.556$인 것으로 분석되었다.

각 원인균별 식중독 발생 건수 예측을 위한 영과잉 음이항회귀모형의 모수 추정값과 표준오차가 Table 3.3에 제시되었다. 병원성 대장균을 제외한 나머지 원인균에 의한 식중독 발생건수는 음이항분포으로 분석하는 것이 적절한 것으로 나타났다. 대부분 기온의 유의도가 높았으며 앞의 분석에서도 언급한 것과 같이 기온이 높아지면 노로바이러스에 의한 식중독이 발생하지 않을 가능성이 높아지고 건수 또한 줄어드는 형태를 가지는 반면 나머지 원인균에 의한 식중독은 높아지거나 늘어나는 추세를 보이고 있다.

영과잉 모형과 제안 모형의 비교를 위해 추정식에서 유도된 적합값과 실제 관측값의 잔차를 비교하였다. 식중독 발생건수의 경우 포아송분포 또는 음이항분포를 기초로 분석이 이루어지기 때문에 일반적인 잔차보다는 다음과 같은 피어슨 잔차나 Anscombe 잔차를 사용할 수 있다.

$$r_P = \frac{y_t - \hat{y}_t}{\hat{y}_t}, \qquad r_A = \frac{3\left(y_t^{\frac{2}{3}} - \hat{y}_t^{\frac{2}{3}}\right)}{2\hat{y}_t^{\frac{1}{6}}}$$

Table 3.4. Comparison of Zero-Inflated Negative Binomial Regression(ZINB) and the Proposed Method

기준	모형	N	C	A	V	A	E
MSE	ZINB	2.011	1.009	0.509	0.754	0.532	4.648
	방법1	1.598	0.834	0.506	0.758	0.519	3.178
	방법2	1.607	0.829	0.505	0.757	0.522	3.160
MAE	ZINB	0.925	0.683	0.513	0.447	0.525	1.735
	방법1	0.802	0.610	0.506	0.436	0.493	1.350
	방법2	0.795	0.608	0.505	0.437	0.492	1.345

하지만 영과잉 모형에 의한 예측에서는 $\hat{y}_t$에 0이 나올 수 있어 위의 잔차를 계산할 수 없다. 이러한 문제점 때문에 이 논문에서는 다음과 같은 평균제곱오차(MSE), 평균절대값오차(MAE), 최대오차(MAX)로 모형을 비교했다.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_t - \hat{y}_t)^2\,, \qquad \text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_t - \hat{y}_t|\,, \qquad \text{MAX} = \max|y_t - \hat{y}_t|\,.$$

Table 3.4는 모형을 비교한 결과이며 표에서 방법1은 Y_t를, 방법2는 Z_t를 이용해 전체 식중독 발생건수를 추정한 것을 의미한다. 비교 결과 노로바이러스(N), 병원성 대장균(C), 기타 원인(E)에 의한 식중독 발생건수에 대한 예측에서 제안방법이 영과잉 음이항회귀모형보다 모두 현저히 우수한 것으로 나타났다. 두 모형 간에는 차이가 거의 없는 것으로 보이며 살모넬라균(S), 비브리오균(V), 황색포도상구균(A)에 의한 식중독 발생 예측에서는 세 모형 모두 비슷한 것으로 조사되었다. 이는 앞에서 본 것처럼 살모넬라균, 비브리오균, 황색포도상구균에 의한 식중독 발생건수가 150 전후로 상대적으로 적고 0 관측값이 많아 영과잉 모형이 잘 적용된 것으로 보인다. 이에 반해 노로바이러스, 병원성 대장균, 기타 원인에 의한 건수는 0 관측값이 많지만 365주 동안 특정 기간들에서 집중되는 것으로 나타나 영과잉 모형보다는 제안방법이 더 예측을 잘 하는 것으로 분석되었다.

예측력을 비교하기 위해 시점 k까지의 자료로 모형을 적합시키고 3시차 후까지의 예측값을 구하여 실제값과 비교해 보았다. 여기서 k는 260부터 364까지 사용되었고 1시차의 경우 105개, 2시차는 104개, 3시차는 103개의 예측오차를 통해 MSE, MAE, MAX를 계산하였다. 각각의 원인균별 예측오차에 대한 MSE, MAE, MAX의 결과는 Table 3.5와 같다.

각 방법에 의한 예측결과에는 예측기간에 따라 차이가 거의 없는 것으로 나타났으며 MSE 관점에서 노로바이러스와 기타원인균에 의한 식중독발생 예측은 제안하고자 하는 두 방법 모두 영과잉모형에 비해 확실히 우수한 것으로 나타났다. MAE에서도 비슷한 결과를 얻었으며 병원성 대장균, 살모넬라균, 황색포도상구균의 경우에도 제안한 방법이 영과잉모형보다 우수한 것을 볼 수 있다. 비브리오균에 의한 식중독 예측에서 MAE에서는 영과잉모형에 의한 예측이 우수했으나 MAX의 관점에서는 제안방법이 유의하게 좋은 것으로 나타났다. 노로바이러스, 살모넬라, 비브리오균에 의한 식중독 예측의 MAX에서 제안방법이 좋은 예측력을 가지는 것으로 분석되었다.

4. 결론

우리나라에서 발생한 주별 식중독 발생건수를 5대 원인균으로 나누어 보면 50% 이상의 주에서 0 관측값이 보고되고 있어 일반적인 포아송회귀모형으로 분석하는 것보다 영과잉모형을 사용하는 것이 더 적절하다. 모형을 선택하거나 개선하고자 할 때 추가적으로 고려해야 하는 성질은 원인균 별로 발생하는 식중독의 발생건수가 이전의 건수에 영향을 받는다는 것이다. 이를 설명할 수 있는 항을 개별 영과잉에

Table 3.5. Comparison of Zero-Inflated Negative Binomial Regression(ZINB) and the Proposed Method

원인균	기준	1주 후			2주 후			3주 후		
		ZINB	방법1	방법2	ZINB	방법1	방법2	ZINB	방법1	방법2
N	MSE	1.613	1.420	1.432	1.626	1.560	1.576	1.651	1.460	1.466
	MAE	0.921	0.887	0.888	0.925	0.906	0.906	0.929	0.888	0.889
	MAX	4.612	4.316	4.282	4.617	4.368	4.347	4.671	4.512	4.502
C	MSE	1.084	1.061	1.054	1.095	1.060	1.059	1.102	1.064	1.061
	MAE	0.697	0.699	0.699	0.701	0.700	0.699	0.702	0.705	0.703
	MAX	4.601	4.555	4.548	4.610	4.419	4.412	4.607	4.258	4.245
S	MSE	0.472	0.457	0.464	0.474	0.482	0.489	0.467	0.441	0.447
	MAE	0.506	0.513	0.520	0.506	0.523	0.528	0.500	0.496	0.498
	MAX	3.236	2.918	2.928	3.232	3.065	3.059	3.224	3.096	3.113
V	MSE	0.395	0.396	0.399	0.404	0.429	0.433	0.405	0.535	0.538
	MAE	0.368	0.381	0.384	0.374	0.397	0.400	0.372	0.434	0.437
	MAX	3.499	3.251	3.264	3.500	3.308	3.308	3.507	3.511	3.518
A	MSE	0.391	0.375	0.380	0.394	0.397	0.401	0.395	0.366	0.369
	MAE	0.517	0.517	0.522	0.518	0.529	0.532	0.517	0.514	0.516
	MAX	1.774	1.754	1.734	1.765	1.738	1.735	1.769	1.732	1.724
E	MSE	4.555	3.856	3.845	4.573	3.748	3.721	4.633	3.822	3.811
	MAE	1.710	1.567	1.569	1.708	1.544	1.542	1.725	1.560	1.558
	MAX	6.682	6.634	6.737	6.672	6.226	6.253	6.679	6.321	6.425

포함시켜 분석할 수 있으나 이 논문에서는 Yeo (2010)의 비교분석에서 우수한 예측력을 가진 AR(4)로 식중독 발생의 종속성을 모형에 반영하고 기준범주로짓모형으로 해당 기상상태에서 전체 발생건수 중 각 원인균이 차지하는 비율을 추정하여 원인균 별로 식중독이 얼마나 발생하는지를 알아보았다. 영과잉 모형과 비교한 결과 살모넬라균, 비브리오균, 황색포도상구균에 의한 식중독 발생 예측에서는 유의한 차이가 없었으나 노로바이러스, 병원성 대장균, 기타 원인에 의한 식중독 발생 예측에서는 제안모형이 유의하게 우수한 것으로 나타났다.

References

Bentham, G. and Langfod, I. H. (1995). Climate change and the incidence of food poisoning in England Wales, *International Journal of Biometeorology*, **39**, 81–96.

Choi, K., Kim, B., Bae, W., Jung, W. and Cho, Y. (2008). Developing the index of foodborne disease occurrence, *The Korean Journal of Applied Statistics*, **21**, 649–658.

Fleury, M. Charron, D. F., Holt, J. D., Allen, O. D. and Maarouf, A. R. (2006). A time series analysis of the relationship of ambient temperature and common bacterial enteric infections in two Canadian provinces, *International Journal of Biometeorology*, **50**, 385–391.

Jung, H. S., Kim, B. J., Cho, S. and Yeo, I. K. (2012). Analysis of food poisoning via zero inflation models, *The Korean Journal of Applied Statistics*, **25**, 859–864.

Lambert, D. (1992). Zero-inflated Poisson regression models with an application to defects in manufacturing, *Technometrics*, **34**, 1–14.

Magny, G. C., Murtugudde, R., Sapiano, M. R. P and Colwell, R. R (2008). A environmental signatures associated with cholera epidemics, *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 17676–17681.

Miller, J. M. (2007). *Comparing Poisson, Hurdle, and ZIP model fit under varying degrees of skew and zero-inflation*, University of Florida, DAI-A 68/06, Dec 2007.

Patrick, M. E., Christiansen, L. E., Steen Ethelberg, M. W., Madsen, H. and Wegener, H. C. (2004). Effects of climate on incidence of Camphylobacter spp. in humans and prevalence in broiler flocks in Denmark, *Applied and Environmental Microbiology*, **70**, 7474–7480.

SAS Institute Inc. (2008). *SAS/ETS User's Guide (Version 9.2, Chap.10, The COUNTREG Procedure)*, SAS Institute Inc., Cary, NC, USA.

Yeo, I. K. (2012). Models for forecasting food poisoning occurrences, *Journal of the Korean Data & Information Science Society*, **23**, 1117–1125.

원인균별 식중독 발생 건수 예측

여인권[a,1]

[a]숙명여자대학교 통계학과

(2013년 8월 19일 접수, 2013년 10월 15일 수정, 2013년 10월 28일 채택)

요 약

이 논문에서는 우리나라에서 발생하는 원인균별 식중독 발생건수를 예측하는 방법을 제안한다. 우리나라에서 보고되는 주별 식중독 발생 건수를 원인균로 나누면 자료에 많은 0의 관측값이 포함되어 있으며 식중독 발생 간에 종속성을 가진다. 이 현상을 모형화하기 위해 이 논문에서는 전체 식중독 건수를 자기회귀모형으로 예측하고 원인균별 식중독 발생 확률을 다범주 로짓모형으로 추정한다. 예측된 식중독 건수와 추정된 원인균별 식중독 발생 확률을 곱하여 원인균별 식중독 발생건수를 예측한다. 제안된 방법의 타당성을 확인하기 위해 평균제곱오차와 평균절대편차를 이용하여 제안 방법과 영과잉모형을 비교해 본다.

주요용어: 자기회귀모형, 다범주 로짓모형, 영과잉 음이항 회귀분석.

이 논문은 2012년도 숙명여자대학교의 교내연구비에 의하여 수행되었음.

[1](142-742) 서울특별시 용산구 청파로47길 100, 숙명여자대학교 통계학과, 교수. E-mail: inkwon@sm.ac.kr