Journal of the Korean Data & Information Science Society
2017, **28**(1), 57–66

http://dx.doi.org/10.7465/jkdi.2017.28.1.57
한국데이터정보과학회지

Reed - Frost 모형을 이용한 전염병 감염 확률 추정[†]

엄은진[1] · 황진섭[2] · 최보승[3]

[1]대구대학교 대학원 통계학과 · [2]대구대학교 전산통계학과 · [3]고려대학교 응용통계학과

접수 2016년 12월 30일, 수정 2017년 1월 5일, 게재확정 2017년 1월 12일

요 약

질병의 확산 과정을 설명하기 위한 모형으로 가장 대표적인 방법은 Kermack과 McKendrick (1927)에 의해 제안된 SIR (susceptible - infectious - recovered) 모형이다. SIR 모형을 구축하기 위해서는 질병의 감염률 (transition rate)과 회복률 (recovery rate)이 주어져 있거나 질병의 전체 확산 과정이 데이터로 주어진 경우 추정을 통하여 구할 수 있다. 하지만 데이터가 제한적으로 관찰된 경우 직접적인 감염률와 회복률의 계산이 불가능 하다. 본 연구에서는 관찰된 자료가 가지는 한계점을 고려하여 질병의 초기 확산과정에서 질병 감염 확률을 추정하기 위하여 리드-프로스트 (Reed-Frost) 모형 (Andersson과 Britton, 2000)을 적용하였다. 리드-프로스트 모형은 질병의 최초 감염자 수, 최종 감염자 수, 그리고 최초 감염대상자의 수가 주어졌을 때 이를 통하여 감염 확률을 추정하기 위한 모형이다. 본 연구에서는 서아프리카의 카메룬 공화국에서 조사된 역학 조사 자료를 이용하여 콜레라의 초기 감염 확률을 추정하였다. 그리고 추정된 결과를 이용하여 다시 SIR 모형에 적용하여 질병의 확산 경로에 대한 예측을 수행하였다. 예측 결과 조사 지역의 주민 가운데 50% 이상이 감염될 것으로 예측되었으며 질병의 전파는 약 한달 정도 지속될 것으로 예측 되었다.

주요용어: 리드-프로스트 모형, 에스아이알 모형, 질병 확산 모형, 콜레라.

1. 서론

전염병은 세균, 기생충, 바이러스와 같은 다양한 병원체에 의해 감염되어 발병하는 질환이다. 병원체에 의한 감염은 다른 사람과의 접촉, 음식과 물의 섭취, 호흡에 의한 병원체의 흡입 등 여러 경로를 통해 발생한다. 전염병의 기원을 살펴보면 인류가 탄생한 이래 끊임없이 인류를 괴롭혀 왔다. 그 중에서도 전 세계의 넓은 지역에 걸쳐 많은 목숨을 앗아 간 광역적 전염병의 시작은 서기 165년에서 180년 로마 제국의 마르쿠스 아우렐리우스 안토니우스 황제 시절로 거슬러 올라간다. 역병으로 인하여 당시 이탈리아 반도 전역에서 500만 명 이상이 목숨을 잃은 것으로 추정된다. 그럼에도 불구하고 인류는 전염병의 정체를 제대로 알지 못해 전염병에 걸려 죽었지만 19세기 후반에 들어서 파스퇴르와 코호 등에 의해 전염병 가운데 대부분의 이유가 박테리아에 의해 생긴다는 것이 확인되었고, 20세기 들어서는 박테리아 이외에 바이러스와 리케챠, 곰팡이 등도 전염병의 원인이라는 사실이 밝혀지면서 인류는 전염병과 싸워 나갔다. 최근에는 광우병, 구제역, 사스 (SARS), 인플루엔자에 이어 에볼라, 메르스 (MERS), 지카 바이러스, 콜레라 등 각종 전염병이 전역을 위협하고 있다. 에볼라는 아프리카를 중심으로 일어나고

[†] 이 논문은 대구대학교 교내 연구비로 지원받아 수행된 연구임 (No.20160257).
[1] (38453) 경상북도 경산시 진량읍 대구대로 201, 대구대학교 대학원 통계학과, 석사과정.
[2] (38453) 경상북도 경산시 진량읍 대구대로 201, 대구대학교 전산통계학과, 조교수.
[3] 교신저자: (30019) 세종특별자치시 세종로 2511, 고려대학 세종캠퍼스 응용통계학과, 조교수.
E-mail: cbskust@korea.ac.kr

있고 메르스는 중동호흡기 질환이다. 에볼라는 치사율이 약 90%가 넘는 전염병으로 질병 발생 시 사망할 확률이 높다. 메르스는 코로나 바이러스에 의해 일어나며 감염 경로가 가벼운 접촉으로도 발병할 수 있기에 감염력이 강하다. 콜레라는 주로 오염된 물, 어패류 등의 음식을 통해 콜레라균이 인간의 몸속에 들어와 소장의 감염에 의해 발생하는 급성 묽은 설사 질병이다.

수인성 전염병은 물과 식품이 전염성 있는 병원성 미생물에 오염되어 발생하는 질환으로 콜레라, 세균성이질, 장티푸스 등의 법정전염병을 의미한다 (Lim, 2007). 감염성 수인성 질환의 집단발생은 물, 식품 등의 공동매개체에 의한다. 오염된 물을 직접 섭취하거나 오염된 물로 조리한 식품에서 콜레라균, 살모넬라균등이 발생할 수 있는 것이다. 우리나라에서는 1996년 9월 경상북도 경주시 가정집 생일잔치에서 음식을 섭취한 59명 중 29명에게 살모넬라증이 발생하였다. 실제로 질병이 발생한 지역의 12곳의 지하수 중 9곳에서 기준치 이상의 일반 세균수와 대장균이 검출되었다. 2004년 9월 경상북도 영천시 초등학교에서도 살모넬라증이 발생했는데 학생 및 교직원이 급식에서 배급되었던 두부계란전을 섭취하고 1,205명 중 설사증 환자가 338명이 발병했다. 집단 질병 발생의 원인은 수도물의 일부 오염으로 인하여 생긴 것이다.문명의 속성 때문에 불행하지만 새로운 전염병이 계속 생겨날 수밖에 없다. 문명의 발달과 전염병의 전파범위와 발생빈도는 비례하기 때문이다. 질병의 역사와 발생의 연구를 보면, 문명화된 인류는 원시시대의 인류보다 훨씬 더 다양하고 복잡한 형태의 질병들에 시달리는 것으로 나타났다. 다시 말하자면 문명화 된 생활이 발달하면 할수록 더 많은 전염병에 시달리게 된다는 것이다.

본 연구의 목적은 역학 조사 자료를 이용하여 수인성 전염병 가운데 하나인 콜레라의 확산과정을 설명하기 위한 모형을 구축하는데 그 목적이 있다. 특히 질병 발생의 초기 시점에서 수집된 자료를 이용하여 질병의 초기 확률을 추정하고 초기 추정값을 활용하여 질병 발생 모형을 구축해 보고자 한다.

전염병의 모형화는 Daley와 Gani (2005)에 의하면 발병의 미래 과정을 예측하고 전염병을 통제하기 위한 전략을 평가하고, 질병이 확산되는 과정을 연구하는데 사용되는 수단이다. 질병 확산의 수학적 모형화의 초기에 실시된 연구자는 1766년 Daniel Bernoulli의 연구이다. 그는 천연두에 대한 질병 확산을 파악하고 예방하기 위해 수학적인 모형을 만들었다. 모형에서의 계산은 천연두에 대한 보편적인 접종이 26년 7개월에서 29년 9개월로 삶의 수명을 증가시킬 수 있음을 보여 주었다 (Bernoulli와 Blower, 2004).

전염병의 전파 과정을 모형화 하는데 있어서 가장 대표적인 방법은 Kermack와 McKendrick (1927)에 의해 제안된 SIR 모형이다. SIR 모형은 모든 구성원들이 접촉 가능하고, 두 구성원들 사이에서 전염이 발생할 가능성이 동일하다고 가정한다. SIR 모형은 전염병 유행의 전체 모집단을 크게 3 집단으로 구분한 후 이 집단간의 이동을 모형화 하여 전염병의 초기 전파과정을 설명하기 위하여 제안된 모형이다. 전체 모집단을 S (susceptible), I (infectious), R (removed or recovered)로 구분한다. 전체 모집단이 고정되어 있다는 가정에서 S는 전체 모집단 가운데 질병에 감염될 가능성이 있는 집단, 즉 감염대상군을 나타낸다. I는 질병에 감염된 집단, 즉 감염군을 의미한다. 마지막으로 R은 질병으로부터 회복되거나 사망하여 벗어나는 집단, 즉 회복군을 의미한다. SIR 모형은 천연두에서 독감에 이르기까지 다양한 질병에 적용할 수 있다. 전염병의 확산 과정을 확률적으로 모형화하기 위한 방법으로 전염병이 유행하는 초기 조건과 전염병의 확산 정도를 추측하기 위해 사용되는 모형이다. $S(t)$, $I(t)$, $R(t)$를 각각 시점 t에서 감염 대상군, 감염군, 회복군의 수를 나타내고 이들의 합은 $N = S(t) + I(t) + R(t)$로 고정되어 있다고 가정할 때 SIR 모형의 식은 다음과 같은 상미분 방정식으로 표현 가능하다.

$$\begin{aligned} N &= S(t) + I(t) + R(t) \\ \frac{dS}{dt} &= -\lambda S(t) I(t) \\ \frac{dI}{dt} &= \lambda S(t) I(t) - \gamma I(t) \\ \frac{dR}{dt} &= \gamma I(t) \end{aligned} \tag{1.1}$$

이 식 (1.1)에서 λ와 γ는 각각 감염률과 회복률을 나타내는 반응 상수이다.

이와 같은 전염병 모형은 크게 결정적 (deterministic) 모형과 확률적 (stochastic) 모형으로 나누어 볼 수 있다. 먼저 결정적 모형은 그 움직임이 확률적이지 않고 식 (1.1)과 같은 수리적 모형으로 표현 가능한 모형이다. 결핵의 경우와 같이 큰 집단을 처리할 때 결정적인 수학적 모형을 사용한다. 결정적 모형에서의 인구에서 개인은 다른 하위 그룹 또는, 구획, 전염병의 특정 단계를 나타내는 각각에 할당된다. 즉 S, I, R 과 같은 문자로 다른 단계를 나타내는 데 사용된다. 하나의 단계에서 다른 단계로 넘어가는 전이확률은 수학적 모형에 따라 미분 방정식을 이용하여 표현된다. 이러한 모형을 구축하는 동안 종 안의 집단 크기는 시간에 대한 미분이고, 역학적 과정은 결정적이라는 점이 가정이 되어야 한다. 즉, 종의 집단 변화는 오직 모형을 개발하는 데에만 기록을 이용하여 계산 될 수 있다 (Brauer와 Castillo-Chavez, 2001).

두 번째는 확률적 (stochastic) 모형이다. 확률적이라는 의미는 임의의 변수를 가진다는 것이다. 확률적 모형은 시간에 따른 하나 이상의 투입에서 임의의 변화를 허용함으로써 잠재적인 결과의 확률 분포를 추정하는 수단이다. 전염병이나 질병의 확산 전파 과정을 설명하는데 있어서 결정적 모형을 이용하는 것은 모형의 단순화 시킬 수 있다. 자료가 가지는 움직임에 반영될 수 있는 임의적 움직임을 반영하기 위해서는 확률적 모형이 보다 적합할 수 있다. 확률 모형은 노출, 질병 및 기타 질병 역학의 위험에 기회의 변화에 따라 달라진다. 모형은 작은 집단에서 변동이 중요할 때 사용된다 (Trottier와 Philippe, 2001). 확률적 움직임을 따르는 전염병 모형은 다음과 같이 확률적 화학 반응 모형 (stochastic kinetic network) 으로 표현가능하다 (Choi와 Rempala, 2012).

$$S + I \xrightarrow{\lambda} 2I : h_1 = h_1(S, I, R) = \lambda SI$$
$$I \xrightarrow{\gamma} R : h_2 = h_2(S, I, R) = \gamma I$$

여기서 h_1, h_2은 각각의 화학 반응식의 발생 정도를 나타내는 위험 함수이다.

Neurirth 등 (2004)은 확률적 방법과 결정적 방법을 이용하여 SIR 모형의 확장 모형인 Susceptible - Exposed - Infectious - Recovered (SEIR) 모형을 구축하였다. SEIR 모형은 SIR에서 E가 추가 된 모형으로써 E 는 Exposed를 나타내고 노출군이라 부른다. 감염 대상군에서 감염군으로 전이되는 과정에서 질병의 잠복기를 추가적으로 고려하는 모형이다. SIR 모형은 SEIR 모형 뿐 만 아니라 SIS, SIRS 모형과 같이 여러 형태로 확장 되었다. SIS 모형은 잠복기가 없는 질병으로 질병에 감염된 뒤 다시 감염 대상군으로 돌아가는 모형이다. SIRS 모형은 SIR 모형에서 회복을 하고 다시 감염 대상군으로 가는 모형이다. SEIR, SEIS 모형은 질병에 감염력이 생기기까지의 잠재기를 가지고 있는 모형, SIRS 모형은 감염 후 회복되어 일시적인 면역력을 갖는 모형이다 (Kim, 2010).

이와 같은 전염병의 확산 과정을 설명하기 위한 SIR류의 모형을 이용하여 실제 전염병 자료에 적합한 연구는 국내외에서 다양하게 진행 되었다. 기존연구를 살펴보면 국내에서는 질병 모형 중에서도 SIR 모형을 가장 많이 사용하고 있다. Hwang 등 (2007)에서는 한국의 말라리아, 신증후군출혈열, 홍역 자료를 이용하여 비선형 회귀식으로 표현되는 SIR 모형을 적용하여 기존 현상에 대해 설명하고 미래를 예측하는 연구를 하였다. Ryu와 Choi (2015)에서는 전염병의 확산 과정인 결정적인 과정과 확률적인 과정에 대한 비교 연구를 하였고, Seo와 Choi (2015)에서는 모형 추정을 위하여 우리나라 신종플루 확진환자 자료를 이용하여 SIR 모형 내에서 각 종의 발생이 포아송 확률 과정을 따른다는 가정 하에 확률적 화학 반응 모형을 이용하여 모형을 구축하여 확산 모형을 연구하였다. Lim 등 (2016)에서는 메르스 역학자료를 이용하여 SIR 모형을 구축하는 연구를 진행하였다. Lee 등 (2009)에서는 국내 쯔쯔가무시증 감염자 자료를 가지고 후향연산식을 이용해 감염자 분포 추정과 질병 확산 모형인 SIRS 모형을 적용하여 유병자수를 추정하였다.

실제 데이터를 이용한 전염병 모형을 구축하기 위해서는 모형을 구성하는 각 집단인 감염대상군, 감염군, 회복군, 혹은 노출군 등의 초기치와 모형의 모수에 해당하는 반응 상수가 주어져야 한다. 각 집단의 초기치는 관찰 혹은 조사를 통하여 구할 수 있다. 반응 상수의 경우 과거의 문헌 연구를 통하여 제공받거나 전염병의 확산과정의 전체 과정 (trajectory)이 관찰되었다면 이를 이용하여 추정할 수 있을 것이다. 그러나 본 연구에서 이용한 자료는 특정지역에서 조사된 전염병의 초기 확산과정의 초기 자료로 제한되어 있다. 전체 확산 및 소멸과정의 데이터가 주어져 있지 않기 때문에 각종 반응 상수를 직접적으로 추정할 수 없다. 본 연구에서는 이러한 한계를 극복하기 위하여 리드-프로스트 (Reed-Frost) 모형 (Anderson과 Britton, 2000, P. 4)을 이용하여 초기시점에서의 전염병 감염 확률을 추정하고자 하였다. 그리고 문헌조사를 통하여 추가적인 모수의 정보를 구한 후 이를 가지고 전염병 모형을 구축하기 위한 SIR 모형을 구축하였다.

본 연구의 이후 진행 과정은 다음과 같다. 2절에서는 본 연구에서 제시하고 있는 리드-프로스트 모형에 대하여 설명하고자 한다. 그리고 3절에서는 실제 조사를 통하여 수집된 아프리카 특정 지역의 콜레라 데이터를 이용하여 모형 구축을 진행하고자 한다. 그리고 마지막 4절에서는 본 연구에서 제시하고 있는 방법의 한계에 대하여 논하고 추구 연구의 진행 과정에 대하여 설명하고자 한다.

2. 리드-프로스트 모형

리드-프로스트 모형은 1928년 Lowell Reed와 Wade Frost에 의해서 제안된 모형으로 한 세대 안에서 전염병의 전파 과정을 설명하고자 하는 모형으로 확률적 변동을 가정하는 모형 가운데 하나이다 (Deijfen, 2011; Abbey, 1952). 주로 가구와 같이 크기가 작은 집단 안에서 질병 전파의 시간이 상대적으로 짧은 감염 전파 과정을 설명하는데 적절한 모형이다. SIR 모형과 기본적으로 유사한 가정으로부터 질병의 확산 과정을 설명한다. 시점 t (여기서 시점이란 반드시 시간흐름에 따른 시점일 필요는 없다) 또는 세대 t에서 감염군의 개인은 독립적으로 감염대상군과 접촉하여 질병을 전파시키는데 이때 감염 확률을 p라 하자. 시점 t에서 감염된 개인은 이제 다음 시점 $t+1$으로 이동하게 되고 시점 $t+1$에서는 감염군이 되고 시점 t에서는 제거된다. 세대 간 감염 확률은 이전 세대의 상태에만 의존한다고 가정하고 연쇄 이항분포를 이용하여 모형을 구축할 수 있다.

X_t와 Y_t가 각각 시점 (또는 세대) t에서 감염대상군의 수와 감염군의 수를 나타낸다고 하자. 다음 시점 $t+1$에서 감염군의 수가 $Y_{t+1} = y_{t+1}$가 될 확률을 계산하기 위하여 리드-프로스트 모형은 연쇄 이항분포를 이용하여 다음과 같이 주어진다.

$$\begin{aligned} P(Y_{t+1} &= y_{t+1} | X_0 = x_0, Y_0 = y_0, \cdots, X_t = x_t, Y_t = y_t) \\ &= P(Y_{t+1} = y_{t+1} | X_t = x_t, Y_t = y_t) \\ &= \binom{x_t}{y_{t+1}} (1-q^{y_t})^{y_{t+1}} (q^{y_t})^{x_t - y_{t+1}}. \end{aligned} \tag{2.1}$$

Y_{t+1}이 결정되면 이에 의하여 $X_{j+1} = X_j - Y_{j+1}$이 성립한다. 즉 시점 t에서 감염되지 않고 감염대상군에 그대로 남아있는 사람은 시점 $t+1$에서도 여전히 감염 대상군으로 남아 있게 되고 이와 같이 한 개인이 여전히 감염되지 않을 확률은 $q = 1-p$가 된다. 시점 t에서 감염을 피하지 못하고 감염된 개인은 다음 시점 $t+1$에서는 감염대상군에서 제외되고 감염군으로 이동하게 된다. 이와 같은 질병의 전파과정이 연속적으로 일어난다고 할 때 질병의 전파는 더 이상 감염군이 남아있지 않고 모두 회복군으로 넘어가 더 이상 전파를 일으키지 못 할 때까지 진행하게 된다. 초기 시점에서 감염대상군과 감염군의 수가 $X_0 = n$와 $Y_0 = m$ 으로 주어지고 더 이상 감염이 일어나지 않는 시점을 $k+1$이라 할 때 각 시점에

서의 감염군의 수는 $y_1, \cdots, y_k, y_{k+1} = 0$로 표시할 수 있으며 이들의 발생 학률은 바로 이전 시점의 값에 의해서만 영향을 받게 되는 마르코프 성질을 따르게 된다. 이러한 성질들로부터 초기 시점 값이 주어졌을 때 감염군의 전체의 확률은 다음과 같이 주어진다.

$$\begin{aligned} & P(Y_1 = y_1, \cdots, Y_k = y_k, Y_{k+1} = 0 | X_0 = n, Y_0 = m) \qquad (2.2) \\ = \; & P(Y_1 = y_1 | X_0 = n, Y_0 = m) \times \cdots \times P(Y_{k+1} = 0 | X_k = x_k, Y_k = y_k) \\ = \; & \binom{n}{y_1} (1-q^m)^{y_1} (q^m)^{n-y_1} \times \cdots \times \binom{x_k}{0} (1-q^{y_k})^0 (q^{y_k})^{x_k} \end{aligned}$$

이러한 질병의 감염 혹은 전파과정에서 최종적인 감염자의 수를 Z라 할 때 이는 $Z = \sum_{t \geq 1} Y_t$가 된다. 이제 식 (2.1)과 (2.2)를 이용하여 감염대상군과 감염군의 초기 수가 주어졌을 때 최종 감염자수 Z가 $Z = z$가 될 확률은 다음과 같이 주어진다.

$$P(Z = z | X_0 = n, Y_0 = m) = \sum_{y : |y| = z} P(Y_1 = y_1, \cdots, Y_k = y_k, Y_{k+1} = 0 | X_0 = n, Y_0 = m) \qquad (2.3)$$

이제 초기 시점의 감염대상군의 수, 감염자의 수, 그리고 최종 감염자수가 주어지면 식 (2.1), (2.2), (2.3)으로부터 감염 확률 p 또는 감염을 피하는 확률 $q = 1 - p$에 대한 우도함수를 구할 수 있고 감염 확률에 대한 추정을 수행할 수 있다.

3. 자료분석

3.1. 리드-프로스트 모형을 이용한 감염 확률 추정

본 연구는 전염병의 초기 감염 확률을 추정하기 위하여 서아프리카에 위치한 카메룬 공화국의 Maroua 지역의 건강 조사 자료를 이용하였다 (Profitos 등, 2014). 그 들의 연구에 의하면 설문조사 데이터는 2013년 6월 1일부터 8월 1일까지 Maroua지역의 네 군데 마을을 대상으로 하여 조사 되었다. 이 기간은 해당 지역의 우기의 시작 시점과 일치하고 네 군데 마을의 선정은 지난 2009년과 2011년에 걸쳐 창궐한 콜레라 발생률을 기준으로 선정하였다. 전체 785명의 개인에 대하여 건강 상태에 대한 자료 수집이 이루어 졌다. 실제 조사는 가구 단위로 총 120개의 가구를 대상으로 조사가 수행되었다. 지도상에 표시된 가구에 대하여 무작위 추출을 통하여 표본 가구가 선택되었고 각 가구로부터 기본적인 인구 통계 뿐만 아니라 각종 건강 상태에 대하여 면접 조사를 수행하였다. 수인성 전염병인 콜레라의 특성을 파악하기 위하여 가정용 저장 용기 및 식수원에서 표본을 수집하였다. 그리고 키우고 있는 가축으로부터 표본을 추출하여 오염 및 감염여부를 함께 조사 하였다. 건강상태에 대해서는 조사 시점 30일 동안의 위장 증상을 경험하였는지에 대한 질문을 수행하였다. 설사, 피 묻은 설사, 복부 경련, 구토, 메스꺼움 또는 발열등 의 현상을 조사를 진행하였다. 전반적으로 문맹률이 낮은 지역이기 때문에 모든 조사는 사전 동의를 통하여 구두로 조사한 후 이를 문서화 하였다. 본 연구에서는 이 가운데 질병의 발생 정보가 모두 제공되고 있는 DOU, DOM 두 마을의 데이터를 이용하여 분석을 진행하였다.

Figure 3.1의 왼쪽 그래프는 조사된 DOU, DOM 두 지역의 각 가구의 가구원 수의 분포를 나타낸 것이다. 가구원 분포를 살펴보면 보면 적게는 1인 가구에서 많게는 15인 가구까지 존재한다. 가구원의 수가 6명인 가구는 총 11가구로 가장 많은 빈도를 보이며 가구원이 1명인 가구부터 9명인 가구까지 대체로 비슷한 분포를 보이다가 10명 이상인 경우 급격하게 감소한다. 이와 함께 오른쪽 그림은 가구별로 발생한 환자의 수에 대한 분포를 나타낸다. 47개의 가구에서는 1명의 환자도 발생하지 않았으며 가장 많은 환자가 발생한 경우는 한 가구에서 6명의 환자가 발생한 경우가 존재한다.

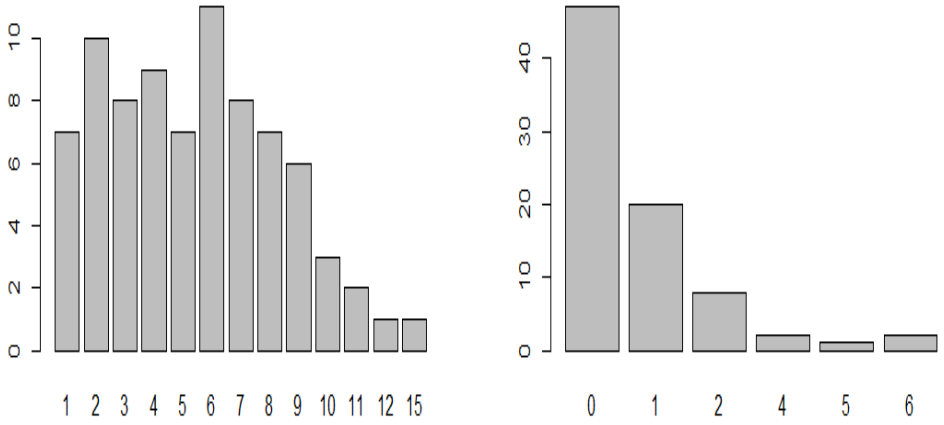

Figure 3.1 Distribution of the number of households (left panel) and the number of disease in each household (right panel)

총 조사기간은 2개월이지만 마을 별로는 각각 한달 동안 조사가 이루어 졌으며 조사 기간에 질병이 발생한 시기와 발생한 사람의 정보를 각각 기록해 놓았다. 각각의 가구를 관찰 단위로 하였을 때 가구 안에서 질병의 발생 및 전파 정도를 정확히 파악할 수 있다. 조사 대상의 모든 가구에 대하여 이러한 질병 확산 정보를 파악할 수 있다. 수집된 데이터의 특성을 살펴보았을 때 2장에서 설명한 리드-프로스트 모형을 이용하여 질병의 발생 확률을 추정하는 것이 매우 적절한 자료라 할 수 있다. 많은 경우에 있어서 질병 확산 모형 구축을 위하여 이용되는 자료들은 광범위한 지역에서 질병의 초기 발생 단계부터 어느 정도 시간이 흐른 후에 확산과정이 진정되고 더 이상 전파가 이루어지지 않은 이후의 자료를 이용하는 것이 보다 일반적이다. 수집된 질병의 확산 자료가 충분한 시간을 가지고 수집되어 전체적인 질병의 확산경로를 파악할 수 있도록 수집되었다면 SIR 모형과 같은 질병 확산 모형에 적용하여 모형의 구축이 가능할 것이다. 그러나 본 연구에서 수집된 자료는 이와는 달리 단기간에 수집된 횡단면 자료라 볼 수 있으며 리드-프로스트 모형을 이용한 발생 확률 추정이 보다 적절할 수 있다. 그리고 추정된 감염 확률과 관찰된 데이터의 전체 인구수와 전염병 감염자수 등과의 정보를 조합하여 SIR 모형의 구축이 가능하다. 관찰된 자료에 이용한 리드-프로스트 모형의 적합 방법을 알아보자.

구체적으로 리드-프로스트 모형으로 감염 확률을 계산하기 위해서는 모든 가구 단위로 감염 확률 식을 계산하여야 한다. 다음 Table 3.1에서 첫 번째 열은 마을 이름을 나타낸다. 두 번째 열은 전체 가구원 수를 나타내고 세 번째 열은 가구원 가운데 증상을 보인 사람의 수를 나타낸다. 이 증상에 따라 감염된 수로 파악할 수 있다. 이 표에는 정리되어 있지 않으나 원 자료로부터 한 달 동안 가구 안에서 증상을 보이는 가구원의 시점을 파악할 수 있다. 두 번째 열의 값은 결국 가구 안에서 감염대상군 초기값 X_0와 감염군 초기값 Y_0의 합을 나타내며 세 번째 열은 최종 감염자의 수인 Z로 나타낸다. 네 번째 열은 같은 수준의 가구원 수와 감염자수를 가지는 가구 수를 나타내고 마지막 열은 수준별로 계산되는 우도 함수값을 나타낸다. 이 우도 함수값들을 모두 곱하여 우도 함수를 구한 후 이로부터 최대 우도 추정치를 구할 수 있다. 계산된 최대우도 추정치는 0.0952이다. 즉 초기 감염확률은 0.0952라 할 수 있다.

Table 3.1 Conditional probability of each household in DOU&DOM region

Region	$X_0 + Y_0$	Z	Number of household	Conditional Probability
DOU	1	0	2	q
	1	1	2	$1-q$
	2	0	4	q^2
	2	1	1	$2(1-q)q^2$
	2	2	1	$(1-q)^2$
	3	0	4	q^3
	4	0	1	q^4
	4	1	1	$4(1-q)q^6$
	4	2	1	$12(1-q)^3q^7$
	4	4	1	$(1-q)^3q^4(1-q^2)$
	5	0	3	q^5
	5	1	3	$5(1-q)q^3$
	5	2	1	$10(1-q)^2q^6$
	6	0	3	q^6
	6	1	1	$6(1-q)^2q^6$
	7	0	2	q^7
	7	1	2	$7(1-q)q^12$
	7	2	1	$(1-q)^2q^6$
	8	0	2	q^8
	8	4	1	$(1-q)^3q^16(1-q^2)$
	8	6	1	$(1-q)^5q^22(1-q^2)$
	9	0	3	q^9
	10	2	1	$(1-q)^2q^117$
	11	0	1	q^{11}
	11	2	1	$55(1-q)^2q^{18}$
	15	5	1	$(1-q)^5q^{60}$
DOM	1	0	1	q
	2	0	3	q^2
	2	1	1	$2(1-q)q^2$
	3	0	3	q^3
	3	2	1	q^3
	4	0	3	q^4
	4	1	1	$4(1-q)q^6$
	6	0	2	q^6
	6	1	2	$6(1-q)q^{10}$
	7	0	1	q^7
	7	1	2	$7(1-q)q^{12}$
	8	0	3	q^8
	9	0	2	q^9
	10	0	1	q^{10}
	12	6	1	$(1-q)^3(1-q^3)^3q^{15}$

3.2. SIR 모형 구축

이제 계산된 초기 질병 발생 확률을 이용하여 SIR 모형을 구축하여 보자. 시뮬레이션을 이용하여 식 (1.1)로부터 SIR모형을 구축할 수 있는데 이를 위해서 필요한 값들은 다음과 같다. 먼저 질병의 감염률을 나타내는 반응 상수 λ (transmission rate)와 회복률을 나타내는 반응 상수 γ (recovery rate)가 주어져야 한다. 또한 감염대상군의 초기값 $S(0)$와 감염군의 초기값 $I(0)$이 함께 주어져야 한다. 이 같은 값들이 주어 졌을 때 초기 발생 확률 p는 다음과 같은 식으로 근사 될 수 있다.

$$p \approx \frac{\lambda S(0)I(0)/N}{\lambda S(0)I(0)/N + \gamma I(0)} \tag{3.1}$$

Table 3.1의 관찰된 데이터로부터 초기 시점의 감염대상군의 수는 $S(0) = 433$이고 감염군의 수는 $I(0) = 61$이다. Eisenberg 등 (2012)연구에 의하면 콜레라에 감염된 후 회복까지 걸리는 대략적인 기간은 3일에서 10일 정도로 볼 수 있으며 이에 가장 짧은 회복기간 3일을 적용하여 회복률 값은 $\gamma = \frac{1}{3} = 0.33$로 하였다. 식 (3.1)로부터 계산되는 감염률은 $\lambda = 0.04$이다. 이제 주어진 값들을 가지고 식 (1.1)의 모형에 대입하여 SIR 모형을 구축할 수 있다.

다음 Figure 3.2는 COPASI (Hoops 등, 2006; Mendes 등, 2009; Schaber, 2012) 프로그램을 이용하여 SIR 모형을 구축한 것이다. COPASI는 생화학 네트워크 모형과 질병 확산 모형의 모의실험과 분석을 위한 응용스프트웨어이다. 상미분 방정식이나 Gillespie 알고리즘 (Gillespie, 1977)에 기반을 확률적 모의실험을 수행할 수 있는 프로그램이다. Figure 3.2에서 점선으로 표시된 부분은 감염대상군을 나타내고 실선으로 표시된 부분은 감염군을 나타낸다. 그 사이에 굵은 점선으로 표시된 부분은 회복군을 나타낸다. 이 모의실험의 결과를 해석하면 조사가 이루어진 Maroua지역에서는 대략적으로 25일 이상 질병이 지속되며 전체 지역 주민가운데 절반 이상에 해당하는 270명 정도가 증상을 보일 것으로 예상된다.

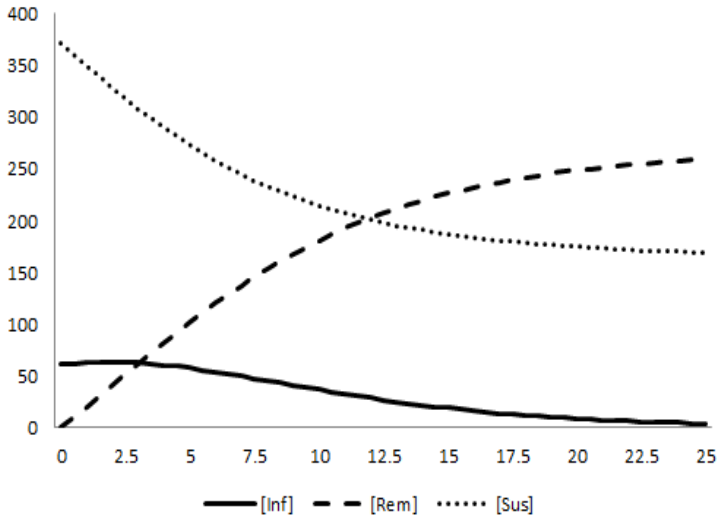

Figure 3.2 Graph of SIR model

4. 결론

본 연구에서는 서아프리카 카메룬 공화국에서 직접 조사된 수인성 전염병인 콜레라 관련 데이터를 이용하여 질병의 초기 시점에서의 감염 확률을 계산한 후 관찰된 데이터와 추가적인 문헌 조사 값을 통하여 SIR 모형을 구축하고자 하였다. 질병의 확산과정의 초기 시점 데이터, 즉 횡단면 자료가 가지는 한계를 극복하고 가구 안에서의 질병 이동 경로를 파악할 수 있다는 장점을 활용하여 리드-프로스트 모형을 적용하여 초기 추정치를 추정하였다. 구축된 우도함수로부터 초기 발생 확률의 최대 우도 추정치를 계산하였다. 리드-프로스트 모형은 상대적으로 간단한 형태의 질병 예측 모형이라 할 수 있다. 하지만 주어진 자료의 한계점과 장점을 최대한 활용하여 적합하는데 그 의의가 있다고 할 수 있다. 따라서 본 연구에서 제시된 방법을 이용하여 초기 발생 확률을 계산한 후 본격적으로 질병 확산 모형을 구축하는데 초기 정보로 활용될 수 있을 것이다.

본 연구에서 적용한 수인성 전염병 가운데 하나인 콜레라 데이터를 이용한 것이다. 수인성 전염병 이기 때문에 사람에 의해서 전파되는 것 뿐 만 아니라 불결한 환경에 의해서 전파의 가능성이 높은 질병이다. 공급되고 저장되는 식수원 뿐만 아니라 기르고 있는 가축에 의해서도 전파가 가능한 질병이다. 따라서 질병의 확산 과정을 구축하는데 있어서 식수원의 오염 정도나 가축의 감염도 인간의 감염에 영향을 미칠 수 있다. 향후 연구를 통하여 이러한 환경 요인을 함께 고려한 질병 확산 모형의 구축을 수행해 볼 수 있을 것이다.

References

Abbey, H. (1952). An examination of the Reed-Frost theory of epidemics. *Human Biology*, **24**, 201-233.

Andersson, H. and Britton, T. (2000). *Stochastic epidemic models and their statistical analysis*, Springer, New York.

Bernoulli, D. and Blower, S. (2004). An attempt at a new analysis of the mortality caused by smallpow and of the advantages of inoculation to prevent it. *Reviews in Medical Virology*, **14**, 275-288.

Brauer, F. and Castillo-Chavez, C. (2001). *Mathematical models in population biology and epidemiology*, Springer, New York.

Choi, B. and Rempala, G. A. (2012). Inference for discretely observed stochastic kinetic networks with applications to epidemic modeling. *Biostatistics*, **13**, 153-165.

Daley, D. J. and Gany, J. (2005). *Epidemic Modeling: An Introduction*, Cambridge University Press, New York.

Deijfen, M. (2011). Epidemics and vaccination on weighted graphs. *Mathematical biosciences*, **232**, 57-65.

Eisenberg, M. C., Robertson, S. L. and Tien, J. H. (2012). Identifiability and estimation of multiple transmission pathways in cholera and waterborne disease. *Journal of Theoretical Biology*, **324**, 84-102.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, **81**, 2340-2361.

Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., Singhal, M., Xu, L., Mendes, P. and Kummer, U. (2006). COPASI-a complex pathway simulator. *Bioinformatics*, **22**, 3067-3074.

Hwang, N. A., Jeong, B. Y., Lim, Y. C. and Park, J. S. (2007). Diseases data analysis using sir nonlinear regression model. *Journal of The Korean Data Analysis Society*, **9**, 49-59.

Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society A: Mathematical Physical and Engineering Sciences*, **115**, 700-721.

Kim, C. S. (2010). *Development and evaluation of influenza confinement strategy using mathematical estimating model*, Final report, Korea Centers for Disease Control & Prevention, Osong, Korea.

Lee, J. H., Murshed, M. S. and Park, J. S. (2009). Estimation of infection distribution and prevalence number of Tsutsugamushi fever in Korea. *Journal of theKorean Data & Information Science Society*, **20**, 149-158.

Lim, H. (2007). Contributing factors of infectous waterborne and foodborne outbreaks in Korea. *Journal of the Korean Medical Association*, **50**, 582-591.

Lim, Y., Do, M. and Choi, B. (2016). A construction of susceptible - infected - removed model using Korean MERS pandemic data. *Journal of the Korean Data Analysis Society*, **18**, 105-115.

Mendes, P., Hoops, S. and Gauges, R. (2009). Computational modeling of bochemical networks using COPASI. *Methods in Molecular Biology -Clifton then Totowa*, **500**, 17-60.

Neurirth, E., Arganbright, D. (2004). *The active modeler: Mathematical modeling with Microsoft Excel*, Tnomson Brooks/Cde, Belmont.

Profitos, M., Mouhaman, A., Lee, S., Garabed, R., Moritz, M., Piperata, B., Tien, J., Bisesi, M. and Lee, J. (2014). Muddying the Waters: A new area of concern for drinking water contamination in Cameroon. *International Journal of Environmental Research and Public Health*, **11**, 12454-12526.

Ryu, S. and Choi, B. (2015). Development of epidemic model using the stochastic method. *Journal of the Korean Data & Information Science Society*, **26**, 301-312.

Schaber, J. (2012). Easy parameter identifiability analysis with COPASI. *Bio systems*, **110**, 183-185.

Seo, M. and Choi, B. (2015). An estimation method for stochastic epidemic model. *Journal of the Korean Data Analysis Society*, **17**, 1247-1259.

Trottier, H. and Philippe, P. (2001). Deterministic modeling of infectious diseases: Theory and methods. *The Internet Journal of Infectious Disease*, **1**.

Journal of the Korean Data & Information Science Society
2017, **28**(1), 57–66

http://dx.doi.org/10.7465/jkdi.2017.28.1.57
한국데이터정보과학회지

An estimation method of probability of infection using Reed - Frost model[†]

Eunjin Eom[1] · Jinseub Hwang[2] · Boseung Choi[3]

[1]Department of Statistics, Daegu University
[2] Department of Computer Science and Statistics, Daegu University
[3]Department of Applied Statistics, Korea University

Received 30 December 2016, revised 5 January 2017, accepted 12 January 2017

Abstract

SIR model (Kermack and McKendrik, 1927) is one of the most popular method to explain the spread of disease, In order to construct SIR model, we need to estimate transition rate parameter and recovery rate parameter. If we don't have any information of the two rate parameters, we should estimate using observed whole trajectory of pandemic of disease. Thus, with restricted observed data, we can't estimate rate parameters. In this research, we introduced Reed-Frost model (Andersson and Britton, 2000) to calculate the probability of infection in the early stage of pandemic with the restriction of data. When we have an initial number of susceptible and infected, and a final number of infected, we can apply Reed - Frost model and we can get the probability of infection. We applied the Reed - Frost model to the Vibrio cholerae pandemic data from Republic of the Cameroon and calculated the probability of infection at the early stage. We also construct SIR model using the result of Reed - Frost model.

Keywords: Epidemic model, Reed - Frost model, SIR model, vibrio cholerae.

[†] This research is supported by Daegu University research grant in 2016 (No.20160257).
[1] Graduate student, Department of Statistics, Daegu University, Gyeongbuk 38453, Korea.
[2] Assistant professor, Department of Computer Science and Statistics, Daegu University, Gyeongbuk 38453, Korea.
[3] Corresponding author: Assistant professor, Department of Applied Statistics, Korea University Sejong Campus, Sejong 30019, Korea. E-mail: cbskust@korea.ac.kr