The Korean Journal of Applied Statistics (2014) DOI: http://dx.doi.org/10.5351/KJAS.2014.27.3.445
27(3), 445–459

Modeling of The Learning-Curve Effects on Count Responses

Minji Choi[a] · Man Sik Park[a,b,1]

[a]Department of Statistics, Sungshin Women's University
[b]Institute of Statistics, Sungshin Women's University

(Received March 3, 2014; Revised April 25, 2014; Accepted June 9, 2014)

Abstract

As a certain job is repeatedly done by a worker, the outcome comparative to the effort to complete the job gets more remarkable. The outcome may be the time required and fraction defective. This phenomenon is referred to a learning-curve effect. We focus on the parametric modeling of the learning-curve effects on count data using a logistic cumulative distribution function and some probability mass functions such as a Poisson and negative binomial. We conduct various simulation scenarios to clarify the characteristics of the proposed model. We also consider a real application to compare the two discrete-type distribution functions.

Keywords: Poisson distribution, negative binomial distribution, learning-curve effects, cumulative distribution function, maximum likelihood method.

1. 서론

일반적으로 특정한 작업에 익숙해진다는 것은 그 작업에 투입되는 노력에 비해 산출되는 성과가 보다 뚜렷해진다는 것을 의미한다. 즉, 동일한 양이나 정도의 노력을 들여 특정한 작업을 유사한 조건이나 상황 하에서 반복적으로 수행하게 되면 초기 시점보다 원하는 성과를 기대 이상으로 얻게 된다는 것을 의미한다. 특정한 작업의 반복을 통해 얻고자 하는 성과는 흔히들 정량적인 측면으로 국한하여 설명한다. 이를 테면, 로봇을 이용한 위암 수술을 예로 들 수 있다. 수술자가 로봇을 이용한 수술을 반복적으로 시행할수록 수술 중, 혹은 수술 이후의 환자에게서 기대할 수 있는 예후에 대한 여러 성과들이 보다 뚜렷하게 나타날 수 있을 것이다. 이러한 성과들은 수술 당시에 측정되는 수술소요시간이나 수술 이후에 관측하게 되는 합병증의 유무, 재수술의 여부 등과 같은 형태로 나타나게 된다. 또한 수술 중 소실되는 혹은 보충되는 혈액량으로도 수술자의 특정 수술에 대한 숙련도를 평가할 수 있다. 부연해서 설명한다면 수술자의 특정 수술방법에 대한 숙련도가 증가할수록 수술소요시간이 일반적으로는 감소하게 되고 수술 이후에 관측하게 되는 환자의 좋지 않은 예후도 줄어들게 될 것이다. 이와 같이 특정한 작업을 수행하는 횟수가 증가할수록 동일한 노력을 들이고도 보다 큰 성과를 얻게 되고 동일한 성과를 얻기 위해 투입되는 노력이 줄어들게 된다는 것을 알 수 있다. 독일의 심리학자 헤르만 에빙하우스는 이 현상을 '학습곡선효과(learning-curve effects)'라고 정의하였다.

This work was supported by the Sungshin Women's University Research Grant of 2012.
[1]Corresponding autor: Sungshin Women's University, Dongseon-dong 3-ga, Seongbuk-gu, Seoul 136-742, Korea. E-mail: mansikpark@sungshin.ac.kr

학습곡선효과는 연구자가 관심을 가지는 성과(혹은 종속변수)를 어떻게 정의하느냐에 따라 숙련도에 비례 혹은 반비례의 관계를 가질 수 있다. 예를 들면, 수술의 합병증의 유무 혹은 입원기간을 통해 수술의 성공 여부를 판단하다면 숙련도와 수술의 성공율은 비례관계를 갖게 되지만 수술소요시간과는 반비례의 관계를 가지게 된다. 이러한 학습곡선효과는 다양한 분야에서 적용되는 분석기법 중 하나이다. 아울러 산업공학분야 및 디자인공학분야 등 다양한 분야에서 학습곡선효과를 통계적 관점에서 규명하고자 하였다 (Back, 2008; Hong, 2007). 최근 학습곡선효과에 대한 의학분야에서의 적용이 활발히 이루어지고 있는데 최첨단 수술기법 및 수술장비를 이용하여 특정 질환자의 생존율을 향상시키기 위한 다방면의 투자 및 지원이 이루어낸 결과라 하겠다. Han 등 (2011)은 비선형회귀모형으로 구현된 학습곡선효과를 통해 수술횟수가 증가하면서 복강경을 이용한 위암절제수술의 수술소요시간이 단축되는지를 밝히고자 하였다. Ferguson 등 (2005) 또한 복강경을 이용한 전립선 절제술(laparoscopic radical prostatectomy)의 수술소요시간을 여러 구간(phase)으로 구분하고 이 구간들의 수술시간을 비교하여 학습곡선효과가 나타났음을 확인하였다. 또한 Park 등 (2012)은 복강경수술에 대한 숙련도가 상이한 여러 수술자에 대해 로봇을 이용한 위암절제술의 학습곡선효과를 비교하고자 하였다. Jeff 등 (2014)은 로봇을 이용한 자궁 절제술의 수술시간에 대해 학습곡선효과를 살펴보았다. 이 외에도 의학분야에서의 학습곡선효과를 적용한 논문에는 Jaffe 등 (2009), Pruthi 등 (2008), Hayn 등 (2010), Schreuder 등 (2010), Chen 등 (2007) 등이 있다.

학습곡선효과를 통계적 모형으로 적합하고자 한 시도 또한 이루어졌는데 Lee와 Park (2012)은 특정한 작업의 실패(혹은 성공) 여부를 반복적으로 관측한 이항 반응자료에 대하여 비선형회귀모형을 가정한 학습곡선효과를 구축하였다. Choi (2013)는 특정한 작업을 통해 얻게 되는 결과의 형태가 이산형인 자료에 대해 학습곡선효과를 모형화하고자 하였다. 이를 위해 포아송분포와 음이항분포 등의 이산형 확률분포를 가정한 통계적 모형을 적합하고 모의실험과 실제자료분석을 통해 각 모형의 성능을 평가하고 비교하였다. 본 연구에서는 Choi (2013)의 확장으로서 다양한 모의실험 및 실제자료 하에서의 모형의 성능을 보다 구체적으로 평가하고자 한다.

본 연구에서의 실제자료분석의 한 예로 Figure 1.1은 1995년 1월부터 2012년 12월까지 사망원인이 인체 면역결핍 바이러스 병인 사망자 수의 추이이다. 본 연구에서 관심을 가지는 변인의 형태는 특정한 작업을 시행한 성과가 이산형인 자료이다. 앞서 언급한 바와 같이 의학분야의 경우 수술 이후에 발병하게 되는 합병증의 수 혹은 수술 중 투여되는 혈액 수 등이 이에 해당될 수 있다. 하지만 현실적으로 이러한 형태의 자료를 구하는 것이 쉽지 않다. 따라서 실제자료분석에서는 일정 기간동안 발생한 이산형 자료들 중 시간이 흘러가면서 발생건수(횟수)가 증가하거나 감소하는 경향을 가진 자료를 고려하였다. 이는 동일한 조건이나 동일한 기간 하에서 관측된 자료라는 측면을 제외하면 작업자의 경험이나 노력 등이 학습곡선효과에 영향을 주는 부분은 고려할 수 없으므로 엄밀한 의미에서는 이러한 자료에 학습곡선효과를 적용하는 것이 문제가 될 수 있음을 미리 밝히는 바이다.

Figure 1.1을 살펴보면 초기 시점에서는 사망자 수가 적다가 시간이 지남에 따라 증가하는 패턴을 보인다. 하지만 일정 시간이 지난 후에는 사망자 수가 일정한 수준을 유지하는 경향을 보이고 있다. 이 연구에서는 이와 같은 추세를 보이는 여러 사회현상이나 자연현상의 자료에 통계적 모형으로 표현된 학습곡선효과를 적합하고자 한다. 이를 위해 저자들이 고려한 모수는 다음과 같다: 1) 초기 시점에서의 평균 발생수, 2) 학습곡선효과가 이루어져 안정화된 시점에서의 평균 발생수, 3) 학습곡선효과가 발생하여 안정기로 접어드는 변곡점, 그리고 4) 학습곡선효과가 일어나는 기간. Figure 1.1의 경우, 학습곡선효과가 발생하게 되는 시점은 대략적으로 2000년 즈음으로 볼 수 있고, 안정기에 접어들게 되는 시점은 2010년 전후로, 초기 시점에서의 평균 발생수는 1–2건으로, 그리고 안정기에 접어든 시점에서의 평균 발생수는 10여건으로 예상할 수 있다. 따라서 학습곡선효과가 발생하여 완성되는 데 걸린 시간은 대략

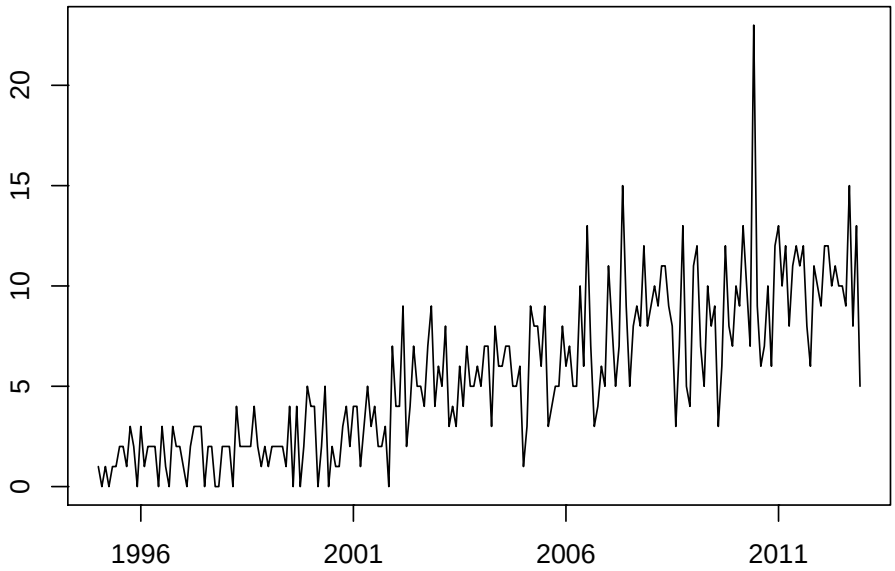


Figure 1.1. The number of death due to human immunodeficiency virus(HIV) from 1995 to 2013(216 months)

적으로 10년 정도이므로 변곡점은 2005년 즈음으로 예상할 수 있다. 주어진 자료에 대한 이와 같은 단순한 예상 혹은 예측을 모형화를 통해 보다 실증적으로 알아보고자 한다.

본 논문의 순서는 다음과 같다. 제 2장에서는 이산형 자료를 이용하여 학습곡선을 모형화하고, 제 3장에서는 모의실험을 진행한다. 제 4장에서는 다양한 사망 원인에 대한 사망자수 자료에 제 2장에서 소개한 통계적 모형을 적합한다. 마지막으로 제 5장에서는 결론 및 앞으로의 연구방향에 대해 알아본다.

2. 이산형 자료를 이용한 학습곡선의 모형화

특정한 작업의 숙련도가 증가될수록 작업을 수행하는데 소요되는 시간이 감소하고 작업의 질 또한 향상된다는 것이 학습곡선효과라 할 수 있다. 따라서 안정기에 접어드는 동안 오직 한 번의 학습효과가 있다는 가정 하에서는 일반적으로 S자 형태 혹은 역S자 형태의 추세를 고려할 수 있다. 본 연구에서는 일정한 시간 간격을 두고 시점이 흘러가면서 발생 수가 증가하다가 안정화되는 경향을 내포한 자료을 모형화하고자 한다. 이를 위해 이산형 확률분포(discrete probability distribution)를 고려하게 되는데 이들 중에서 포아송분포(Poisson distribution; PO)와 음이항분포(Negative binomial distribution; NB)를 토대로 로지스틱분포(logistic distribution)의 누적분포함수(cumulative distribution function)를 이용하여 학습곡선효과의 통계적 모형를 적합하고자 한다.

2.1. 이산형 자료의 확률 분포

단위시간 동안 희귀한 사건의 발생수는 포아송분포를 따른다. 관찰 시점을 $t = 1, \ldots, T$라 하자. 시점 t에서의 발생수 X_t는 평균 발생수가 μ_t인 포아송분포를 따르며 다음과 같이 표현된다.

$$X_t \sim \mathrm{PO}(\mu_t), \quad X_t = 0, 1, 2, \ldots, \ 0 < \mu_t < \infty.$$

그리고 포아송분포의 확률질량함수(probability mass function)는 다음과 같다.

$$g(x_t|\mu_t) = \frac{e^{-\mu_t}\mu_t^{x_t}}{x_t!}, \quad x_t = 0, 1, 2, \ldots. \tag{2.1}$$

음이항분포에서 m번 성공이 일어날 때까지의 실패 횟수를 X_t라 하면 전체 시행횟수인 Y_t는 $Y_t = X_t + m$이 된다. X_t를 사용한 음이항분포의 확률질량함수는 다음과 같이 나타낸다. $0 < p_t < 1$에 대하여

$$g(x_t|p_t, m) = \frac{\Gamma(x_t + m)}{\Gamma(m)\Gamma(x_t + 1)} p_t^m (1 - p_t)^{x_t}, \quad x_t = 0, 1, 2, \ldots. \tag{2.2}$$

여기서, p_t는 시점 t에서의 성공확률이다. 식 (2.2)과 같이 표현된 음이항분포는 발생수를 모수로 가지는 포아송분포(식 (2.1))와는 달리 성공확률을 모수로 가지게 된다. 본 연구에서는 포아송분포의 모수인 평균 발생수와 동일한 형태의 모수를 음이항분포에서도 사용하기 위해 음이항분포의 평균을 μ_t로 재표현하고자 한다. 평균 발생수를 $E(X_t) = \mu_t$라 하면, 성공확률은 $p_t = m/(m + \mu_t)$로 표현될 수 있다. 따라서 μ_t를 이용한 음이항분포의 확률질량함수는 식 (2.3)과 같이 표현되고 $X_t \sim \text{NB}(\mu_t)$로 나타낸다.

$$g(x_t|\mu_t, m) = \frac{\Gamma(x_t + m)}{\Gamma(m)\Gamma(x_t + 1)} \left(\frac{m}{m + \mu_t}\right)^m \left(\frac{\mu_t}{m + \mu_t}\right)^{x_t}, \quad x_t = 0, 1, 2, \ldots. \tag{2.3}$$

2.2. 학습곡선의 모형화를 위한 모형 구축

제1장에 제시된 S자 형태의 추세를 가지는 자료 (Figure 1.1 참조)에서 최종적으로 추정하고자 하는 것은 임의의 시점 t에서의 평균 발생수, μ_t로서 Choi (2013)는 μ_t를 다음의 4개 모수들로 이루어진 함수형태로 간주하였다: 1) 학습효과가 발생하기 전, 초기시점에서의 평균 발생수(θ_1), 2) 학습곡선의 효과가 완성된 이후, 즉 안정기로 접어든 이후 시점에서의 평균 발생수(θ_2), 3) 초기 시점에서의 평균 발생수와 안정화 시점에서의 평균 발생수의 평균값인 $(\theta_1 + \theta_2)/2$를 지날 때의 시점(θ_3), 그리고 4) 학습곡선효과의 기간과 관련한 변동성(θ_4). 따라서 평균발생수는 $\mu_t(\boldsymbol{\theta})$와 같이 나타낼 수 있고 이를 위해 $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)'$를 추정하게 된다. 여기서, $0 \leq \theta_1 < \theta_2$이고 평균 발생수인 $\mu_t(\boldsymbol{\theta})$를 정의하면 식(2.4)로 표현된다.

$$\mu_t(\boldsymbol{\theta}) = \theta_1 + (\theta_2 - \theta_1)F(t|\theta_3, \theta_4). \tag{2.4}$$

여기서, $F(\cdot)$은 임의의 시점 t의 누적분포함수이다. 본 연구에서는 Choi (2013)에서 제안한 대로 로지스틱분포의 누적분포함수를 고려하고자 한다. 이를 적용하면 다음과 같은 $\mu_t(\boldsymbol{\theta})$을 모수로 가지는 포아송분포 혹은 음이항분포를 고려할 수 있다.

$$\mu_t(\boldsymbol{\theta}) = \theta_1 + (\theta_2 - \theta_1)F(t|\theta_3, \theta_4) = \theta_1 + (\theta_2 - \theta_1)\left[1 + \exp\left(-\frac{t - \theta_3}{\theta_4}\right)\right]^{-1}. \tag{2.5}$$

2.3. 모수 추정

학습곡선효과를 구체화하기 위한 통계적 모형을 구축하고 이를 통해 임의의 시점에서의 평균 발생수 $\mu_t(\boldsymbol{\theta})$를 추정하고자 한다. 즉, 식 (2.5)에서 제시된 $\boldsymbol{\theta}$를 추정해야 한다. 이를 위해 본 연구에서는 최대우도추정법(maximum likelihood estimation method)을 사용한다. 최대우도추정법은 확률변수의 확률분포를 결정한 후에 자료의 재연성을 극대화하기 위한 모수를 찾는 방법이고 우도함수(likelihood function; L) 또는 로그우도함수(log-likelihood function; l)를 이용하여 모수를 추정한다. 실제적으로 (로그)우도함수를 이용하여 모수를 추정하기 위해 다음과 같은 뉴튼-랩슨방법(Newton-Raphson

Table 3.1. Scenarios considered in the simulation studies

$\boldsymbol{\theta}$	Scenario								
	1.	2.	3.	4.	5.	6.	7.	8.	9.
θ_1	1	1	1	1	1	1	1	1	1
θ_2	5	5	5	10	10	10	20	20	20
θ_3	30	40	60	30	40	60	30	40	60
θ_4	5	5	5	5	5	5	5	5	5

method)을 사용하고자 한다. 뉴튼-랩슨방법은 먼저 초기값을 지정한 후, 다음 단계의 추정값, 즉 r단계의 $\widehat{\boldsymbol{\theta}}^{(r)}$과 이전 단계인 $(r-1)$단계의 $\widehat{\boldsymbol{\theta}}^{(r-1)}$의 차이가 거의 없어질 때까지 동일한 과정을 반복적으로 시행한다.

$$\widehat{\boldsymbol{\theta}}^{(r)} = \widehat{\boldsymbol{\theta}}^{(r-1)} + \left[\mathbf{I}^{(r-1)}\right]^{-1} \mathbf{u}^{(r-1)}.$$

여기서, $\mathbf{u}$는 스코어 통계량(score statistic)이고, $\mathbf{I}$는 피셔의 정보행렬(Fisher's information matrix)이다. $\mathbf{x} = (x_1, x_2, \ldots, x_T)'$라 할 때, 로그함수는 다음과 같다.

$$l(\boldsymbol{\theta}|\mathbf{x}) \equiv \ln L(\boldsymbol{\theta}|\mathbf{x}) = \ln \prod_{t=1}^{T} g(x_t|\mu_t(\boldsymbol{\theta})) = \sum_{t=1}^{T} \ln g\left(x_t|\mu_t(\boldsymbol{\theta})\right).$$

델타방법(Delta method)을 이용하여 모수의 추정량에 대한 근사적 평균(mean)과 분산(variance)을 구하고자 한다. $\mu_t(\widehat{\boldsymbol{\theta}})$의 평균과 분산은 다음과 같다.

$$\begin{aligned} E\left[\mu_t\left(\widehat{\boldsymbol{\theta}}\right)\right] &\approx \mu_t(\boldsymbol{\theta}) + \sum_{i=1}^{4} \frac{\partial \mu_t}{\partial \theta_i} E\left(\widehat{\theta}_i - \theta_i\right) = \mu_t(\boldsymbol{\theta}), \\ \mathrm{Var}\left(\mu_t\left(\widehat{\boldsymbol{\theta}}\right)\right) &\approx E\left[\left(\mu_t\left(\widehat{\boldsymbol{\theta}}\right) - \mu_t\left(\boldsymbol{\theta}\right)\right)^2\right] \\ &= \sum_{i=1}^{4} \left(\frac{\partial \mu_t}{\partial \theta_i}\right)^2 \mathrm{Var}\left(\widehat{\theta}_i\right) + 2\sum_{i>j}^{4} \left(\frac{\partial \mu_t}{\partial \theta_i}\right)\left(\frac{\partial \mu_t}{\partial \theta_j}\right) \mathrm{Cov}\left(\widehat{\theta}_i, \widehat{\theta}_j\right). \end{aligned}$$

임의의 시점 $t = 1, 2, \ldots, T$에 대해 μ_t의 $100 \times (1-\alpha)\%$ 근사적 신뢰구간(asymptotic confidence interval)은 다음과 같이 표현된다.

$$\mu_t\left(\widehat{\boldsymbol{\theta}}\right) \pm Z_{\frac{\alpha}{2}} \sqrt{\widehat{\mathrm{Var}}\left(\mu_t\left(\widehat{\boldsymbol{\theta}}\right)\right)}.$$

3. 모의실험

제 2장에서 설명한 학습곡선효과의 통계적 모형의 성능을 평가하고 모수의 역할을 확인하기 위해 모의실험을 실시하였다. Table 3.1은 본 연구에서 고려한 여러 모의실험 시나리오이다. 이를 통해 알 수 있듯이, 안정화 단계에 접어든 시점에서의 평균 발생수인 θ_2와 학습효과의 변곡점을 의미하는 θ_3를 여러 값들로 지정하였다. Choi (2013)에서는 학습효과가 발생하기 이전인 초기시점에서의 평균 발생수인 θ_1와 학습곡선의 효과가 변동폭을 담당하는 θ_4를 여러 상황으로 가정하였다. 따라서 이 연구에서 모의실험 결과와 이전 연구의 모의실험 결과를 종합하여 모형의 성능과 모수의 영향력 등을 설명하고자 한다.

3.1. 모의실험 절차

모의실험의 구체적인 절차는 다음과 같다. 편의상 모수의 참값을 $\boldsymbol{\theta}_0 = (\theta_{10}, \theta_{20}, \theta_{30}, \theta_{40})'$로 표현하겠다.

1) 임의의 시점 $t = 1, \ldots, T$에 대해 $\mu_t(\boldsymbol{\theta}_0)$를 계산한다. 즉,

$$\mu_t(\boldsymbol{\theta}_0) = \theta_{10} + (\theta_{20} - \theta_{10})F(t|\theta_{30}, \theta_{40}).$$

2) 단계 1)에서 계산한, 시점 $t = 1, \ldots, T$에 대한 $\mu_t(\boldsymbol{\theta}_0)$를 이용하여 특정 확률분포로부터 표본크기가 T인 B개의 자료, $(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(B)})$를 임의생성한다. 즉, $b = 1, \ldots, B$에 대하여

$$\mathbf{x}^{(b)} = \left(x_1^{(b)}, \ldots, x_t^{(b)}, \ldots, x_T^{(b)}\right)'.$$

여기서, $x_t^{(b)} \sim \text{Poisson}(\mu_t(\boldsymbol{\theta}_0))$ 혹은 $x_t^{(b)} \sim \text{NB}(\mu_t(\boldsymbol{\theta}_0))$이다.

3) 단계 2)에서 생성한 B개의 자료에 학습곡선효과의 통계적 모형을 적용하여 모수를 추정한다. 각 자료로부터 얻은 추정결과를 바탕으로 평균, 중위수(median), 표준편차(standard deviation; SD)를 계산한다. $i = 1, 2, 3, 4$에 대해 추정량의 표준편차는 아래와 같이 나타낸다.

$$\text{SD}_i = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\theta}_i^{(b)} - \widehat{\theta}_i^{(\cdot)}\right)^2}.$$

여기서, $\widehat{\theta}_i^{(b)}$는 b번째 자료로부터 얻은 θ_i의 추정값이고, $\widehat{\theta}_i^{(\cdot)} = (1/B)\sum_{b=1}^{B}\widehat{\theta}_i^{(b)}$이다.

4) 단계 3)의 결과를 이용하여 μ_t의 $100 \times (1-\alpha)\%$ 경험적 신뢰구간(empirical confidence interval)을 계산한다. 경험적 신뢰구간은 매 시점마다 얻은 B개의 $\widehat{\mu}_t$를 정렬하여 사용하는 방법으로 다음과 같이 표현된다.

$$\left(\{\widehat{\mu}_t\}_{\left(\frac{(B+1)\cdot\alpha}{2}\right)}, \{\widehat{\mu}_t\}_{\left((B+1)\left(1-\frac{\alpha}{2}\right)\right)}\right). \tag{3.1}$$

여기서, $\{\widehat{\mu}_t\}_{(j)}$는 $\widehat{\mu}_t$의 j번째 순서통계량이다. 또한, 근사적 신뢰구간은 다음과 같이 구할 수 있다.

$$\left(\mu_t(\widehat{\boldsymbol{\theta}}) - Z_{\frac{\alpha}{2}}\sqrt{\widehat{\text{Var}}(\mu_t(\widehat{\boldsymbol{\theta}}))},\ \mu_t(\widehat{\boldsymbol{\theta}}) + Z_{\frac{\alpha}{2}}\sqrt{\widehat{\text{Var}}(\mu_t(\widehat{\boldsymbol{\theta}}))}\right). \tag{3.2}$$

Table 3.1에 제시된 9개의 모의실험 시나리오에 대하여 $T = 100$, $B = 199$로, 음이항분포에서의 성공횟수는 $m = 100$으로 지정하였다.

3.2. 모의실험 결과

모의실험의 결과를 설명하기에 앞서 몇 가지 용어들을 정리하고자 한다. 포아송분포에서 임의생성된 자료를 포아송자료로, 음이항분포에서 임의생성된 자료는 음이항자료로 각각 정의한다. 또한 포아송 모형은 관측된 발생수가 포아송분포를 따르는 것으로 가정한 통계적 모형으로, 음이항 모형은 관측된 발생수가 음이항분포를 따르는 것으로 가정한 통계적 모형으로 각각 정의한다. Table 3.2에 표기된 SEM은 각 모수 추정값의 평균 표준오차를 의미한다. Table 3.2는 포아송자료를 포아송 모형에, 음이항자료를 음이항 모형에 적합한 결과를 싣고 있다. Figure 3.1와 Figure 3.2는 그에 따른 신뢰구간을 나타내고 있다.

Table 3.2. Estimation results obtained from the true models

S.	$\boldsymbol{\theta}$	$\boldsymbol{\theta}_0$	Poisson model				Negative binomial model			
			Mean	Med	SD	SEM	Mean	Med	SD	SEM
1	$\theta 1$	1	0.947	0.973	0.321	0.366	0.927	0.917	0.352	0.357
	$\theta 2$	5	5.028	4.990	0.327	0.318	5.029	5.026	0.341	0.327
	$\theta 3$	30	30.177	30.143	3.018	3.527	30.276	30.289	2.979	3.227
	$\theta 4$	5	4.846	4.822	2.196	2.897	4.772	4.400	2.723	2.865
2	$\theta 1$	1	0.946	0.971	0.261	0.249	0.937	0.940	0.279	0.274
	$\theta 2$	5	4.994	4.999	0.345	0.354	5.061	5.047	0.358	0.382
	$\theta 3$	40	39.877	39.820	2.933	3.047	40.187	40.033	3.162	3.121
	$\theta 4$	5	4.826	4.763	2.538	2.520	5.370	4.775	3.210	2.618
3	$\theta 1$	1	1.000	1.007	0.190	0.166	0.967	0.965	0.182	0.166
	$\theta 2$	5	5.105	5.048	0.568	0.541	5.075	4.983	0.536	0.524
	$\theta 3$	60	60.825	60.434	3.453	3.258	59.963	59.497	3.445	3.190
	$\theta 4$	5	4.841	4.509	2.755	2.406	4.731	4.302	2.976	2.330
4	$\theta 1$	1	0.960	0.970	0.342	0.368	0.931	0.952	0.329	0.377
	$\theta 2$	10	10.017	9.997	0.439	0.439	10.037	10.036	0.449	0.462
	$\theta 3$	30	30.023	29.852	1.744	1.699	29.954	30.069	1.781	1.722
	$\theta 4$	5	5.017	4.797	1.565	1.497	5.100	5.083	1.504	1.540
5	$\theta 1$	1	0.988	0.979	0.245	0.248	1.014	1.016	0.255	0.252
	$\theta 2$	10	10.042	10.007	0.472	0.484	10.023	10.030	0.509	0.503
	$\theta 3$	40	39.975	40.056	1.937	1.652	39.553	39.827	2.481	1.692
	$\theta 4$	5	4.911	4.745	1.476	1.338	4.802	4.837	1.434	1.371
6	$\theta 1$	1	0.991	0.979	0.152	0.166	0.989	0.992	0.173	0.167
	$\theta 2$	10	10.117	10.113	0.640	0.671	10.016	9.998	0.648	0.693
	$\theta 3$	60	60.162	60.061	2.009	1.844	60.001	59.764	1.819	1.890
	$\theta 4$	5	5.036	4.875	1.356	1.314	4.949	4.826	1.429	1.330
7	$\theta 1$	1	1.002	1.023	0.365	0.381	0.952	0.944	0.423	0.388
	$\theta 2$	20	20.003	19.976	0.601	0.613	19.991	20.014	0.685	0.670
	$\theta 3$	30	29.993	30.051	1.048	1.062	30.000	29.919	1.123	1.125
	$\theta 4$	5	4.952	4.878	0.906	0.876	5.030	4.976	0.961	0.916
8	$\theta 1$	1	0.953	0.944	0.257	0.251	0.946	0.977	0.265	0.251
	$\theta 2$	20	20.032	20.007	0.640	0.677	20.014	20.005	0.736	0.739
	$\theta 3$	40	40.052	40.004	1.043	1.074	39.969	39.935	1.095	1.142
	$\theta 4$	5	5.031	5.067	0.826	0.802	4.999	4.939	0.893	0.836
9	$\theta 1$	1	0.998	0.998	0.184	0.167	0.986	0.988	0.173	0.168
	$\theta 2$	20	20.095	20.020	0.853	0.907	20.099	20.147	0.995	0.995
	$\theta 3$	60	60.039	60.081	1.115	1.159	59.996	59.989	1.290	1.251
	$\theta 4$	5	4.907	4.869	0.733	0.762	4.985	4.895	0.785	0.807

S., scenario; $\boldsymbol{\theta}$, parameter vector; $\boldsymbol{\theta}_0$, true parameter vector; Med, Median; SD, standard deviation; SEM, mean of standard errors.

각 자료에 대해 참 모형(true model)을 적용한 경우, 시나리오에 상관없이 평균과 중위수가 모수의 참값과 상당히 유사하다는 것을 확인할 수 있다. 또한 모수 추정값의 변동성을 살펴보면, 추정값의 표준편차와 평균 표준오차가 매우 유사한 결과를 나타내고 있다. 그리고 θ_3의 변동성이 다른 모수들에 비해 큰 것으로 나타나 변곡점의 위치가 민감하게 영향을 받고 있음을 알 수 있다. 포아송자료를 포아송 모형에 적합시킨 추정결과에 대해, 시나리오 1–3은 θ_2가 5일 때 θ_3가 30, 40, 60인 경우인데, θ_3가 증가함

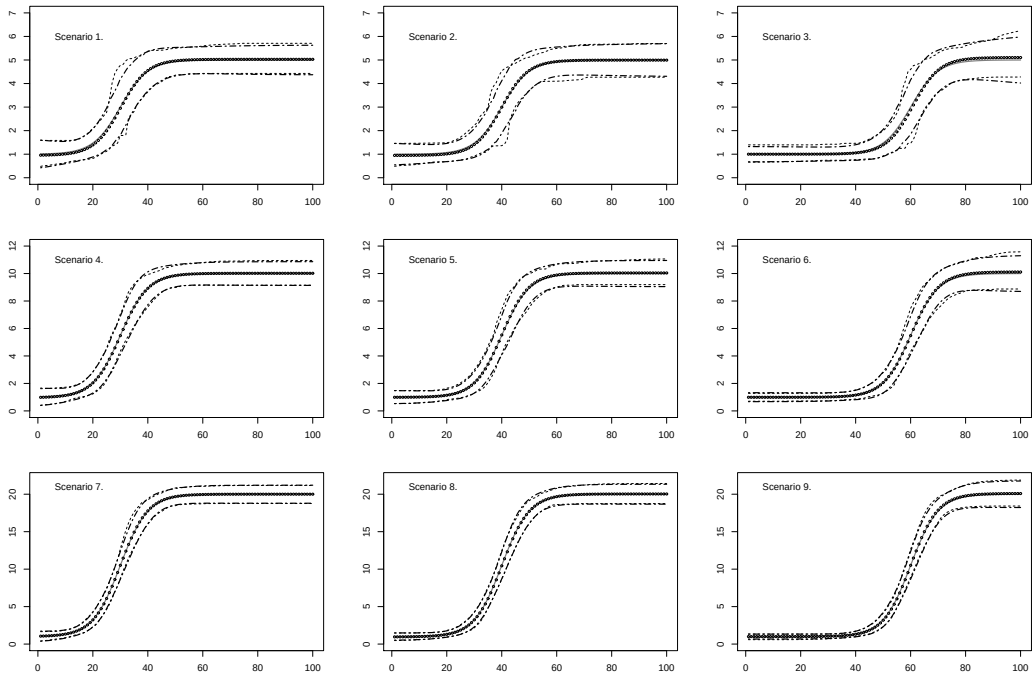


Figure 3.1. Plots of the 95% confidence intervals obtained from Poisson model(solid line, true value; points joined by lines, estimate; dotted line, asymptotic confidence interval; thick dashed line, empirical confidence interval)

에 따라 θ_1의 변동성은 줄어들게 되나, θ_2의 변동성은 증가하는 경향을 보인다. 그리고 θ_3를 고정시킨 경우(시나리오 1, 4, 7)를 살펴보면, θ_1과 θ_2가 증가함에 따라 θ_2의 변동성은 증가하나, θ_3와 θ_4의 변동성은 감소한다. 음이항 모형을 음이항자료에 적합시킨 결과를 살펴보자. 우선 θ_2를 고정시킨 경우(시나리오 4–6), θ_3가 증가함에 따라 θ_1과 θ_4의 변동성은 감소하나, θ_3의 변동성은 증가하게 된다. 다음으로 θ_3를 고정시킨 경우(시나리오 2, 5, 8)에 대해 θ_2가 증가함에 따라 θ_2의 변동성은 증가하나 θ_3와 θ_4의 변동성은 감소하는 경향을 보인다.

이전 연구의 결과와 함께 살펴보면, 두 연구 모두 각 자료에 대해 참 모형을 적용한 경우, 시나리오에 상관없이 평균과 중위수가 모수의 참값과 매우 유사하게 나타났다. 또한 대체적으로 θ_3의 변동성(분산)이 가장 크게 나타났다. 먼저 포아송자료를 포아송 모형에 적합시킨 결과, θ_1과 θ_4가 증가함에 따라 모든 모수의 분산이 증가하는 경향을 보였다. 반면 θ_2가 증가하는 경우에는 θ_1과 θ_2의 분산은 증가하였으나, θ_3과 θ_4의 분산은 감소하는 경향을 나타냈다. 또한 θ_3가 증가하는 경우에는 θ_1의 분산은 감소하는 반면, θ_2의 분산은 감소하는 것을 확인할 수 있었다. 다음으로 음이항자료를 음이항 모형에 적합시킨 결과에 대해 확인해 보면, θ_1이 증가함에 따라 θ_1, θ_3, 그리고 θ_4의 분산은 증가하였고, θ_4가 증가함에 따라 모든 모수의 분산은 증가하는 경향을 나타냈다. θ_2가 증가하는 경우에는 θ_2의 분산은 증가하나, θ_3와 θ_4의 분산은 감소하는 것을 확인할 수 있다. 또한 θ_3가 증가하는 경우에는 θ_1과 θ_4의 분산은 감소하는 반면 θ_3의 분산은 증가하는 경향을 나타냈다.

이제 포아송자료와 음이항자료를 각각 참모형으로 적합한 후 얻은 신뢰구간들에 대해 설명하고자 한다. Figure 3.1과 Figure 3.2는 각 시점마다 얻은 $B = 199$개의 $\widehat{\mu}_t$의 경험적 신뢰구간(식 (3.1) 참조)과 추정된 모수의 분산-공분산행렬로부터 계산한 근사적 신뢰구간(식 (3.2) 참조)을 평균 발생수의 평균과 그

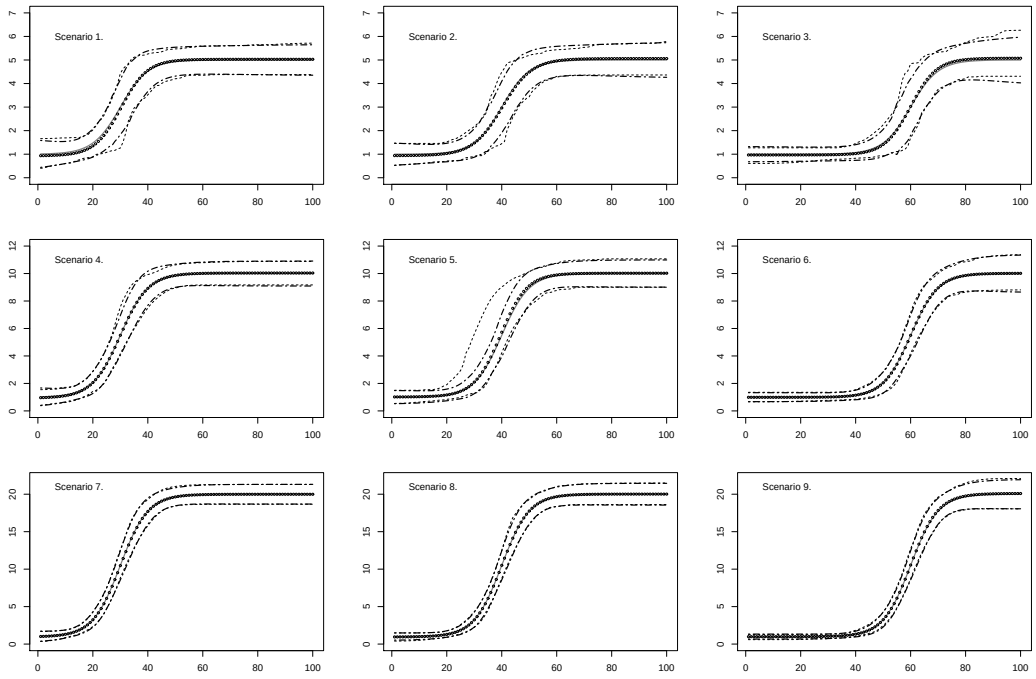


Figure 3.2. Plots of the 95% confidence intervals obtained from Negative binomial model(solid line, true value; points joined by lines, estimate; dotted line, asymptotic confidence interval; thick dashed line, empirical confidence interval)

참값을 보여주고 있다. 먼저 포아송 모형의 결과를 Figure 3.1로 살펴보면 전체적으로 평균 발생수가 참값과 구분이 되지 않을 정도로 매우 유사하게 추정되었음을 알 수 있다. 안정화단계에 접어든 이후 시점의 평균 발생수가 작은 경우(시나리오 1–3)에 변곡점 부근에서 경험적 신뢰구간의 폭이 근사적 신뢰구간보다 크다는 것을 제외하면 모든 시나리오 하에서 경험적 신뢰구간과 근사 신뢰구간이 매우 유사한 경향을 보이는 것으로 나타났다. Figure 3.2 또한 유사한 경향을 보이고 있다.

지금까지의 모의실험 결과는 포아송분포 혹은 음이항분포에서 임의생성한 자료를 각각의 참 모형에 적합한 것이다. 하지만 실제자료분석에서는 주어진 자료의 확률분포가 정확히 무엇인지 알 수 없으므로 현실적으로 적용가능한 예비 모형들을 모두 적용할 수 밖에 없고 여러 모형들 중에서 최선의 모형을 선택해야 할 것이다. 결국 참 모형에 대한 정보가 부족할 경우 어떠한 모형이 강건(robust)한지를 살펴볼 필요가 있다. 따라서 포아송자료를 포아송 모형(참 모형) 뿐만 아니라 음이항 모형(거짓 모형)에, 음이항자료 또한 음이항 모형(참 모형)과 포아송 모형(거짓 모형)에 모두 적합시킬 필요가 있다.

참 모형 선택비율 평가에서는 참 모형과 거짓 모형을 자료에 적용하여 각 모형의 AIC(Akaike information criterion)를 기준으로 어떠한 통계적 모형이 다른 모형에 비해 강건한지를 살펴보았다. Table 3.3은 $B = 199$개의 자료에 대해 특정 분포로부터 생성된 자료가 참 모형에 의해 얼마나 잘 설명되는지를 모형선택 비율의 형태로 나타내었다. 예를 들면, 시나리오 1의 포아송자료의 AIC를 살펴보자. AIC에 의한 정분류율이 63.3%로서 이는 전체 199개의 포아송자료 중 126개의 자료가 음이항 모형(거짓 모형)보다는 포아송 모형(참 모형)이 더 작은 AIC값을 가졌음을 의미한다. 즉, 약 63.3%가 자료의 분포와 동일한 모형(참 모형)에 보다 잘 적합하였다는 것을 알 수 있다. 포아송자료의 경우를 살펴보면,

Table 3.3. Empirical percentages of choosing the true models under each of the scenarios

Dist.	Scenario								
	1.	2.	3.	4.	5.	6.	7.	8.	9.
PO	126 (63.3)	130 (65.3)	127 (63.8)	126 (63.3)	142 (71.4)	133 (66.8)	152 (76.4)	152 (76.4)	148 (74.4)
NB	89 (44.7)	77 (38.7)	69 (34.7)	105 (52.8)	92 (46.2)	89 (44.7)	123 (61.8)	114 (57.3)	112 (56.3)

Data are expressed as number of data sets (percentage); Dist., distribution; PO, Poisson; NB, Negative binomial.

모든 시나리오에서 AIC에 의한 정분류율이 매우 유사하게 나타났고 최소 63%에서 최대 77%인 것으로 나타났다. 음이항자료의 경우 또한 모든 시나리오에서 AIC에 의한 정분류율은 큰 차이를 보이지 않았다. 하지만 대체적으로 θ_2와 θ_1의 차이가 작을수록 참 모형보다는 거짓 모형이 선택될 가능성이 55%를 상회하는 것으로 나타났다.

Choi (2013)의 결과와 본 연구의 결과를 종합적으로 판단해보자. 두 연구 모두 참 모형에 대해서 모든 시나리오에서 AIC에 의한 정분류율은 큰 차이는 보이지 않았다. θ_1과 θ_4의 값에 변화를 준 연구에서는 포아송자료의 정분류율이 최소 72% 에서 최대 85%인 것으로 나타났고, 음이항자료의 정분류율은 최소 51% 에서 최대 65% 로 나타났다. 결국 음이항자료의 AIC에 의한 정분류율이 두 연구에서 모두 작게 나타났다. 따라서 두 통계적 모형들 중 포아송 모형이 음이항 모형에 비해 참 분포의 정보가 부족한 상태에서도 적용가능하다고 판단된다.

4. 실제자료분석

실제자료분석을 통해 학습곡선효과의 통계적 모형을 평가하고자 한다. 분석에 사용된 자료는 국가통계포털에 공시되어 있는 자료로서, 특정 사망원인에 대한 사망자수이다. 여러 사망원인들 중 시간이 지남에 따라 사망자 수가 증가하거나 혹은 감소하는 패턴을 보이는 자료를 본 연구에서 사용하였다. 분석에 사용된 자료는 다음의 2개 자료로서, 1995년 1월부터 2012년 12월(전체 216개월)까지 특정질환으로 인한 사망자 수이다: Data 1. 인체 면역결핍 바이러스로 인한 사망자 자료(범위: 0−23명); Data 2. 자궁목의 악성신생물인 사망자 자료(범위: 26−114명). 각 자료에 포아송 모형과 음이항 모형을 적용하여 최대우도추정법으로 모수를 추정하고 근사적 신뢰구간을 계산하였다. 또한 교차타당법(leave-one-out cross-validation method; CV(1))을 적용하여 모형의 신뢰성을 평가하였다. CV(1)를 이용하여, 전체 시점(T) 중 임의의 한 시점(t)를 제외한 $(T-1)$개의 시점에서 관측한 자료만으로 모형을 적합하고 이를 이용하여 모형적합에 제외된 t 시점의 예측값을 산출한다. 예를 들어, 시점 $t=1$의 자료를 제외한, 시점 2부터 시점 T까지의 자료를 이용하여 모수를 추정한 후, 시점 1에서의 예측값을 계산한다. 모든 시점에 대해서 동일한 방법으로 추정값을 구한 후, 학습곡선효과를 모형화한다. 교차타당법에서의 각 추정값에 대한 변동성 평가를 위한 표준오차는 다음과 같이 나타낸다.

$$\mathrm{SE}_i = \sqrt{\frac{T-1}{T}\sum_{t=1}^{T}\left(\widehat{\theta}_{i(-t)} - \widehat{\theta}_{i(\cdot)}\right)^2}.$$

여기서, $\widehat{\theta}_{i(-t)}$는 시점 t의 자료를 제외하고 분석한 결과의 i번째 모수 추정값이고 $\widehat{\theta}_{i(\cdot)} = \sum_{t=1}^{T}\widehat{\theta}_{i(-t)}$이다. Table 4.1에서의 SEM는 Table 3.2의 그것과 마찬가지로, 교차검증에 사용된, 표본크기가 $T-1$인 자료

Table 4.1. Estimation results of real application

			Poisson model			Negative binomial model		
		Parm.	Est.	SE(SEM)	P-value	Est.	SE(SEM)	P-value
Data 1		θ_1	0.680	0.595	0.255	0.692	0.600	0.250
		θ_2	11.005	1.347	<.001	10.977	1.382	<.001
		θ_3	116.920	10.944	<.001	116.740	11.250	<.001
		θ_4	38.388	10.749	<.001	38.139	10.948	<.001
		AIC	913.8			914.9		
		BIC	927.3			928.4		
	CV	θ_1	0.679	0.541 (0.597)		0.691	0.536 (0.602)	
		θ_2	11.007	1.302 (1.352)		10.979	1.289 (1.387)	
		θ_3	116.927	10.536(10.982)		116.748	10.446(11.289)	
		θ_4	38.397	10.368(10.785)		38.148	10.292(10.984)	
		RMS	2.312			2.312		
Data 2		θ_1	53.130	1.129	<.001	53.006	1.413	<.001
		θ_2	83.462	0.827	<.001	83.467	1.122	<.001
		θ_3	76.386	2.390	<.001	76.102	3.078	<.001
		θ_4	6.895	1.480	<.001	7.191	1.921	<.001
		AIC	1602.5			1617.1		
		BIC	1616.0			1630.6		
	CV	θ_1	53.130	1.494(1.132)		53.006	1.503(1.416)	
		θ_2	83.462	0.976(0.829)		83.467	0.977(1.124)	
		θ_3	76.385	3.427(2.396)		76.102	3.416(3.085)	
		θ_4	6.894	1.902(1.484)		7.190	1.920(1.926)	
		RMS	10.256			10.259		

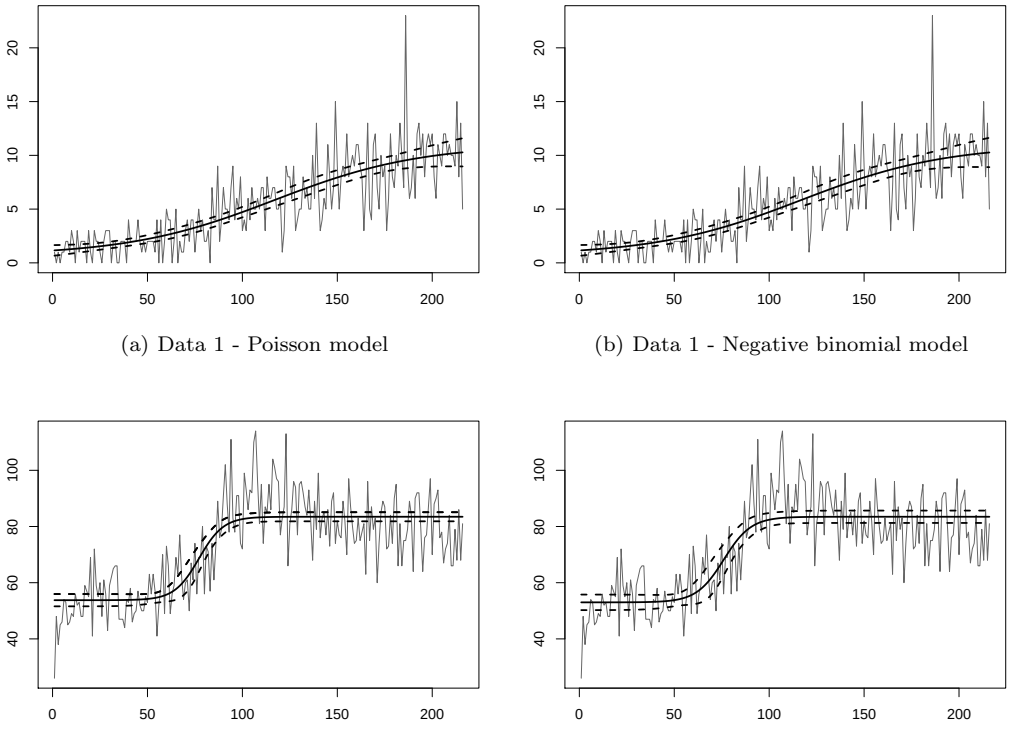
Parm., parameter; Est., estimate; CV, leave-one-out cross-validation; SE, standard error; SEM, mean of standard errors; C.I., confidence interval; RMS, RMSECV.

들로부터 얻은 각 추정값의 평균 표준오차로 계산하였다. 이와 더불어 교차타당법에 의한 평균제곱근오차(root mean squares of error of the cross-validation; RMSECV)를 고려하였고 이는 다음과 같이 정의한다.

$$\text{RMSECV} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(X_t - \widehat{\mu}_{(-t)}\right)^2}.$$

여기서, $\widehat{\mu}_{(-t)}$는 t 시점의 자료를 제외하고 구한 모수 추정값, $\{\widehat{\theta}_{i(-t)}\}$을 이용한 적합값이다.

Table 4.1는 2개의 실제자료에 대한 분석결과이다. 먼저 최대우도추정법에 대해서, 모든 자료에 있어서 통계적 모형에 따른 모수 추정값들은 매우 유사한 결과를 보이고 있다. 그리고 추정값의 표준편차와 평균 표준오차가 매우 유사한 결과를 나타내고 있다. 하지만 모든 자료에 대해 AIC와 BIC 관점에서 포아송 모형이 음이항 모형에 비해 설명력이 뛰어나고 각 모수의 추정값에 대한 변동성 역시 작음을 확인하였다. 인체 면역결핍 바이러스로 인한 사망자 자료(Data 1)에 대해 살펴보면, 포아송 모형의 경우 초기 시점에서의 평균 사망자수는 0.680명으로, 안정화단계에서의 평균 사망자수는 11.005명으로 추정되었고, 음이항 모형을 적용한 경우에도 초기 시점에서의 사망자수는 평균 0.692명으로, 안정기에 접어든 이후 시점에서의 사망자수는 평균 10.977명으로 나타나 모형들 간의 차이는 없다고 판단된다. 사망원인이 자궁목의 악성신생물인 사망자 자료(Data 2)에 포아송 모형을 적용한 결과 초기 사망자수는 53.130명,



(a) Data 1 - Poisson model

(b) Data 1 - Negative binomial model

(c) Data 2 - Poisson model

(d) Data 2 - Negative binomial model

Figure 4.1. Plots of the 95% asymptotic confidence intervals from real data analysis

안정화 이후 시점에서의 사망자수는 83.462명으로 추정되었다. 그리고 음이항모형에 적용한 결과 초기 사망자수와 안정화 된 이후의 사망자수 역시 거의 동일한 값으로 추정되었다.

교차타당법에 의한 분석결과, 모든 자료에 있어서 통계적 모형에 따른 모수 추정값들은 매우 유사한 결과를 보이고 있으나 RMSECV의 관점에서 인체 면역결핍 바이러스로 인한 사망자 자료(Data 1) 에 대해서는 동일한 값이 나타났고, 사망원인이 자궁목의 악성신생물인 사망자 자료(Data 2)에 대해서는 포아송 모형이 음이항 모형에 비해 더 작은 값이 나타났다. 인체 면역결핍 바이러스로 인한 사망자 자료(Data 1)에 대해 살펴보면, 포아송 모형의 경우 학습곡선효과가 발생하여 안정기로 접어드는 변곡점이 116.927번째 개월로, 학습곡선효과가 일어나는 기간이 38.397개월로 나타났다. 사망원인이 자궁목의 악성신생물인 사망자 자료(Data 2)에 대해 살펴보면, 음이항 모형의 경우 학습곡선효과가 발생하여 안정기로 접어드는 변곡점이 76.102번째 개월로, 학습곡선효과가 일어나는 기간이 7.190개월로 나타났다. 표준오차 관점에서 살펴보면, 인체 면역결핍 바이러스로 인한 사망자 자료(Data 1)의 표준오차는 분석 방법과 모형에 상관없이 비슷하게 나타났다. 또한 사망원인이 자궁목의 악성신생물인 사망자 자료(Data 2)의 표준오차 역시 분석 방법과 모형에 상관없이 유사한 결과가 나타났다.

Figure 4.1는 최대우도추정법을 이용하여 구한 포아송모형과 음이항모형의 평균 사망자수와 95% 신뢰구간이다. 95% 신뢰구간은 최대우도추정법을 통하여 구한 분산-공분산행렬로부터 계산되었다. Figure 4.1(a)와 Figure 4.1(b)는 인체 면역결핍 바이러스로 인한 사망자 자료(Data 1)를 각각 포아송 모형과 음이항 모형에 적용시킨 결과이다. 평균 사망자수가 실제자료에 잘 적합하는 것을 확인할 수 있고, 안정

화 된 이후 시점에서는 초기 시점에 비해 신뢰구간이 넓어지는 것을 확인할 수 있다. 다음으로 Figure 4.1(c)와 Figure 4.1(d)는 사망원인이 자궁목의 악성신생물인 사망자 자료(Data 2)를 각각 포아송 모형과 음이항 모형에 적용시킨 결과이다. 이 자료에 있어서도 평균 사망자수가 잘 적합되었음을 확인할 수 있다. 또한 자료에 따라 학습곡선효과의 형태에 차이가 있음을 확인할 수 있는데, Data 1에서는 초기 시점에서의 사망자수와 안정화 이후 시점에서의 사망자수가 크게 차이 나지 않으며 학습곡선 효과가 일어나는 기간이 길어 완만하게 증가하지만 Data 2의 경우에는 초기 시점에서의 사망자수와 안정화 이후 시점에서의 사망자수가 차이가 나며 학습곡선효과가 일어나는 기간이 짧아 Data 1에 비해 가파르게 증가하는 것을 확인할 수 있다.

5. 결론

본 연구에서는 일정한 시간 간격을 두고 시점이 흘러가면서 특정 사건의 발생수가 증가하다가 안정화되는 S자 형태에 대해 모형화하였다. 이산형 분포로는 포아송분포와 음이항분포를 이용하였고, 로지스틱분포의 누적분포함수를 이용하여 학습곡선효과의 통계적 모형을 소개하였다. 모의실험에서는 안정화 단계에 접어든 시점에서의 평균 발생수인 θ_2와 학습효과의 변곡점을 의미하는 θ_3의 값을 변형하여 모형의 성능과 영향력 등을 평가하였다. 모수 추정 방법으로는 최대우도추정법를 이용하였다.

모의실험 결과, 각 자료에 대해 참 모형을 적용한 경우 시나리오에 상관없이 평균과 중위수가 모수의 참 값과 상당히 유사하게 나타났으며, 대체적으로 θ_3의 변동성(표준편차와 평균 표준오차)이 가장 크게 나타났다. 참 모형에 자료를 적합시킨 결과, θ_2가 증가함에 따라 포아송자료와 음이항자료 모두 θ_3와 θ_4의 변동성이 감소하였고, θ_3가 증가하는 경우에는 θ_1의 변동성은 감소하였으나 θ_2와 θ_3의 변동성은 증가하였다.

다음으로 AIC에 의한 정분류율을 평가한 결과, 포아송자료의 경우 모든 시나리오에서 AIC에 의한 정분류율이 매우 유사하게 나타났고 최소 63%에서 최대 77%로 나타났다. 음이항자료의 경우 역시 AIC에 의한 정분류율은 큰 차이를 보이지 않았으나 θ_1 와 θ_2의 차이가 작을수록 참 모형보다는 거짓 모형이 선택될 가능성이 상대적으로 높게 나타났다. 따라서 두 통계적 모형들 중 포아송 모형이 음이항 모형에 비해 참 분포의 정보가 부족한 상태에서도 적용가능하다고 판단된다.

실제자료분석에서는 인체 면역결핍 바이러스로 인한 사망자 자료(Data 1)와 사망원인이 자궁목의 악성신생물인 사망자 자료(Data 2)를 이용하였다. 포아송 모형과 음이항 모형을 적용하여 최대우도추정법으로 모수를 추정하였고, 교차타당법을 적용하여 모형의 신뢰성을 평가하였다. 실제자료분석 결과, 분석 방법과 모형에 상관없이 추정값과 표준오차가 유사하게 나타났다. 하지만 AIC와 BIC 관점에서 포아송 모형이 음이항 모형에 비해 설명력이 뛰어났고, RMSECV 관점에서 Data 2의 경우 포아송 모형이 음이항 모형에 비해 포아송 모형이 더 좋음을 확인할 수 있었다.

실제자료분석에서 사용된 자료는 특정기간동안 발생한 이산형 자료들 중 시간이 흘러가면서 발생건수(횟수)가 증가하는 경향을 가졌다. 이는 동일한 조건이나 동일한 기간 하에서 관측된 자료라는 측면을 제외하면 작업자의 경험이나 노력 등이 학습곡선효과에 영향을 주는 부분은 고려할 수 없으므로 엄밀한 의미에서는 이러한 자료에 학습곡선효과를 적용하는 데에 다소 문제가 될 수 있다는 것이 이 연구의 제한점이다.

References

Back, W.J. (2008). Cost effective analysis of the fuel cell with a learning curve, *Graduate School of Dongshin University*, master's thesis.

Chen, W., Sailhamer, E., Berger, D. L., and Rattner, D. W. (2007). Operative time is a poor surrogate for the learning curve in laparoscopic colorectal surgery, *Surgical Endoscopy*, **21**, 238–243.

Choi, M. J. (2013). Modeling of the learning curve on the count responses, *Graduate School of Sungshin Women's University*, master's thesis.

Ferguson, G. G., Ames, C. D., Weld, K. J., Yan, Y., Venkatesh, R., and Landman, J. (2005). Prospective evaluation of learning curve for laparoscopic radical prostatectomy: identification of factors improving operative times, *Adult urology*, **66**, 840–844.

Han, H. J., Choi, S. B., Park, M. S., Lee, J. S., Kim, W. B., Song, T. J., and Choi, S. Y. (2011). Learning curve of single port laparoscopic cholecystectomy determined using the non-linear ordinary least squares method based on a non-linear regression model: An analysis of 150 consecutive patients, *Journal of hepato-biliary-pancreatic sciences.*, **18**, 510–515.

Hayn, M. H., Hussain, A., Mansour, A. M., Andrews, P. E., Carpentier, P., Castle, E., Dasgupta, P., Rimington, P., Thomas, R., Khan, S., Kibel, A., Kim, H., Manoharan, M., Menon, M., Mottrie, A., Ornstein, D., Peabody, J., Pruthi, R., Redorta, J. P., Richstone, L., Schanne, F., Stricker, H., Wiklund, P, handrasekhar, R., Wilding, G. E., and Guru, K. A. (2010). The learning curve of robot-assisted radical cystectomy: results from the International Robotic Cystectomy Consortium, *European Urology*, **58**, 197–202.

Hong, J. I. (2007). A study on user's learnability evaluation method using learning curve model, *Graduate School of Korea University of Technology and Education*, master's thesis.

Jaffe,J., Castellucci, S., Cathelineau, X., Harmon, J., Rozet, F., Barret, E., and Vallancien, G. (2009). Robot-assisted laparoscopic prostatectomy: a single-institutions learning curve, *Urology*, **73**, 127–133.

Jeff, F. L., Melissa F., and Huang J. Q. (2014). Learning curve analysis of the first 100 robotic-assisted laparoscopic hysterectomies performed by a single surgeon, *International Journal of Gynecology and Obstetrics*, **124**, 88–91.

Korean Statistical Information Service. http://kosis.kr/statisticsList

Lee, S. J. and Park, M. S. (2012). Statistical modeling of learning curves with binary response data, *Journal of the Korean Statistical Society*, **19**, 433–450.

Park, S. S., Kim, M. C., Park, M. S., and Hyung, W. J. (2012). Rapid adaptation of robotic gastrectomy for gastric cancer by experienced laparoscopic surgeons, *Surgical Endoscopy*, **26**, 60–67.

Pruthi, R. S., Smith, A., and Wallen, E. M. (2008). Evaluating the learning curve for robot-assisted laparoscopic radical cystectomy, *Journal of Endourology*, **22**, 2469–2474.

Schreuder, H. W., Zweemer, R. P., van Baal, W. M., van de Lande J., Dijkstra, J. C., and verheijen, R. H. (2010). From open radical hysterectomy to robot-assisted laparoscopic radical hysterectomy for early stage cervical cancer: aspects of a single institution learning curve, *Journal of Gynecologic Surgery* , **7**, 253–258.

개수형 자료에 대한 학습곡선효과의 모형화

최민지[a] · 박만식[a,b,1]

[a]성신여자대학교 통계학과, [b]성신여자대학교 통계연구소

(2014년 3월 3일 접수, 2014년 4월 25일 수정, 2014년 6월 9일 채택)

요 약

일반적으로 특정한 작업에 익숙해진다는 것은 그 작업에 투입되는 노력에 비해 산출되는 성과가 보다 뚜렷해진다는 것을 의미한다. 동일한 양이나 정도의 노력을 들여 특정한 작업을 반복적으로 수행하게 되면 초기 시점보다 원하는 성과를 기대 이상으로 얻게 된다는 것을 의미한다. 이를 학습곡선효과(learning-curve effects)'라고 한다. 본 연구에서는 특정한 작업을 반복시행한 결과가 개수형인 형태로 측정되는 변수에 대해 (역)S자 형태를 가지는 통계적 모형을 적용하고자 한다. 다양한 모의실험 하에서의 모형의 성능을 평가하고 특정질환으로 인한 사망자 자료에 적합하였다.

주요용어: 포아송분포, 음이항분포, 학습곡선, 누적분포함수, 최대우도추정법.

이 논문은 2012년도 성신여자대학교 학술연구조성비 지원에 의하여 연구되었음.

[1]교신저자: (136-742) 서울특별시 성북구 보문로 34다길 2, 성신여자대학교 자연과학대학 통계학과.
E-mail: mansikpark@sungshin.ac.kr