

결측값이 있는 가산 자료에서 이산형 와이블 회귀 모형에 관한 연구[†]

유한나¹

¹부산외국어대학교 컴퓨터 소프트웨어학부

접수 2018년 12월 18일, 수정 2019년 1월 7일, 게재 확정 2019년 1월 11일

요약

이산형 와이블 회귀모형 (discrete Weibull regression model) 은 셀 수 있는 가산 자료 (discrete count data) 의 분포 형태와 상관없이 적용시킬 수 있는 모형이다. 자료의 산포 종류에 상관없이 모두 적용 가능하다는 장점이 있지만 가산 자료에 대해 이를 적용한 논문들이 많지 않다. 본 연구에서는 이산형 와이블 회귀모형을 결측치가 있는 가산 자료에 적용하여 보았다. 국민건강영양조사 제 7기 1차년도 (2016) 자료를 바탕으로 단일 대체법 (single imputation)을 이용해 결측치들을 대체한 후 이산형 와이블 회귀모형과 영과잉 포아송 모형 (zero-inflated Poisson model)을 비교한 결과 두 모형 중 이산형 와이블 회귀모형이 더 잘 적합이 되었다. 또한 모의 실험을 통하여 여러 다양한 세팅 하에 결측치를 대체한 경우와 대체하지 않은 경우의 편향 (bias) 을 비교해 단일 대체법을 사용하는 것이 편향정도를 줄일 수 있는 것을 확인하였다. 따라서 결측치가 있는 가산 자료에 결측치를 대체한 후 이산형 와이블 회귀모형에의 적합은 검정력은 물론 모형의 적합도도 높일 수 있을 것으로 사료된다.

주요용어: 가산 자료, 국민건강영양조사, 대체 방법, 이산형 와이블 회귀모형.

1. 서론

일정한 시간 혹은 공간에서 발생한 사건의 수로 정의되는 셀 수 있는 가산 자료 (discrete count data) 는 여러 분야에서 발생한다. 이러한 가산 자료에 대해 일반적으로 가장 널리 적용하는 모형은 포아송 모형 (Poisson model) 이다. 포아송 모형은 적용하기 쉽다는 장점이 있는 반면 기댓값과 분산이 동일해야 한다는 강한 가정 때문에 실제 데이터에 잘 맞지 않는 경우가 많다. 자료가 과대 산포 (over dispersion) 혹은 과소 산포 (under dispersion) 인 경우 포아송 모형의 적용은 추정치의 편향 (bias)을 가져오고 잘못된 결론으로 이어질 수 있다. 음이항 모형 (negative Binomial model) 은 자료가 과대 산포되어 있을 때 많이 사용되지만 자료가 심하게 치우쳐져 있는 경우에는 적합이 잘 되지 않는다. 영 (zero)을 과도하게 포함하고 있는 영과잉 자료 (zero inflated data) 도 이산형 자료에 많이 쓰는 모형 중 하나이다. Chun (2017) 은 보험 자료에 대해 영 과잉 음이항 회귀분석을 적용하였고 Kim (2003) 은 영 과잉 자료에 대해 영과잉 모형과 포아송 회귀모형 그리고 의사결정나무 모형을 비교 분석하였다. 또한 Kim 등 (2017) 에서는 영과잉 허들 모형을 이용하여 그러나 이러한 영과잉 포아송 모형은 자료에 영이 과도하게 많은 경우에 적용이 되고 과소 산포인 경우에는 적합이 잘 되지 않는다. 반대로 과소 분포 되어 있는

[†] 이 논문은 2018년도 부산외국어대학교 학술연구조성비의 재원으로 수행된 연구임.

¹ (46234) 부산광역시 금정구 금샘로 485, 부산외국어대학교 컴퓨터 소프트웨어학부, 조교수.
Email: pinkcan78@buefs.ac.kr

자료의 경우 일반화 포아송 회귀모형 (generalised Poisson regression) (Efron, 1986; Famoye, 1993) 또는 Conway-Maxwell Poisson (COM-Poisson) 회귀모형 (Sellers 등, 2010) 이 많이 적용되기도 하는데 모형이 다소 복잡하고 계산과정이 복잡하다는 단점이 있다 (Chanalidis 등, 2017). 본 연구에서 사용한 이산형 와이블 모형 (discrete Weibull model)은 이산 자료가 과대 산포 혹은 과소 산포인 경우에 모두 적용할 수 있는 모형으로 Nakagawa 등 (1975)에서 처음 소개되었다. 이후 이산형 와이블 분포에 대한 모수 추정에 대한 논문들 (Khan 등, 1989; Kulasekera 등 1994)이 몇 편 발표되었고 최근에 Klakattawi 등 (2018)에서는 이산형 와이블 모형을 실제 여러 데이터들에 적용하여 데이터가 과대 산포., 과소 산포를 보이는 경우 다른 모형들에 비해 적합이 잘 됨을 보였다. Peluso 등 (2018)에서는 이산형 와이블 분포에 기반한 일반화 가법 모형 (discrete Weibull generalised additive model)을 적용하여 각 모수들을 공변량에 대한 선형, 비선형인 형태로 모델링하여 모의실험결과를 통해 모형의 다양한 확장 가능성을 보여주었다. 이렇듯 이산형 와이블 모형에 대한 소개는 오래 전에 있었지만 이 모형을 기반으로 한 논문은 많지 않고 더더욱 자료에 결측치들이 있을 때 이를 고려한 논문은 없다.

본 연구 논문에서는 결측치가 있는 자료에서 단일 대체 방법 (single imputation)을 이용해 결측치들을 대체한 후 이산형 와이블 회귀모형을 이용하여 셀 수 있는 가산 자료에 대한 분석을 하였다. 결측치 대체를 통해 분석의 검정력을 높일 수 있고 이산형 와이블 모형을 통해 자료의 다양한 분포에 적용 가능한 것을 보였다. 2절에서는 Klakattawi 등 (2018)에서 설명한 이산형 와이블 회귀모형을 소개하고 본 연구에 사용한 단일 결측치 대체 방법을 설명하였다. 3절에서는 국민건강영양조사 제 7기 1차년도 (2016) 자료를 바탕으로 분석한 결과를 기술하였다. 4절에서는 모의실험 결과를 설명하였고 연구 결과에 대한 결론과 향후 연구의 방향은 5절에서 다루었다.

2. 결측값이 있는 이산형 와이블 회귀 모형

2.1. 이산형 와이블 분포

확률변수 Y 가 (type 1) DW 분포를 따를 때 기호로는 다음과 같이 나타낸다.

$$Y \sim DW(q, \beta),$$

여기서 모수들은 $0 < q < 1$ 와 β 이고 Y 의 누적분포함수는 아래와 같이 나타낼 수 있고,

$$F(y; q, \beta) = \begin{cases} 1 - q^{(y+1)^\beta} & y = 0, 1, 2, \dots \\ 0 & y < 0. \end{cases}$$

확률 질량함수는 다음과 같다.

$$f(y; q, \beta) = q^{y^\beta} - q^{(y+1)^\beta} \quad \text{for } y = 0, 1, 2, \dots,$$

$f(0) = 1 - q$ 이므로 모수 q 는 확률변수 Y 가 0이 아닌 값을 갖을 확률을 의미하고 모수 β 는 Y 가 가질 수 있는 범위의 값들을 조절하여 분포의 왜도를 결정한다. 특히 β 는 이산형 변수에 대해서 자주 언급되는 VR (variance rate)로 표현되는 산포 (dispersion) (Cameron과 Trivedi, 2013)와 연관되어 있다. VR은 다음과 같이 정의된다.

$$VR = \frac{\text{observed variance}}{\text{theoretical variance}}.$$

$VR > 1$ 이면 과대 산포 (over dispersion), $VR < 1$ 이면 과소 산포 (under dispersion) 라 하며 $VR = 1$ 인 경우에는 등 산포 (equal dispersion)라 한다. 특히 $0 < \beta \leq 1$ 인 경우와 $\beta \geq 1$ 인 경우에는 q 값과 상관없이 각각 과대 산포와 과소 산포가 되며 $1 < \beta < 3$ 인 경우에는 q 값에 따라 과대 혹은 과소 산포가 된다. 포아송 분포는 대표적인 등 산포에 해당되고 음이항 분포는 과대 산포 모형의 대표적인 경우이다. DW 모형은 과대, 등 산포, 과소 산포 모두에 대해 다룰 수 있고 이것이 이산형 자료에 적용되는 여러 다른 모형들에 비해 갖는 가장 큰 장점이라고 볼 수 있다.

모수들 q, β 에 대한 최대우도추정량은 아래와 같이 주어지는 로그우도 함수

$$l(y_1, \dots, y_n) = \sum_{i=1}^n \log((q)^{y_i^\beta} - (q)^{(y_i+1)^\beta})$$

를 최대화하여 쉽게 구할 수 있다.

Y 의 기댓값 $E(Y)$ 과 분산 $Var(Y)$ 는 각각 다음과 같이 나타낼 수 있는데 닫힌 형식 (closed form)이 없어서 수치 근사 방법을 이용해야 하는 단점이 있다.

$$E(Y) = \sum_{y=1}^{\infty} (q)^{y^\beta}, \quad Var(Y) = \sum_{y=1}^{\infty} (2y-1)(q)^{y^\beta}.$$

그러나 τ 분위수인 즉, $P(Y \leq \mu^\tau) = 1 - q^{(y+1)^\beta} \geq \tau$ 를 만족하는 μ^τ 에 대해서는 아래와 같이 구할 수 있다.

$$\mu^\tau = \left[\left(\frac{\log(1-\tau)}{\log(q)} \right)^{\frac{1}{\beta}} - 1 \right].$$

이를 바탕으로 DW 분포의 중위수는 $\mu^{0.5} = \left[\left(\frac{\log(2)}{\log(q)} \right)^{\frac{1}{\beta}} - 1 \right]$ 와 같다.

2.2. 이산형 와이블 회귀 모형

다음으로 독립변수들을 고려한 이산형 와이블 회귀모형 (discrete Weibull regression model) 을 생각해볼 수 있다. 독립변수들을 고려한 이산형 와이블 분포를 따르는 확률 변수 Y 에 대해서는 아래와 같이 나타낼 수 있다.

$$Y|X \sim DW(q(X), \beta(X)).$$

이때 $X = (1, X_1, \dots, X_p)$ 은 p 개의 독립변수들로 이루어진 벡터이다. 독립변수들은 각각 모수 $q(X)$ 와 $\beta(X)$ 를 통해 연결될 수 있는데 본 연구에서는 q 에 대하여만 독립변수들을 연결한 $Y|X \sim DW(q(X), \beta)$ 를 고려하였다. Y 의 누적분포함수는 아래와 같고

$$F(y; q(x), \beta) = \begin{cases} 1 - q(x)^{(y+1)^\beta} & y = 0, 1, 2, \dots \\ 0 & y < 0. \end{cases}$$

확률 질량함수는 다음과 같이 나타낼 수 있다.

$$f(y; q(x), \beta) = q(x)^{y^\beta} - q(x)^{(y+1)^\beta}, \quad \text{for } y = 0, 1, 2, \dots, \quad 0 < q(x) < 1, \quad \beta > 0.$$

독립변수들을 고려한 이산형 와이블 회귀모형식은 아래와 같이 나타낼 수 있고

$$\log(-\log(q(X))) = X\theta, \quad \log(\beta) = \vartheta,$$

이때 $\theta = (\theta_0, \theta_1, \dots, \theta_p)'$ 는 회귀계수들로 이루어진 벡터이고 ϑ 는 β 에 대한 모수이다. 이 외에도 $q(X)$ 에 대하여 로짓 링크인 $\log(\frac{q(X)}{1-q(X)})$ 를 사용할 수도 있다.
모수들에 대한 최대우도추정량은 아래의 우도함수를 이용하여

$$L(y, x, \theta, \vartheta) = \prod_{i=1}^n f(y_i | x_i) = \prod_{i=1}^n ((q(x_i))^{y_i^\beta} - (q(x_i))^{(y_i+1)^\beta}).$$

다음의 로그우도함수를 최대화 시키는 값으로 구해진다.

$$l(y, x, \theta, \vartheta) = \sum_{i=1}^n \log((q(x_i))^{y_i^\beta} - (q(x_i))^{(y_i+1)^\beta}).$$

모수들이 추정이 된 후 $E(Y|X)$ 는 닫힌 형식이 아니어서 수치 근사방법을 이용해야하지만 독립변수들을 고려한 조건부 τ 분위수 μ^τ 는 아래의 식과 같이 쉽게 구할 수 있다.

$$\mu^\tau = \left[\left(\frac{\log(1-\tau)}{\log(q(x))} \right)^{\frac{1}{\beta}} - 1 \right].$$

$\tau = 0.5$ 일 때 즉, 중위수 $\mu^{0.5}$ 에 로그를 씌우게 되면 아래 식처럼 나타낼 수 있고

$$\log(\mu^{0.5} + 1) = \frac{1}{\beta} \log(\log(2)) - \frac{1}{\beta} X\theta,$$

이러한 형태는 모수들에 대한 해석을 용이하게 한다. 즉, $\frac{1}{\beta} \log(\log(2) - \theta_0)$ 은 독립변수값들이 모두 0일 때의 조건부 중위수와 관련되어있고 $-\frac{\theta_p}{\beta}$ 는 X_p 변수가 1단위 증가할 때 중위수의 변화정도를 의미한다. 중위수에 대해 로그를 씌워서 독립변수들과 연결시키는 모형의 형태는 평균에 대해서 로그 연결 함수를 사용하는 포아송 분포와 음이항 분포에서의 회귀모형식과 유사하다.

2.3. 결측값 대체 방법

공변량들에 결측치 (missing value)가 있을 때 이를 무시하고 분석을 할 경우 표본수의 감소로 인해 검정력이 낮아지게 된다. 따라서 결측치들을 단순히 제거하기보다는 그럴듯한 (plausible) 값으로 대체를 하여 분석하는 방법들이 많이 사용되고 있다. 결측치를 처리하는 방법은 크게 단일 대체법 (single imputation)과 다중 대체법 (multiple imputation)이 있다. 본 연구에서는 단일 대체법 중 PMM (predictive mean matching) 방법을 사용한 후 본 분석을 실시하였다. 이 방법은 회귀 모형에서의 예측값을 이용하여 결측치를 대체하는 방법이다 (Heitjan과 Little, 1991; Schenker와 Jeremy, 1996).

$\mathbf{Y} = (Y_1, \dots, Y_n)$ 을 n 개의 개체와 p 개의 변수들로 이루어져 있는 행렬이라 하고 $Y_i = (Y_{i1}, \dots, Y_{ip})$ 를 i 번째 개체에 대한 자료 벡터로 가정하자. 또한 $Y^{obs} = (Y_1^{obs}, \dots, Y_p^{obs})$ 와 $Y^{mis} = (Y_1^{mis}, \dots, Y_p^{mis})$ 를 각각 자료 행렬 \mathbf{Y} 에서 관찰된 데이터와 결측 데이터라고 가정하자. 그러면 각 개체에 대해 조건부 기대값 $\hat{\mu} = E(Y^{mis}|Y^{obs})$ 이 추정되고 결측치는 가장 가까운 개체의 값으로 대체 된다.

3. 국민건강영양조사 (KNHANES) 자료 분석

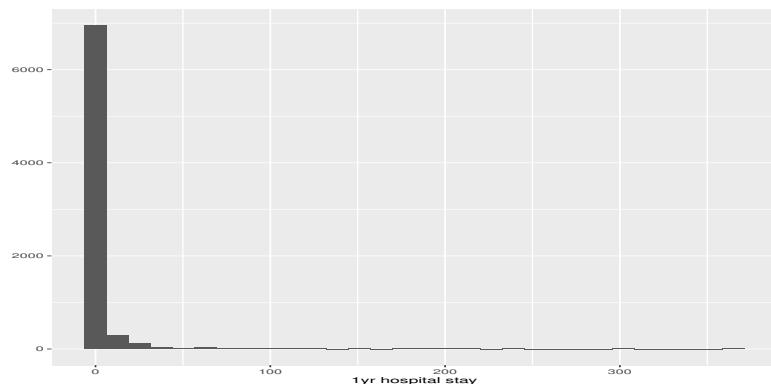
3.1. 데이터 설명

국민건강영양조사 (KNHANES)는 1995년에 제정된 국민건강증진법 제 16조에 근거하여 시행되고 있는 전국 규모의 건강 및 영양조사이다. 2007년부터 국가통계의 시의성 향상을 위해 2007년 이후 매년 시행되고 있고 보건 정책의 기초자료로 활용 되어오고 있다. Lee 등 (2010)은 국민건강영양조사 대상자들의 민간의료보험 가입여부에 따른 입원 이용 횟수를 비교 분석하였다. 또한 Kim 등 (2008)은 소득계층에 따른 의료이용 격차 분석을 목적 국민건강영양조사 2005년도 자료를 바탕으로 1년간 의료서비스 (외래이용, 입원) 이용 횟수에 영향을 미치는 요인들을 찾고자 다중 선형 회귀분석을 실시하였다. 그러나 입원 횟수 또는 의료 서비스 횟수에 대해서는 선형 회귀식을 이용하는 것보다 이산형 와이블 모형을 적용시키는 것이 더 적합할 것으로 사료된다. 본 연구에서 사용한 제 7기 1차년도 (2016) 자료는 가장 최근 자료로서 이 자료에 들어있는 8150 명의 1년간 입원일수에 영향을 미치는 요인들을 찾아내기 위해 이산형 와이블 회귀모형을 적합하였다. 1년간 입원일수에 영향을 미치는 요인들로 고려한 변수들은 성별 (sex), 나이 (age), 동/읍면 (town), 가구 소득 사분위수 (home income), 교육수준 (education)과 직업 (job) 이었다. 제 7기 1차년도 자료에서 분석에 사용한 변수 구성 및 각 변수들의 결측률은 아래 Table 3.1에 나타나있다. 성별과 나이는 결측치가 존재하지 않았고 가구 소득 사분위수는 0.4%의 낮은 결측률을 보였다. 교육수준은 8.4%의 결측률을 보였고 직업의 결측률은 25.1%로 상대적으로 높은 결측률을 보였다. 1년간 입원일수는 8.2%의 결측률을 보였고 다섯 수치료약값은 (0, 0, 0, 0, 365)로 나타나 결측값을 제외한 전체 대상자의 72.8% 가 1년간 입원일수가 0으로 나타났다.

1년간 입원일수에 대한 분포는 Figure 3.1에 나타나있는데 그림에서 보는 바와 같이 그 분포가 상당히 왼쪽으로 치우쳐있음을 확인할 수 있다. 또한 자료에 결측치도 존재하여 8.2%의 결측률을 나타났다. 각 개체에 결측치가 하나라도 존재하면 이를 개체들을 제거하고 분석하는 list-wise deletion method을 사용할 경우 데이터의 25.6% 가 제거 되고 전체 데이터의 74.4% 만이 분석에 사용되어 결과적으로 검정력이 떨어질 수 있다. 따라서 결측치들을 단일 대체 (single imputation) 방법 중의 하나인 predictive mean matching (PMM) 방법을 이용하여 대체한 후 이산형 와이블 모형을 적용하여 1년간 입원 일수에 영향을 미치는 변수들을 알아보고자 하였다.

Table 3.1 Descriptive statistics and missing rates for the KHNHANES data

| variables | categories | n (%) or mean (SD) | missing rate |
|-------------|-------------------------------------|--------------------|--------------|
| sex | male | 3665 (45) | 0% |
| | female | 4485 (55) | |
| age | | 41.81 (22.98) | |
| town | town | 6604 (81.0) | 0% |
| | country | 1546 (19.0) | |
| home income | low | 1407 (17.3) | 0.4% |
| | mid-low | 2047 (25.1) | |
| | mid-high | 2316 (28.4) | |
| | high | 2346 (28.8) | |
| education | under elementary | 2665 (32.7) | 8.4% |
| | middle School | 820 (10.1) | |
| | high School | 1881 (23.1) | |
| | university | 2101 (25.8) | |
| job | management, expert, service etc. | 2134 (26.2) | 25.1% |
| | agriculture, assembler, etc. | 1328 (16.3) | |
| | no job | 2638 (32.4) | |

**Figure 3.1** Distribution of 1 year hospital stay in KHNHANES data

3.2. 결측치값을 대체한 이산형 와이블 모형 적용 결과

1년간 입원일수에 영 (zero) 값이 상대적으로 많이 나타나 영과잉 포아송 모형과 이산형 와이블 회귀 모형과의 적합도를 비교를 해보았다. 아래 Table 3.2는 결측치값들을 단일 대체법인 PMM 으로 대체한 후 이산형 와이블 모형과 영과잉 포아송 모형 (zero-inflated Poisson model) 을 각각 적용하여 분석한 결과이다.

분석 결과 두 모형간의 유의한 변수들이 상이하게 나왔고 각 범주형 변수 안에서도 유의한 범주들도 다르게 나타났다. 두 모형 모두 성별, 나이, 가구 소득, 직업이 유의한 변수로 나타난 반면 교육수준과 동/읍면 구분 변수는 영과잉 포아송 모형에서만 유의하게 나타났다. 직업 변수의 경우에는 이산형 와이블 모형과 영과잉 포아송 모형에서 유의한 범주가 상이하게 나타났다.

Table 3.2 Maximum likelihood estimates, AIC of the KNHANES data using discrete Weibull regression and zero inflated Poisson models

| | variables | DW | ZIP |
|-------------|-----------------------------|----------|----------|
| sex | female | -0.199* | 0.110* |
| age | | -0.003* | 0.017* |
| town | country | 0.0119 | -0.137* |
| home income | mid-low | 0.181* | -0.332* |
| | mid-high | 0.240* | -0.727* |
| | high | 0.220* | -0.917* |
| education | middle School | 0.025 | -0.272* |
| | high School | 0.036 | 0.011 |
| | university | 0.044 | -0.097* |
| job | agriculture, assembler etc. | -0.006 | 0.312* |
| | no job | -0.088* | 0.736* |
| other | | =0.290* | |
| AIC | | 18280.94 | 59062.88 |

*: p-value<0.05

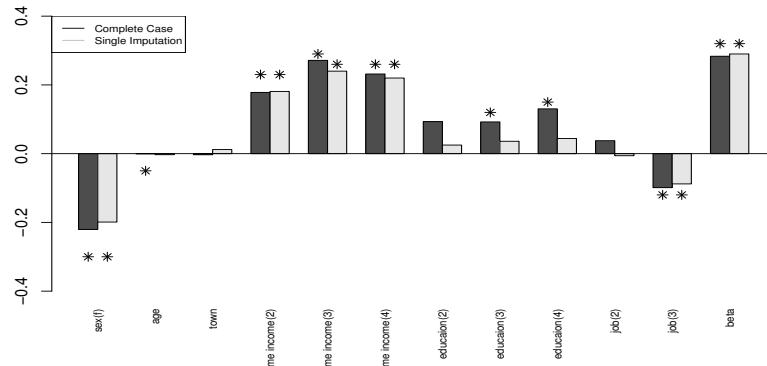
모형의 선택 기준은 AIC 값을 기준으로 하였고 이때 이산형 와이블 모형의 AIC 값이 로 영과잉 포아송 모형의 값 보다 월등히 작아 1년간 입원일수에 대한 모형은 이산형 와이블 모형이 적합이 더 잘 된다는 것을 알 수 있다. 이산형 와이블 모형에 근거하여 1년간 입원일수에 성별 (sex) 의 경우 남성에 비해서 여성이 1년간 입원일수의 중위수가 유의하게 높았고, 연령 (age)은 높을수록 1년간 입원 일수의 중위수가 높게 나타났다. 소득 사분위수 (home income) 의 경우에는 하위 25%에 비해 나머지 소득 분위수에 해당되는 사람들의 입원일수의 중위수가 유의하게 낮게 나타났다. 직업의 경우에는 전문직 (management, expert, service etc.) 에 비해서 무직인 경우가 1년간 입원일수의 중위수가 유의하게 높게 나타났다.

아래 Figure 3.2 은 결측치를 단일 대체법을 사용하여 대체한 후 와이블이산형 모형과 결측치들을 모두 제거한 complete case (CC) 에 대한 추정치값들을 비교한 그래프이다. 결측치들을 대체한 경우와 단순히 제거하여 완전 자료 (complete case)에 대하여 분석한 경우에 있어서 유의한 변수들이 상이함을 확인할 수 있다. 즉, 나이 변수는 단일 대체법인 경우 유의하게 나타났지만 완전 자료인 경우에는 유의하지 않게 나타난 반면, 교육 수준의 범주가 3 (high school)와 4 (university) 인 경우에는 CC 분석에서만 유의하게 나타났다. 결측치를 모두 제거한 경우에는 전체 데이터의 74.4% 만이 사용되므로 검정력이 떨어질 수 있다.

3.3. 모의 실험

모의 실험을 통하여 결측치가 있는 자료에서 단일 대체법을 이용하여 이산형 와이블 모형을 적용한 경우와 대체법을 이용하지 않고 complete case (CC) 로 분석한 경우에 대해서 여러 다양한 세팅 하에 추정치들의 편향 (bias)을 비교하였다. $Y|X \sim \text{Discrete Weibull} (q(X), \beta)$ 를 가정하고 $q(X) = \exp(-\exp(\theta_0 + \theta_1 X))$ 에서 각 모수들은 $\theta_0 = -2.6, \theta_1 = 0.9, \beta = 4$ 로 놓았다. 한 개의 독립변수 $X \sim U(0, 1)$ 는 균일분포를 따르도록 하였고 missing mechanism은 임의 결측 (missing at random; MAR) 을 가정하였다. 전체 표본의 크기를 $n = 50, 300, 500$ 3가지로 설정하고 결측률은 20%, 50%, 75%로 하여 각 시뮬레이션 조합 안에서 단일대체법 (SI) 과 CC 의 편향정도를 비교하였다. Table 3.3은 모의실험 결과를 정리한 것이고 Figure 3.3은 수치로 나타난 값들을 보기 쉽게 그림으로 그려본 것이다.

Figure 3.3을 보면 표본의 크기가 50인 경우에 모든 결측률에서 단일 대체법이 CC 보다 편향정도가 작게 나왔고 CC 는 결측률 값에 따라 편향 정도가 영향을 받지만 단일 대체법은 편향정도가 크게 변하



*: p-value<0.05

Figure 3.2 Discrete Weibull coefficient for using single imputation and complete case

Table 3.3 Bias of θ_0 , θ_1 and β of SI and CC under various missing rates with $n = 50, 300, 500$

| | | | θ_0 | θ_1 | β |
|-----------------|-----|----|------------|------------|---------|
| missing rate | 20% | CC | 0.8661 | -0.0990 | -1.5940 |
| | | SI | -0.1548 | 0.0747 | -0.3775 |
| | 50% | CC | 0.6605 | -0.5602 | -0.5764 |
| | | SI | -0.1691 | 0.17064 | -0.3766 |
| | 75% | CC | 5.3064 | -5.9265 | -2.2404 |
| | | SI | -0.9578 | 1.1041 | 0.1397 |
| $n = 300$ | | | | | |
| missing rate | 20% | CC | 0.3183 | 0.0619 | -0.4216 |
| | | SI | -0.0584 | 0.0362 | 0.0800 |
| | 50% | CC | 0.1017 | -0.1789 | 0.4649 |
| | | SI | 0.0072 | -0.0956 | 0.0720 |
| | 75% | CC | 1.6090 | -1.1642 | -0.1298 |
| | | SI | 0.5496 | -0.3422 | -0.3298 |
| $n = 500$ | | | | | |
| missing rate | 20% | CC | 0.7699 | -0.0018 | -1.033 |
| | | SI | 0.0551 | -0.0071 | -0.0472 |
| | 50% | CC | 0.4215 | -0.0735 | -0.1081 |
| | | SI | 0.0177 | -0.0410 | -0.0051 |
| | 75% | CC | 2.1260 | -1.490 | -0.8070 |
| | | SI | -0.2137 | 0.0731 | 0.2324 |

지 않는 것을 확인할 수 있다. 결측률이 75%로 증가한 경우에는 CC 와 SI 모두 편향정도가 증가하지만 CC 의 경우 증가 폭이 상대적으로 훨씬 크다는 것을 알 수 있다. 이러한 현상은 표본의 크기가 300인 경우에도 SI 가 CC 보다 편향정도가 월등히 낮은 것을 확인할 수 있었고 결측률이 50%로 증가하여도 편향정도가 크게 증가하지 않음을 보였다. 결측률이 75%인 경우에는 CC 의 편향정도가 상당히 증가한 것을 볼 수 있다. 표본의 크기가 500인 경우에는 SI 의 경우 대부분의 결측률에 대해서 표본의 크기가 작을 때 보다 편향정도가 훨씬 작아지지만 CC 의 경우에는 표본의 크기가 으로 커져도 편향정도가 작아

지지 않는다는 것을 확인할 수 있다. 모의실험 결과를 통해서 자료에 결측치가 있는 경우에는 이를 대체한 후 분석을 하는 것이 추정치가 덜 편향됨을 알 수 있다.

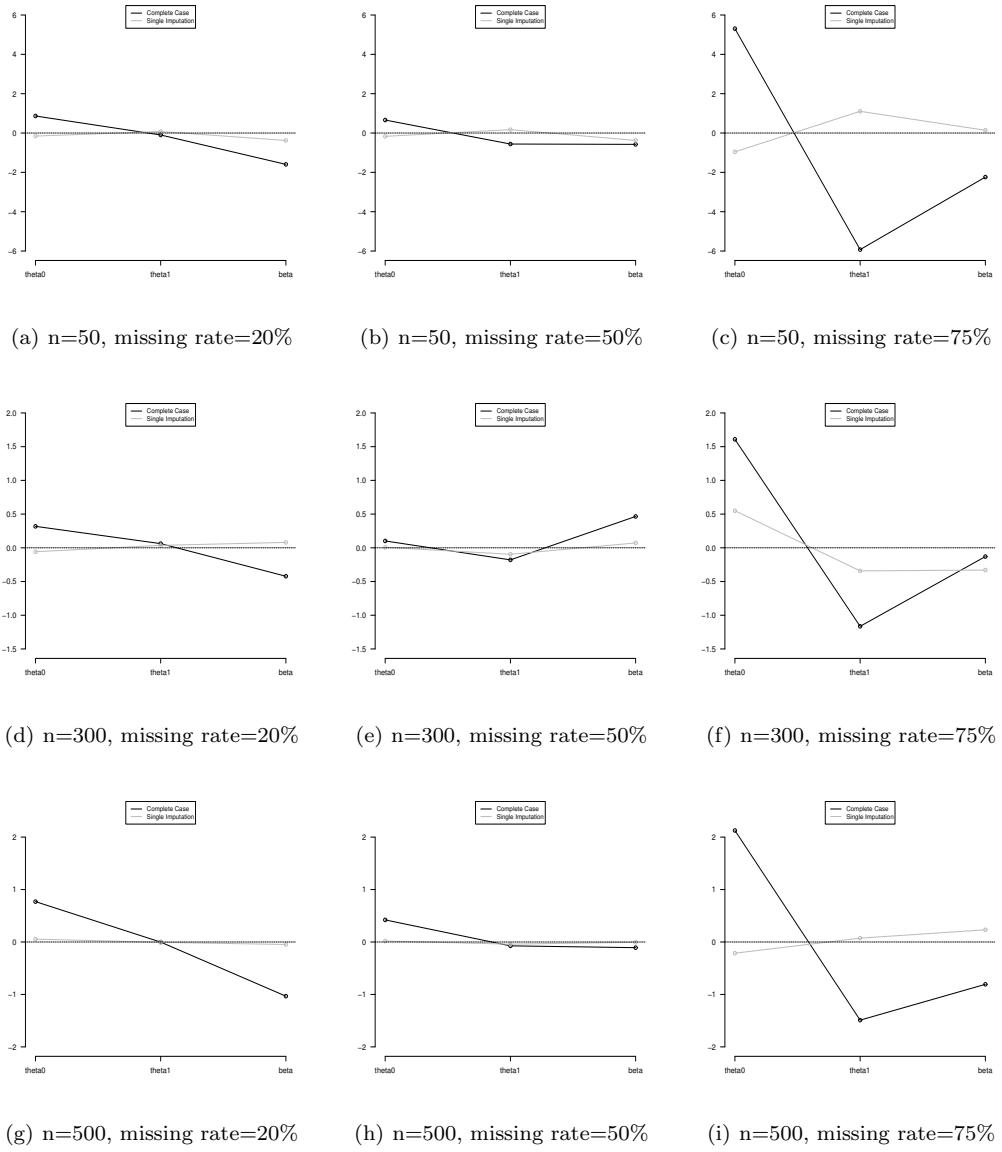


Figure 3.3 Comparison of bias of θ_0 and θ_1 of SI and CC under various missing rates with $n = 50, 300, 500$

4. 결론

본 연구에서는 셀 수 있는 가산 자료에 존재하는 결측치들을 단일 대체법을 이용하여 대체하여 정보 손실을 최소화 한 후 이산형 와이블 회귀 모형을 적용하였다. 데이터의 분포가 과대 산포 (over dispersed), 과소 산포 (under dispersed) 또는 등산포 (equal dispersed) 인 경우 이들에 모두 적용할 수 있다는 점이 이산형 와이블 회귀모형의 큰 장점이다. 제 7기 1차년도 (2016) 국민건강영양조사 자료에 있는 결측치들을 대체한 후 이산형 와이블 모형과 영과잉 포아송 모형을 적용하여 비교한 결과 이산형 와이블 모형의 AIC 값이 월등히 작게 나와 모형이 더 잘 적합됨을 알 수 있었다.

또한 모의실험 연구를 통하여 결측값을 대체하는 경우가 결측치들을 모두 제거하는 경우 보다 추정치들의 편향 정도가 낮다는 것을 확인하였다.

본 연구에서는 결측치 대체 방법으로 단일 대체법을 사용하였는데 향후 연구계획으로는 다중 대체법 (multiple imputation)을 이용하여 결측치들을 대체해보고 더 다양한 모의실험을 통하여 분포가 잘못 정의 (miss-specification) 되어 있을 때의 이산형 와이블 모형의 추정의 정확도를 보고자 한다. 또한 본 분석에 사용한 국민건강영양조사 자료는 복합 데이터 (complex survey data)로 복합표본설계 (complex sampling) 요소를 반영해서 비교해야 더 정확한데 본 논문에서는 이산형 와이블 모형의 적합에 더 초점을 맞추어 이러한 데이터의 특성을 정확히 반영하지는 못했다. 따라서 차후 후속 연구는 복합 데이터 내의 충화 변수, 집락변수 그리고 가중치를 고려하여 이산형 와이블 모형을 확대 적용하고자 한다.

References

- Cameron, A. C. and Trivedi, P. K. (2013). *Regression analysis of count data*, Cambridge University Press.
- Chanalidis, C., Evers, L., Neocleous, T. and Nobile, A. (2018). Efficient Bayesian inference for COM-Poisson regression models. *Statistics and Computing*, **28**, 595-608.
- Chun, H. (2017). Fit of the number of insurance solicitor's turnovers using zero-inflated negative binomial regression. *Journal of the Korean Data & Information Science Society*, **28**, 1087-97.
- Efron, B. (1986). Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association*, **81**, 709-721.
- Famoye, F. (1993). Restricted generalized Poisson regression model. *Communications in Statistics-Theory Methods*, **22**, 1335-1354.
- Heitjan, D. F. and Little, R. J. A. (1991). Multiple imputation for the fatal accident reporting system. *Applied Statistics*, **40**, 13-29.
- Khan, M. A., Khalique, A. and Abouammoh, A. (1989) On estimating parameters in a discrete Weibull distribution. *IEEE Transactions on Reliability*, **38**, 348-350.
- Kim, J., Bang, S. and Kwon, O. (2017). Analysis of scientific military training data using zero-inflated and Hurdle regression. *Journal of the Korean Data & Information Science Society*, **28**, 1511-1520.
- Kim, K. M. (2003). An application to zero-Inflated Poisson Regression Model. *Journal of the Korean Data & Information Science Society*, **14**, 45-53.
- Kim, T. I., Choi, Y. Y. and Lee, K. H. (2008). Analysis on the differences in medical service usage in terms of income levels. *Korean Social Security Studies*, **24**, 53-75.
- Klakattawi, H. S., Vinciotti, V. and Yu, K. (2018). A simple and adaptive dispersion regression model for count data. *Entropy*, **20**.
- Kulasekera, K. (1994) Approximate MLEs of the parameters of a discrete Weibull distribution with type I censored data. *Microelectronics Reliability*, **34**, 1185-1188.
- Lee, Y. C., Im, B. H. and Park, Y. H. (2010). The determinants and comparison of health behavior and health service by private medical insurance on national health-nutrition survey. *Journal of the Korea Contents Association*, **10**, 190-204.
- Peluso, A. and Vinciotti, V. (2018). Discrete weibull generalised additive model: An application to count fertility data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, Unpublished manuscript.

- Schenker, N. and Jeremy, M. G. T. (1996). Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis*, **22**, 425-446.
- Sellers, K. F. and Shmueli, G. (2010). A flexible regression model for count data. *The Annals of Applied Statistics*, **4**, 943-961.
- Nakagawa, T. and Osaki, S. (1975). The discrete Weibull distribution. *IEEE Transactions on Reliability*, R-24.

A study of discrete Weibull regression model with missing data[†]

Hanna Yoo¹

¹Department of Computer Software, Busan University of Foreign Studies

Received 18 December 2018, revised 7 January 2019, accepted 11 January 2019

Abstract

Discrete Weibull regression model can be adapted to discrete count data with different type of dispersions. It can be adopted to various types of dispersion however it is not used widely in discrete data and there isn't much research papers that deal with discrete Weibull regression model. In this paper, discrete Weibull regression model is adapted to data that has missing values. Single imputation method is used to impute the missing values. We analyzed the seventh Korea National Health and Nutrition Examination Survey (KNHANES VII), 2016 to assess the factors for 1 year hospital stay. We compare the results using discrete Weibull regression model with zero-inflated Poisson model and shown that discrete Weibull regression model provided better fit. We also performed simulation studies to show the accuracy of the discrete Weibull regression with using single imputation under various missing rates and sample size. Through simulation studies, it was shown that using imputation methods yield better results than deleting the missing values. Using imputation with discrete Weibull regression model to discrete data will increase the power and enables the wide applicability to various types of dispersion data.

Keywords: Discrete count data, discrete Weibull regression model, imputation method, KNHANES.

[†] This work was supported by the research grant of the Busan University of Foreign Studies in 2018.

¹ Assistant professor, Department of Computer Software, Busan University of Foreign Studies, 65 Geumsaem-ro 485 beon-gil, Geumjeong-gu, Busan, Korea. Email: pinkcan78@bufs.ac.kr