저항적 포아송 회귀와 활용

허명회* · 성내경** · 임용빈***†
* 고려대학교 통계학과
** 고려대학교 통계연구소
*** 이화여자대학교 통계학과

Resistant Poisson Regression and Its Application

Myung-Hoe Huh* · Nae Kyung Sung** · Yong Bin Lim***†
* Dept. of Statistics Korea University
** Institute of Statistics Korea University
*** Dept. of Statistics Ewha Woman's University

Key Words : Resistant Poisson Regression, Outliers

Abstract

For the count response we normally consider Poisson regression model. However, the conventional fitting algorithm for Poisson regression model is not reliable at all when the response variable is measured with sizable contamination. In this study, we propose an alternative fitting algorithm that is resistant to outlying values in response and report a case study in semiconductor industry.

1. 서 론

종속변수를 y, 이에 대한 p개의 설명변수를 x_1, …, x_p로 표기하기로 하자. y가 계수형인 경우 흔히 고려되는 회귀 모형은 포아송 회귀 모형(Poisson regression model)이다. 즉

$$y \mid (x_1, \cdots, x_p) \sim \text{Poisson}\, \theta(x_1, \cdots, x_p; a)$$

여기서

$$\theta(x_1, \cdots, x_p; a) = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p),$$

a는 주어진 양의 상수, $\beta_0, \beta_1, \cdots, \beta_p$는 미지의 계수이다. [$\log(a)$를 오프셋(offset)이라 하는데 관측 개체의 크기를 고려할 필요가 있는 경우 적용된다. 만약 그럴 필요가 없는 데이터 셋에 대하여는 $a \equiv 1$로 볼 수 있다.]

† 교신저자 yblim@ewha.ac.kr

예를 들어 반도체 제조 공정에서 웨이퍼에서 관측되는 결점 수(defect count)를 y, 이에 대하여 영향을 줄 수 있는 제 공정요인을 $x_1, \cdots, x_p$, 원판의 크기를 a라고 하면 자연스럽게 이와 같은 포아송 회귀 모형을 가정할 수 있을 것이다. [원판의 크기가 모두 동일한 경우엔 당연히 a를 고려할 필요가 없다.]

포아송 회귀 모형의 최대 가능도(maximum likelihood) 적합을 위하여 일반화 선형모형(generalized linear model)에서 범용적으로 적용되는 반복 가중최소제곱(IWLS ; iterative weighted least squares) 알고리즘을 활용할 수 있는데 (McCullagh and Nelder, 1983), 이를 포아송 회귀에 한정하여 정리하면 다음과 같다 (이하 n은 관측 수를 나타낸다).

[단계 0] 회귀계수 $\beta_0, \beta_1, \cdots, \beta_p$에 적절한 초기값 $\beta_0^0, \beta_1^0, \cdots, \beta_p^0$을 준다. 예컨대

$$\beta_0^0 = \log\left[\sum_{i=1}^{n} y_i / \sum_{i=1}^{n} a_i\right]$$

$$\beta_1^0 = 0, \cdots, \beta_p^0 = 0$$

여기서 y_i와 a_i는 종속변수 y의 i번째 관측값과 a의 i번째 값이다 ($i=1, \cdots, n$).

[단계 1] 각 관측 i에서 수정종속변량(adjusted dependent variate)

$$z_i = \log(\theta_i^0) + (y_i - \theta_i^0)/\theta_i^0$$

를 산출한다. 여기서 $\theta_i^0 = a_i \exp(\beta_0^0 + \beta_1^0 x_{i1} + \cdots + \beta_p^0 x_{ip})$이고 x_{ij}는 설명변수 x_j의 i번째 관측값이다 ($i=1, \cdots, n$; $j=1, \cdots, p$).

[단계 2] 각 관측 i에 대한 가중치 w_i를 θ_i^0로 놓는다 ($i=1, \cdots, n$).

[단계 3] 회귀계수 벡터 $\beta=(\beta_0, \beta_1, \cdots, \beta_p)^t$에 대한 추정치를 업데이트 한다:

$$\hat{\beta}^0 = (X^t W X)^{-1} X^t W (z - \log a)$$

를 산출한다. 여기서 $X=(x_{ij})_{i=1, \cdots, n; j=0, 1, \cdots, p}$ $X=(x_{ij})_{i=1, \cdots, n; j=0, 1, \cdots, p}$, $W=\mathrm{diag}(w_1, \cdots, w_n)$, $z=(z_1, \cdots, z_n)^t$, $a=(a_1, \cdots, a_n)^t$이며 $x_{i0}=1$이다.

[단계 4] β의 추정치에 변화가 없을때까지 [단계 1]부터 [단계 3]을 충분히 반복한다.

이 알고리즘은 Frome, Kutner와 Beauchamp (1973)에 의하여 최초 제안되었다. 통상적인 포아송 회귀는 IWLS 알고리즘이 구현된 상용화된 팩키지 소프트웨어, 예컨대 SAS의 GENMOD Procedure를 사용하면 별다른 수고 없이 추정모형을 얻을 수 있다(SAS Institute, 1997).

문제는 종속변수 y의 관측이 비제어 및 미지의 원인에 의하여 오염되는 상황에서 앞의 IWLS 추정치가 불안정한 행태를 보일 수밖에 없다는 점이다. 연구자들이 반도체 제조 공정에서 경험한 바에 의하면 결점 계수기(defect counter)는 일시적인 부유 먼지의 영향을 받거나 매칭 패턴 틀과의 부정합, 그리고 이외의 알 수 없는 원인에 의하여 꽤 빈번하게 전혀 합당하지 않은 수치를 산출하여 데이터베이스에 저장한다. 따라서 결점 수가 상당한 경우 오히려 기록된 검사결과를 믿을 수 없게 된다.

그러므로 관측된 계수형 반응치 y의 일부 기록에 오류가 의심스러운 경우, 즉 반응치 y의 관측에 오염이 예상되는 상황에서 포아송 회귀 모형을 안정적으로 추정하는 방법에 대하여 연구할 필요가 있다. 본 연구의 목적은 안정적인 추정방법인 저항적 포아송 회귀모형의 추정 방법을 개발하고 반도체 제조공정에서의 활용 사례를 제시하는 데 있다.

2. 저항적 모형적합

통계적 절차가 일부 특이한 관측값에 의해 큰 영향을 받는 것은 바람직하지 않기에 일반적으로 둔감성이 추천된다. 이런 성질을 저항성(resistance)이라고 하는데, 문제는 지나치게 떨어져 있는 관측값(outliers, 특이점)을 어떻게 정의하느냐 하는 것이다. 흔히 특이점은 자료의 주 군집으로부터 상당히 떨어져 있는, 즉 시각적으로 분리되는 관측값을 말하므로 운영적 정의화(operational definition) 하는데 문제가 된다. 따라서 본 연구에서는 특이점을 분포의 극단적인 α 분위수 부분에서 발생한 관측값으로, 구체적으로 포아송 분포에서의 특이점을 상위 0.1% 분위수와 같거나 큰 값으로 정의하기로 한다. 예컨대 평균이 10인 포아송 분포의 경우 $P(X \geq 22) = 0.0007$, 평균이 40인 포아송 분포의 경우 $P(X \geq 62) = 0.0008$, 평균이 100인 포아송 분포의 경우 $P(X \geq 133) = 0.0009$이므로 22 이상, 62 이상, 133 이상인 관측값을 각각 Poisson(10) 분포, Poisson(40) 분포, Poisson(100) 분포의 특이점으로 규정한다.

각 포아송 분포에서 개별적으로 상위 0.1% 분위수를 일일이 계산하는 것은 비효율적일 것이다. 대신, 중심극한정리에 근거하여, 포아송 파라미터 θ가 상당히 큰 경우 정규근사를 활용하여 보다 쉽게 포아송 분포의 상위 0.1% 분위수를 근사적으로 구하는 방법을 생각해보자. 즉,

$$(y_{0.001} - \theta)/\sqrt{\theta} \simeq z_{0.001}$$

로부터(여기서 $z_{0.001}$는 표준정규분포의 상위 0.1%

분위수, 즉 $z_{0.001}=3.09$) Poisson(θ) 분포의 상위 0.1% 분위수 $y_{0.001}$에 대한 근사값을 산출해 볼 수 있을 것이다. 이 식에 따라 $\theta=10, 40, 100$에 대한 근사적 상위 0.1% 분위수를 구하여 보면 각각 $y_{0.001}=20, 60, 131$이 나온다. 이 값을 앞에서 구한 포아송 분포의 상위 0.1% 분위수와 비교하면, 실제 상위 0.1% 분위수보다 다소 작은 것을 확인할 수 있다. 정규 근사를 보다 정확히 하기 위해서는 왜도 및 첨도를 고려하여 Fisher-Cornish 근사를 하는 것이지만(Cox and Hinkley, 1974) 여기서는 각 θ 별로 $z_{0.001}$을 보다 나은 값으로 대치하는 방법으로 하고자 한다. 예컨대 $\theta=100$에서는 $P(X\geq 133)=0.0009$임을 주목하여,

$$(y_{0.001}-\theta)/\sqrt{\theta}=(133-100)/\sqrt{100}=3.3$$

이므로 3.09($=z_{0.001}$) 대신 3.3을 사용하자는 것이다. 일반적으로

θ	$y_{0.001}$	$(y_{0.001}-\theta)/\sqrt{\theta}(=\tilde{z}_{0.001})$
0.1	3	9.2
1	6	5.0
10	22	3.8
40	62	3.5
100	133	3.3
400	464	3.2

임에 근거하여, 이 연구에서는 $z_{0.001}$ (=3.09) 대신 이보다 다소 크며 θ에 의존하는 $\tilde{z}_{0.001}$에 의거하여 특이점 판정을 할 것이다. 그러나 모든 θ에서 $\tilde{z}_{0.001}$을 정하는 것은 역시 비효율적이므로 θ가 기준점인 0.1, 1, 10, 100, 400이 아닌 경우에는 보다 보수적인 관점에서 θ보다 작되 가장 가까운 기준점에서의 $\tilde{z}_{0.001}$을 사용하기로 한다. 예컨대 $\theta=60$인 경우 $\theta=40$에 해당하는 $\tilde{z}_{0.001}$(= 3.5)로 한다. 이에 따라 $\theta=60$에서는 $60+3.5\sqrt{60}\simeq 87$ 이상의 값을 특이점으로 판정한다. $\theta=60$에서 정확한 상위 0.1% 분위수는 86이다.

이와 같은 준비 하에서 포아송 회귀 모형 적합을 위하여 다음의 저항적 반복가중최소제곱 (Resistant IWLS 이하 RIWLS) 알고리즘을 제안한다.

[단계 0] 회귀계수 $\beta_0, \beta_1, \cdots, \beta_p$에 저항성이 강한 초기값 $\beta_0^0, \beta_1^0, \cdots, \beta_p^0$ 을 준다:

$$\beta_0^0=\log[\,median_i\, y_i/\,median_i\, a_i]$$
$$\beta_1^0=0,\ \cdots,\ \beta_p^0=0,$$

여기서 y_i와 a_i는 종속변수 y의 i번째 관측값과 a의 i번째 값이다.

[단계 1] 각 관측 i에 대하여 수정종속변량 (adjusted dependent variate)

$$z_i=\log(\theta_i^0)+(y_i-\theta_i^0)/\theta_i^0$$

를 산출한다. 여기서 $\theta_i^0=a_i\exp(\beta_0^0+\beta_1^0 x_{i1}+\cdots+\beta_p^0 x_{ip})$이고 y_i와 x_{ij}는 종속변수 y와 설명변수 x_j의 i번째 관측값이다.

[단계 2] 각 관측 i에 대한 가중치 w_i를 다음과 같이 θ_i^0로 놓는다.

w_i $=\theta_i^0$, y_i가 Poisson (θ_i^0) 분포에서 특이값이 아닌 경우.

$=0$, y_i가 Poisson (θ_i^0) 분포에서 특이값인 경우.

[단계 3] 회귀계수 벡터 $\beta=(\beta_0, \beta_1, \cdots, \beta_p)^t$에 대한 추정치를 업데이트 한다:

$$\hat{\beta}^0=(X^tWX)^{-1}X^tW(z-\log a)$$

를 산출한다. 여기서 $X=(x_{ij})_{i=1,\cdots,n;j=0,1,\cdots,p}$, $X=\mathrm{diag}(w_1,\cdots,w_n)$, $z=(x_1,\cdots,x_n)^t$, $a=(a_1,\cdots,a_n)^t$이며 $x_{i0}=1$이다.

[단계 4] β의 추정치에 변화가 없을때까지 [단계 1]부터 [단계 3]을 충분히 반복한다.

RIWLS 알고리즘은 IWLS 알고리즘과 단계 0, 단계 2에서 일부 다르다. 단계 0에서는 저항성이 강한 초기값에서 시작하도록 한 것이고 단계 2에서는 특이값으로 판단되는 관측에 가중치 0을 부여함으로써 모형적합에 포함되지 않도록 한 것이다.

3. 모의실험

2장에서 제안한 저항적 알고리즘 RIWLS가 제대로 작동하는지를 확인하고자 여러 유형의 모의실험을 시행하였다.

첫째 모의실험은 1,000개 관측으로 구성된 일변량 자료의 경우인데 750개 관측은 Poisson($e^{2.3026}$)으로부터 생성되었고 250개 관측은 Poisson ($e^{3.3698}$)으로부터 생성되었다($e^{2.3026}=10$, $e^{3.3698}=40$). 이 실험에서의 관심은 표본의 주 분포인 Poisson(10)을 부 분포인 Poisson(40)에 영향 받지 않고 잘 탐지해내는가에 있다. 알고리즘을 수행한 결과, 3회 반복으로 250개 부 분포 관측 (특이값 관측) 중 248개를 정확히 걸러냈다. 추정된 포아송 회귀계수는 2.3058($=\hat{\beta}_0$)이었다(s.e.는 0.0115). 이어서 주 분포로부터의 관측 수를 650개로 줄이고 부 분포로부터의 관측 수를 350개로 늘여서 실험해본 결과 350개 부 분포 관측(특이값 관측) 중 348개를 정확히 걸러냈다. 계속 이어서 주 분포로부터의 관측 수를 550개로 줄이고 부 분포로부터의 관측 수를 450개로 늘여서 실험해본 결과 450개 부 분포 관측(특이값 관측) 중 449개를 정확히 걸러냈다. 이것은 RIWLS 알고리즘이 일변량 자료의 경우 자료 오염률이 상당히 높더라도 주 분포를 비교적 정확히 찾아낼 수 있음을 뜻한다.

둘째 모의실험은 설명변수가 1개인 포아송 회귀모형의 경우이다. 각 관측에서 x는 정규분포 N(0,1)으로부터 생성되었는데, y는 x에 조건화하여 750개 관측값이 Poisson($e^{2.3026+0.5x}$)으로부터 생성되었고 250개 관측값이 Poisson($e^{3.3689+0.5x}$)으로부터 생성되었다. 그 결과 750개 주 분포 관측은 모두 옳게 분류되었다. 그러나 250개 부 분포 관측 중 10개 관측을 정상적 관측으로 잘못 분류되었다. 760개 관측으로 7회 반복후 추정된 모형은 2.3145 ($=\hat{\beta}_0$)과 0.4963($=\hat{\beta}_1$)로 나타났다(각각에 대한 s.e.는 0.0120과 0.0105). 일부 부 분포 관측에 대한 분류 오류에도 불구하고 추정은 제대로 된 것을 알 수 있다. 이어서 주 분포로부터의 관측 수를 650개로 줄이고 부 분포로부터의 관측 수를 350개로 늘여서 실험해본 결과 350개 부 분포 관측(특이값 관측) 중 328개를 정확히 걸러냈다(22개를 오분류함). 계속 이어서 주 분포로부터의 관측 수를 550개로 줄이고 부 분포로부터의 관측 수를 450개로 늘여서 실험해본 결과 450개 부 분포 관측(특이값 관측) 중 387개를 정확히 걸러냈다(63개를 오분류함). 이것은 주 분포와 부 분포의 겹침이 실체화됨에 따라 일부 오염자료로부터 RIWLS 알고리즘이 완전히 자유로울 수 없음을 의미하는데 그것은 어쩔 수 없는 문제일 것이다.

4. 반도체 사례

이 장에서는 연구자들이 반도체 제조공정 분석에서 경험한 한 실제 사례를 제시하기로 한다. 종속변수 Y는 1개 원판에서 계수(count)된 결점 수이며 이에 대하여 고려하는 설명변수는 X1, X2, X3, X4, X5 등 5개이고 관측 수 n은 1,192이다. 변수 내용은 기업 비밀상 자세히 밝히기 어렵지만 정전기량 및 오븐 온도에 관한 실시간 기록등이다. 결점 수 Y의 최소값, 25% 분위수, 50% 분위수, 75% 분위수, 최대값은 각각 0, 21, 31, 46, 8393이다. 언뜻 보아도 일부 관측값이 심하게 오염되어 있음을 알 수 있다.

특이점에 대한 아무런 고려 없이 모든 관측 자료로써 포아송 회귀모형을 IWLS 알고리즘으로 적합한 결과는 다음과 같다.

Parameter	Estimate	S.E.	Chi^ 2	P-Value
Intercept	588.6541	37.9455	240.66	0.0000
X1	5.7954	3.2792	3.12	0.0772
X2	-11.4349	2.8968	15.58	0.0000
X3	-1.4841	0.2244	43.73	0.0000
X4	-0.1053	0.1573	0.45	0.5030
X5	-3.4093	0.2793	149.00	0.0000

한편, 저항적 알고리즘 RIWLS에 의한 포아송 회귀모형 적합 결과는 다음과 같다. n = 1,192개의 관측 중에서 약 25%인 298개가 절단되었고 나머지 894개 관측이 사용되었다.

Parameter	Estimate	S.E.	Chi^ 2	P-Value
Intercept	32.5309	7.2510	20.1278	0.0000
X1	-0.2908	0.4229	0.4728	0.4917
X2	1.5287	0.3593	18.1037	0.0000
X3	-0.2388	0.0433	30.4010	0.0000
X4	-0.0395	0.0218	3.2686	0.0706
X5	0.0628	0.0437	2.0616	0.1511

IWLS 알고리즘에 의한 포아송 회귀모형 적합결과와 비교하여 통계적으로 각 계수의 유의성 여부, 유의한 계수의 부호 등에서 상당한 차이가 있음을 볼 수 있다. 예컨대 변수 X2에 있어서 두 적합결과에서 모두 통계적으로 유의하게 나타났지만 IWLS 알고리즘에 의한 적합 결과로는 계수 부호가 음인데 반하여 RIWLS 알고리즘에 의한 적합 결과로는 계수 부호가 양으로 나타났다. 계수 부호가 정반대인 것이다. 이와 같이 부호가 바뀐 회귀계수가 총 5개 중 3개나 된다.

최종적으로 어느 것이 나은지는 공학적 지식과 확인 실험으로 판정될 것이다. 다만, 기존의 포아송 회귀모형이 일부 특이점에 과다하게 의존하고 있을 가능성이 있으므로 대안적으로 저항적 포아송 회귀모형 적합을 시도해볼 필요가 있다는 것이 이 연구의 소결론이다.

참 고 문 헌

[1] Cox, D. R. and Hinkley, D. V.(1974). *Theoretical Statistics*. Chapman and Hall, London.

[2] Frome, E. L., Kutner, M. H. and Beauchamp, J. J.(1973). "Regression analysis of Poisson-distributed data", *Journal of the American Statistical Association*, Vol. 68. pp. 935-940.

[3] McCullagh, P. and Nelder, J. A.(1983). *Generalized Linear Models*. Chapman and Hall, London.

[4] SAS Institute(1997). *SAS/STAT Software: Changes and Enhancements through Release 6.12*. Cary, NC.