범주형 자료분석
Count Models

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

June 9, 2025

진행 순서

1. 포아송 회귀모형
2. 회귀계수의 해석
3. 유의성 검정과 모형 적합도
4. 음이항 회귀모형
5. 영팽창 회귀모형

포아송 회귀모형

포아송 회귀모형

포아송 모형을 통해 종속변수가 가산 자료인 경우에 대응할 수 있다.

- 사회과학과 보건의료 분야에서 비음 정수(non-negative integer)인 (사건 따위에 관한) 가산 자료(count data)를 분석하게 될 수 있다.
- 이런 경우는 몇몇 분야에서는 제법 흔한 상황이지만, 어떤 분야에서는 그다지 일반적이지는 않을 수도 있다.
- 시간을 들여 흥미로운 가산형 종속변수(e.g., 프러시아 군인이 말의 뒷발에 채여 나가떨어지는 횟수, 출판 후 3년 간 논문의 피인용횟수, 지난 30일간 대중교통 이용 횟수, 지난 5년간 병원 방문 횟수)에 대해 상상해보자.

포아송 회귀모형

포아송 회귀모형은 나름의 수학적 구조를 가지고 있다.

- 사건이 발생하는 레이트(rate) λ를 발생율(incidence rate)이라고 부르는데, 이것은 확률과 미묘하게 개념이 다르다.
- 가령 레이트는 1.5 사람-해(person-years)처럼 1보다 클 수도 있고, 확률과는 구별되는 단위로 측정된다.
- 포아송 회귀모형은 바로 이 레이트 λ에 대해 회귀식을 세우게 된다.

$$\lambda_i = e^{\beta_0 + \beta_1 X_i}$$

$$\ln \lambda_i = \beta_0 + \beta_1 X_i$$

- 로짓 회귀모형과 이 부분을 비교해보자. 그때 종속변수는 로짓(logit)이었다.

$$\ln \frac{P(Y=1|X)}{1-P(Y=1|X)} = \beta_0 + \beta_1 X_i$$

포아송 회귀모형

- 사건 발생의 기대 횟수(expected count) $\mathbb{E}(Y|X)$는 다음과 같다. 이때 E는 노출(exposure)이고 시간, 면적, 인구 따위가 전형적이다.

$$\begin{aligned}\mathbb{E}(Y_i|X_i) &= E_i \cdot \lambda_i \\ &= E_i \cdot e^{\beta_0+\beta_1 X_i} \\ &= e^{\ln(E_i)+\beta_0+\beta_1 X_i}\end{aligned}$$

- 종종 실제 분석에서는 $E_i = 1$를 설정하기도 한다. 그러면 레이트 λ와 기대 횟수 $\mathbb{E}(Y|X)$가 같아지고, $\ln(E) = 0$이 된다.

포아송 회귀모형

- 노출이 다른 사건 발생은 상호 독립적이라면, Y_i는 포아송 분포를 따른다는 것을 유도할 수 있다(증명 생략). 이때 $\mu_i = \mathbb{E}(Y_i|X_i)$ 임을 기억하자.

$$Y_i \sim \text{Poisson}(\mu_i)$$

- 포아송 분포의 확률밀도함수(PDF)는 다음과 같다.

$$P(Y_i = y|X_i) = \frac{e^{-\mu_i}\mu_i^y}{y!}$$

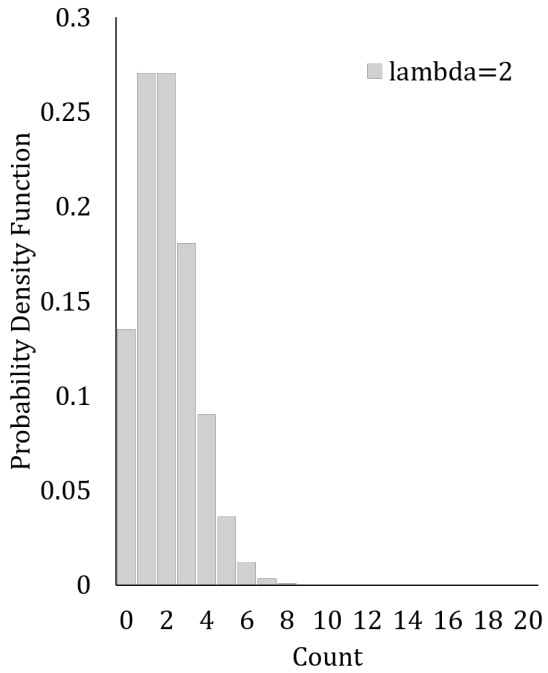
- 이때 μ는 평균 사건 발생 횟수이다($mathbbE(Y_i|X_i)$와 같은 것이다).

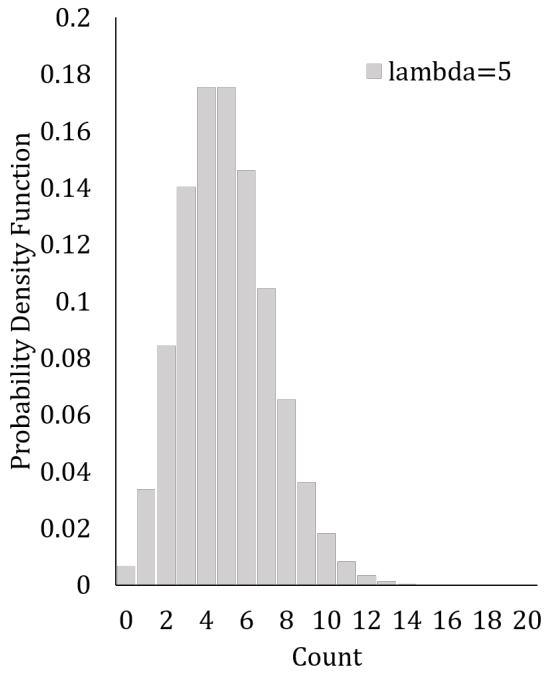
$$\mathbb{E}(Y_i|X_i) \equiv \mu_i = E_i \cdot \lambda_i = E_i \cdot e^{\beta_0+\beta_1 X_i}$$

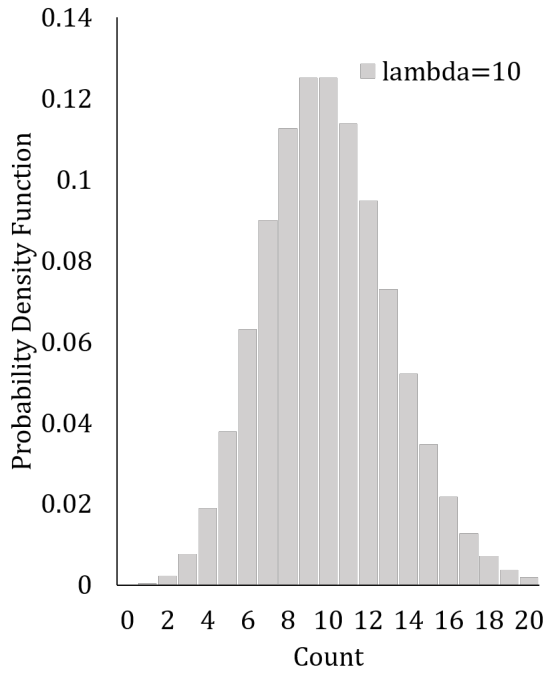
포아송 회귀모형

포아송 분포는 독특한 형태를 가지고 있기 때문에 기억해두어야 한다.

- 사건 발생 횟수의 평균이 작다면 사건발생 횟수가 0일 확률은 오히려 낮아진다.
- 사건 발생 횟수의 평균이 충분히 커지면 정규분포에 근사한다(approximate).







포아송 회귀모형

최대우도법을 활용하여 포아송 회귀모형의 회귀계수를 추정한다.

- 우도함수(likelihood function)의 직관적 의미를 잊지 말고 떠올려 보자.
- 이것은 앞서 설명한 확률밀도함수에서 유도될 수 있다. 그다지 어렵지 않으므로 성급하게 포기하지 말자.
- 아래 로그 우도함수를 극대화하는 β를 찾는 것이 바로 포아송 회귀모형인 셈이다.

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} P(Y = y|X)dX$$

$$\ln \mathcal{L}(\beta) = \sum_{i=1}^{n} [-\mu_i + y_i \log \mu_i - \log(y_i!)]$$

회귀계수의 해석

회귀계수의 해석

포아송 회귀모형의 회귀계수 해석은 매우 직관적이다.

- E가 노출, λ가 레이트라면 그 곱은 사건 발생의 기대 횟수 $\mathbb{E}(Y_i|X_i)$이다.

$$\mathbb{E}(Y_i|X_i) = E_i \cdot \lambda_i = e^{\ln(E_i)+\beta_0+\beta_1 X_i}$$

- 이제 e^{β}는 X가 한 단위 증가했을 때, 사건 발생의 기대 횟수가 증가하는 비율로 곧장 해석될 수 있다(Why?).

$$\frac{\mathbb{E}(Y_i|X_i+1)}{\mathbb{E}(Y_i|X_i)} = \frac{e^{\ln(E_i)+\beta_0+\beta_1(X_i+1)}}{e^{\ln(E_i)+\beta_0+\beta_1 X_i}} = e^{\beta}$$

- "X가 한 단위 증가하면, 사건 발생 횟수는 e^{β_1} 배 만큼 증가한다. 또는 $(e^{\beta_1}-1)\times 100\%$ 만큼 증가한다."

회귀계수의 해석

- 그런데 만일 E가 노출, λ가 레이트라면 그 곱은 사건 발생의 기대 횟수 $\mathbb{E}(Y_i|X_i)$이다.

$$\mathbb{E}(Y_i|X_i) = E_i \cdot \lambda_i = e^{\ln(E_i)+\beta_0+\beta_1 X_i}$$

- 이제 e^{β}는 X가 한 단위 증가했을 때, 사건 발생의 기대 횟수가 변화하는 비율로 해석될 수 있다(Why?).

$$\frac{\mathbb{E}(Y_i|X_i+1)}{\mathbb{E}(Y_i|X_i)} = \frac{e^{\ln(E_i)+\beta_0+\beta_1(X_i+1)}}{e^{\ln(E_i)+\beta_0+\beta_1 X_i}} = e^{\beta}$$

- "X가 한 단위 증가하면, 사건 발생 횟수는 e^{β_1} 배 만큼 증가한다. 또는 $(e^{\beta_1}-1)\times 100\%$ 만큼 증가한다."

회귀계수의 해석

- 위 식에서 양변을 노출 E로 나누어주면, 이제부터 단위 노출 당 사건 횟수 λ의 비율로 해석할 수 있게 된다.

$$\mathbb{E}\left(\frac{Y_i}{E} \mid X_i\right) = \frac{\lambda_i}{E_i} = e^{\beta_0+\beta_1 X_i}$$

- 이제 e^{β}는 X가 한 단위 증가했을 때, 사건 발생 레이트(incidence rate)가 변화하는 비율로 해석될 수 있다(Why?).

$$\frac{\mathbb{E}(Y_i/E|X_i+1)}{\mathbb{E}(Y_i/E|X_i)} = \frac{e^{\beta_0+\beta_1(X_i+1)}}{e^{\beta_0+\beta_1 X_i}} = e^{\beta}$$

- "X가 한 단위 증가하면, 사건 발생율은 e^{β_1} 배 만큼 증가한다. 또는 $(e^{\beta_1}-1)\times 100\%$ 만큼 증가한다."

회귀계수의 해석

- 우리는 e^{β_1}를 특별히 사건 발생률 변화율(Incidence-rate ratio; IRR)이라고 부른다 (번역이 몹시 이상하다).
- 사건 횟수 해석법에서는 노출이 커지면 사건 횟수도 자연스럽게 증가한다.

$$\mathbb{E}(Y_i|X_i) = E_i \cdot \lambda_i = E_i \cdot e^{\beta_0+\beta_1 X_i}$$

- 단위 노출당 사건 횟수(또는 레이트 해석법)에서는 사건 횟수가 노출에 따라 표준화된다.

$$\mathbb{E}(Y_i/E_i|X_i) = \lambda_i = e^{\beta_0+\beta_1 X_i}$$

- 두 가지 해석법은 사실 수학적으로 같지만, 실제 의미에서는 약간 다르게 들리므로 잘 구분하자.

회귀계수의 해석

- 예측 횟수(predicted count), 예측 확률(predicted probabilities), 또는 평균한계효과(average marginal effect)도 해석에 사용할 수 있다.
- 예측 횟수는 다음과 같다.

$$\widehat{Y}_i = \mathbb{E}(Y_i \mid X_i) = E_i \cdot e^{\widehat{\beta}_0 + \widehat{\beta}_1 X_i}$$

- 예측 확률, 즉 i번째 관측치에서 사건이 정확히 y번 발생할 확률은 다음과 같다.

$$P(Y_i = y|X_i) = \frac{e^{-\widehat{\mu}_i}\widehat{\mu}_i^y}{y!} \qquad \left(\widehat{\mu}_i = E_i \cdot e^{\widehat{\beta}_0 + \widehat{\beta}_1 X_i}\right)$$

- AME는 다음과 같다.

$$\frac{1}{n}\sum \frac{\partial \mathbb{E}(Y_i|X_i)}{\partial X_i} = \frac{1}{n}\sum \frac{\partial}{\partial X_i} E_i \cdot e^{\beta_0 + \beta_1 X_i} = \frac{1}{n}\sum E_i \cdot e^{\beta_0 + \beta_1 X_i}\beta_1$$

- 포아송 회귀모형에서 AME는 잘 안쓰이고 IRR이 주로 쓰인다.

유의성 검정과 모형 적합도

유의성 검정과 모형 적합도

유의성 검정과 모형적합도 역시 이항 로짓 회귀모형과 똑같다.

- 순서형 로짓 회귀모형에서도 회귀계수의 유의성 검정을 위해 왈드 검정을 사용한다.

$$\chi_1^2 = \left(\frac{\hat{b} - \beta}{\mathrm{SE}_{\hat{b}}}\right)^2 = \left(\frac{\hat{b}}{\mathrm{SE}_{\hat{b}}}\right)^2$$

- χ_1^2 값이 충분히 크면 통계적으로 유의하게 귀무가설($H_0 : \beta = 0$)을 기각할 수 있다.
- Stata에서는 Z 검정을 한다. 표본 크기가 충분히 커지면 왈드 검정과 t 검정, Z 검정은 결국 수렴한다.

$$Z = \frac{\hat{b}}{\mathrm{SE}_{\hat{b}}}$$

유의성 검정과 모형 적합도

- 모형적합도를 살펴보기 위해 (1) 우도비 검정(likelihood-ratio test), (2) 유사 결정계수(pseudo R^2), (3) 정보 기준(information criteria), (4) Pearson's chi^2 기준을 주로 확인한다.
- 첫째, 우도비 검정의 검정통계량 G는 χ^2 분포를 따른다.

$$G = -2 \ln \frac{\mathcal{L}_{\text{null}}}{\mathcal{L}_{\text{model}}} = -2 \ln(\mathcal{L}_{\text{null}} - \mathcal{L}_{\text{model}}) \sim \chi^2_k$$

- 우도비 검정의 귀무가설은 다음과 같다. 이를 기각하지 못하면 "이 모형은 아무 짝에도 쓸모가 없다"라는 의미로 받아들여진다(Why?).

$$H_0 : \mathcal{L}_{\text{null}} = \mathcal{L}_{\text{model}}$$

유의성 검정과 모형 적합도

- 똑같은 원리를 사용하여 두 모형을 비교할 때도 우도비 검정을 사용할 수 있다.

$$G = -2 \ln \frac{\mathcal{L}_{\text{restricted}}}{\mathcal{L}_{\text{full}}} = -2(\ln \mathcal{L}_{\text{restricted}} - \ln \mathcal{L}_{\text{full}}) \sim \chi^2_{\Delta k}$$

- $\mathcal{L}_{\text{restricted}}$는 독립변수가 다소 적게 들어간 모형의 우도 함수값이고, $\mathcal{L}_{\text{full}}$은 그보다 독립변수가 좀 더 들어간 모형의 우도 함수값이다.
- 이때 제한모형(restricted model)은 완전모형(full model) 안에 내포되어(nested) 있어야 한다.
- χ^2로 검정하는 귀무가설은 아래와 같다. 이를 기각하지 못하면 "완전모형이 제한모형보다 나은 구석이 없다"라는 의미로 받아들여진다.

$$H_0 : \mathcal{L}_{\text{restricted}} = \mathcal{L}_{\text{full}}$$

유의성 검정과 모형 적합도

- 둘째, 유사결정계수는 선형회귀모형에서 사용되는 결정계수 R^2를 비선형모형에서 흉내낸 것이다.
- Stata에서는 McFadden's R^2를 사용한다.

$$\text{McFadden's } R^2 = 1 - \frac{\ln \mathcal{L}_{\text{model}}}{\ln \mathcal{L}_{\text{null}}}$$

- 값이 클수록 모형의 설명력이 높다고 말할 수 있지만, 선형회귀모형의 R^2처럼 설명된 분산의 비율로 해석하지 않도록 주의해야 한다.

유의성 검정과 모형 적합도

- 셋째, 아카이케 정보기준(AIC)또는 베이즈 정보기준(BIC)을 보고할 수 있다.

$$AIC = -2\ln \mathcal{L}_{\text{model}} + 2k$$
$$BIC = -2\ln \mathcal{L}_{\text{model}} + k\ln(n)$$

- 당연히 AIC와 BIC 둘 다 그 값이 작을수록 좋다(Why?).
- 정보기준은 우도비 검정과는 달리 내포성(nestedness) 여부를 따지지 않는다. 그러나 유의성 검정 단계가 없으므로 모형의 개선 여부 판단이 약간 애매모호하다.

유의성 검정과 모형 적합도

- 넷째, Pearson's χ^2 GOF 지표는 예측된 횟수와 실제 횟수 간의 차이를 χ^2 검정을 통해 확인한다.
- Pearson 검정통계량 χ^2는 다음과 같이 정의된다.

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

- 물론 y_i는 관측된 사건 수이고, $\hat{\mu}_i = \mathbb{E}(Y_i \mid X_i)$는 모형에 기반한 기대값이다.
- 귀무가설은 자료가 포아송 분포한다는 것이므로, 어지간하면 기각하고 싶지 않다 (Why?).

음이항 회귀모형

음이항 회귀모형

포아송 회귀모형은 무너지기 쉬운 가정을 전제로 한다.

- 이 모형은 수학적으로 아름답고 많은 자연 현상을 잘 설명하지만, 사회 현상을 설명할 때는 다소 문제점을 안고 있다.
- 포아송 회귀모형은 다음의 가정을 필요로 한다(증명 생략).

$$Var(Y|X) = \mathbb{E}(Y|X)$$

- 사회적 현실에서는 $Var(Y|X) > \mathbb{E}(Y|X)$ 가 매우 빈번하게 발생하며 이런 경우를 과대산포(overdispersion)라고 부른다.
- 특히 관측치 간에 독립성(independence)이 보장되어 있지 않을 때 이런 문제가 발생하기 쉬운 것으로 알려져 있다.

음이항 회귀모형

- 다행히 유력한 대안이 여럿 존재하며, 특히 음이항(negative binomial) 회귀모형이 폭넓게 사용된다.
- 음이항 회귀모형은 이른바 산포 모수(dispersion parameter)를 하나 더 사용하여 분산에 대해서도 다음과 같이 모형화한다.

$$Y_i \sim \text{Poisson}(\mu_i)$$

$$\mu_i = \mathbb{E}(Y_i|X_i) = e^{\ln(E_i)+\beta_0+\beta_1 X_i+\ln(\nu)}$$

음이항 회귀모형에서는 μ_i이 확률적으로 변할 수 있도록 모형화한다.

$$\mu_i^* = \mu_i\nu_i, \quad \nu_i \sim \text{Gamma}\left(\frac{1}{\alpha}, \alpha\right)$$

- 포아송 회귀모형과의 차이점에 주목하자.

음이항 회귀모형

- 감마 분포의 특성을 통해 다음을 알 수 있다(증명 생략).

$$\mu_i^* \sim \text{Gamma}\left(\frac{1}{\alpha}, \alpha\mu_i\right)$$

- 최종적으로 우리는 Y의 조건부 분산을 계산할 수 있다(증명 생략).

$$\begin{aligned} Var(Y_i|X_i) &= \mu_i\left(1+\alpha\mu_i\right) \\ &= \mathbb{E}(Y_i|X_i)\left(1+\alpha\mathbb{E}(Y_i|X_i)\right) \end{aligned}$$

- α가 바로 산포 모수이고, 유의성 검정의 대상이 된다.

음이항 회귀모형

- 만약 $H_0 : \alpha = 0$ 이라는 귀무가설을 기각하지 못하면 포아송 회귀분석으로 축약된다 (Why?).

$$Var(Y|X) = \mathbb{E}(Y|X)\left(1 + 0 \cdot \mathbb{E}(Y_i|X_i)\right) = \mathbb{E}(Y|X)$$

- 반대로 $H_a : \alpha > 0$라는 대립가설을 채택하게 되면 음이항 회귀모형을 선택한다.
- 검정통계량 χ^2를 통해 귀무가설을 검정한다.

$$\chi^2_1 = 2\left(\ln \mathcal{L}_{\text{NB}} - \ln \mathcal{L}_{\text{Poisson}}\right)$$

- 음이항 회귀모형은 포아송 회귀모형보다 좀 더 일반화된 형태라고 할 수 있다(Why?).

음이항 회귀모형

음이항 회귀모형이 포아송 회귀모형보다 많이 쓰이는 것 같다.

- 잘라 말하긴 어렵지만, 사회 현상은 대부분 과대산포 경향이 있다(Why?). 그로 인해 포아송과 음이항의 결과가 제법 다른 경우가 많다.
- 사회과학도의 응용 연구의 입장에서 일단 반드시 음이항 회귀모형을 꼭 적용해 보는 것이 좋다.
- 만일 포아송 회귀모형과 결과가 크게 다르지 않다면 포아송으로 가도 좋다. 그렇지 않다면 음이항 회귀모형 쪽이 선호된다.

영팽창 회귀모형

영팽창 회귀모형

유독 자료에 0이 많다면 영팽창 회귀모형을 고려해야 한다.

- 왜 0이 많을까? 특별한 이론적 고려가 필요없다면 음이항 회귀모형이 적절할 수도 있다.
- 그러나 자료생성 과정(data-generating process)에 있어서 0을 만드는 상태와 그렇지 않은 상태가 혼합되어 있다면, 이를 고려한 모형이 필요하다.
- 영팽창 포아송 회귀모형은 다음의 두 가지 프로세스로 구성된다(Why?).

  (1) F_i의 가중치로 항상 0이 나오는 상태(structural zero)
  (2) $1 - F_i$의 가중치로 포아송 분포에 따라 생성된 상태

$$P(Y_i = 0|X) = F_i + (1 - F_i)e^{-\mu_i}$$

$$P(Y_i = y_i > 0|X) = (1 - F_i) \cdot \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}$$

영팽창 회귀모형

- 이때, F_i은 로짓 또는 프로빗 모형을 사용할 수 있다.

$$F_i = \frac{1}{1+\exp(-Z_i\gamma)}$$

$$1 - F_i = \frac{\exp(-Z_i\gamma)}{1+\exp(-Z_i\gamma)}$$

- 좋은 영팽창 회귀모형을 자료에 적합하기 위해서는 좋은 Z를 찾아내야 한다(이것은 생각보다 쉽지 않다).
- 그러므로 영과잉(excessive zeroes)을 만들어내는 구조적 요소가 무엇인지 이론적으로 잘 포착해야 한다.

영팽창 회귀모형

과대산포까지 있다면 영팽창 음이항 회귀모형을 사용할 수 있다.

- 구조는 사실상 매우 유사하다.

$$P(Y_i = 0) = F_i + (1 - F_i) \cdot P_{\text{NB}}(0|\mu_i, \alpha)$$
$$P(Y_i = y_i > 0) = (1 - F_i) \cdot P_{\text{NB}}(Y = y_i \mid \mu_i, \alpha)$$

- 마찬가지로 F_i은 로짓 또는 프로빗 모형을 통해 추정한다.
- 이 식이 본질적으로는 영팽창 포아송 회귀모형과 같음을 이해하자.

영팽창 회귀모형

- ZIP과 ZINB는 서로 내포되어 있지 않아 우도비 검정을 사용할 수 없다. 대신 AIC나 BIC를 사용하자.
- $H_0 : \alpha = 0$ 귀무가설을 기각할 수 있으면 음이항 회귀모형 또는 영팽창 음이항 회귀모형을 선택한다.
- 원론적으로 보면, 과대산포 문제를 일으키는 매커니즘이 영팽창 회귀모형을 지지할 수도 있고, 음이항 회귀모형을 지지할 수도 있다(Why?).
- 따라서 네 가지 모형을 모두 적합시키고 비교해보는 것도 괜찮은 전략이 된다.