Negative Binomial Regression

Jieun Shin

2023-04-29

1. 음이항 회귀모형

y를 음이항 분포를 따르는 계수형 (count) 값이라 하자. 포아송 분포에서는 평균과 분산이 같지만, 음이항 분포에서는 평균이 분산보다 작다고 가정한다. 음이항 모형은 (1)반응변수 y가 어떤 사건이나 현상에 대한 계수값을 가지고 (2) 음이항 분포를 구성하는 모수를 가지는 일반화 선형모형 (GLM; generalized linear model)이다. y의 평균 μ와 산포모수 $\alpha > 0$를 갖는 음이항 분포는 다음과 같이 정의된다.

$$f(y) = \mathbb{P}(Y = y) = \frac{\Gamma(y + \frac{1}{\alpha})}{\Gamma(y+1)\Gamma(\frac{1}{\alpha})} \left(\frac{1}{1 + \alpha\mu} \right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu}{1 + \alpha\mu} \right)^{y}$$

연결함수 (link function)에 의해 $\ln \mu = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$로 표현되고 여기서 $X_1, X_2, \cdots, X_p$는 독립변수, $\beta_1, \beta_2, \cdots, \beta_p$는 회귀계수이다. 각 변수는 n개의 관측값 $(x_{1j}, x_{2j}, \cdots, x_{nj})^T$을 가지고, $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)^T$는 모수벡터라 하자. 그러면 설계행렬 (design matrix) $\boldsymbol{X}$는

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{12} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

와 같이 나타내어진다. 음이항 분포를 $i = 1, 2, \ldots, n$번째 관측치에 대하여

$$\begin{aligned} f(y_i) = \mathbb{P}(Y = y_i) &= \frac{\Gamma(y_i + \frac{1}{\alpha})}{\Gamma(y_i+1)\Gamma(\frac{1}{\alpha})} \left(\frac{1}{1 + \alpha\mu_i} \right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu_i}{1 + \alpha\mu_i} \right)^{y_i} \\ &= \frac{\Gamma(y_i + \frac{1}{\alpha})}{\Gamma(y_i+1)\Gamma(\frac{1}{\alpha})} \left(\frac{1}{1 + \alpha\exp(X_i\boldsymbol{\beta})} \right)^{\frac{1}{\alpha}} \left(\frac{\alpha\exp(X_i\boldsymbol{\beta})}{1 + \alpha\exp(X_i\boldsymbol{\beta})} \right)^{y_i} \end{aligned}$$

와 같이 정리할 수 있다.

2. 로그가능도 (log-likelihood) 함수

음이항 분포의 로그 가능도함수는 다음과 같다.

$$\log L(\alpha, \boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ y_i \log\alpha + \left(y_i + \frac{1}{\alpha} \right) \log(1 + \alpha\exp(X_i\boldsymbol{\beta}) + \log\Gamma\left(y_i + \frac{1}{\alpha} \right) - \log(y_i + 1) - \log\Gamma\left(\frac{1}{\alpha} \right) \right\}$$

3. 음이항분포의 유도

λ와 u가 주어졌을 때, y의 분포를 $f(y_i;\lambda,u)=\frac{e^{-\lambda_i u_i}(\lambda_i u_i)^{y_i}}{y_i!}$라 하자.

감마분포에 의해 $g(u)$를 정의하는 방법으로부터 y의 분포는 $u=\exp(\epsilon)$이 된다. 여기서 $\log\mu_i = x_i\boldsymbol{\beta}+\epsilon_i$이고 평균은 감마분포의 평균과 같다. 감마분포와 u가 주어졌을 때의 평균을 갖는 포아송분포의 혼합분포의 평균은 u가 주어졌을 때 y의 평균이며 분산은 u가 주어졌을 때 y의 분산이다. 포아송-감마 혼합분포는 다음과 같이 정의된다.

$$
\begin{aligned}
f(y_i;\lambda,u) &= \int_0^\infty \frac{e^{-\lambda_i u_i}(\lambda_i u_i)^{y_i}}{y_i!}\frac{v^v}{\Gamma(v)}u_i^{v-1}e^{-vu_i}du_i \\
&=\frac{\lambda_i^{y_i}}{\Gamma(y_i+1)}\frac{v^v}{\Gamma(v)}\int_0^\infty e^{-\lambda_i u_i}\cdot u_i^{(y_i+v)-1}du_i \\
&=\frac{\Gamma(y_i+v)}{\Gamma(y_i+1)\Gamma(v)}\cdot\left(\frac{v}{\lambda_i+v}\right)^v\cdot\left(\frac{\lambda_i}{\lambda_i+v}\right)^{y_i} \\
&=\frac{\Gamma(y_i+v)}{\Gamma(y_i+1)\Gamma(v)}\cdot\left(\frac{1}{1+\frac{\lambda_i}{v}}\right)^v\cdot\left(1-\frac{1}{1+\frac{\lambda_i}{v}}\right)^{y_i}
\end{aligned}
$$

감마분포의 척도모수 (scale parameter) α가 v의 역수 꼴이며, 음이항 분포에서 과산포 (overdispersion)모수 혹은 이질성 (heterogeneity)모수라 부른다. 최종적으로 음이항 분포가 다음과 같이 정의된다.

$$
f(y_i;\mu,\alpha)=\frac{\Gamma(y_i+\frac{1}{\alpha})}{\Gamma(y_i+1)\Gamma(\frac{1}{\alpha})}\cdot\left(\frac{1}{1+\lambda_i\alpha}\right)^{\frac{1}{\alpha}}\cdot\left(1-\frac{1}{1+\lambda_i\alpha}\right)^{y_i}
$$

4. 추정방법

α와 $\boldsymbol{\beta}$는 IRLS (iteratively reweighted least square)로 추정한다. 이 방법은 피셔 스코어 함수 (Fisher score function)를 이용하는 방법이며, 이는 1차 미분한 행렬로 선형모형을 추정하기 위해 사용되는 최대 우도 (ML; maximum likelikhood) 추정의 일부이다.

지수족 분포는 $f(y_i;\theta,\phi)=\exp\left\{\frac{y_i\theta_i-b(\theta_i)}{\alpha_i(\phi)}+c(y_i;\phi)\right\}$와 같이 나타내어지는데, 여기서 θ_i는 정준 모수 (canonical parameter) 또는 연결함수이고, $b(\theta_i)$는 누적량 (cumulant), $\alpha(\phi)$는 척도모수, $c(y_i;\phi)$는 정규화 상수 (normalization term)이다. 지수족 분포의 장점은 특정 분포의 θ에 대한 1차, 2차미분으로부터 유일한 (unique)한 평균과 분산을 알 수 있다는 것이다:

$$
b'(\theta_i)=\text{mean},\quad b''(\theta_i)=\text{variance}
$$

일반화 선형모형의 pdf는 $f(y_i;\theta,\phi)$이고 여기서 y_i는 반응변수 (response variable), θ_i는 위치모수 (location parameter), ϕ는 척도모수이다. 로그 가능도함수를 $L(\theta_i,\phi;y_i)$라 하면 IRLS는 테일러 전개 (Taylor expansion)를 기반으로 유도된다:

$$
0=f(y_0)+f(y_1-y_0)f'(y_0)+\frac{(y_1-y_0)^2}{2!}f''(y_0)+\cdots
$$

처음 두 번째 항까지만 고려하면,

$$
\begin{aligned}
&0=f(y_0)+f(y_1-y_0)f'(y_0) \\
&\Rightarrow y_1=y_0-\frac{f(y_0)}{f'(y_0)}
\end{aligned}
$$

의 관계식을 얻는다.

로그 가능도 함수는 최대점 (peak)이 존재한다. ML추정은 그래디언트 (gradient) 혹은 피셔 스코어 (로그 가능도함수의 β에 대한 1차 도함수)를 0으로 놓고 푸는 것이다. 로그 가능도 함수의 2차 도함수 행렬을 정보행렬 (information matrix) 또는 헤시안 행렬 (Hessian matrix)라고 부르며 분산-공분산 행렬이 된다. 분산-공분산 행렬의 대각원소로부터 추정량의 표준오차를 알 수 있다.

IRLS 알고리즘의 설명을 위해 로그 가능도 함수의 1차 도함수 행렬을 U로, 2차 도함수 행렬을 H로 표기하자:

$$U = \partial L, \quad H = \partial^2 L$$

뉴튼-랩슨 (Newton-Raphson) 알고리즘을 이용한 모수 추정값은

$$\boldsymbol{\beta}^+ \to \boldsymbol{\beta} - H^{-1}U$$

를 반복함으로써 얻는다. 로그 가능도 함수는 지수족의 형태로 $L(\theta_i; y_i, \phi) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i;\phi)$이고, β_j에 대하여 L을 풀기위해 연쇄법칙 (chain rule)을 이용하면

$$\frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial L}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}$$

와 같이 전개할 수 있다. 각 term 을 차례대로 풀면,

$$\frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i\theta_i - b'(\theta_i)}{a_i(\phi)} + \sum_{i=1}^{n} \frac{y_i - \mu_i}{a_i(\phi)}$$

이고, $b'(\theta_i) = \mu_i$임을 이용하여 두 번째 term

$$\begin{aligned} \frac{\partial \mu_i}{\partial \theta_i} &= \frac{\partial b'(\theta_i)}{\partial \theta_i} = b''(\theta_i) = V(\mu_i) \\ \Rightarrow \frac{\partial \theta_i}{\partial \mu_i} &= \frac{1}{V(\mu_i)} \end{aligned}$$

을 얻을 수 있다.

그리고 $\eta_i = X_i^T \beta_j$이므로 네 번째 term,

$$\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial (X_i^T \beta_j)}{\partial \beta_j} = X_{ij}$$

을 얻을 수 있다. 그리고 세 번째 term,

$$\frac{\partial \mu_i}{\partial \eta_i} = [g^{-1}(\eta_i)]' = \frac{1}{\frac{\partial \eta_i}{\partial \mu_i}} = \frac{1}{g'(\mu_i)}$$

을 얻는다. 또한 μ_i의 η_i에 대한 미분은 연결함수의 역수임을 이용한다:

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_i}{a_i(\phi)V(\mu_i)g'(\mu_i)} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_i}{a_i(\phi)V(\mu_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right) = 0$$

여기서 y_i는 반응변수, μ_i는 적합된 변수 (fitted variable)이다. 정보행렬을 $I = E\left[\frac{\partial^2 L}{\partial \beta_j \partial \beta_k}\right] = E\left[\frac{\partial L}{\partial \beta_j}\frac{\partial L}{\partial \beta_k}\right]$라 놓으면 다음과 같이 전개할 수 있다:

$$
\begin{aligned}
I &= \frac{\partial}{\partial \beta_j}\left[\frac{(y_i-\mu_i)x_j}{a_i(\phi)V(\mu_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)\right]\cdot \frac{\partial}{\partial \beta_k}\left[\frac{(y_i-\mu_i)x_k}{a_i(\phi)V(\mu_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)\right] \\
&= \frac{(y_i-\mu_i)^2 x_j x_k}{\{a_i(\phi)V(\mu_i)\}^2}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2
\end{aligned}
$$

이 때, $(y_i-\mu_i)^2 = a_i(\phi)V(\mu_i)$이므로 $V(y_i) = a_i(\phi)V(\mu_i) = (y_i-\mu_i)^2$라 하면

$$
I = \frac{x_j x_k}{V(y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 = \frac{x_j x_k}{V(y_i)g'^2}
$$

이 성립한다. 따라서 뉴튼-랩슨 알고리즘은

$$
\boldsymbol{\beta}^+ \leftarrow \boldsymbol{\beta} - \left[\frac{x_j x_k}{V(y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2\right]^{-1}\cdot\left[\frac{(y_i-\mu_i)x_k}{V(y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)\right]
$$

이 된다.

계속해서 역함수 term을 양 변에 곱해보자:

$$
\left[\frac{x_j x_k}{V(y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2\right]\boldsymbol{\beta}^+ = \left[\frac{x_j x_k}{V(y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2\right]\boldsymbol{\beta} + \left[\frac{(y_i-\mu_i)x_k}{V(y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)\right]
$$

여기서 $W = \frac{1}{V(y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$라 놓으면 선형 예측자 (linear predictor) $\eta_i = X_i\beta$에 대하여 왼쪽 항은

$$
\left[\frac{x_j x_k}{V(y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2\right]\boldsymbol{\beta}^+ = (X^T W X)\boldsymbol{\beta}^+
$$

로 표현되며, 오른 쪽 첫번째 항은

$$
\left[\frac{x_j x_k}{V(y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2\right]\boldsymbol{\beta} = X^T W \eta_i
$$

이 되고, W와 $V(y_i)$의 정의에 의하여 오른쪽 두 번째 항은

$$
\frac{(y_i-\mu_i)x_k}{V(y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right) = \frac{(y_i-\mu_i)x_k}{\frac{1}{W}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2}\left(\frac{\partial \mu_i}{\partial \eta_i}\right) = x_k W (y_i-\mu_i)\left(\frac{\partial \mu_i}{\partial \eta_i}\right)
$$

으로 표현할 수 있다.

정리하면

$$
\begin{aligned}
(X^T W X)\boldsymbol{\beta}^+ &= X^T W \eta_i + x_k W (y_i-\mu_i)\left(\frac{\partial \mu_i}{\partial \eta_i}\right) \\
&= X^T W \eta_i + \frac{(y_i-\mu_i)x_k}{\frac{1}{W}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)
\end{aligned}
$$

이고, $z_i = \eta_i + (y_i-\mu_i)\left(\frac{\partial \mu_i}{\partial \eta_i}\right)$라 놓으면 최종적으로

$$
\begin{aligned}
&(X^TWX)\boldsymbol{\beta}^+ = X^TWZ \\
\Rightarrow &\boldsymbol{\beta}^+ = (X^TWX)^{-1}WZ
\end{aligned}
$$

를 얻는다.

5. 예시

5-1. 로지스틱 회귀

확률변수 Y_i가 $B(n_i, p_i)$를 따른다고 가정하고 성공의 비율 Y_i/n_i을 모델링하는 것이 목적으로 두자. 그러면 비율의 평균과 분산은 각각 $\mathbb{E}(Y_i/n_i) = p_i$, $V(Y_i/n_i) = p_i(1-p_i)/n_i$이다. 따라서 분산 함수는 $V(\mu_i) = \mu_i(1-\mu_i)$이다. 연결함수는 $\eta_i : (0,1) \mapsto (-\infty, \infty)$이며, 다음과 같은 형태를 갖는다:

$$
\eta_i = \text{logit } \mu_i = \log \frac{\mu_i}{1-\mu_i}
$$

로지스틱 회귀에서의 연결함수 η_i, working response z_i와 weights w_i는 다음과 같이 정리된다.

$$
\begin{aligned}
\eta_i &= \text{logit } \mu_i, \\
z_i &= \eta_i + (y_i - \mu_i)\frac{\partial \mu_i}{\partial \eta_i} = \eta_i + \frac{y_i - \mu_i}{\mu_i(1-\mu_i)}, \\
w_i &= \frac{1}{V(\mu_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 = \mu_i(1-\mu_i)
\end{aligned}
$$

5-2. 포아송 회귀

확률변수 Y_i가 $\text{Poisson}(\mu_i)$를 따른다고 가정하자. 그러면 Y_i의 평균과 분산은 각각 $\mathbb{E}(Y_i) = \mu_i$, $V(Y_i) = \mu_i$이다. 따라서 분산 함수는 $V(\mu_i) = \mu_i$이다. 연결함수는 $\eta_i : (0,\infty) \mapsto (-\infty, \infty)$이며, 다음과 같은 형태를 갖는다:

$$
\eta_i = \log \mu_i
$$

포아송 회귀에서의 연결함수 η_i, working response z_i와 weights w_i는 다음과 같이 정리된다.

$$
\begin{aligned}
\eta_i &= \log \mu_i, \\
z_i &= \eta_i + (y_i - \mu_i)\frac{\partial \mu_i}{\partial \eta_i} = \eta_i + \frac{y_i - \mu_i}{\mu_i}, \\
w_i &= \frac{1}{V(\mu_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 = \mu_i
\end{aligned}
$$

5-3. 음이항 회귀

확률변수 Y_i가 $NB(\mu_i, \phi)$를 따른다고 가정하자. 그러면 Y_i의 평균과 분산은 각각 $\mathbb{E}(Y_i) = \mu_i$, $Var(Y_i) = \mu_i + \frac{\mu_i^2}{\phi}$이다. 따라서 분산 함수는 $V(\mu_i) = \mu_i + \frac{\mu_i^2}{\phi}$이다. 연결함수는 $\eta_i : (0,\infty) \mapsto (-\infty, \infty)$이며, 다음과 같은 형태를 갖는다:

$$
\eta_i = \log \mu_i
$$

음이항 회귀에서의 연결함수 η_i, working response z_i와 weights w_i는 다음과 같이 정리된다.

$$
\begin{aligned}
\eta_i &= \log \mu_i, \\
z_i &= \eta_i + (y_i - \mu_i)\frac{\partial \mu_i}{\partial \eta_i} = \eta_i + \frac{y_i - \mu_i}{\mu_i}, \\
w_i &= \frac{1}{V(y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 = \frac{1}{\mu_i + \frac{\mu_i^2}{\phi}}\frac{1}{\mu_i^2} = \frac{1}{1 + \mu_i \phi^{-1}}
\end{aligned}
$$

6. 시뮬레이션

가능도함수는 다음과 같이 정의된다:

$$
L(\boldsymbol{\beta}, \alpha; y_1, \dots, y_n) = \prod_{i=1}^{n} f(y_i; \mu, \alpha)
$$

이에 따라 로그 가능도함수는 다음과 같이 정의된다:

$$
\log L = \sum_{i=1}^{n} \left[\sum_{k=1}^{y_i - 1} \log(1 + \alpha y_i - \alpha k) - y_i \log(1 + \alpha \mu_i) - \frac{\log(1 + \alpha \mu_i)}{\alpha} + y_i \log \mu_i - \log y_i \right]
$$

모수벡터를 $\theta = (\boldsymbol{\beta}^T, \alpha)^T$로 정의할 때, score function은 다음과 같이 정의된다:

$$
S(\theta) = \begin{bmatrix} \frac{\partial \log L}{\partial \boldsymbol{\beta}} \\ \frac{\partial \log L}{\partial \boldsymbol{\alpha}} \end{bmatrix}
$$

여기서 구체적인 값은 다음과 같이 계산되는 값이다:

$$
\begin{aligned}
\frac{\partial \log L}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} \left[\frac{x_i (y_i - \mu_i)}{1 + \alpha \mu_i} \right] \\
\frac{\partial \log L}{\partial \boldsymbol{\alpha}} &= \sum_{i=1}^{n} \left[\sum_{k=1}^{y_i} \frac{y_i - k}{1 + \alpha y_i - \alpha k} - \frac{y_i \mu_i}{1 + \alpha \mu_i} - \frac{\mu_i}{\alpha(1 + \alpha \mu_i)} + \frac{\log(1 + \alpha \mu_i)}{\tau^2} \right]
\end{aligned}
$$

그리고 2차 도함수 행렬인 Hessian matrix는 다음과 같이 구성되며,

$$
H(\theta) = \begin{bmatrix} \frac{\partial^2 \log L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} & \frac{\partial^2 \log L}{\partial \boldsymbol{\beta} \partial \alpha} \\ \frac{\partial^2 \log L}{\partial \alpha \partial \boldsymbol{\beta}} & \frac{\partial^2 \log L}{\partial \alpha^2} \end{bmatrix}
$$

구체적으로 계산되는 값 역시 다음과 같다:

$$
\begin{aligned}
\frac{\partial^2 \log L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} &= \sum_{i=1}^{n} \left[-\frac{\mu_i (1 + \alpha y_i) x_i x_i^T}{(1 + \alpha \mu_i)^2} \right] \\
\frac{\partial^2 \log L}{\partial \boldsymbol{\beta} \partial \alpha} &= \sum_{i=1}^{n} \left[-\frac{(y_i - \mu_i) x_i}{(1 + \alpha \mu_i)^2} \right] \\
\frac{\partial^2 \log L}{\partial \alpha^2} &= \sum_{i=1}^{n} \left[-\sum_{k=1}^{y_i} \frac{(y_i - k)^2}{(1 + \alpha y_i - \alpha k)} + \frac{\mu_i^2 y_i}{(1 + \alpha \mu_i)^2} + \frac{\mu_i (1 + 2\alpha \mu_i)}{\alpha^2 (1 + \alpha \mu_i)^2} + \frac{\mu_i}{\alpha^2 (1 + \alpha \mu_i)} - \frac{2 \log(1 + \alpha \mu_i)}{\alpha^3} \right]
\end{aligned}
$$

다음으로 시뮬레이션에 사용될 함수를 지정하자.ㄴ 여기서 Dvec은 가능도함수의 1차 도함수 벡터, Hmat은 가능도함수의 2차 도함수 행렬이다.

```
Dvec = function(x, y, mu, tau){
  n = dim(x)[1]
  p = dim(x)[2]

  dbeta = c()
  for(j in 1:p){
    dbeta[j] = sum( x[,j] * (y-mu) / (1 + tau * mu) )
  }

  vv = c()
  for(i in 1:n){
    vv[i] = 0
    if(y[i] != 0) for(v in 0:(y[i]-1)) vv[i] = vv[i] + 1/(v+tau^{-1})
  }

  dtau = sum(tau^{-2} * (log(1+tau*mu) - vv) + (y-mu)/tau/(1+tau*mu))

  return(c(dbeta, dtau))
}


###

Hmat = function(x, y, mu, tau){
  n = dim(x)[1]
  p = dim(x)[2]

  ddbeta = matrix(0, p, p)
  for(j in 1:p){
    for(k in 1:p){
      ddbeta[j, k] = sum( mu*(1+tau*y)*x[,j]*x[,k] / (1+tau*mu)^2 )
    }
  }

  ddbetatau = c()
  for(j in 1:p){
    ddbetatau[j] = sum( mu*(y-mu)*x[,j]/(1+tau*mu)^2 )
  }


  vv = c()
  for(i in 1:n){
    vv[i] = 0
    if(y[i] != 0) for(v in 0:(y[i]-1)) vv[i] = vv[i] + (v/1+tau*v)^2
  }

  ddtau = sum( vv + 2*tau^{-3}*log(1+tau*mu)-(2*tau^{-2}*mu)/(1+tau*mu)
               - (y+tau^{-1}*mu^2)/(1+tau*mu)^2)

  Hessian = cbind(rbind(ddbeta, ddbetatau), c(ddbetatau, ddtau))
  rownames(Hessian) = NULL
```

```r
  return( Hessian )
}
```

모수벡터 θ의 추정은 뉴튼-랩슨알고리즘으로 한다:

$$\theta^{(t+1)} = \theta^{(t)} + [-H(\theta^{(t)})]^{-1}[S(\theta^{(t)})], \quad t = 0, 1, 2, 3, \dots$$

뉴튼랩슨 알고리즘도 함수로 정의해준다:

```r
# 뉴튼랩슨 알고리즘
NewtonRaphson = function(x, y){
  n = dim(x)[1]
  p = dim(x)[2]

  max_iter = 200
  eps = 1e-5

  theta = runif(p+1, 0.3, 0.5)  # (beta, tau)
  theta_new = 100

  t = 0
  while(sum((theta-theta_new)^2) > eps || t < max_iter){
    t = t + 1

    if(t != 1){  # update
      theta = theta_new
    }
    mu_new = exp(x %*% theta[1:p])  # theta[1:p] is mu
    tau_new = theta[p+1]            # theta[p+1] is tau

    D = Dvec(x, y, mu_new, tau_new)
    H = Hmat(x, y, mu_new, tau_new)

    theta_new = theta + solve(H) %*% D

    # cat("iteration =", t, "|| theta_new =", theta_new, '\n')
    # cat("loss =", norm(theta-theta_new), '\n')
  }

  rownames(theta_new) = c(paste0("beta",(1:p)-1), "tau")

  return(theta_new)
}
```

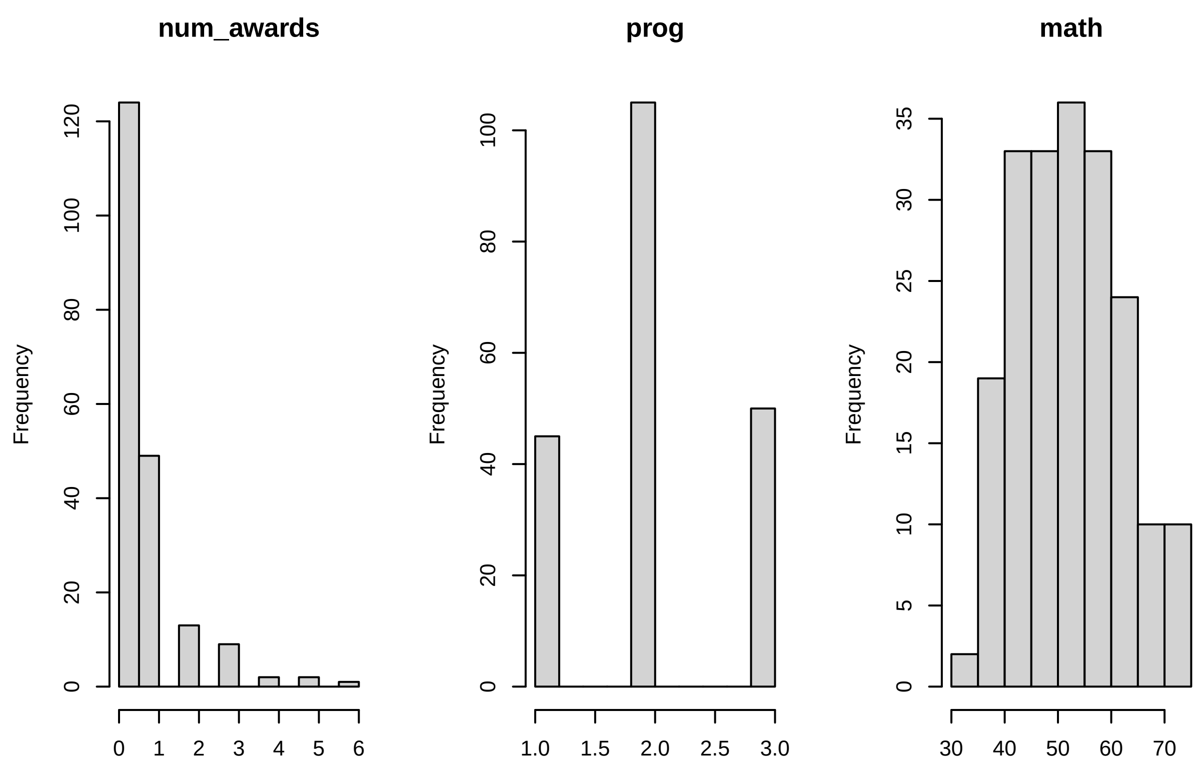
분석에 사용할 데이터는 UCLA에서 다운받은 예시 데이터를 사용하였다. `num_awards`는 계수형의 결과 변수이며 고등학교에서 1년에 받은 상 수를 나타내고, `math`는 연속형 예측 변수로 수학 기말고사에서 학생의 점수를 나타낸다. 마지막으로 `prog`는 범주형 예측 변수로 학생들이 등록한 프로그램의 유형을 나타내는 세 가지 수준 1(General, 일반), 2(Academic, 학업), 3(Vocational, 직업)을 의미한다. 연속형인 `math` 변수는 scaling을 해주었고, 범주형인 `prog` 변수는 General을 baseline으로 하여 나머지 Academic과 Vocational을 더미변수로 바꾸어주었다. 반응변수인 `num_awards`는 평균이 0.63, 분산이 1.109로 과산포가 존재하기에 음이항 회귀무형을 사용하기에 적합하다고 할 수 있다. 더 큰 마지막으로 `NewtonRaphson`함수에 넣기 위해 해당 데이터를 design matrix로 바꾸어주었다.

```r
dat = read.csv("https://stats.idre.ucla.edu/stat/data/poisson_sim.csv")
dat$id = NULL
```

```r
summary(dat)
```

```
##    num_awards        prog            math      
##  Min.   :0.00   Min.   :1.000   Min.   :33.00  
##  1st Qu.:0.00   1st Qu.:2.000   1st Qu.:45.00  
##  Median :0.00   Median :2.000   Median :52.00  
##  Mean   :0.63   Mean   :2.025   Mean   :52.65  
##  3rd Qu.:1.00   3rd Qu.:2.250   3rd Qu.:59.00  
##  Max.   :6.00   Max.   :3.000   Max.   :75.00
```

```r
par(mfrow = c(1, 3))
for(i in 1:ncol(dat)) hist(dat[,i], main = paste0(colnames(dat)[i]), xlab = '')  # 히스토그램
```




```r
cat('num_awards의 평균 =', mean(dat$num_awards), '\n')
```

```
## num_awards의 평균 = 0.63
```

```r
cat('num_awards의 분산 =', var(dat$num_awards), '\n')
```

```
## num_awards의 분산 = 1.108643
```

```r
# 데이터 형식 변환
dat$math = as.numeric(dat$math) # int -> numeric으로 변환
dat$prog = factor(dat$prog, levels=1:3, labels=c("General", "Academic", "Vocational"))

# prog변수를 dummy variable로 변환 (base를 General로 함)
dat$Academic = ifelse(dat$prog == 'Academic', 1, 0)
dat$Vocational = ifelse(dat$prog == 'Vocational', 1, 0)
```

```r
# design matrix
X = dat
X$prog = NULL
X$num_awards = NULL
X$math = scale(X$math)
X = as.matrix(X)
X = cbind(1, X)

y = dat$num_awards
head(X)
```

```
##                math Academic Vocational
## [1,] 1 -1.2430021        0          1
## [2,] 1 -1.2430021        0          0
## [3,] 1 -0.9227783        0          1
## [4,] 1 -1.1362608        0          1
## [5,] 1 -1.3497433        0          1
## [6,] 1 -1.1362608        0          0
```

```r
head(y)
```

```
## [1] 0 0 0 0 0 0
```

뉴튼-랩슨 알고리즘의 결과와 glm.nb의 결과를 비교해보자. beta 추정치들은 비슷하며, 뉴튼랩슨에서의 tau 추정값의 역수는 glm.nb의 추정값과 같으므로 알고리즘이 잘 작동함을 확인하는 것으로 글을 마친다.

```r
library(MASS)
NewtonRaphson(X, y) # 방법 1
```

```
##              [,1]
## beta0 -1.5518980
## beta1  0.6658039
## beta2  1.0750711
## beta3  0.3669593
## tau    0.1635413
```

```r
summary(glm.nb(y ~ X[,-1]))  # 방법 2
```

```
##
## Call:
## glm.nb(formula = y ~ X[, -1], init.theta = 6.114661779, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0322  -0.8343  -0.5039   0.2276   2.3222
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -1.5519     0.3397  -4.568 4.92e-06 ***
## X[, -1]math           0.6658     0.1080   6.165 7.05e-10 ***
## X[, -1]Academic       1.0751     0.3670   2.929   0.0034 **
## X[, -1]Vocational     0.3670     0.4523   0.811   0.4172
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(6.1147) family taken to be 1)
```

```
##
##     Null deviance: 257.93  on 199  degrees of freedom
## Residual deviance: 169.76  on 196  degrees of freedom
## AIC: 373.81
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  6.11
##           Std. Err.:  5.58
##
##  2 x log-likelihood:  -363.811
```