The Korean Journal of Applied Statistics (2017) DOI: https://doi.org/10.5351/KJAS.2017.30.5.775
30(5), 775–791

Binary regression model using skewed generalized *t* distributions

Mijeong Kim[a,1]

[a]Department of Statistics, Ewha Womans University

(Received August 22, 2017; Revised September 3, 2017; Accepted September 5, 2017)

Abstract

We frequently encounter binary data in real life. Logistic, Probit, Cauchit, Complementary log-log models are often used for binary data analysis. In order to analyze binary data, Liu (2004) proposed a Robit model, in which the inverse of cdf of the Student's t distribution is used as a link function. Kim *et al.* (2008) also proposed a generalized t-link model to make the binary regression model more flexible. The more flexible skewed distributions allow more flexible link functions in generalized linear models. In the sense, we propose a binary data regression model using skewed generalized t distributions introduced in Theodossiou (1998). We implement R code of the proposed models using the glm function included in R base and R sgt package. We also analyze Pima Indian data using the proposed model in R.

Keywords: skewed generalized t distribution, binary regression model, logistic model, generalized linear model

1. 서론

이진(binary) 데이터는 두 가지 범주를 갖는 데이터로써, 의학, 사회 과학 등 다양한 분야에서 이진 데이터를 분석하는 경우가 많다. 이진 데이터를 회귀 분석하는 방법으로 로지스틱(Logistic), 프로빗(Probit), Cauchit, Complementary log-log 모형이 주로 쓰인다. 예를 들어, 사망과 생존이라는 두 가지 경우를 갖는 변수에 관심이 있을 경우, 사망한 경우를 1, 생존한 경우를 0으로 놓고, 사망할 확률에 대한 회귀분석을 하는 것이다. 이진 데이터 회귀 분석에서 반응 변수는 특정 사건이 일어날 확률이 되고, 이 값은 0과 1 사이의 값을 가지므로, 공변량(covariates)을 포함한 선형 회귀식을 적절히 변형시켜 0과 1 사이의 값을 갖도록 하는 것이 이진 데이터 회귀 분석의 기본 개념이다. 확률은 0과 1의 값을 취하므로, 특정 분포의 누적확률값을 관심 있는 사건의 확률과 연결하면 이진 데이터의 회귀 분석이 가능해진다. 로지스틱 분포를 이용하면 로지스틱 회귀 분석이 되고, 정규 분포를 이용하면 프로빗 회귀 분석이 된다. 로지스틱 분포와 정규분포는 대칭 분포이므로, 두 분포를 이용하였을 때에는 설명변수가 1단위 증가하고 감소할 때, 특정 사건이 일어날 확률이 0에 근접하는 속도와 1에 근접하는 속도가 같은 결과를 얻게 된다. 현실적으로 그러한 경우는 드물기 때문에, 대칭 분포를 이용하는 것보다 비대칭 분포

This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean Government (NRF-2017R1C1B5015186).

[1]Department of Statistics, Ewha Womans University, 52 Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, Korea. E-mail: m.kim@ewha.ac.kr

를 이용하는 것이 좀 더 현실성 있는 분석이 가능해 질 수 있다. 이러한 점 때문에, 비대칭 확률 분포를 이용한 이진 데이터 분석 또한 제안되었다. Liu (2004)에서는 t 분포를 이용한 방법을 제시하였고, Kim 등 (2008)은 기운 t 분포를 이용한 일반화 t-link 모형을 제시하였다. 이진 데이터 회귀 분석에 분포를 이용하는 방법을 고려하여, 지금까지 이용된 확률 분포 함수보다 더 유연한 분포를 찾고, 그러한 분포를 이용한다면 좀 더 유연한 이진 데이터 회귀 분석 모형을 만들 수 있을 것이다. 기운 일반화 t 분포(skewed generalized t distribution)는 정규 분포, t 분포, 일반화 t 분포, 기운 t 분포, 기운 라플라스 분포 등 여러 다양한 분포의 일반화된 형태를 갖춘 분포로써 아주 유연한 분포로 알려져 있다. 본 연구에서는 기운 일반화 t 분포를 소개하고 이진 데이터 분석에 이용할 수 있는 방법에 대하여 논하는 것을 목표로 한다. 2장에서는 기존에 제안된 대칭 분포, 비대칭 분포를 이용한 이진 데이터 분석 방법을 설명한다. 또한 Complementary log-log 방법과 Stukel (1988)이 제안한 일반화 로지스틱 모형과 같이 누적확률 분포를 이용하지 않은 모형도 함께 설명하도록 한다. 3장에서는 기운 일반화 t 분포에 대한 설명을 하고, 4장에서는 기운 일반화 분포를 R에서 이용할 수 있는 방법을 제안하고 5장에서는 데이터 분석 결과를 보여주도록 한다.

2. 기존 방법

이진 데이터의 회귀식은 반응 변수 $Y = 0, 1$과 관련 있는 공변량 $\boldsymbol{X} = (X_1, \ldots, X_p)$의 선형 회귀식을 연결시킴으로써 모형화할 수 있다. 주어진 $\boldsymbol{x}$에 대하여 $Y = 1$일 확률을 $\pi(\boldsymbol{x}) = P(Y = 1|\boldsymbol{X} = \boldsymbol{x})$라고 하면 연결 함수(link function) $g(\cdot)$를 이용하여 다음과 같은 모형을 만들 수 있다.

$$\begin{aligned} g\{\pi(\boldsymbol{x})\} &= \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}, \\ \pi(\boldsymbol{x}) &= g^{-1}\left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}\right). \end{aligned} \tag{2.1}$$

$\pi(\boldsymbol{x})$은 0과 1 사이의 값을 취하므로, $g^{-1}(\cdot)$에 대해서 특정 분포 함수의 누적확률 분포를 이용할 수 있다. 이것은 잠재 변수(latent variable)로도 설명할 수 있다.

$$y = \begin{cases} 1, & \text{if } \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} + \epsilon > 0, \\ 0, & \text{otherwise.} \end{cases}$$

위의 식으로부터 다음과 같은 관계가 성립한다.

$$P(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = P\left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} + \epsilon > 0\right) = F\left(-\epsilon < \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}\right) = g^{-1}\left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}\right).$$

이 때, F는 $-\epsilon$의 누적 확률 분포이고, ϵ의 분포에 따라 g가 결정된다. ϵ이 대칭 분포를 가질 경우 $-\epsilon$와 ϵ은 같은 분포를 가지므로, ϵ이 로지스틱 분포를 따르면 식 (2.1)은 로지스틱 모형, ϵ이 정규 분포를 따르면 식 (2.1)은 프로빗 모형이 된다. ϵ은 관찰되지 않지만, 반응 변수 y에 영향을 주므로 잠재 변수 또는 잠재 독립 변수라고 한다. ϵ 대신에 $Y_1^* = \boldsymbol{X}\boldsymbol{\beta} + \epsilon$을 잠재 변수라고 보기도 한다. 반면에, 분포가 아닌 함수 g를 이용하여 식 (2.1)을 만족시킬 수 있다. 위의 관계식으로부터 n개의 데이터 $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, n$이 주어졌을 때, 로그 우도 함수를 다음과 같이 구할 수 있다.

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \log\{\pi(\boldsymbol{x}_i)\} + \left(n - \sum_{i=1}^{n} y_i\right) \log\{1 - \pi(\boldsymbol{x}_i)\}.$$

이 로그 우도 함수를 최대로 하는 최대 우도 추정치 $\hat{\boldsymbol{\beta}}$를 찾는 방식으로 $\boldsymbol{\beta}$를 추정한다.

2.1. 대칭 분포를 이용한 방법

2.1.1. 로지스틱 모형 로지스틱 분포의 확률 밀도함수 $f(w)$과 누적 분포 함수 $F(w)$는 다음과 같다.

$$f(w) = \frac{\exp\{-(w-\mu)/s\}}{s\left[1+\exp\{-(w-\mu)/s\}\right]^2},$$
$$F(w) = \frac{\exp\{-(w-\mu)/s\}}{1+\exp\{-(w-\mu)/s\}}.$$

위의 식에서 μ는 위치 모수(location parameter)이고, $s>0$는 척도 모수(scale parameter)이다. 위의 식으로부터 다음과 같은 식을 얻는다.

$$\log\left\{\frac{F(w)}{1-F(w)}\right\} = -\frac{(w-\mu)}{s}.$$

이 식에서 $F(w)$의 오즈(odds)의 로그 변형이 w의 선형식으로 표현된 것을 확인할 수 있다. 비슷한 방법으로 로지스틱 회귀식은 다음과 같이 정의할 수 있다.

$$\text{logit}\left\{P(Y=1|\boldsymbol{X}=\boldsymbol{x})\right\} = \log\left\{\frac{P(Y=1|\boldsymbol{X}=\boldsymbol{x})}{1-P(Y=1|\boldsymbol{X}=\boldsymbol{x})}\right\} = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}.$$

이 때, $P(Y=1|\boldsymbol{X}=\boldsymbol{x})/\{1-P(Y=1|\boldsymbol{X}=\boldsymbol{x})\} = P(Y=1|\boldsymbol{X}=\boldsymbol{x})/P(Y=0|\boldsymbol{X}=\boldsymbol{x})$이고, 이것은 $P(Y=1|\boldsymbol{X}=\boldsymbol{x})$의 오즈이다. $\boldsymbol{x}=(1,x_1,\ldots,x_p)$이고, $\boldsymbol{\beta}=(\beta_0,\beta_1,\ldots,\beta_p)^{\mathrm{T}}$라고 하자. 이 때, 오즈는 다음과 같이 표현할 수 있다.

$$\text{odds} = \exp\left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}\right) = \exp(\beta_0+\beta_1x_1+\cdots+\beta_px_p).$$

위의 식으로부터, x_j가 한 단위 증가할 때, log(odds)는 β_j $(j=1,\ldots,p)$만큼 증가한다고 해석할 수 있다. 오즈를 이용하여 해석할 수 있다는 장점 때문에 로지스틱 모형이 이진 데이터의 분석에 자주 쓰이고 있다.

2.1.2. 프로빗 모형 프로빗 모형은 식 (2.1)에서 g^{-1}가 정규 분포의 누적 확률 분포인 모형이다. 정규 분포의 확률 밀도 함수 $f(w)$과 누적분포 함수 $F(w)$은 다음과 같다.

$$f(w) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(w-\mu)^2}{2\sigma^2}\right\},$$
$$F(w) = \Phi\left(\frac{w-\mu}{\sigma}\right) = \frac{1}{2}\left\{1+\text{erf}\left(\frac{w-\mu}{\sigma\sqrt{2}}\right)\right\}. \tag{2.2}$$

이 때 $\Phi(\cdot)$는 표준 정규 분포의 누적확률 분포이고, erf$(\cdot)$은 오차 함수(error function)이다. 식 (2.2)에서 좌변은 확률 값, $\Phi(\cdot)$의 변수(argument)가 w에 대한 선형식이다. 위의 모형에 기초한 프로빗 모형은 다음과 같이 정의된다.

$$\text{Probit}\{P(Y=1|\boldsymbol{X}=\boldsymbol{x})\} = \Phi^{-1}\{P(Y=1|\boldsymbol{X}=\boldsymbol{x})\} = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}.$$

2.1.3. Cauchit 모형 코쉬 분포(Cauchy distribution)의 확률 밀도 함수 $f(w)$와 누적분포 함수 $F(w)$는 다음과 같다.

$$f(w) = \frac{1}{\pi\gamma\left\{1+\left(\frac{w-w_0}{\gamma}\right)^2\right\}},$$
$$F(w) = \frac{1}{\pi}\arctan\left(\frac{w-w_0}{\gamma}\right)+\frac{1}{2},$$

이 때 w_0는 위치 모수, γ는 척도 모수이다. 위의 식으로부터 Cauchit 모형은 다음과 같은 형태가 된다.

$$\text{Cauchit}\{P(Y=1|\boldsymbol{X}=\boldsymbol{x})\} = \tan\left[\pi\left\{P(Y=1|\boldsymbol{X}=\boldsymbol{x}) - \frac{1}{2}\right\}\right] = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}.$$

2.1.4. 로빗(Robit) 모형 Pregibon (1982)는 이상점(outlier)이 있는 데이터의 경우에는 로지스틱과 프로빗 회귀 모형의 최대 우도 추정치가 robust 하지 않은 점을 지적하였다. 이러한 점을 개선하기 위해 Liu (2004)는 로빗 회귀 모형을 제안하였다. 로빗 회귀 모형은 식 (2.1)에서 $g^{-1}(\cdot)$가 t 분포의 누적 분포 함수인 경우를 지칭한다.

$$f_\nu(w) = \frac{\Gamma\{(\nu+1)/2\}}{\sqrt{\pi\nu}\Gamma(\nu/2)}\left(1+\frac{w^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$
$$F_\nu(w) = \frac{1}{2} + x\Gamma\left(\frac{\nu+1}{2}\right)\frac{{}_2F_1\left\{1/2,(\nu+1)/2,3/2,-w^2/\mu\right\}}{\sqrt{\pi\nu}\Gamma(\nu/2)}.$$

이 식에서 ν는 자유도이고, ${}_2F_1$은 초기하 함수(hypergeometric function)이다. 로빗 모형은 다음과 같이 정의된다.

$$P(Y=1|\boldsymbol{X}=\boldsymbol{x}) = F_\nu\left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}\right).$$

Liu (2004)는 자유도 ν을 알고 있을 때와 그렇지 않은 경우에 대해서 EM 알고리즘 방법을 이용하여 $\boldsymbol{\beta}$를 추정하는 방법을 제시했다. Cauchit 모형은 $\nu = 1$인 로빗 모형의 특별한 경우이고, 또한 자유도 ν가 클수록 프로빗 회귀 모형에 근사한다. 로지스틱 모형은 로빗 모형에 속하지는 않지만, Liu (2004)에서 자유도가 7인 t 분포를 이용하였을 때, 로빗 모형이 로지스틱 회귀 모형에 근사함을 보였다.

2.1.5. 대칭 분포를 이용한 모형의 비교 Figure 2.1은 정규 분포, 로지스틱 분포, 코쉬 분포, 자유도 3인 t 분포의 확률 밀도 함수와 누적확률 분포를 보여주고 있다. 이 확률 분포들을 중앙값의 확률 밀도가 큰 순서로 나열하면, 로지스틱 분포, 정규 분포, 자유도 3인 t 분포, 코쉬 분포 순이다. 절대적 기준으로 어느 방법이 더 낫다고 판단할 수는 없지만, 데이터를 더 적합할 수 있는 모형이 가장 좋은 모형이 될 것이다. 정규 분포를 이용한 프로빗 모형과 로지스틱 분포를 이용한 로지스틱 모형은 확률 밀도 함수가 거의 비슷한 값을 알 수 있다. 즉, 두 모형의 예측치가 비슷한 값으로 계산된다. 이 경우에는, 오즈로 해석이 용이한 로지스틱 방법이 해석 측면에서 유용하게 쓰일 수 있다. 실제로 프로빗 모형보다는 로지스틱이 모형이 더 자주 사용되고 있다.

식 (2.1)에서 $g^{-1}(\cdot)$을 대칭 분포의 누적 확률 분포를 이용할 경우, 다음과 같은 관계가 성립한다.

$$g\{\pi(\boldsymbol{x})\} = -g\{1-\pi(\boldsymbol{x})\}. \tag{2.3}$$

여기서 $\pi(\boldsymbol{x}) = 0.50$을 중심으로 $\pi(\boldsymbol{x})$의 반응 곡선이 대칭이 된다. 즉, $\pi(\boldsymbol{x}) = P(Y=1|\boldsymbol{X})$가 0과 1로 접근하는 속도가 같게 된다. Figure 2.1은 제시된 분포는 모두 대칭 분포로써, 누적 확률 분포의 확률값이 0으로 근접하는 기울기와 1로 근접하는 기울기의 절대값이 $\epsilon = 0$을 기준으로 대칭임을 확인할 수 있다. 하지만, 실제로는 $\boldsymbol{x}$가 증가 또는 감소함에 따라 특정 사건이 일어날 확률이 0으로 근접하는 속도와 1로 근접하는 속도가 다른 경우가 일반적이다. 이러한 면에서, 대칭 분포를 이용한 모든 이진 데이터 회귀 분석은 데이터를 적합하는 데 한계를 갖고 있다.

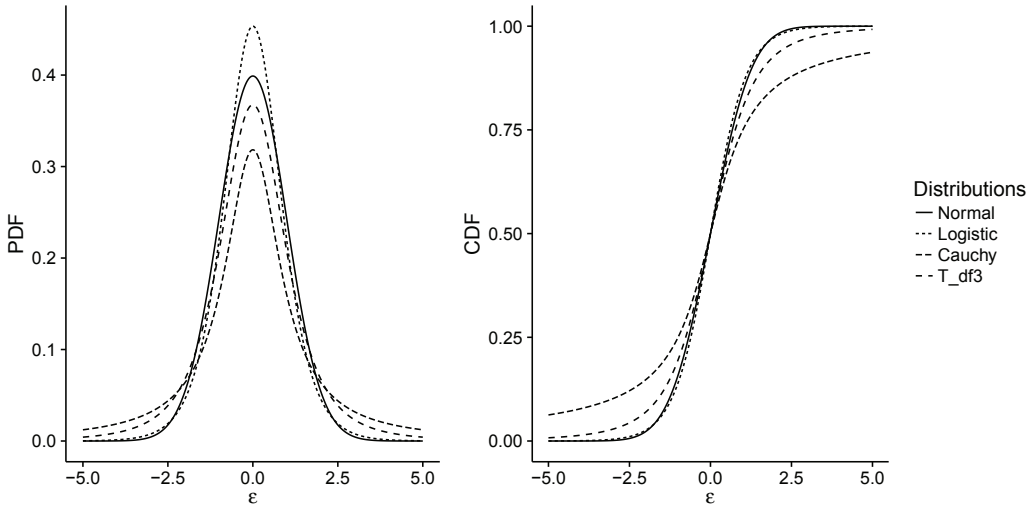

Figure 2.1. Probability distribution functions (PDF) and cumulative distribution functions (CDF) of Normal, Logistic, Cauchy, Student's t with degree of freedom 3 (T_df3).

2.2. 비대칭 분포를 이용한 모형

실제로, $\pi(\boldsymbol{x}) = P(Y = 1|\boldsymbol{X} = \boldsymbol{x})$가 0으로 접근하는 속도와 1로 접근하는 속도가 다를 가능성이 크다. 이 경우에는 $g^{-1}(\cdot)$에 대해 비대칭 분포를 이용하는 것이 적합하다.

2.2.1. 기운 분포를 이용한 모형 Chen 등 (1999)는 기운 정규 분포를 이용하여 다음과 같은 혼합 모형을 제안하였다. $\boldsymbol{w} = (w_1, w_2, \ldots, w_n)^{\mathrm{T}}$를 잠재 독립 변수라고 하자.

$$y_i = \begin{cases} 0, & \text{if } \ w_i < 0, \\ 1, & \text{if } \ w_i \geq 0, \end{cases}$$
$$w_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \delta z_i + \epsilon_i, \quad i = 1, \ldots, n, \tag{2.4}$$

여기서 z_i는 확률 변수 Z의 i번째 관측치고, Z는 누적 확률 분포 G, 밀도 함수 g를 갖는다. 이 때, G와 g는 각각 기운 분포의 누적 확률 분포와 확률 밀도 함수이고, F는 대칭 분포의 누적 확률 분포이다. Z와 ϵ이 독립이라고 가정한다면, 위의 분포로부터 다음과 같은 모형을 얻는다.

$$P(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = \int_{-\infty}^{\infty} F\left(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \delta z_i\right) g(z_i) dz_i,$$

식 (2.4)는 절편을 포함하고 있지 않음을 유의하자. Kim 등 (2008)에서는 식 (2.4)에 절편을 포함할 경우, 절편과 δ이 교락되어 δ와 절편이 식별이 불가능한(unidentifiable) 문제가 발생함을 지적하였다.

2.2.2. 일반화 t 연결 함수 모형(generalized t-link model) Kim 등 (2008)은 Arellano-Valle 등 (1995)와 Azzalini와 Valle (2006)에서 제시된 기운 t 분포를 이용하여 다음과 같은 혼합 모형에 이

용 가능한 분석 방법을 제안했다.

$$y_i = \begin{cases} 0, & \text{if } w_i < 0, \\ 1, & \text{if } w_i \geq 0, \end{cases}$$
$$w_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \delta\{z_i - E(Z)\} + \epsilon_i, \quad i = 1, \ldots, n. \tag{2.5}$$

z_i는 확률 변수 Z의 i번째 관측치이다. Kim 등 (2008)은 연결함수 g가 잘못 선택되면 편향 추정(biased estimation)될 가능성이 있기 때문에 유연한 분포를 통해 연결함수 g를 더 잘 추정할 수 있음을 언급하였다. Kim 등 (2008)에서 ϵ에 이용된 분포는 다음과 같다.

$$p_{gt,\nu_1,\nu_2}(w) = \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\frac{\nu_1+1}{2}\right)}{\sqrt{\nu_2}\Gamma\left(\frac{\nu_1}{2}\right)} \left(1 + \frac{w^2}{\nu_2}\right)^{-\frac{1}{2}(\nu_1+1)}.$$

이 때, ν_1은 형태 모수, ν_2는 척도 모수이다. W가 p_{gt,ν_1,ν_2}를 따르는 경우, $\nu_1 > 1$일 때 $E(W) = 0$이고, $\nu_1 > 2$일 때, $\mathrm{var}(W) = \nu_2/(\nu_1 - 2)$이다. 특별히, $\nu_1 = \nu_2 = \nu$일 때, 위의 분포는 자유도 ν인 Student t 분포가 된다. ν_1이 작을 수록, 두꺼운 꼬리를 가진 분포가 되는 점은 Student t 분포와 같은 특성을 갖는다. Kim 등 (2008)의 모형은 식 (2.5)이 절편을 포함하더라도 δ가 식별이 가능함(identifiable)을 보이고, 베이지안 방법을 이용하여 모수를 추정하였다.

2.2.3. 비대칭 분포를 이용한 모형의 비교 Chen 등 (1999)과 Kim 등 (2008)은 혼합 모형을 포함한 이진 데이터 분석 방법을 제안하였다. 두 모형은 기운 분포를 이용하여, 특정 현상이 일어날 확률이 1에 접근하는 속도와 0에 접근하는 속도를 다르게 할 수 있으므로 대칭 분포를 이용했을 때보다 이진 데이터를 더 잘 적합할 수 있다. Kim 등 (2008)의 모형은 연결함수의 역수가 기운 t 분포라는 점에서 Chen 등 (1999)이 이용한 기운 정규보다 유연한 연결함수를 이용하였다. Chen 등 (1999)의 모형은 혼합 모형 부분에서 절편과 랜덤 요인이 둘 다 포함되었을 때, 랜덤 요인의 계수와 절편이 교락될 수 있으나, Kim 등 (2008)의 모형은 절편과 랜덤요인이 식별 가능한 방법을 도출하였다.

2.3. 확률 분포를 이용하지 않은 모형

2.3.1. log-log 모형과 Complementary log-log 모형 2.1장에서 언급했듯이 대칭 분포를 이용한 이진 데이터 회귀 모형은 $\pi(\boldsymbol{x})$가 0으로 근접하는 속도와 1로 근접하는 속도가 같은 특성을 갖는다. 그러나, 데이터에 따라서는 이러한 특성을 따르지 않는 경우가 발생할 수 있다. Yates (1955)는 다음과 같이 두가지 변환 식 (2.6)과 식 (2.7)을 이용하여 $\pi(\boldsymbol{x})$가 0으로 근접하는 속도와 1로 근접하는 속도가 다른 비대칭성을 가진 모형을 제안하였다.

- Log-log 모형

$$\pi(\boldsymbol{x}) = \exp\left\{-\exp\left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}\right)\right\}, \tag{2.6}$$
$$\log\left[-\log\{\pi(\boldsymbol{x})\}\right] = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}.$$

- Complementary log-log 모형

$$\pi(\boldsymbol{x}) = 1 - \exp\left\{-\exp\left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}\right)\right\}, \tag{2.7}$$
$$\log\left[-\log\{1 - \pi(\boldsymbol{x})\}\right] = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}.$$

Log-log 모형은 $\pi(\boldsymbol{x})$가 0으로 더 빠르게 접근하고, complementary log-log 모형은 1로 더 빠르게 접근한다. O'hagan와 Leonard (1976)에 의해 기운 정규 분포가 처음 제안되었으므로 Yates (1955)의 complementary log-log 모형은 그 당시에 비대칭성을 가진 이진 데이터 분포로써 활용가치가 있었을 것으로 생각된다.

2.3.2. 일반화 로지스틱 모형(generalized logistic models) Stukel (1988)은 로지스틱 모형을 변형한 다음과 같은 일반화 로지스틱 모형을 제안했다.

$$\mu_\alpha(\eta) = \frac{\exp\{h_\alpha(\eta)\}}{1+\exp\{h_\alpha(\eta)\}},$$

이 때, $h_\alpha(\eta)$ 단조 증가하는 η의 비선형 함수이다. $\alpha = (\alpha_1, \alpha_2)^{\mathrm{T}}$는 이 함수의 형태를 결정하는 모수이다. $\eta \geq 0$ ($\mu \geq 0.5$)일 때,

$$h_\alpha = \alpha_1^{-1}\left\{\exp\left(\alpha_1|\eta|\right)-1\right\} I(\alpha_1 > 0) + \eta I(\alpha_1 = 0) - \alpha_1^{-1}\log\left(1-\alpha_1|\eta|\right) I(\alpha_1 < 0).$$

$\eta \leq 0$ ($\mu \leq 0.5$)일 때,

$$h_\alpha = -\alpha_2^{-1}\left\{\exp\left(\alpha_2|\eta|\right)-1\right\} I(\alpha_2 > 0) + \eta I(\alpha_2 = 0) - \alpha_2^{-1}\log\left(1-\alpha_2|\eta|\right) I(\alpha_2 < 0).$$

Chen 등 (1999)는 Stukel (1988)이 제시한 일반화 로지스틱 모형에 대해서 베이지안 추정 시, 부적절 사전 분포를 이용할 경우 부적절 사후 분포를 얻게 된다는 점을 지적했다. 베이지안 추정에서는 $\pi(\theta) = 1$과 같이 분포가 아닌 상수 함수를 사전 분포로 이용하더라도, 사후 분포가 확률 분포가 되는 경우는 사전 분포로 이용할 수 있다. 그러나 일반화 로지스틱 모형의 경우에는, 상수 함수를 사전 분포로 이용할 경우, 부적절 사후 분포를 얻게 되므로 추정에 제약이 생긴다. 예를 들면, 최고사후밀도구간(highest posterior density interval)를 구할 수 없거나 베이지안 가설 검정을 할 수 없다.

2.3.3. 확률 분포를 이용하지 않은 모형에 대한 고찰 Complementary log-log 모형은 Yates (1955)에서 제안된 모형으로써, 그 당시까지는 기운 분포가 제안되기 전이었기 때문에, 기운 분포를 이용하지 않고 비대칭성을 가진 모형을 제안했다는 점에서 의미가 있다. 그러나 Chen 등 (1999), Kim 등 (2008)에 의해 기운 분포를 이용한 분석 방법이 제안되었기 때문에, 기운 분포를 이용한 모형을 이용하여 log-log 모형과 complementary log-log 모형의 장점인 비대칭성을 가지면서 더 유연한 분석이 가능해졌다. 로그 우도 함수를 이용한 추정 방법을 이용하기 어렵거나, 또는 추정치의 점근적인(asymptotic) 특성을 파악하기 어려울 때, 베이지안 방법을 이용하는 경우가 있다. 분포를 이용하지 않은 일반화 로지스틱 모형의 경우에는 부적절 사전 분포 이용시 부적절 사후 분포를 도출한다는 점에서 베이지안 방법에 제약이 생기는 단점이 있다.

3. 기운 일반화 t 분포(skewed generalized t distribution)의 이용

이 장에서는 이진 데이터의 기본 모형 식 (2.1)에 연결함수로 기운 일반화 t 분포를 이용하여 이진 데이터를 분석할 수 있는 방법을 제안한다. Kim 등 (2008)은 연결함수에 기운 t 분포를 이용하였는데, 이 방법이 지금까지 이진 데이터 분석에 적용된 분포 중 가장 유연한 분포이다. 2장에서 설명한 잠재 변수의 분포가 어떤 분포를 갖는지 미리 가정하기는 어려우므로, 지금까지 이진 데이터 분석에 도입한 모든 분포를 다 포함하면서도 더 유연한 분포를 이용할 수 있다면 이진 데이터를 더 잘 적합시킬 수 있을 것이다. Theodossiou (1998)가 제시한 기운 일반화 t 분포(skewed generalized t distribution)는 Kim 등

(2008)가 이용한 기운 t 분포를 포함하는 모형으로써, McDonald와 Newey (1988)가 제시한 일반화 t 분포(generalized t (GT) distribution)를 확장하여 만들어진 분포이다. 이 분포는 환율, 주식 시장 수익률, 귀금속 가격과 같은 금융 데이터를 분석하기에 적합하도록 왜도(skewness)와 첨도(kurtosis)를 자유롭게 바꿀 수 있도록 만들어졌다. 기운 일반화 t 분포는 다음과 같다.

$$f_{\mathrm{SGT}}(x;\mu,\sigma,\lambda,p,q)=\frac{p}{2v\sigma q^{\frac{1}{p}}B\left(\frac{1}{p},q\right)\left[\frac{|x-\mu+m|^p}{q(v\sigma)^p\{\lambda\mathrm{sign}(x-\mu+m)+1\}^p}+1\right]^{\frac{1}{p}+q}}. \tag{3.1}$$

이 때, $B(a,b)=\Gamma(a)\Gamma(b)/\Gamma(a+b)$이고, 모수는 다음과 같은 조건을 만족해야 한다.

$$\{(\sigma,\lambda,p,q):\sigma>0,-1<\lambda<1,p>0,q>0\}$$

위의 식에서 m은 다음과 같은 값을 갖는다.

$$m=\frac{2v\sigma\lambda q^{\frac{1}{p}}B\left(\frac{2}{p},q-\frac{1}{p}\right)}{B\left(\frac{1}{p},q\right)}.$$

기운 일반화 t 분포는 정규 분포, t분포, 기운 t 분포, 일반화 t 분포, 극단값분포(extreme value distribution) 등을 포함한다. 따라서, 기운 일반화 t 분포를 통해 프로빗 모형, 로빗 모형, Chen 등 (1999)과 Kim 등 (2008)가 제시한 모형을 만드는 것이 가능해진다. Davis (2015)는 모수 μ, σ, λ, p, q를 바꿈으로써 생성되는 분포를 정리하였다. μ는 중심, σ는 왜도, p와 q는 첨도를 결정하는 모수이다. 또한 Hansen 등 (2010)은 Figure 4.1을 통하여 각 분포들간의 관계를 정리하였다. 기운 일반화 t 분포는 SGT로 표시하였고, 다른 분포는 아래에 분포의 영문 명과 함께 약자를 괄호 안에 표기하였다.

1. 기운 일반화 오차 분포(skewed generalized error distribution; SGED) : $q=\infty$
2. 일반화 t 분포(generalized t distribution; GT): $\lambda=0$
3. 기운 t 분포(skewed t distribution; ST): $p=2$
4. 기운 라플라스 분포(skewed Laplace distribution; SLaplace): $p=1$, $q=\infty$
5. 일반화 오차 분포(generalized error distribution; GED): $\lambda=0$, $q=\infty$
6. 기운 정규 분포(skewed normal distribution; SNormal): $p=2$, $q=\infty$
7. t 분포(student t Distribution; T): $\lambda=0$, $p=2$
8. 기운 코시 분포(skewed Cauchy distribution; SCauchy): $p=2$, $q=1/2$
9. 라플라스 분포(Laplace distribution; Laplace): $\lambda=0$, $p=1$, $q=\infty$
10. 균등 분포(uniform distribution; Uniform): $p=\infty$
11. 정규 분포(normal distribution; Normal): $\lambda=0$, $p=2$, $q=\infty$
12. 코시 분포(Cauchy distribution; Cauchy): $\lambda=0$, $p=2$, $q=1/2$

위에 나열한 것처럼 기운 일반화 t 분포를 이용하면 다양한 분포를 만들 수 있으므로, 데이터를 더 잘 적합할 수 있다. 하지만, 기운 일반화 t 분포를 이용하기 위해 모수 다섯 개를 모두 추정할 경우에는 프로빗 모형과 같이 표준 정규 분포를 이용할 때보다 자유도가 떨어지게 되고, 로지스틱 모형처럼 오즈를 직접 이용할 수 없기 때문에 해석 측면에서는 로지스틱 모형보다 활용도가 떨어질 수 있다. 원래 데이터에

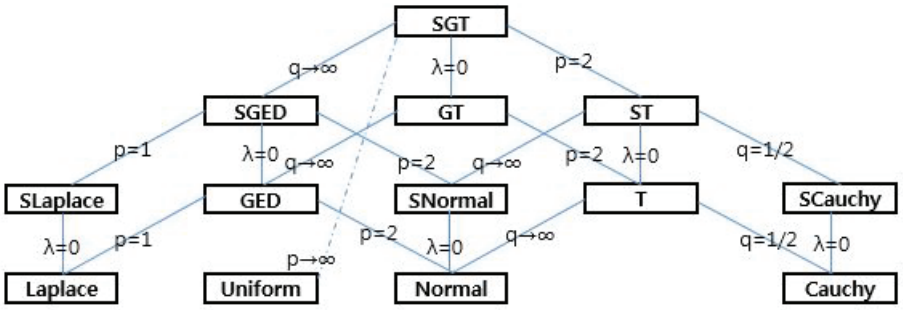

Figure 4.1. The skewed generalized t distribution tree.

서 종속 변수는 0, 1 두 가지 값을 갖지만, 분석을 위해 데이터를 다음과 같이 정리할 수 있다. 독립 변수 $\boldsymbol{x}$는 같은 값이 여러 개 일 수 있다. 원래 데이터의 개수가 아닌, 중복된 $\boldsymbol{x}$를 하나의 경우로 봤을 때, $\boldsymbol{x}$의 경우의 수를 N으로 한다. $\boldsymbol{x}_i$ $(i = 1, \ldots, N)$가 주어졌을 때 $\boldsymbol{x}_i$를 갖는 데이터의 수를 n_i라고 할 때, y_i는 n_i개의 데이터 중에서 1을 갖는 데이터의 수라고 하자. $x_{i0} = 1$로 두고 j는 j번째 독립변수를 의미하고, β_j는 j번째 독립변수의 계수를 뜻한다. 본 연구에서는 고정 요인만 고려한 간단한 모형만 고려하도록 한다. 기운 일반화 t 분포의 누적 확률 밀도 함수 $F_{\mathrm{SGT}}(\mu, \sigma, \lambda, p, q)(\cdot)$를 이용하여 다음과 같은 회귀 모형을 만들 수 있다.

$$P(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = F_{\mathrm{SGT}}(\mu, \sigma, \lambda, p, q)\left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}\right).$$

로그 우도 함수는 다음과 같다.

$$L(\boldsymbol{\beta}) = \log\left[\prod_{i=1}^{N}\left\{F_{\mathrm{SGT}}(\mu, \sigma, \lambda, p, q)\left(\sum_j \beta_j x_{ij}\right)\right\}^{y_i}\left\{1 - F_{\mathrm{SGT}}(\mu, \sigma, \lambda, p, q)\left(\sum_j \beta_j x_{ij}\right)\right\}^{n_i - y_i}\right] \quad (3.2)$$

로그 우도 함수를 최대로 하는 $\boldsymbol{\beta}$와 함께 모수 μ, σ, λ, p, q를 추정함으로써 회귀식을 추정할 수 있다. $p_i = P(Y_i = 1|\boldsymbol{X} = \boldsymbol{x}_i)$의 추정치를 $\hat{p_i}$라고 할 때, 편차(deviance)는 포화 모형(saturated model)과 추정치를 이용했을 때의 우도의 차이로써 이진 데이터의 경우는 다음과 같이 계산된다.

$$2\sum_i\left[y_i \log\left(\frac{y_i}{\hat{p_i}}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n - \hat{p_i}}\right)\right].$$

4. R glm을 이용한 방법

3장에서 최대 우도 방법을 이용하여 이진 데이터의 회귀 분석을 추정할 수 있음을 언급하였다. 식 (3.2)를 $\boldsymbol{\beta}$로 미분하면 점수 함수(score function)을 얻을 수 있고, 점수 함수의 근을 찾으면 최대 우도값을 얻을 수 있다. 점수 함수는 명시적 형태(explicit form)으로 나오지 않기 때문에, 수치 해석(numerical)적으로 근을 구해야 한다. McCullagh와 Nelder (1989), Wood (2006)에 수치적 방법 중 한 방법인 iterative reweighted least squares (IRWLS) 방법이 잘 설명되어 있다. R glm 함수에서도 디폴트로 IRWLS 방법이 이용되고 있다. Koenker (2006)에 R glm에서 제공하지 않은 연결 함수(link function)을 glm 함수를 이용하여 분석할 수 있는 방법을 제공하였다. 이 방법을 이용하면 R에서 기운 일반화 t 분포의 누적 분포의 역함수를 연결 함수로 이용하여 glm 함수를 이용하는 것이 가능해진다. R 패키지 sgt (skewed generalized t distribution tree)를 이용하면 기운 일반화 t 분포의 확률 밀도 함수,

누적 확률 함수, 분위수함수(quantile function) 등을 이용할 수 있다. sgt 패키지에 대한 설명은 Davis (2015)에서 확인할 수 있다. R에서 이용할 수 있는 sgt 연결 함수를 다음과 같이 코딩하였다.

```
library(sgt)
sgt<-function(m,sig,lambda,p,q,mean.cent, var.adj){
  linkfun <- function(mu)
    qsgt(mu,m,sig=sig,lambda=lambda,p=p,q=q,mean.cent, var.adj)
  linkinv <- function(eta){
    thresh <- -qsgt(.Machine$double.eps,m,sig=sig,lambda=lambda,p=p,q=q, mean.cent,
                    var.adj)
    eps <- .Machine$double.eps^.2
    eta <- pmin(thresh, pmax(eta,-thresh))
    psgt(eta,m,sig=sig,lambda=lambda,p=p,q=q,mean.cent, var.adj)
  }
  mu.eta <- function(eta)
    pmax(dsgt(eta,m,sig=sig,lambda=lambda,p=p,q=q,mean.cent, var.adj),.Machine$
         double.eps^.5)
  valideta <- function(eta) TRUE
  name <- "sgt"
  structure(list(linkfun=linkfun, linkinv=linkinv,
                 mu.eta=mu.eta, valideta=valideta, name=name),
            class = "link-glm")
}
```

R에서 glm 함수에 sgt link를 이용하여 다음과 같이 입력하면 IRWLS 방법으로 이진 데이터의 분석이 가능해진다.

```
m=0; sig=1; lambda=0; p=2; q=Inf; mean.cent=1; var.adj=sqrt(2);
glm(y~.,family=binomial(link=sgt(m,sig,lambda,p,q,mean.cent, var.adj))
```

위에서 m,sig,lambda,p,q는 각각 기운 일반화 t 분포 식 (3.1)의 모수 $\mu, \sigma, \lambda, p, q$이므로, 이 모수를 바꿈으로써 다양한 기운 일반화 t 분포의 활용이 가능하다. 참고로 위에서 이용한 모수는 정규분포에 해당되는 값이다. mean.cent와 var.adj는 R sgt 패키지에 포함된 기운 일반화 t 분포 함수와 관련된 옵션으로 mean.cent가 가질 수 있는 값은 1(true)과 0(false)이다. mean.cent가 1이면 μ는 분포의 평균이 되고, 0이면 μ는 분포의 최빈값으로 설정한다. var.adj가 가질 수 있는 값은 1, 0과 양수로서 분포의 분산값을 조정하는 값이다. 여기서 유의할 것은 σ^2이 항상 분포의 분산이 되는 것은 아니라는 점이다. 분산 계산은 σ 외에 다른 모수와도 관련이 있기 때문에 자세한 내용은 Davis (2015)를 참조하도록 하자.

R glm의 link는 다음과 같은 요소를 포함하고 있다.

1. 분위수 함수 (4번째 줄): 기운 일반화 t 분포의 분위수 함수 qsgt를 입력한다.

   ```
   qsgt(mu,m,sig=sig,lambda=lambda,p=p,q=q,mean.cent, var.adj)
   ```

   위의 코드에서 mu는 $\pi(\boldsymbol{x}) = P(Y = 1|\boldsymbol{X} = \boldsymbol{x})$가 입력되는 변수이고, 다른 변수(argument)는 sgt 연결 함수 코드 다음 부분에 설명되어 있다.

Table 5.1. The description of the variables of Pima Indian data

Variables	Description
pregnant	number of times pregnant
glucose	plasma glucose concentration at 2 hours in an oral glucose tolerance test
diastolic	diastolic blood pressure (mm Hg)
triceps	triceps skin fold thickness (mm)
insulin	2-Hour serum insulin (mu U/ml)
bmi	body mass index (weight in kg/(height in metres squared))
diabetes	diabetes pedigree function
age	age (years)
test	test whether the patient shows signs of diabetes (coded 0 if negative, 1 if positive)

2. 누적 확률 분포 함수 (10번째 줄): 기운 일반화 t 분포의 누적 확률 분포 함수 psgt를 입력한다.

```
psgt(eta,m,sig=sig,lambda=lambda,p=p,q=q,mean.cent, var.adj)
```

위의 코드에서 eta는 선형 식 $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}$에 해당하는 값이 입력되는 변수이다.

3. 확률 밀도 함수 (13번째 줄): 기운 일반화 t 분포의 확률 밀도 함수 dsgt를 입력한다.

```
pmax(dsgt(eta,m,sig=sig,lambda=lambda,p=p,q=q,mean.cent, var.adj),.Machine$double.
     eps^.5)
```

위의 코드에서 eta는 선형 식 $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}$에 해당하는 값이 입력되는 변수이다. .Machine$double.eps는 R에서 다음을 만족하는 최소의 double로써, $2.220446e^{-16}$과 같은 값이다.

```
1+x!=1
```

위의 코드는 확률 밀도로 이용되는 값이 .Machine$double.eps의 5제곱보다 작은 값이 되지 않도록 설정해 주는 부분이다.

이 sgt 링크를 이용하여 특정 분포를 이진 데이터 분석의 연결함수로 이용하는 몇 가지 경우를 부록에서 제시하도록 한다.

5. 데이터 분석

이 논문에서 이용한 데이터는 피마 인디언의 당뇨와 관련된 설문 조사 자료를 이용하여 회귀 분석 결과를 비교하고자 한다. 피마 인디언(Pima Indian)은 1950년대까지 비만인 사람들이 없었으나 20세기 중반 이후 인스턴트 섭취가 늘어나 최근에는 그들 중 80%가 비만, 60%가 당뇨 판정을 받았다고 한다. 다른 종족과 달리 피마 인디언의 급격한 건강 악화가 학자들의 관심사가 되기도 하였다. 이 데이터는 UCI Machine Learning Repository에서 제공하고 있고, R 패키지 faraway에서도 pima라는 이름으로 구할 수 있다. 이 데이터에 포함된 변수는 다음과 같다. pregnant(임신 횟수), glucose(경구 포도당 내성 검사에서 2 시간 째의 혈장 포도당 농도), diastolic(확장기 혈압, 단위는 mm Hg), triceps(삼두근, 단위는 mm), insulin(2 시간 혈청 인슐린, 단위는 mu U/ml), bmi(체질량 지수)는 Body mass index로써 단위는 weight in kg/(height in metres squared)이다. diabetes(당뇨병의 유전적 요소를 측정한 값), age(나이)와 test(당뇨병이면 1, 아니면 0)를 변수로 포함하고 있다. Table 5.1에 이 데이터에 포함된 변수를 정리하였다. 설명은 R 패키지 faraway에 있는 pima 데이터의 설명을 가져왔다.

기운 일반화 t 분포로 정규분포를 만들 수 있으므로, sgt 연결함수를 이용한 프로빗 모형의 구현이 다음과 같이 가능해진다.

```
> library(sgt)
> glm(test ~. ,data=pima, family=binomial(link=sgt(0,1,0,2,Inf,1,sqrt(2))))

Call:  glm(formula = test ~ ., family = binomial(link = sgt(0, 1, 0,
    2, Inf, 1, sqrt(2))), data = pima)

Coefficients:
(Intercept)     pregnant      glucose    diastolic      triceps      insulin
 -4.8637528    0.0722842    0.0198836   -0.0079255    0.0012370   -0.0007415
        bmi     diabetes          age
  0.0523174    0.4982427    0.0101975

Degrees of Freedom: 767 Total (i.e. Null);  759 Residual
Null Deviance:     993.5
Residual Deviance: 725.6  AIC: 743.6
```

위의 결과는 probit 연결함수를 이용했을 때와 결과가 같음을 다음에서 확인할 수 있다.

```
> glm(test ~ . ,data=pima, family=binomial(link=probit))

Call:  glm(formula = test ~ ., family = binomial(link = probit), data = pima)

Coefficients:
(Intercept)     pregnant      glucose    diastolic      triceps      insulin
 -4.8637528    0.0722842    0.0198836   -0.0079255    0.0012370   -0.0007415
        bmi     diabetes          age
  0.0523174    0.4982427    0.0101975

Degrees of Freedom: 767 Total (i.e. Null);  759 Residual
Null Deviance:     993.5
Residual Deviance: 725.6  AIC: 743.6
```

로그 우도값이 최대가 되는 것은 편차가 최소가 될 때이다. 따라서 직접 로그 우도값을 계산하는 코딩을 하지 않아도, R에서 sgt 연결 함수, glm 함수와 optim function을 이용하여 편차가 최소가 되는 모수의 추정치를 구할 수 있다. 이 데이터 분석에 이용한 목적함수는 부록에 제시하였다.

Table 5.2에 sgt 연결함수를 이용한 데이터 분석 결과를 제시하였다. t 분포, 기운 t 분포, 일반화 t 분포, 기운 일반화 t 분포를 이용한 결과를 비교할 수 있다. t 분포를 이용한 모형은 자유도가 200.00이고 프로빗 모형과 아주 비슷한 결과를 보여 준다. 이 때, 위치 모수는 선형 식의 절편값을 바꾸는 역할을 하므로 위치 모수 μ은 0으로 고정하고 실행하였다. 각각 다른 연결함수를 이용하였으므로, 모형 간에 회귀 계수를 직접 비교하기는 어렵지만, 같은 모형 내에서는 변수들 간의 계수를 비교하는 것이 의미가

Table 5.2. Results of Pima Indian data using sgt link

	t (df = 200.00)			Skewed-t		
	$(\mu=0,\sigma=1,\lambda=0,p=2,q=100.00)$			$(\mu=0,\sigma=1,\lambda=0.38,p=2,q=0.89)$		
	Estimate (ste.err)	p-value		Estimate (ste.err)	p-value	
(Intercept)	−4.883 (0.391)	$< 2e^{-16}$	***	−6.081 (0.670)	$< 2e^{-16}$	***
pregnant	0.073 (0.019)	$9.96e^{-5}$	***	0.093 (0.026)	0.000	***
glucose	0.020 (0.002)	$< 2e^{-16}$	***	0.029 (0.003)	$< 2e^{-16}$	***
diastolic	−0.008 (0.003)	0.009	**	−0.009 (0.004)	0.0385	*
triceps	0.001 (0.004)	0.765		−0.002 (0.005)	0.752	
insulin	−0.001 (0.001)	0.161		−0.001 (0.001)	0.215	
bmi	0.052 (0.009)	$9.88e^{-10}$	***	0.068 (0.013)	$1.02e^{-7}$	***
diabetes	0.502 (0.171)	0.003	**	0.924 (0.245)	0.0002	***
age	0.010 (0.005)	0.064	·	0.012 (0.007)	0.109	
Deviance	725.47			718.21		
AIC	743.47			736.21		
	Generalized t			Skewed generalized t		
	$(\mu=0,\sigma=2,\lambda=0,p=22.38,q=0.07)$			$(\mu=0,\sigma=1,\lambda=0.43,p=17.79,q=0.07)$		
	Estimate (ste.err)	p-value		Estimate (ste.err)	p-value	
(Intercept)	−10.812 (0.959)	$< 2e^{-16}$	***	−5.375 (0.556)	$< 2e^{-16}$	***
pregnant	0.149 (0.040)	0.000	***	0.075 (0.023)	0.001	***
glucose	0.046 (0.005)	$< 2e^{-16}$	***	0.026 (0.003)	$< 2e^{-16}$	***
diastolic	−0.015 (0.007)	0.021	*	−0.007 (0.004)	0.061	·
triceps	−0.001 (0.009)	0.943		−0.001 (0.005)	0.779	
insulin	−0.001 (0.001)	0.202		−0.001 (0.001)	0.147	
bmi	0.111 (0.019)	$9.54e^{-9}$	***	0.061 (0.011)	$3.85e^{-8}$	***
diabetes	1.335 (0.384)	0.000	***	0.814 (0.212)	0.000	***
age	0.019 (0.012)	0.116		0.011 (0.007)	0.082	·
Deviance	719.65			714.51		
AIC	737.65			732.51		

있다. 모형을 비교하기 위해서는 AIC를 비교하는 것이 합리적이다. AIC가 가장 작은 모형은 모수를 $\mu=0$, $\sigma=1$, $\lambda=0.43$, $p=17.79$, $q=0.07$로 갖는 기운 일반화 t 분포이다. 이러한 모수를 갖는 기운 일반화 t 분포의 누적 확률 분포는 Figure 5.1에서 확인할 수 있다. 비교를 위해 정규분포도 함께 제시하였다. sgt 연결함수를 이용하는 것이 다른 모형보다 AIC가 작고 프로빗 모형이나 t 분포를 연결함수의 역수로 쓴 모형보다 해석 측면에서 의미있는 결과를 보여 주고 있다. 피마 인디언이 다른 종족에 비해 빠른 속도로 비만이 된 것은 유전적으로 저장 능력이 뛰어나기 때문인데, 위의 결과에서 diabetes가 유의하고 계수 또한 다른 계수의 추정치에 비해 큰 값이 나왔다. 근육이 적을 수록 비만일 가능성이 높고 또한 당뇨가 될 가능성이 큼으로 triceps의 계수가 음수가 나오는 것이 상식적으로 이해할 수 있는 결과이다. 위의 결과에서는 triceps의 부호가 음수가 나왔지만, 프로빗 모형과 t-link 모형에서는 양수임을 확인할 수 있다.

6. 결론

이진 데이터의 분석으로 주로 쓰이는 모형으로써 로지스틱, 프로빗, Cauchit, Complementary log-log 모형을 들 수 있다. 로지스틱, 프로빗, Cauchit 모형은 확률이 0과 1에 접근하는 속도가 같은 모형이

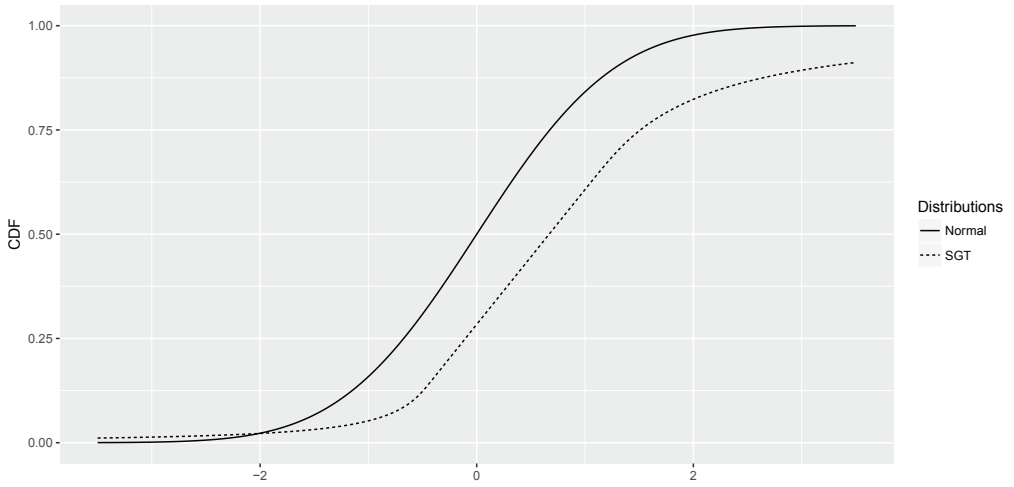

Figure 5.1. Cumulative distribution functions (CDF) of Normal and skewed generalized t with parameters $(\mu, \sigma, \lambda, p, q) = (0, 1, 0.43, 17.79, 0.07)$.

다. 이것은 대칭인 분포에 기반하기 때문이다. Complementary log-log 모형는 확률이 0과 1에 접근하는 속도가 다른 대표적인 모형이다. 비대칭인 분포를 이용하여 Complementary log-log 모형보다 좀 더 유연한 이진 데이터 분석 모형을 만들 수 있다. Theodossiou (1998)에서 제시된 기운 일반화 t 분포는 잘 알려진 정규 분포, t 분포 뿐만 아니라, 극단값 분포까지도 만들 수 있는 아주 유연한 분포이며, 이 분포를 이용하여 좀 더 유연한 이진 데이터 회귀 모형이 가능해진다. 우도값을 최대로 하여 잘 적합(fitting)된 모형을 통해 의미있는 회귀 계수를 도출하고자 할 때에는 기운 일반화 t 분포에 근거한 이진 데이터 회귀 분석이 기존에 제시된 다른 모형보다 유용하게 이용될 수 있을 것으로 기대한다. 로빗 모형에서는 이상치에 영향을 적게 받는 최대 우도 추정치를 구하는 방법으로 EM 알고리즘을 이용하였는데, 후속 연구로 기운 일반화 t 분포를 이용하여 EM 알고리즘 개발을 제안할 수 있고, 일반화 t 분포를 이용한 혼합 모형에 대해서도 베이지안 추정을 적용할 수 있을 것이다.

부록

부록에서는 기운 일반화 t로 만들 수 있는 몇가지 분포를 이용하여 피마 인디언 데이터를 분석하는 코드를 제공하도록 한다.

- t 분포 ($\lambda = 0$, $p = 2$)

```
df=1;
m=0; sig=1; lambda=0; p=2; q=df/2;
glm(test ~ . , data=pima,
    family=binomial(link=sgt(m, sig, lambda, p, df/2,mean.cent = FALSE,var.adj =
           sqrt(2))))
```

첫번째 줄 df에 t 분포의 자유도를 입력하면 t분포의 자유도를 조절하여 이진데이터 분석의 연결 함수로 이용할 수 있다. 위의 경우는 t 분포의 자유도가 1인 경우이므로 다음과 같이 Cauchit 모형을 이

용한 것과 같다. Koenker (2006)에 Gosset 링크를 통해 t분포를 이용한 glm을 제시하였는데, 위의 코드 결과와 다음 Gosset 링크를 통해 얻은 결과가 같다.

```
g.cauchit <- glm(test ~ . ,data=pima, family=binomial(link="cauchit"))
g.t <- glm(test ~ . ,data=pima, family=binomial(link=Gosset(df)))
```

- 기운 정규 (skew-normal) 분포 ($p = 2$, $q = \infty$)

```
m=0; sig=1; lambda=0; p=2; q=Inf;
glm(test ~ . , data=pima,
    family=binomial(link=sgt(m, sig, lambda, p, q,mean.cent = FALSE,var.adj = 1)))
```

- 기운 t (skew-t) 분포 ($p = 2$)

```
m=0; sig=1; lambda=0.5; p=2; q=1;
glm(test ~ . , data=pima,
    family=binomial(link=sgt(m, sig, lambda, p, q,mean.cent = FALSE,var.adj = 1)))
```

- 일반화 t (generalized t) 분포 ($\lambda = 0$)

```
m=0; sig=1; lambda=0; p=1.5; q=1;
glm(test ~ . , data=pima,
    family=binomial(link=sgt(m, sig, lambda, p, q,mean.cent = FALSE,var.adj = 1)))
```

다음은 5장 데이터 분석에 이용되었던 목적함수 obj.function이다. glm 결과에서 deviance를 최소로 하도록 코딩하였다.

```
obj.function <- function(par){
  tryCatch({
    sig=par[1]; lambda=par[2]; p=par[3]; q=par[4];
    g<- glm(test ~ . ,data=pima, family=binomial(link=sgt(0,sig,lambda,p,q,0,1)))
    deviance(g)
  }, error=function(e){})
}
```

R 패키지 sgt에 포함된 함수 dsgt, psgt, qsgt는 모수가 var.adj = 1이고, p와 q의 곱이 2보다 작거나 같으면 경고 메시지가 뜨도록 만들어져 있다. optim 함수를 이용하여 모수를 찾을 때, 이러한 모수의 조합이 생길 수 있으므로, 이 경우에는 경고를 무시하고 optim을 계속 이용하도록 tryCatch 함수를 이용하였다. R에서 optim 함수를 이용하여 위의 목적함수를 최소로 만드는 기운 일반화 t분포의 모수를 다음과 같은 방법으로 찾을 수 있다.

```
options(warn = -1)
library(sgt)
data(pima,package="faraway")
min.dev<-optim(par=c(1,0,3,1), fn=obj.function, lower=c(0, -1, 0, 0),
        upper=c(Inf, 1,Inf,Inf), method="L-BFGS-B")
```

목적함수 obj.function를 최소로 하는 모수는 min.dev$par에 저장하였고, 이 모수를 이용하여 glm을 실행하는 코드는 다음과 같다.

```
m=0; sig=min.dev$par[1]; lambda=min.dev$par[2]; p=min.dev$par[3]; q=min.dev$par[4];
g<- glm(test ~ . ,data=pima, family=binomial(link=sgt(0,sig,lambda,p,q,0, 1)))
summary(g)
```

References

Arellano-Valle, R. B. and Bolfarine, H. (1995). On some characterizations of the t-distribution, *Statistics & Probability Letters*, **25**, 79–85.

Azzalini, A. and Valle, A. D. (1996). The multivariate skew-normal distribution, *Biometrika*, **83**, 715–726.

Chen, M. H., Dey, D. K., and Shao, Q. M. (1999). A new skewed link model for dichotomous quantal response data, *Journal of the American Statistical Association*, **94**, 1172–1186.

Davis, C. (2015). *The Skewed Generalized T Distribution Tree Package Vignette*, Available from: https://cran.r-project.org/web/packages/sgt/vignettes/sgt.pdf

Hansen, C., McDonald, J. B., and Newey, W. K. (2010). Instrumental variables estimation with flexible distributions, *Journal of Business & Economic Statistics*, **28**, 13–25.

Kim, S., Chen, M. H., and Dey, D. K. (2008). Flexible generalized t-link models for binary response data, *Biometrika*, **95**, 93–106.

Koenker, R. (2006). *Parametric links for binary response.* The Newsletter of the R Project Volume 6/4, October 2006, 32.

Liu, C. (2004). Robit regression: a simple robust alternative to logistic and probit regression. In *Applied Bayesian Modeling and Casual Inference from Incomplete-Data Perspectives*, 227–238.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models* no. 37 in Monograph on Statistics and Applied Probability.

McDonald, J. B. and Newey, W. K. (1988). Partially adaptive estimation of regression models via the generalized t distribution, *Econometric Theory*, **4**, 428–457.

O'hagan, A., and Leonard, T. (1976). Bayes estimation subject to uncertainty about parameter constraints, *Biometrika*, **63**, 201–203.

Pregibon, D. (1982). Resistant fits for some commonly used logistic models with medical applications, *Biometrics*, **38**, 485–498.

Stukel, T. A. (1988). Generalized logistic models, *Journal of the American Statistical Association*, **83**, 426–431.

Theodossiou, P. (1998). Financial data and the skewed generalized t distribution, *Management Science*, **44**(12-part-1), 1650–1661.

UCI Machine Learning Repository *http://archive.ics.uci.edu/ml/index.php*

Wood, S. N. (2006) *Generalized Additive Models: An Introduction with R*, CRC Press, Boca Ranton, FL.

Yates, F. (1955). The use of transformations and maximum likelihood in the analysis of quantal experiments involving two treatments, *Biometrika*, **42**, 382–403.

기운 일반화 t 분포를 이용한 이진 데이터 회귀 분석

김미정[a,1]

[a]이화여자대학교 통계학과

(2017년 8월 22일 접수, 2017년 9월 3일 수정, 2017년 9월 5일 채택)

Abstract

이진 데이터는 일상 생활에서 자주 접할 수 있는 데이터이다. 이진 데이터를 회귀 분석하는 방법으로 로지스틱(Logistic), 프로빗(Probit), Cauchit, Complementary log-log 모형이 주로 쓰이는데, 이 방법 이외에도 Liu (2004)가 제시한 t 분포를 이용한 로빗(Robit) 모형, Kim 등 (2008)에서 제시한 일반화 t-link 모형을 이용한 방법 등이 있다. 유연한 분포를 이용하면 유연한 회귀 모형이 가능해지는 점에 착안하여, 이 논문에서는 Theodossiou (1998)에서 제시된 기운 일반화 t 분포 (Skewed Generalized t Distribution)의 이용하여 우도 함수를 최대로 하는 이진 데이터 회귀 모형을 소개한다. 기운 일반화 t 분포를 R glm 함수, R sgt 패키지를 연결하여 이 논문에서 제시한 방법을 R로 분석할 수 있는 방법을 소개하고, 피마 인디언(Pima Indian) 데이터를 분석한다.

주요용어: 기운 일반화 t 분포, 이진 데이터 회귀 분석, 로지스틱 모형, 일반화 선형 모형

이 논문은 2017년도 연구재단 연구 과제 NRF-2017R1C1B5015186에 의하여 수행되었음.

[1](03760) 서울특별시 서대문구 이화여대길 52, 이화여자대학교 통계학과. E-mail: m.kim@ewha.ac.kr