Estimation and Application of Skew-normal Data for Generalized Linear Regression

Wenjun Lyu and Zhaoqing Feng*
School of Economics, Shanghai University, Shanghai, China
*Corresponding author

***Abstract*—Generalized linear models are generally applied in statistical researches. Since a lot of real data reveal nonnormality especially skew-normality, new assumption is proposed that error terms follow skew-normal distribution to increase the adaptability of GLMs, which forms GLMSNs. To estimate the parameters in the linear part in models, penalized expectation maximization algorithm is extended. This paper focuses on the combination of skew-normal data and GLMs to get more robust results. Several applications and empirical analyses are given to fit GLMSNs and models selection is presented by Bayesian information criterion.**

***Keywords*—skew-normal distributions; generalized linear models; EM-algorithm**

I. INTRODUCTION

Generalized linear models including a nonparametric component of some covariate into the linear regressor is one kind of semiparametric regression analysis approaches. Recent works proved asymmetry response in GLMs, whereas Relvas and Paula [5] derived an iterative process and some diagnostic procedures with AR(1) symmetric errors.

Critically, the error terms are assumed to be skew-normal, not usually normality. In many fields such as biostatistics, econometrics, epidemiology and quantitative social research, data from related experiments show numerical characteristics like leptokurtosis and clustering, which are called skewed data. In this case, normal distribution is rare and traditional linear structure model is no longer applicable. The nonparametric explanatory part in regression models increased applicability. The expectation maximization algorithm is also improved and will be evaluated.

Ferreira and Paula develop the diagnostic analysis via local influence as well as generalized leverage for partial linear models of skew-normal errors. An application about pollen-related allergy [1] is conducted. Z.Zhou and Z.Lin (2017) discussed the varying coefficient nonlinear models considering nonstationary regressors, whose work is inspiring.

Through simulation study and real data empirical analysis, the extended GLMs can improve the fitting results. In the simulation study, the efficiency of the EM algorithm can be tested. Using Bayesian information criterion and penalized log-likelihood function to select an appropriate model. Under some mild conditions, the asymptotic distribution theory for the resulting estimators is established [7].

II. THE STATISTICAL TECHNIQUE

A. Skew-normal GLMs

When $\varphi(\cdot;\mu,\sigma^2)$ presents the probability density function (pdf) of the normal distribution with mean μ and variance σ^2, $\Phi(\cdot)$ the cdf of the standard normal distribution, the pdf of skew-normal distribution is (Azzalini [3])

$$f(y;\mu,\sigma^2,\lambda) = 2\varphi(y;\mu,\sigma^2)\Phi(\frac{\lambda(y-\mu)}{\sigma}),\ y \in \mathbb{R} \quad (1)$$

where λ controls the normality of the regressor as λ=0 or not. The denotation of the random variable Y is simplified as $SN(\mu,\sigma^2,\lambda)$. The expectation and variance of Y are derived from its stochastic form by

$$E(Y) = \mu + \sqrt{\frac{2\sigma^2\lambda^2}{\pi(1+\lambda^2)}},\ Var(Y) = \sigma^2\left(1 - \frac{2\lambda^2}{\pi(1+\lambda^2)}\right) \quad (2)$$

Following analysis will present the GLM model under skew-normal error terms as well as the penalized maximum likelihood estimation using EM algorithm. Model is proposed as $y_i = x_i'\beta + f(t_i) + \varepsilon_i,\ i = 1,\ldots,n$, where y_i are the explained variables of the ith experimental unit distributed to $SN(\mu_i,\sigma^2,\lambda)$, x_i explanatory variables with a p-dimensional column vector and β the coefficient vector. Error term ϵ_i is independently distributed to zero-mean skew-normal distribution, ie $SN(\mu_i,\sigma^2,\lambda)$, which is the core component of the GLMSN. The model can be simplified as $Y = X\beta + Nf + \varepsilon$ which is the matrix form where there are n observed units and p dimensions for design matrix X. N is an (n×k) incidence matrix with (i, j) matching the indicator function $I(t_i=t_{j0})$, j=1,…,k. f equals $(f(t_{10}),\ldots,f(t_{k0}))T$ where $(t_{10},\ldots,t_{k0})$ are the values of ti. Generalized linear regression models of skew-normal data (GLMSN) indicates $y_i \sim SN(\mu_i,\sigma^2,\lambda)$ inferring the mean $\mu_i = x_i'\beta + n_i'f$, which derives the log-likelihood function of $\theta = (\beta', f', \sigma^2, \lambda)$ as

$$l(\theta) \propto -\frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i)^2 + \sum_{i=1}^{n}\ln\left(\Phi\left(\frac{\lambda(y_i - \mu_i)}{\sigma}\right)\right), l_p(\theta, \alpha) = l(\theta) - \frac{\alpha}{2}J(f) \quad (3)$$

In consideration of the calculability and identifiability to coefficient matrix, a practical method of log-likelihood penalization is introduced, which uses a penalty function to avoid regression problems between over-fitting and non-identification.

B. EM Algorithm

Expectation-maximization (EM) algorithm, introduced by Dempster et al (1977), is an efficient approach to maximum likelihood estimation (MLE). The EM algorithm works by two steps: first E-step for calculating the conditional expectation of log-likelihood as well as parameters' estimation and next M-step for maximizing the result from the first step. Estimating parameters for GLM by using EM algorithm can be summarized in 2 main steps in detail [1][8]:

Step 1: E-step: *compute* $\hat{z}_i^{(k)} = \hat{\lambda}^{(k)}\hat{\gamma}_i^{(k)} + \hat{\sigma}^{(k)}W_\Phi$ *where the* $\hat{\gamma}_i^{(k)} = y_i - x_i{}'\hat{\beta}^{(k)} - \mathrm{n}_i{}'\hat{f}^{(k)}$

Step 2: M-step: *calculate estimations* $\hat{\beta}^{(k+1)}$, $\hat{f}^{(k+1)}$, $\hat{\sigma}^{2^{(k+1)}}$, $\hat{\lambda}^{(k+1)}$ *as*

$$\hat{\beta}^{(k+1)} = (X'X)^{-1}X'\left(y_i - n_i'\hat{f}^{(k)} - \frac{\hat{\lambda}^{(k)}}{1+\hat{\lambda}^{(k)^2}\hat{z}_i^{(k)}}\right) \quad (4)$$

$$\hat{f}^{(k+1)} = \left(N'N + \frac{\alpha\hat{\sigma}^{(k)^2}}{(1+\hat{\lambda}^{(k)^2})K}\right)N'\left(y_i - x_i'\hat{\beta}^{(k)} - \frac{\hat{\lambda}^{(k)}}{1+\hat{\lambda}^{(k)^2}\hat{z}_i^{(k)}}\right) \quad (5)$$

$$\overset{\wedge}{\sigma^2}{}^{(k+1)} = \left(1'\overset{\wedge}{z^2}{}^{(k+1)} - 2\overset{\wedge}{\lambda}{}^{(k)}\hat{z}{}^{(k)'}\left(y - \overset{\wedge}{\mu}{}^{(k)}\right) + \left(1 + \hat{\lambda}^{(k)^2}\right)S\right) / 2n \quad (6)$$

$$\hat{\lambda}^{(k+1)} = \hat{z}^{(k)'}(y - \hat{\mu}^{(k)})/S \quad (7)$$

then based on the cdf of Gaussian distribution, we have pdfs:

$$f(y; \mu, \sigma^2, \lambda) = 2\varphi(y; \mu, \sigma^2)\int_0^{+\infty}\Phi(z; \lambda(y - \mu), \sigma^2)dz \quad (8)$$

$$f(y, z; \mu, \sigma^2, \lambda) = 2\varphi(y; \mu, \sigma^2)\Phi(z; \lambda(y - \mu), \sigma^2) \quad (9)$$

Ferreira and Paula (2017) [1] developed complete log-likelihood function with missing data, which calculates results of E-step:

$$Q(\theta \mid \hat{\theta}) = E\left[l_c(\theta \mid y_c) \mid y, \hat{\theta}^{(k)}\right] \propto -n\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum\hat{z}_i^{2^{(k)}} \quad (10)$$
$$+\frac{\lambda}{\sigma^2}\sum\hat{z}_i^{(k)}(y_i - \mu_i) - \frac{1+\lambda^2}{2\sigma^2}\sum(y_i - \mu_i)^2$$

(10) is called Q-function. According to the work of Green (1990) and Ibacache and Paula (2011) [5], the EM algorithm for penalized likelihood estimation can be inferred as

$$Q_p(\theta \mid \hat{\theta}) = Q(\theta \mid \hat{\theta}) - \frac{\alpha}{2}J(f), J(f) = \int_a^b[f''(t)]^2dt \quad (11)$$

where J(f) presents the penalty of pdf and smoothing parameter is negative. The Akaike information criterion (AIC) is applied for model selection[1]. When training the model and increasing the number of parameters, namely, increasing the complexity of the model, AIC tends to leads to overfitting. We consider Bayesian information criterion (BIC) if the dimension is too large and the training sample data is relatively small, the dimension disaster phenomenon can be effectively avoided. More effective model selection through BIC from EM algorithm is expected in empirical study. Penalized function is included in log-likelihood function.

$$AIC(\alpha) = 2\left[2 + p + df(\alpha)\right] - 2l_p(\hat{\theta},\ \alpha) \quad (12)$$

$$BIC(\alpha) = \ln(n)\left[2 + p + df(\alpha)\right] - 2l_p(\hat{\theta},\ \alpha) \quad (13)$$

III. Simulation Study

In this section, a simulation study and real data application are presented to assess the performance of the proposed EM-Algorithm. Two studies are implemented through R programming language using built-in data set fossil.

A. Simulation Study

A simple simulation example is conducted in this section using the proposed methodology and generate data with equations mentioned before. We assume the f(t) following doppler effect which is a cosine function and regressors distributed uniformly. Inspired by the work of Paula[1], we generate 500 samples for n=100, 150 and 200. Replicated studies present the decreasing bias of maximum likelihood estimation and empirical standard errors closing to standard deviations.

TABLE I. EM ESTIMATES AND EMPIRICAL STANDARD ERROR ESTIMATES (N=500)

parameter	*true value*	*mean*	*SD*	*SE*
β	5	4.99	0.09	0.06
σ^2	1	0.96	0.10	0.08
λ	3	2.94	0.54	0.39

Mean, standard deviations and standard error estimates after 500 iterations become very close to the true value, which reveal that the nonparametric estimators are consistent.

B. Application——California Air Polution Data & Fossil

We use R software built-in data California air pollution and fossil data set to demonstrate the difference between normal regression and skew-normal regression. The *calif.air.poll* data frame has 345 sets of observations ozone level and meteorological variables in Upland, California, U.S.A., in 1976, containing columns as follows:

- ozone_level: Ozone concentration (ppm) at Sandburg Air Force Base.
- pressure: Pressure gradient at Daggett, California.
- height: Inversion base height, feet.
- temperature: Inversion base temperature, degrees Fahrenheit.

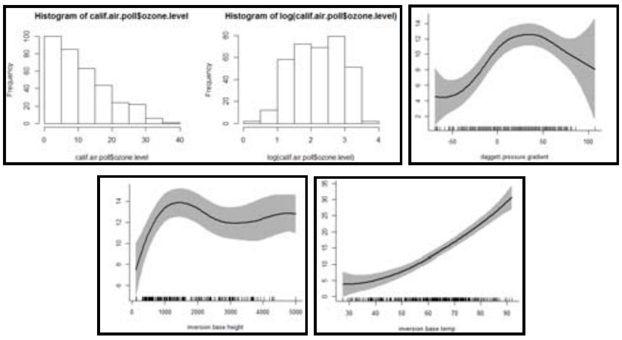

FIGURE I. HISTOGRAMS AND GLM FITTED CURVE OF POLLUTION

Histograms show the nonnormality of ozone levels in the original scale while approximate normality appears in log-scale. Considering a skew-normal distribution to this data set seems reasonable. As covariates included, GLMSN for ozone levels is $\sqrt{ozone_i} = \beta_1 pressure_i + \beta_2 height_i + \beta_3 temperature_i + f(t_i) + \varepsilon_i$ for i in (1: 345), t_i denotes the number of the ith observation of the ozone level which is ordinal variable. ε_i is independently distributed. Thus covariate matrix has a (345 × 3) dimension, N for (345×345) and $f = (f(t_1), \dots, f(t_{345}))'$.

Semiparametric regressions of ozone levels separately to pressure, height and temperature. Plots show the nonlinear relation between explained variable and explanatory variable generating smooth curves and confidence interval. Trend graphs of ozone level separately about pressure, height and temperature vary if one of the covariates is removed. For example, if height is removed, the monotonically increasing relationship between ozone level and temperature will be broken then a new first-rise-then-fall relation will be established. Widths of confidence intervals indicate the stability of the regression for each explanatory variable. In left and middle subgraphs, intervals are wider than right subgraph which illustrates proposed estimation capturing the tendency.

Results of regression using the mixed model representation of penalized spines is given. The degrees of freedom and knots of spline show the good fitness.

The *fossil* data frame has 106 sets of observations on fossil shells, containing columns as follows:

- strontium.ratio: ratios of strontium isotopes.
- age: fossil age in millions of years

TABLE II. GLM REGRESSION RESULTS

non-linear components	**summary**		
	df	*spar*	*knots*
pressure	4.697	88.8	31
height	4.198	2741.0	39
temperature	3.248	58.0	38

Skew-normality again appears in fossil data from histograms and boxplot. Points scattering evenly around the fitted curve shows that the linear model is not applicable but partially linear model can generate a good fitting. It is noticed that the intensity of points is related to the width of confidence interval that intervals tighten where points gather, which reflects the advantage of generalized linear models than linear models.

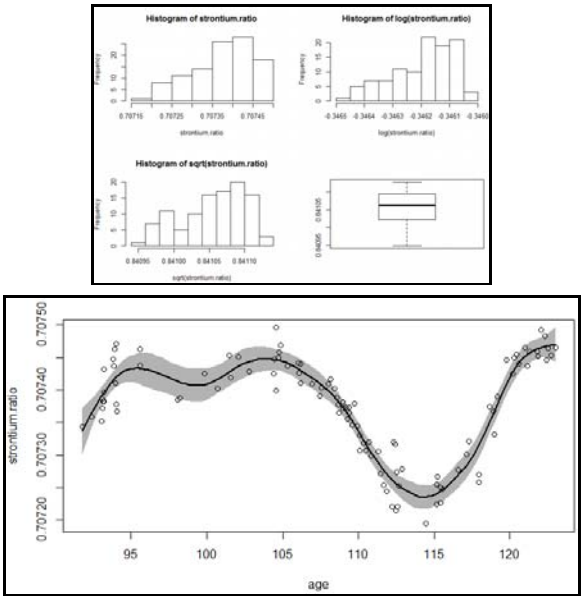

FIGURE II. HISTOGRAMS AND GLM FITTED CURVE OF FOSSIL

The regression curve verified the nonlinearity and there exist several extreme points. Concavity and convexity both appear in the fitted curve and the intensity of the points scattering infers the accuracy of generalized linear model to this skew-normal kind of data.

C. *Application——Annual Household Income*

We extract data from Chinese Household Income Project (2013) to analyze the influencing factors to annual income of rural families whose members have completed high school education. The explanatory variables consist of:

- inc: annual household income of 2013.
- page: family members' average age.
- phealth: family members' average health level (1~5 stands from best to bad)
- ptime: family members' average working time in a year (months)

The income data are not distributed normally form the histograms and boxplot drawn below. From correlation matrix and VIF, multicollinearity is excluded. Comparison analysis will be performed between normal and skew-normal error terms.

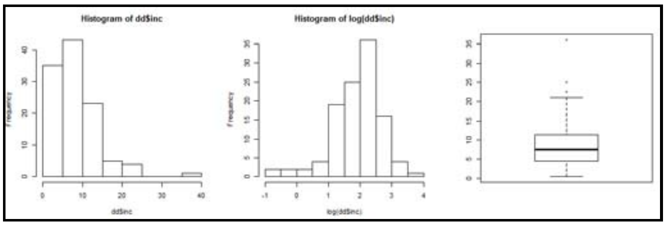

FIGURE III. HISTOGRAMS OF INCOME

Table 3 reports the values of the estimated parameters and the BIC and AIC are given for each model. By comparing two estimations we may notice inferential conclusion differs completely opposite for the sign of the coefficients. Values of Bayesian and Akaike information criterion are also presented for each model that two criterions are both smaller in generalized linear model than linear model indicating its better performance. Especially the pressure turns zero in generalized linear model although that is close to zero in linear model. This reflects the linear part of GLM regression relatively underestimates the effects of regressors. Residuals analysis for two regression models is given in the form of median value, the latter model being closer to zero. We use Bayesian information criterion with penalized spline in GLM and the value is smaller and it shows a little bit better goodness for fit which means the model selection tends to GLM.

TABLE III. COEFFICIENTS OF NORMAL AND GLMSN COMPARISION

covariate	LM		GLM	
	coefficient	***standard errors***	***coefficient***	***standard errors***
intercept	1.665	4.670	0.297	0.081
pressure	-0.031	0.077	0.000	0.001
height	-0.832	0.751	0.013	0.010
temperature	0.901	0.278	-0.019	0.006
AIC	699.6		660.6	
BIC	713.2		674.1	
Residuals (median)	-0.592		-0.061	

IV. CONCLUSION

From the estimation and applications, the proposed skew-normal generalized linear models are promising and the extended EM algorithm with penalized function works better in term of information criteria. A simple simulation study and empirical analysis give the comparison between linear and generalized linear regression in skew-normal data. The results of rural household income regression illustrate the applicability of the method. To estimate the parameters in the linear part in models, the extended penalized expectation maximization algorithm is used. The combination of skew-normal data and GLMs gives more robust results according to empirical analysis of household income.

REFERENCES

[1] Clécio S. Ferreira & Gilberto A. Paula (2017) Estimation and diagnostic for skew-normal partially linear models, Journal of Applied Statistics, 44:16, 3033-3053

[2] Chris Fraley & Adrian E Raftery (2002) Model-Based Clustering, Discriminant Analysis, and Density Estimation, Journal of the American Statistical Association, 97:458, 611-631

[3] Xia H, Härdle W (2006) Semi-parametric estimation of partially linear single-index models. J Multivar Anal 97:1162–1184

[4] A. Azzalini, A class of distributions which includes the normal ones,Scand.J.Statist.12(1985), pp. 171–178

[5] C.E.M. Relvas and G.A. Paula, Partially linear models with first-order autoregressive symmetric errors,Statist.Papers57(2016), pp. 795–825

[6] Fang, J., Liu, W. & Lu, X, Empirical likelihood for heteroscedastic partially linear single-index models with growing dimensional data, Metrika (2018) 81: 255

[7] Zhiyong Zhou, Zhengyan Lin,Varying coefficient partially nonlinear models with nonstationary regressors,Journal of Statistical Planning and Inference,Volume 194,2018,pp 47-64,ISSN 0378-3758

[8] A. Ganjavi, E. Christopher, C. M. Johnson and J. Clare, A study on probability of distribution loads based on expectation maximization algorithm, 2017 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), Washington, DC, 2017, pp. 1-5.

[9] D. Ruppert, M.P.Wand, andR. Carrol, Semiparametric Regression,CambridgeUniversityPress, New York, 2003.

[10] H. Zhu and S. Lee, Local influence for generalized linear mixed models,Canad.J.Statist.31 (2003), pp. 293–309.

[11] S.Y. Lee and L. Xu, Influence analysis ofnonlinear mixed-effects models, Comput. Statist. Data Anal. 45 (2004), pp. 321–341.

[12] Emre Dünder, Serpil Gümüştekin, Naci Murat and Mehmet Ali Cengiz, Variable selection in linear regression analysis with alternative Bayesian information criteria using differential evaluation algorithm, Communications in Statistics - Simulation and Computation, 47 :2 (2017), pp. 605-614