Communications of the Korean Statistical Society
2012, Vol. 19, No. 6, 793–798

DOI: http://dx.doi.org/10.5351/CKSS.2012.19.6.793

Hidden Truncation Normal Regression

Sungsu Kim[1,a]

[a]Applied Statistics Unit, Indian Statistical Institute

Abstract

In this paper, we propose regression methods based on the likelihood function. We assume Arnold-Beaver Skew Normal(ABSN) errors in a simple linear regression model. It was shown that the novel method performs better with an asymmetric data set compared to the usual regression model with the Gaussian errors. The utility of a novel method is demonstrated through simulation and real data sets.

Keywords: Arnold-Beaver skew normal distribution, asymmetric errors, goodness of fit test, simple linear regression.

1. Introduction

Many distributions encountered in practice are not symmetric but are skewed to some extent (Arnold and Beaver, 2000; Azzalini, 1986), as differently from what is usually known to be symmetric and unimodal, which can be modeled using a Gaussian distribution. SenGupta and Ugwuowo (2006) showed that many circular distributions are also asymmetrically distributed. Regression models using asymmetric error distributions appear in Bianco *et al.* (2005) and Marazzi and Yohai (2002). Bianco *et al.* (2005) introduces a regression model that is a natural extension of the method of moment estimates for ordinary regression and discusses their asymptotic properties. Marazzi and Yohai (2002) consider robust estimation of the linear regression model with symmetric and and asymmetric error distribution, where they introduce truncated maximum likelihood estimators. In this paper, we consider a regression model that employs a distribution called hidden truncation normal distribution or Arnold-Beaver Skew Normal(ABSN) distribution (Arnold and Beaver, 2000), which is suitable to model asymmetric or bimodal distributions. Useful properties of ABSN distribution are well studied in Arnold and Beaver (2000) and Azzalini (1986), and we refer our readers to the detailed results on ABSN distribution. A multivariate version of ABSN distribution is proposed in Azzalini and Dalla Valle (1996). In this paper, a linear - linear regression refers to a model that uses a linear dependent variable and a linear independent variable. Hidden truncated regression refers to a linear - linear regression with the assumption of ABSN errors. In the proposed hidden truncation regression model, we will consider a linear link function that relates the mean of the dependent variable to the independent variable through the following equation (Kutner, 2004)

$$\mu_{Y|X=x} = \beta_0 + \beta_1 x.$$

[1] Visiting Professor, Applied Statistics Unit, Indian Statistical Institute, 203 Barrackpore Trunk Road, Kolkata, 700108, India. E-mail: dr.sungsu@gmail.com

2. Hidden Truncation Normal Regression

Suppose $Y_1, \ldots, Y_n$ are independent variables, which are observed for fixed concomitant values denoted by $x_1, \ldots, x_n$. We assume that the conditional distribution of Y_i given $X_i = x_i$ is an ABSN distribution. This means that $Y_i = y_i$ is assumed to be observed for fixed x_i, and there exists a hidden covariable Z_i, which is left truncated at a value, h, where (Y_i, Z_i) has a joint bivariate normal distribution. The density is then given by

$$f_Y(y|\lambda_0, \lambda_1, \sigma) = \frac{\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)\Phi\left(\lambda_0 + \lambda_1\frac{(y-\mu)}{\sigma}\right)}{\sqrt{2\pi\sigma^2}\,\Phi\left(\frac{\lambda_0}{\sqrt{1+\lambda_1^2}}\right)},$$

where values of λ_0 and λ_1 are determined by values of h and σ_{YZ}, the covariance between Y_i and Z_i, and $\phi(\cdot)$ and $\Phi(\cdot)$ represent the pdf and the cdf of the standard normal distribution. Now, allowing the conditional mean to vary for different values of X, we write

$$f_{Y|x}(y|x, \beta_0, \beta_1, \lambda_0, \lambda_1, \sigma) = \frac{\exp\left(-\frac{(y-\mu(x))^2}{2\sigma^2}\right)\Phi\left(\lambda_0 + \lambda_1\frac{(y-\mu(x))}{\sigma}\right)}{\sqrt{2\pi\sigma^2}\,\Phi\left(\frac{\lambda_0}{\sqrt{1+\lambda_1^2}}\right)}, \tag{2.1}$$

where we use the linear mean link as $\mu(x) = \beta_0 + \beta_1 x$.

Proposition 1. *The moment generating function of the above density is given by*

$$M_{Y|x}(t) = \frac{\exp\left(\mu(x)t + \frac{\sigma^2 t^2}{2}\right)\Phi\left(\frac{\lambda_0 + \lambda_1\sigma t}{\sqrt{1+(\lambda_1)^2}}\right)}{\Phi\left(\frac{\lambda_0}{\sqrt{1+(\lambda_1)^2}}\right)}.$$

For a proof, our readers can refer to Arnold and Beaver (2000).
Using the linear mean link function and the additive property of the ABSN distribution, we write our model as

$$Y = \beta_0 + \beta_1 x + \sigma\epsilon, \quad \epsilon \sim \text{ABSN}(\lambda_0, \lambda_1),$$

where

$$f(\epsilon|\lambda_0, \lambda_1, \sigma) = \frac{\exp\left(-\frac{\epsilon^2}{2\sigma^2}\right)\Phi\left(\lambda_0 + \lambda_1\frac{\epsilon}{\sigma}\right)}{\sqrt{2\pi\sigma^2}\Phi\left(\frac{\lambda_0}{\sqrt{1+\lambda_1^2}}\right)}. \tag{2.2}$$

Using the above model, it is found that the errors have a non-zero conditional mean and a constant conditional variance (Arnold and Beaver, 2000). As a consequence, we have a bias given by

$$E\left(Y - \beta_0 - \beta_1 x|\lambda_0, \lambda_1, \sigma\right) = \frac{\sigma\lambda_1}{\sqrt{1+\lambda_1^2}}\,\Lambda\left(\frac{\lambda_0}{\sqrt{1+\lambda_1^2}}\right),$$

where $\Lambda(\cdot)$ is called the inverse Mill's ratio and represents $\phi(\cdot)/\Phi(\cdot)$. The bias is caused because the ABSN density is not centered at 0. This bias can be removed by shifting the model by the amount of the absolute difference between 0 and the mean of an ABSN distribution. This means that we add or subtract the same bias term so that the errors have 0 conditional mean direction.

Proposition 2. *If λ_1 becomes larger, the bias becomes closer to $\sigma/\sqrt{\pi}$.*

Proof:

$$\lim_{\lambda_1\to\infty} \frac{\sigma\lambda_1}{\sqrt{1+\lambda_1^2}}\Lambda\left(\frac{\lambda_0}{\sqrt{1+\lambda_1^2}}\right) = \lim_{\lambda_1\to\infty} \frac{\sigma\lambda_1}{\sqrt{1+\lambda_1^2}} \lim_{\lambda_1\to\infty} \Lambda\left(\frac{\lambda_0}{\sqrt{1+\lambda_1^2}}\right)$$
$$= \sigma\cdot\Lambda(0) = \sigma\frac{\phi(0)}{\Phi(0)} = \sigma\frac{\frac{1}{2\sqrt{\pi}}}{\frac{1}{2}} = \frac{\sigma}{\sqrt{\pi}},$$

where we use the Slutsky's product limit theorem. □

Therefore, it is seen that the amount of bias is asymptotically proportional to the unknown standard error of the regression model.

Proposition 3. *The conditional variance of the density shown in (2.2) is given by*

$$V(\epsilon|\lambda_0,\lambda_1,\sigma) = \sigma^2\left[1+\frac{\sigma^2\lambda_1^2}{1+\lambda_1^2}\Lambda\left(\frac{\lambda_0}{\sqrt{1+\lambda_1^2}}\right)\left(1-\Lambda\left(\frac{\lambda_0}{\sqrt{1+\lambda_1^2}}\right)\right)\right].$$

A proof is a straightforward exercise, therefore omitted here. The MLEs of the 5 parameters are obtained by simultaneously maximizing the log likelihood function shown below, with respect to the 5 parameters.

$$\frac{1}{2n}\sum_{i=1}^{n}\left[\log\frac{\exp\left(-\frac{(y-\beta_0-\beta_1x)^2}{2\sigma^2}\right)\Phi\left(\lambda_0+\lambda_1\frac{(y-\beta_0-\beta_1x)}{\sigma}\right)}{\sqrt{2\pi\sigma^2}\Phi\left(\frac{\lambda_0}{\sqrt{1+\lambda_1^2}}\right)}\right].$$

Throughout this paper, we have applied R function called 'optim' to obtain the MLEs, where it employs Newton-Raphson method of numerical optimization technique.

2.1. Asymptotic properties of the MLEs

One can easily check that an ABSN density satisfies the regularity conditions for consistency and asymptotic normality of the MLE (Kim, 2009). Thus, the MLEs, defined to be solutions of the likelihood equations: $n^{-1}\partial L_n(\theta)/\partial\theta = 0$, are consistent for θ_0, and

$$\sqrt{n}\left(\hat{\theta}_{ML}-\theta_0\right)\xrightarrow{d} N\left(0,(I[\theta_0])^{-1}\right),$$

where $I[\theta_0] = E\left[(\partial L_n(\theta)/\partial\theta)(\partial L_n(\theta)/\partial\theta')\right]_{\theta_0}$. The resulting asymptotic distribution of the MLE is given by

$$\hat{\theta}_{ML}\overset{a}{\sim} N\left(\theta_0,\frac{(I[\theta_0])^{-1}}{n}\right).$$

Table 1: ML estimates and standard errors from a simulation

	Parameter [True Value]			
	β_0 **[1.5]**	β_1 **[2]**	λ_0 **[−1.1]**	λ_1 **[−2.2]**
Estimates	1.42	1.92	−0.93	−2.07
(Standard Error)	(0.42)	(0.55)	(0.87)	(0.66)

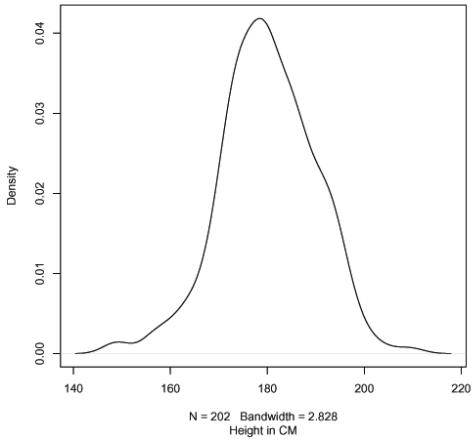

Figure 1: *Density plot of 202 heights*

In our context, $\theta_0 = (\beta_0, \beta_1, \sigma, \lambda_0, \lambda_1)$. The estimated MLEs are asymptotically normal with the estimated variance-covariance matrix given as the inverse of the observed information matrix.

2.2. Simulation and real examples

Suppose we fix 5 levels of X, an independent variable, with 10 replicates each, and the data generating process is such that $\epsilon \sim$ABSN(λ_0, λ_1) and $y = \beta_0 + \beta_1 x + \epsilon$. 100 such samples of size 50 were simulated to compute the ML estimates for each of the 100 samples. We picked 0.5, 1, 1.5, 2, 2.5 for the 5 levels of X. The corresponding 5000 values of Y were generated using values -1.1, -2.2, 1.5, and 2 for λ_0, λ_1, β_0, and β_1, respectively. The estimation results are shown in Table 1. It is shown that the ABSN model adequately fits the simulated data set and contains true parameter values within 95% confidence limits.

Next, we illustrate maximum likelihood estimation using a data set consisting of height and weight data on 102 male and 100 female athletes collected at the Australian Institute of Sport, courtesy of Richard Telford and Ross Cunningham (Cook and Weisberg, 1994). A smoothed density plot of the height data is presented in Figure 1. Slight right skewness of data set is evident in the figure. For a comparison purpose, we fit both models using normal errors and ABSN errors to the height and weight data set. The fit result is shown in Table 2. Testing hypotheses,

$$H_0 : \lambda_0 = \lambda_1 = 0 \quad vs. \quad H_1 : \text{ at least one of them is not equal to zero,}$$

with 0.1 level of significance, the likelihood ratio test statistic is given as

$$2\ln(535.1 - 523.9) = 4.83,$$

Table 2: Weight (y) *vs.* Height (x) of 202 Australian athletes.

	Parameter					Goodness of Fit
	β_0	β_1	σ	λ_0	λ_1	−log likelihood
Normal	139.3	0.54	8.58			535.1
ABSN	139.5	0.55	9.93	26.2	−19.4	523.9

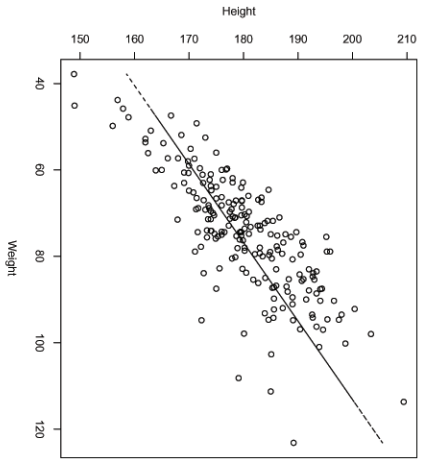

Figure 2: *Fitted line for weight vs. height*

which follows asymptotically the chi-square distribution with 2 degrees of freedom. Since the critical value, $\chi^2_{2,0.1}$, is 4.61, there is enough evidence to say that the data set has an ABSN distribution with λ_0 and λ_1 equal to 26.2 and −19.4, respectively, at 0.1 significant level. The exact p-value is given by 0.089. The fact that it was not significant at a higher level such as 0.05 seemed to be attributed to having a data set which is not severely asymmetric.

The mean value of Y when $X = x$ is given by

$$E(Y|x) = 139.5 + 0.55x - 9.92\Lambda(1.35) = 137.8 + 0.55x. \tag{2.3}$$

The estimated intercept value does not have a practical meaning; however, it is estimated that the weight increases 0.55 pounds for unit increase of inch in height. The fitted line (2.3) is shown on the scatter plot in Figure 2. It is shown that the ABSN model is a good fit for the weight versus height data set.

3. Conclusion

Many distributions encountered in practice are not symmetric but are skewed to some extent; subsequently, the use of regression models with asymmetrically distributed errors (such as ABSN errors) are in demand. Simulation results show that a maximum likelihood estimation is suitable for the parameters of the novel regression model. Fitting the weight and height data set shows that the hidden truncation regression model fits better (even with slight skewness in the data set) than the usual normal regression model.

References

Arnold, B. C. and Beaver, R. (2000). Hidden truncation models, *Sankhyā: The Indian Journal of Statistics*, **62**, 23–35.

Azzalini, A. (1986). A class of distributions which includes the normal ones, *Scandinavian Journal of Statistics*, **12**, 171–178.

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

Bianco, A., Ben, M. and Yohai, V. (2005). Robust estimation for linear regression with asymmetric errors, *Canadian Journal of Statistics*, **33**, 511–528.

Cook, R. D. and Weisberg, S. (1994). *An Introduction to Regression Graphics*, John Wiley and Sons, Inc., New York.

Kim, S. (2009). *Inverse Circular Regression with Possibly Asymmetric Error Distribution*, PhD Dissertation, University of California, Riverside.

Kutner, M. H. (2004). *Applied Linear Regression Models*, 4th ed., New York.

Marazzi, A. and Yohai, V. (2002). Adaptive truncated maximum likelihood regression with asymmetric errors, *Journal of Statistical Planning and Inference*, **122**, 271–291.

SenGupta, A. and Ugwuowo, F. (2006). Asymmetric circular-linear multivariate regression models with applications to environmental data, *Environmental and Ecological Statistics*, **13**, 299–309.

Received August 29, 2012; Revised October 10, 2012; Accepted October 12, 2012