Other Distributions for a Continuous Response Aside the Normal Distribution in a Linear Regression Model.

Felix Boakye Oppong
Hasselt University, Agoralaan Gebouw D, BE-3590 Diepenbeek, Belgium

Abstract

In many instances, when one encounters a continuous response in model building, the normal distribution is often the preferred choice for the distribution of the response given the predictors. In particular, to some statisticians, the normal distribution is seen as the only distribution for a continuous response. Even when the assumption of normality is not met, various transformations are applied on the data so that it appears to be more nearly normal. This is sometimes not pleasant since, the model may no longer apply directly to the original scale of measurement, which is in most cases of interest. Likewise, in doing so, one tries to force the model framework and distributional assumption that may not be best for the data at hand. Aside transformations, other distributions exist and can equally (or even better) be used for a continuous response in a linear regression model. The theory in GLM extends the linear regression theory such that, a much broader family of distributions can be used for the error terms other than the normal distribution. In this paper, other continuous distributions are used to illustrate how they outperform the normal distribution in some instances. It is also shown that, occasionally (for a continuous response), the normal distribution does not seem to be a choice unless transformations are applied. As a tool for assessing which of the distributions provides the best fit, both AIC and BIC are used.

To fit a GLM in SAS, the GENMOD procedure is used. In R, this can be accomplished by using the *glm* function. With these tools, only a handful of distributions can be used for the error terms. However, with the GAMLSS package in R, a number of distributions can be utilized. In using GAMLSS, the distribution of the response variable does not necessarily have to belong to the exponential family.

Keywords: GLM, Exponential Family, AIC, BIC, GAMLSS.

1. Introduction

Linear models are used to study how a quantitative variable depends on one or more predictors/covariates/explanatory variables. In a linear model, the predictors can either be quantitative or qualitative. A linear regression model can be written as

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i$$

Where y is the outcome, $x_1, x_2, \ldots, x_p$ are the set of predictors, $\beta_1, \beta_2, ..\beta_p$ are the set of parameters to be determined and ε_i are the random errors (residuals). The model makes the assumptions that $\varepsilon_i \sim N(0, \sigma^2)$. That is, the random errors are independent of each other and are normally distributed with a zero mean and a constant variance $, \sigma^2$ (Neter *et al*. 2005). With these assumptions, the expected value of the response, y, in the linear model is

$$E(y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}$$

This is called a linear model because, the mean response is a linear function of the unknown parameters, $\beta_1, \beta_2, ..\beta_p$. The popularity of linear regression model is ascribed to several factors. Linear regression models are natural approximating polynomials for more complex functional relationships. Secondly, the parameters, $\beta_1, \beta_2, ..\beta_p$, of the linear regression model are straightforward to estimate. Furthermore, there is a well-developed statistical theory for linear regression models. Provided the assumptions of the linear regression model are satisfied, statistical tests on the model parameters, confidence and predictive intervals for the mean response, etc, can easily be obtained and used for inferences (Myers *et al*, 2012). Likewise, linear regression models can be fitted in almost all statistical packages.

Like with any other statistical model, the assumptions underlying the use of the linear regression model should be met before it can be used in making inference (Oppong and Agbedra, 2016). Occasionally, in a linear regression model, the assumption of normally distributed errors is violated. In such instances, data transformations are often applied so that in the end, the random errors will appear to be more nearly normal. Nonetheless, data transformation comes with some consequences. First, the model no longer applies directly to the original scale of measurement, which is in most cases of interest. Secondly, in applying transformations, one tries to force the model framework and distributional assumption that may not be best for the data.

2. Methodology

The linear regression model can be extended such that, other distribution aside the normal distribution can also be used. The theory in generalized linear model (GLM) is a significant development beyond linear regression theory in which a much broader family of distributions can be used for the error terms other than the normal distribution as used in linear regression (McCullagh and Nelder, 1989). In GLM, three components are distinguished namely, the random component, systematic component and the link function. The random component specifies a probability distribution for the response variable (Y), the systematic component identifies the set of predictors used and the link function specifies a function that maps $E(Y)$ to the systematic component (Nelder and Wedderburn, 1972). In this paper, interest is in the random component of GLM. More specifically, the random component in GLM allows the inclusion of distributions from the *Exponential Family.* The idea of the random component in GLM is to allow one to include more suitable probability models, rather than to try and make things fit into the usual (often not appropriate) normal-based methods. In the absence of a rigorous treatment of GLM, McCullagh and Nelder (1989) can be referred to for more detail. In this paper, a real data set is used to illustrate an instance in which the normal distribution is not a better distribution for the continuous response. To aid in the selection of a best fitting model (distribution), Akaike Information Criterion (AIC) together with Bayesian Information Criterion (BIC) is used.

2.1 AIC and BIC

AIC and BIC are the most common information-theoretic criteria used in model selection (Lesaffre and Lawson, 2012). In particular, information criteria techniques for model selection emphasize minimizing the amount of information required to express the data and the model. This leads to the selection of models that are efficiently represented by the data (Acquah, 2010). That is, AIC and BIC balance model fit with model complexity (number of parameters). Unlike with other model selection tools, AIC and BIC can be used for both nested and non-nested models although, they are mostly used for non-nested models. By definition,

$$AIC = -2\log\left(\mathcal{L}(\hat{\beta})\right) + 2K \text{ and } BIC = -2\log\left(\mathcal{L}(\hat{\beta})\right) + (\log n)K$$

$\mathcal{L}$ is the likelihood function, $\widehat{\beta}$ is the maximum likelihood estimate of β, K is the number of estimated parameters (including the variance) and n is the sample size. The distinction between AIC and BIC is in the second term which depends on the sample size, n, in the case of BIC. AIC tends to select more complicated models whereas BIC often leads to the selection of simpler models (Kuha, 2004). With these criteria, models with smaller AIC or BIC are considered to provide a better fit. Ideally, a difference of more than 5 is considered a substantial evidence for the model with smaller AIC or BIC whereas, a difference of more than 10 is often considered a strong evidence (Lesaffre and Lawson, 2012). In this paper both criteria will be used in the selection of the “best” fitting model.

2.2 Other distributions for continuous response

With GLM, continuous distributions of the *Exponential Family* can equally be used in a regression model aside the normal distribution (McCullagh and Nelder, 1989). For a continuous response, members of the one parameter or multi-parameter exponential family of distributions can be implemented. However, preference for one distribution over the other depends on the range of support of the distribution, that is, $(-\infty, \infty)$, $(0, \infty)$ or $(0,1)$. Some of these distributions include normal, inverse normal, exponential, gamma, inverse gamma, beta, Pareto, Weibull, chi-squared and many others. All of these distributions take the form

$$f(y; \theta, \phi) = \exp\{\big(y\theta - b(\theta)\big)/a(\phi) \ + c(y, \phi)\}$$

$E(Y) = b'(\theta), Var(Y) = b''(\theta)a(\phi)$, ϕ is called the dispersion parameter and θ is the natural parameter (Agresti, 2012).

As a way of illustrating situations in which other continuous distributions are more practical compared to the normal distribution, the salmonella data (Ribbens e*t al*, 2007) is used in a regression model. The data set used consist of measurements of three variables from 1402 pigs. The measured variable are Sample to Positive ratios (SP-ratios), weight and sampling time in days. SP-ratios is a continuous variable calculated as the ratio of the differences between Optical density (OD) and the mean optical density of the positive and negative controls ($\overline{OD}pos$ and $\overline{OD}neg$). Weight is put into four categories, that is, <40kg, 40-59kg, 60-80kg and >80kg and Sampling time (in days) records the day of the year (day 3 to day 363) on which measurements are taken. Here, SP ratio is regressed on weight and sampling time. That is, it is of interest to study the effect of weight and sampling time on SP ratio.

2.3 Model selection.

For illustration sake, a simple model which assumes a linear relationship between the predictors (weight, sampling time) and the response variable (SP ratio) is considered. Also, for simplicity, a model without interaction is used. The fitted model is expressed as:

$$f(x; \beta) = \beta_0 + \beta_1 \text{Weight}_1 + \beta_2 \text{Weight}_2 + \beta_3 \text{Weight}_3 + \beta_4 \text{Sampling Time},$$

Where $f(x; \beta)$ specifies the distribution of the response, that is, gamma, inverse gamma, lognormal, inverse Gaussian, Weibull or normal (Gaussian) distribution.

$$\text{Weight}_1 = \begin{cases} 1 & \text{weight} < 40 \\ 0 & \text{Otherwise} \end{cases}, \text{Weight}_2 = \begin{cases} 1 & \text{weight } 40 - 59 \\ 0 & \text{Otherwise} \end{cases}, \text{Weight}_3 = \begin{cases} 1 & \text{weight is } 60 - 80 \\ 0 & \text{Otherwise} \end{cases}$$

To illustrate how to fit this model in practical applications, SAS® 9.4 and R version 3.1.3 are used. In SAS, the GENMOD procedure is used to fit a generalized linear model. However, not all distributions of the *exponential family* are available in the GENMOD procedure. On the other hand, one can easily define a distribution through DATA step programming statements used within the procedure. The available distributions are gamma, geometric, inverse Gaussian, multinomial, negative binomial, normal, Poisson and the zero-inflated Poisson (SAS Institute Inc., 2008). All the aspects of the GENMOD procedure will not be considered in this paper. Here, the procedure is used to illustrate how other continuous distributions can be incorporated in the linear regression model. For our model, the following SAS program is used.

```
proc genmod data =pigData;
class weight;
model SP_Ratio=weight day /dist=NORMAL;
run;
```

With the *dist* option which stand for distribution, other distributions can be included. For example, dist= GAMMA uses the gamma distribution instead of the normal distribution for the distribution of the response. The *class* option allows SAS to treat weight as a classification variable and not a continuous variable.

To fit a GLM in R, the *glm* function can be used. With this function, the family of distributions in use are the normal (Gaussian), binomial, Poisson, gamma, inverse Gaussian and the quasi family which allows fitting user-defined models by maximum quasi-likelihood. To fit the model in R, the following program can be used.

1. glm(SP ~ Weight_1 + Weight_2 + Weight_3 + day, family=gaussian)
2. glm(SP ~ Weight_1 + Weight_2 + Weight_3 + day, family=Gamma)

With the *family* option, other distributions can be included aside the normal distribution. For instance, model 2 makes use of the gamma distribution.

With the Generalized Additive Models for Location Scale and Shape (GAMLSS) package in R, a number of distributions can be included in the regression model. “GAMLSS is a general framework for fitting regression type models where the distribution of the response variable does not necessarily have to belong to the exponential family” (Stasinopoulos and Rigby, 2007). With GAMLSS, distribution that belong to the exponential family as well as those that do not belong to the exponential family of distributions can equally be used in the regression model. For the distributions considered in this paper, we illustrate how they can be used in a linear regression model using the GAMLSS package.

gamlss(SP~ Weight_1 + Weight_2 + Weight_3 + day, family=NO)

The family=NO option uses the normal distribution as the conditional distribution of the response given the predictors. For the other distributions, we replace with family=LOGNO for the lognormal distribution, family= GA for gamma, family= IGAMMA for the inverse gamma distribution, family= IG for inverse Gaussian and family= WEI for the Weibull distribution.

3. Results

To gain insight in the distribution of the response, a histogram of SP ratio is obtained (Figure 1). The histogram depicts a positively skewed distribution. It should be emphasized that the response (SP ratio), can only take positive values $(0, \infty)$. SP ratio is a continuous variable which usually ranges from 0 to 4 but even higher values can be observed. With the data at hand, SP ratio ranges from 0.005 to 3.44. Hence, the family of distribution that will be more appropriate are those distribution that take non-negative values. Hence, in this case, the normal distribution will not be a good choice since it will stand the chance of predicting a negative SP ratio. However, for the purpose of illustration, the normal distribution is also used, just to show how inappropriate it is in such situations. For the positively skewed distributions, gamma, inverse gamma, lognormal, inverse Gaussian and the Weibull distributions are considered. Also the symmetric nature of the log(SP ratio) suggests that a regression model with a normal distribution and logarithm of SP ratio as the response could be considered as well. However, given the fact that the linear regression model is conditional on the covariates and since the normality assumption is on the residuals and not on the response, the choice of the model cannot be entirely based on this plot.

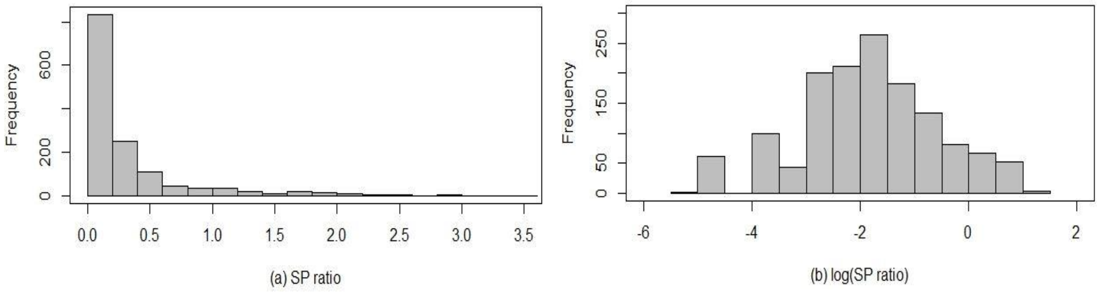

Figure 1. Histogram of SP ratio and log(SP ratio)

Since this is not an actual data analysis, not all output will be presented. Here, interest is in choosing a distribution that provides the best fit in terms of AIC and BIC.

In fitting the specified model with these distributions, the results in Table 1 is obtained. From the output, the normal distribution is seen to perform poorly among all the distributions considered. This was indeed to be expected since all the other distribution have support between $(0, -\infty)$ whereas the normal distribution has support form $(-\infty, \infty)$. SP ratio can only take positive values hence, distributions with support between $(0, -\infty)$ will provide better fit. It should be emphasized that, in practical applications, either AIC or BIC is used and not both. In this paper, both are used to show that it many situation (not always), they lead to the selection of the same model. Here, the distribution that provides the best fit is the lognormal distribution. Compared to all the other distributions, the lognormal distribution has the lowest AIC and/or BIC

Table 1. AIC and BIC for models with different distributions for the response.

Distribution	AIC	BIC
gamma	-369.23	-348.24
inverse gamma	-384.03	-363.05
lognormal	-652.14	-631.16
inverse Gaussian	-430.32	-409.33
Weibull	-426.21	-405.23
normal	1850.92	1871.91

To further provide evidence that the normal distribution is not the only ultimate distribution for continuous responses, some attention is given to this distribution. From the symmetric nature of the histogram in Figure 1 (b), a model with log(SP ratio) as response is also considered for the normal distribution. Without delving much into the model assumptions, it is observed in Figure 2 that the error (residuals) are normally distributed as expected. However the AIC and BIC associated with this model is still very high compared to the other distribution considered in Table 1. The AIC for the model with a normal distribution and log(SP ratio) as response is 4524.403 and a BIC of 4555.877. Compared to the model with the normal distribution on the original response scale, the model with the log transformed response performs even worse.

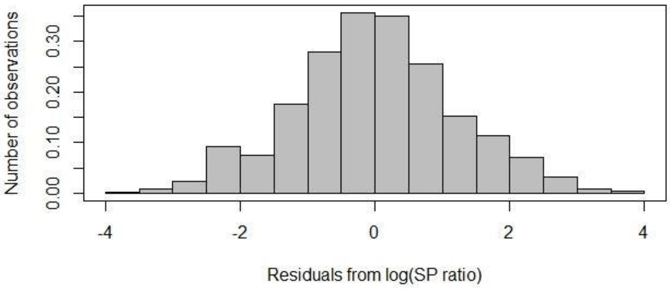

Figure 2. Histogram of residuals of model with log(SP ratio) as response

4. Conclusion

Although the normal distribution is often seen as an “omnipresent” distribution in many statistical application (Oppong and Agbedra, 2016) particularly in modelling a continuous response, it is not by far the only available distribution. In some instances, when a continuous response is encountered in a regression model, the normal distribution is often not a plausible distribution for the response given the predictors. With GLM, a much broader family of distributions can be used for the error terms other than the normal distribution as used in linear regression (McCullagh and Nelder, 1989).

In this paper, it has been shown that, many other continuous distribution can be used in a linear regression model and often, some of these distributions provide a better fit compared to the normal distribution. Likewise, it has been demonstrated that, the normal distribution is sometimes not a choice at all even though transformations are possible alternatives. However, transformations to near normality comes with a price. First, the model no longer applies directly to the original scale of measurement, which is mostly of interest. Likewise, in applying transformations, one forces the model framework and distributional assumption that may not be best for the data at hand.

In illustrating how GLM is implemented in practice, the GENMOD procedure in SAS as well as the *glm* function in R is employed. With these procedures, a handful of distributions are allowed for the distribution of the response. However, with the GAMLSS package in R, a substantial number of distributions can be used. In GAMLSS, the distribution of the response variable does not necessarily have to belong to the exponential family. Hence distributions of the exponential family together with distributions that do not belong to the exponential family can be used in the regression model (Stasinopoulos and Rigby, 2007).

References

Acquah, H. D. G. (2010). Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship. Journal of Development and Agricultural Economics, 2(1), 1-6.

Agresti, A. (2012). *Categorical Data Analysis.* (2nd Edition). John Wiley & Sons.

Kuha, J. (2004). AIC and BIC: Comparison of assumptions and performance. Sociological Methods and Research, 33, 188-229.

Lesaffre, E. and Lawson A. (2012). *Bayesian Biostatistics*. John Wiley and Sons.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models.* (2nd Edition). Chapman and Hall/CRC

Myers, R. H., Montgomery, D. C., Vining, G. G. and Robinson, T .J. (2010). *Generalized linear models: with Applications in Engineering and the Sciences.* (2nd Edition). John Wiley & Sons.

Nelder, J., and Wedderburn, R. W. M. (1972). Generalized linear models. J. Roy. Statist. Soc. Ser. A 135: 370-384

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (2005). *Applied Linear Statistical Models*. (5th Edition). New York: McGraw-Hill Education.

Oppong, F. B. and Agbedra, S. Y. (2016). Assessing Univariate and Multivariate Normality, A Guide For Non-Statisticians. Mathematical Theory and Modeling, 6(2), 26-33.

Ribbens, S., Dewulf, J., Maes, D., Koenen, F., Mintiens, K., Desadeleer, L. and Kruif, A. (2007) A survey on biosecurity in Belgian pig herds.Prev. Veter. Med., doi: 10.1016/j.prevetmed.2007.07.009.

SAS Institute Inc. (2008). SAS/STAT® 9.2 User’s Guide. Cary, NC: SAS Institute Inc.

Stasinopoulos, M. D. and Rigby, R. A. (2007). Generalized Additive Models for Location Scale and Shape (GMLSS) in R. Journal of Statistical Software , 23(7), 1-46.