Unique Regression Model for both Symmetric and Asymmetric Regressed Variable

Mian Arif Shams Adnan[1], Silvey Shamsi[2], Rahmatullah Imon[3]
[1]Department of Computer Science, Ball State University
[2]Center for Business and Economic Research (CBER), Ball State University
[3]Department of Mathematical Sciences, Ball State University
Muncie, IN 47304

Abstract
All previously derived Regression Models were either based on the normality assumption(s) of the regressed variable and their non-normality counterparts on their asymmetric distributional assumption of the explained variable. The present paper addressed a new sort of common regression model that is suitable for error(s) either following a symmetric distribution or an asymmetric distribution. Attempt for estimation of parameters has also been addressed.

Key word: Explained Variable ; Invariance.

1. Introduction

One of the basic problems in regression analysis is the normality assumption of the distribution of the explained variable. The requirement is not pragmatically met in several situations. In real life the explained variable may have asymmetric pattern or symmetric-non-normal pattern or symmetric-normal pattern. So, it is better to develop a regression model which is fit to any of the aforesaid situations. So, if the distribution of the explained variable is symmetric-normal of symmetric-non-normal or asymmetric, we can use our regression model and we can analyze the data to test several hypotheses.

Adnan and Kiser (2011) developed a Generalized Exponential distribution which is the core probability distribution that can generate a couple of other probability distributions such as Generalized Gamma, Generalized Chi-square, Generalized t, Generalized F, etc. Generalized Beta 1st kind and 2nd kind distributions are also addressed from the generalized distribution. The two parameter generalized double exponential distribution was defined as a distribution of a random variable X having the probability density function

$$f(x) = \frac{b\sqrt[b]{a}}{2\Gamma(\frac{1}{b})}\, e^{-|ax^b|} \;;\; -\infty \le x \le \infty. \qquad (1)$$

where a is the scale-parameter and b is the shape-parameter such that $a, b > 0$.

Attempts have been made here to develop a regression model which is invariant to the type or shape of the explained variable. Here the explained variable follows the distribution addressed in equation (1).

2. Least Square Method for Simple Linear Regression with Double Exponential Regressed Variable

Let the simple linear regression model is

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{2.1}$$

where the intercept β_0 and the slope β_1 are unknown constant known as regression coefficients and ε is a random error component. Here,

$$\varepsilon \sim DE\left(0, \frac{\Gamma\left(\frac{3}{b}\right)}{a^{2/b}\Gamma\left(\frac{1}{b}\right)}\right)$$

and

$$y \sim DE\left(\beta_0 + \beta_1 x, \frac{\Gamma\left(\frac{3}{b}\right)}{a^{2/b}\Gamma\left(\frac{1}{b}\right)}\right)$$

The errors are assumed to have mean zero and unknown variance $\frac{\Gamma\left(\frac{3}{b}\right)}{a^{2/b}\Gamma\left(\frac{1}{b}\right)}$. Here the errors are uncorrelated. There is a Generalized Double Exponential probability distribution for y at each possible value for x such that

$$E(y|x) = \beta_0 + \beta_1 x$$

and

$$V(y|x) = V(\beta_0 + \beta_1 x + \varepsilon) = \frac{\Gamma\left(\frac{3}{b}\right)}{a^{2/b}\Gamma\left(\frac{1}{b}\right)}.$$

Although the mean of y is a linear function of x that is the conditional mean of y depends on x, but the conditional variance of y does not depend on x. Moreover, the responses y are uncorrelated since the errors ε are uncorrelated.

Since the parameters β_0 and β_1 are unknown, they should be estimated using sample data. Suppose that we have n pairs of data, say $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$ obtained from a controlled experimental design or from an observational study or from existing historical records. Least Square method estimates β_0 and β_1 so that the sum of squares of differences between the observations y_i and the straight line is minimum. From equation 2.1 we can write

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i; i = 1, 2, \ldots, n \tag{2.2}$$

Equation 2.1 presents the Population Regression Model and equation 2.2 expresses the Sample Regression Model. Now the sum of squares of deviations from the true line is

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{2.3}$$

Now the least square estimates of β_0 and β_1 must satisfy

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \tag{2.4}$$

and $$\frac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = 0. \quad (2.5)$$

After simplification the two normal equations are generally found such that

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \quad (2.6)$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i \quad (2.7)$$

The solution to the normal equations is

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (2.8)$$

and $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \quad (2.9)$$

where $$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \text{ and } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Therefore, $\hat{\beta}_0$ and $\hat{\beta}_1$ are the Least Square estimates of the intercept and slope respectively.

The fitted Simple Linear Regression Model is

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x. \quad (2.10)$$

3. Statistical Properties of Least Square Estimators

The estimators of the regression parameters are linear and unbiased.

3.1 Linearity

The estimators of the regression parameters are linear.

$$\hat{\beta}_1 = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} = \frac{\sum(x_i-\bar{x})y_i - \sum(x_i-\bar{x})\bar{y}}{\sum(x_i-\bar{x})^2} = \frac{\sum[(x_i-\bar{x})y_i]}{\sum(x_i-\bar{x})^2}$$

Assuming, $\frac{(x_i-\bar{x})}{\sum(x_i-\bar{x})^2} = k_i$, $\hat{\beta}_1$ gives the following form such that

$$\therefore \hat{\beta}_1 = \sum_{i=1}^{n} k_i y_i \quad (3.1)$$

Similarly, $$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{1}{n}\sum y_i - \bar{x}\sum k_i y_i$$

$$\therefore \hat{\beta}_0 = \sum\left[\frac{1}{n} - \bar{x}k_i\right] y_i \quad (3.2)$$

Thus, both $\hat{\beta}_0$ and $\hat{\beta}_1$ are expressed as linear functions of $y's$.

3.2 Unbiasedness

The estimators of the regression parameters are unbiased.

$$\hat{\beta}_1 = \sum_{i=1}^{n} k_i y_i = \sum k_i (\beta_0 + \beta_1 x_i + \varepsilon_i)$$

$$\hat{\beta}_1 = \beta_0 \sum k_i + \beta_1 \sum k_i x_i + \sum k_i \varepsilon_i \quad (3.3)$$

$$\therefore E(\hat{\beta}_1) = \beta_0 \sum E(k_i) + \beta_1 \sum E(k_i x_i) + \sum E(k_i \varepsilon_i) = 0 + \beta_1 \sum E(k_i x_i) + \sum k_i E(\varepsilon_i)$$

$$= \beta_1 \sum E\left[\frac{(x_i-\bar{x})}{\sum(x_i-\bar{x})^2} x_i\right] + \sum \frac{(x_i-\bar{x})}{\sum(x_i-\bar{x})^2} E(\varepsilon_i) = \beta_1 \sum E\left[\frac{(x_i-\bar{x})}{\sum(x_i-\bar{x})^2} x_i\right] + 0 = \beta_1 \sum E\left[\frac{{x_i}^2 - x_i\bar{x}}{\sum(x_i-\bar{x})^2}\right]$$

$$= \beta_1 \sum \left[\frac{{x_i}^2 - x_i\bar{x}}{\sum(x_i-\bar{x})^2}\right] = \beta_1 \left[\frac{\sum {x_i}^2 - \bar{x}\sum x_i}{\sum(x_i-\bar{x})^2}\right] = \beta_1 \left[\frac{\sum {x_i}^2 - n\bar{x}^2}{\sum {x_i}^2 - \bar{x}n\bar{x}}\right] = \beta_1$$

$$\therefore E(\hat{\beta}_1) = \beta_1$$

Again, $$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\therefore E(\hat{\beta}_0) = E(\bar{y}) - \bar{x}E(\hat{\beta}_1) = E\left(\frac{1}{n}\sum y_i\right) - \bar{x}\beta_1 = \frac{1}{n}\sum E(y_i) - \bar{x}\beta_1$$

$$= \frac{1}{n}\sum(\beta_0 + \beta_1 x_i) - \bar{x}\beta_1 = \beta_0 + \bar{x}\beta_1 - \bar{x}\beta_1 = \beta_0$$

$$\therefore E(\hat{\beta}_0) = \beta_0.$$

$\therefore$ $\hat{\beta}_1$ and $\hat{\beta}_0$ are the unbiased estimators of β_1 and β_0.

3 Maximum Likelihood Estimators

For n pairs of data, $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$ where $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i; i = 1, 2, \ldots, n$

and

$$y \sim DE(\beta_0 + \beta_1 x, \frac{\Gamma(\frac{3}{b})}{a^{2/b}\Gamma(\frac{1}{b})}),$$

the Maximum Likelihood Function will be

$$L = \prod_{i=1}^{n} f\left(y_i | x_i, \beta_0, \beta_1, \frac{\Gamma(\frac{3}{b})}{a^{2/b}\Gamma\left(\frac{1}{b}\right)}\right)$$

$$= \frac{b\sqrt[b]{a}}{2\lceil\frac{1}{b}} e^{-|a(y_1-\beta_0-\beta_1 x_1)^b|} \times \frac{b\sqrt[b]{a}}{2\lceil\frac{1}{b}} e^{-|a(y_2-\beta_0-\beta_1 x_2)^b|} \times \ldots \times \frac{b\sqrt[b]{a}}{2\lceil\frac{1}{b}} e^{-|a(y_n-\beta_0-\beta_1 x_n)^b|}$$

$$= \prod_{i=1}^{n} \frac{b\sqrt[b]{a}}{2\lceil\frac{1}{b}} e^{-|a(y_i-\beta_0-\beta_1 x)^b|}$$

$$\therefore L = \left(\frac{b\sqrt[b]{a}}{2\lceil\frac{1}{b}}\right)^n e^{-|a\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x)^b|}$$

where, $a\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x)^b$ is positive or negative.

For, positive $a\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x)^b$

$$\therefore L=\left(\frac{b\sqrt[b]{a}}{2\left\lceil\frac{1}{b}\right\rceil}\right)^n e^{-|a\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x)^b|}$$

Taking logarithm on both sides we find the log likelihood function as

$$\log_e L=\log_e\left[\left(\frac{b\sqrt[b]{a}}{2\left\lceil\frac{1}{b}\right\rceil}\right)^n e^{-a\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x)^b}\right]$$

$$\Rightarrow \log_e L=\ln\left(\frac{b\sqrt[b]{a}}{2\left\lceil\frac{1}{b}\right\rceil}\right)^n+\ln\left[e^{-a\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x)^b}\right]$$

$$\Rightarrow \log_e L=n\ln b+\frac{n}{b}\ln a-n\ln 2\left\lceil\frac{1}{b}-a\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x)^b\right.$$

Now the maximum likelihood estimator can be obtained as a solution of the following equations

$$\frac{\partial}{\partial(\beta_0,\beta_1)}\log_e L\,[a,b;\ y_1,y_2,\dots,y_n\,]=0$$

$$\frac{\partial}{\partial\beta_1}\log_e L\,[a,b;\ y_1,y_2,\dots,y_n\,]=0$$

$$=\frac{\partial}{\partial\beta_1}\left[n\ln b+\frac{n}{b}\ln a-n\ln 2\left\lceil\frac{1}{b}-a\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x)^b\right.\right]$$

$$0=0-a\sum_{i=1}^{n}b(y_i-\beta_0-\beta_1x)^{b-1}(-x_i)$$

$$\sum_{i=1}^{n}x_i(y_i-\beta_0-\beta_1x)^{b-1}=0 \qquad (4.1)$$

Again,
$$\frac{\partial}{\partial\beta_0}\log_e L\,[a,b;\ y_1,y_2,\dots,y_n\,]=0$$

$$=\frac{\partial}{\partial\beta_0}\left[n\ln b+\frac{n}{b}\ln a-n\ln 2\left\lceil\frac{1}{b}-a\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x)^b\right.\right]$$

$$0=0-a\sum_{i=1}^{n}b(y_i-\beta_0-\beta_1x)^{b-1}(-1)$$

$$\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x)^{b-1}=0 \qquad (4.2)$$

Solving the two equations (4.1) and (4.2) by Numerical Integration say Newton Raphson Method we can estimate the β_0 and β_1. If we put b = 2, the equation (4.1) and (4.2) become normal equations for the regression model where the error term is distributed as Normal.

However, before solving the equations (4.1) and (4.2), we need to estimate the value of a and b. Adnan and Kiser (2011) addressed a classification of distributions based on various

specification of the parameters of a and b. The various specifications of a and b along with the classification of distributions are addressed below.

Table: The various specifications of a and b along with the classification of distributions for the response variable (y).

Sl.	Name of the distribution	$f(x)$	Support	a	b	y	Mean	Variance
1	Generalized Double Exponential	$\frac{b\sqrt[b]{a}}{2\Gamma(\frac{1}{b})}\, e^{-\lvert a(y-\beta_0-\beta_1 x)^b\rvert}$	$-\infty \le y \le \infty$	a	b	y	$\beta_0 + \beta_1 x$	$\frac{\Gamma(\frac{3}{b})}{a^{2/b}\Gamma\left(\frac{1}{b}\right)}$
2	Std. Laplace	$\frac{1}{2}e^{-\lvert y\rvert}$	$-\infty \le y \le \infty$	1	1	y	0	2
3	Laplace	$\frac{1}{2\lambda}\, e^{-\left\lvert\frac{y-\beta_0-\beta_1 x}{\lambda}\right\rvert}$	$-\infty \le y \le \infty$	$\frac{1}{\lambda}$	1	y	$\beta_0 + \beta_1 x$	$2\lambda^2$
4	Standard normal	$\frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}(y)^2}$	$-\infty \le y \le \infty$	$\frac{1}{2}$	2	y	0	1
5	Normal	$\frac{1}{\sqrt{2\pi}\sigma}\, e^{-\frac{1}{2}(\frac{y-\beta_0-\beta_1 x}{\sigma})^2}$	$-\infty \le y \le \infty$	$\frac{1}{2\sigma^2}$	2	y	$\beta_0 + \beta_1 x$	σ^2
6	Log-normal	$\frac{e^{-\frac{1}{2}(\frac{\log y-\beta_0-\beta_1 x}{\sigma})^2}}{y\sqrt{2\pi}\sigma}$	$0 \le y \le \infty$	$\frac{1}{2\sigma^2}$	2	y	$e^{\beta_0+\beta_1 x+\frac{1}{2}\sigma^2}$	$e^{2(\beta_0+\beta_1 x)+\sigma^2}\{e^{\sigma^2}-1\}$
7	Exponential	$ae^{-a(y)}$	$0 \le y \le \infty$	a	1	y	$\frac{1}{a}$	$\frac{1}{a^2}$
8	Gamma	$\frac{y^{\frac{1}{b}-1}e^{-(y)}}{\Gamma(\frac{1}{b})}$	$0 \le y \le \infty$	1	b	y	$\frac{1}{b}$	$1/b$
9	Rayleigh	$\frac{2y}{\lambda^2}e^{-\frac{(y)^2}{\lambda^2}}$	$0 \le y \le \infty$	$\frac{1}{\lambda^2}$	1	y	0.886λ	$0.215\lambda^2$

If for drawn the random sample the variable y is distributed to any distribution other than the aforesaid one, it is required to estimate a and b to classify its pattern. The pattern can be unfolded from the mean variance or higher order moments relationship of the distribution.

Conclusion

The authors are trying to develop the inference procedure for the generalized regression model. The link with the generalized linear model is being thought to be developed.

References

(i). Adnan, M. A. S. (2011). *A Generalization of the Family of Exponential and Beta Distributions*. Proc. Joint Statistical Meetings organized by American Statistical Association, held in Florida, USA during July 30-August 4, 2011.

(ii). *A Two Parameter Generalized Exponential Distribution.* Presented in the 2010 Joint Statistical Meetings organized by American Statistical Association, held in Vancouver, Canada during July 31-August 5, 2010.