Thailand Statistician
July 2023; 21(3): 510-515
http://statassoc.or.th
Contributed paper

Identifiability of Longitudinal Regression Models with Response Variables Observed on $(0, 1)$ or $[0, 1]$

Elham Tabrizi [a] and Ehsan Bahrami Samani*[b]

[a] Department of Mathematics, Faculty of Mathematics and Computer Science, Kharazmi University, Tehran, Iran

[b] Department of Statistics, Faculty of Mathematical Science, Shahid Beheshti University, Tehran, Iran

*Corresponding author; e-mail: ehsan_bahrami_samani@yahoo.com

Received: 30 December 2020
Revised: 14 July 2021
Accepted: 30 September 2021

Abstract

The identifiability problem is treated for mixtures involving two specific $(0, 1)$ and $\{0, 1\}$ supported distributions. It is proved and elaborated that two conditions are needed to achieve the model identifiability in a longitudinal beta regression model or a longitudinal zero and one inflated beta regression model: the model parameter vector does not contain any intercept, with all the domains of the covariates containing at least one interval in the real numbers and zero. The techniques for establishing identifiability used here may be applied to other statistical models.

Keywords: Beta regression model, inflation, random effect.

1. Introduction

In many areas in science and engineering, beta regression model is widely used when these studies are aimed at studying how different variables affect some proportion (observed on the open interval $(0, 1)$). Some different types of regression for modelling data restricted to the unit interval $(0, 1)$ are given in Kieschnick and McCullough (2012). They compared the mentioned models using two common samples and recommended the researchers to use either a beta or a quasi-likelihood regression model. Ferrari and Cribari-Neto (2004) proposed a suitable parameterization of beta distribution in terms of its mean and a precision parameter, followed by introducing a beta regression model for response variables confined within the standard unit interval $(0, 1)$. Longitudinal beta regression model was studied by Verkuilen and Smithson (2012). When zeros or ones appear in the data, response variables of this kind (called fractions observed on the closed interval $[0, 1]$) are often difficult to analyze using beta theory models since their distributions may be quite poorly modeled by beta distribution in which a positive probability is not allowed at values 0 and 1 for some parameter values. Ospina and Ferrari (2010) have considered a mixed continuous-discrete distribution based on a mixture of the beta and a Bernoulli distribution. Using the work of Ospina and Ferrari (2010), Ospina and Ferrari (2012) extended the beta regression to zero-or-one-inflated beta regression model. Model identifiability refers to the question whether the parameters of a candidate model can be uniquely determined by data. There are several reasons why such a question could be of utmost importance. The most important reasons are (1) it is at all possible to determine the true parameters regardless of the number of observations, and (2) numerical search procedures for the parameter estimates may have problems.

Definition 1 Let $P_{\boldsymbol{\theta}}^{y}$ be a the distribution for a random object, Y, associated with parameter vector $\boldsymbol{\theta}$ for a candidate model. This model is identifiable if the equations in $\boldsymbol{\theta}$ arisen from the equation

$$P_{\boldsymbol{\theta}}^{y} = P_{\boldsymbol{\theta}^*}^{y}, \qquad \text{for all } y$$

has the only solution $\boldsymbol{\theta} = \boldsymbol{\theta}^*$.

There is a substantial literature on identifiability. It is the prerequisite for consistent and asymptotically unbiased estimation. Therefore, before attempting to perform parameter estimation (fitting a model), it is important to check model identifiability. The covariance matrix of response variable in a linear mixed effect model is equal to the sum of two parameter matrices. Hence, non-identifiability of parameters may occur. Some theorems about non-identifiability have been noticed for the linear mixed effect models in Wang (2013). Other situations in which fitting a model to data does not yield unique parameter estimates can be generalized to linear models with random effects, the models with non-ignorable missing-data mechanism, and mixture models. Miao et al. (2016) investigated the model identifiability of the normal, normal mixture, and t mixture models with non-monotone missing mechanisms. Cui et al. (2017) extended the work of Miao et al. (2016) to the exponential family. Culpepper (2019) discussed some identifiability conditions for structured multinomial mixture models with binary attributes. Some early references on statistical models under missing not at random mechanism are **?** and **?**. **?** presented three theorems to check model identifiability of a general form of the joint model for two types of responses. Based on the three theorems, **?** found some sufficient conditions for the identifiability of some types of the beta regression models such as inflated beta regression model with non-ignorable missing values and longitudinal joint modelling of Poisson and inflated beta responses with non-ignorable missing values. Since investigating identifiability issue in longitudinal generalized linear models is an interesting topic and the structure of the beta regression model is similar to that of the generalized linear models, we address the problem of identifiability in the longitudinal beta regression (LB) model and the longitudinal zero and one inflated beta (LIB) regression model.

The paper is organized as follows. In Section 2, we establish some sufficient conditions for the identifiability of a longitudinal beta regression model and Section 3 refers to such conditions for the longitudinal zero and one inflated beta regression model.

2. Modelling the Proportional and Fractional Outcomes

In many studies, LB and LIB regression models have been some top choices to model proportional and fractional outcomes. In Sections 2.1 and 2.2, an overview of LB and LIB models will be given, respectively.

2.1. Longitudinal beta regression for modelling proportions

Let Y_{it} denote the response variable that given a random effect, follows the beta distribution with mean μ_{it} and precision ϕ presented in Ferrari and Cribari-Neto (2004) for the $i-th$ individual at time t for $i = 1, ..., I$ and $t = 1, ..., T$. Suppose that the outcome model is a longitudinal beta regression model as:

$$\begin{aligned} &Y_{it}|b_i \sim \mathit{Beta}(\phi\mu_{it}, \phi(1-\mu_{it})), \quad i = 1, ..., I, \quad t = 1, ..., T, \\ &logit(\mu_{it}) = \mathbf{x}_{it}{}'\boldsymbol{\gamma}^{\mu} + b_i, \end{aligned} \tag{1}$$

where $b_i \overset{iid}{\sim} \mathit{Normal}(0, \sigma^2)$ is the random intercept, $\mathbf{x}_{it} = (x_{it}^1, x_{it}^2, ...x_{it}^K)'$ represent the whole set of K covariates, and $\boldsymbol{\theta} = (\boldsymbol{\gamma}^{\mu\prime}, \phi, \sigma)'$ is the model parameter vector. The joint density function for the joint model (1) is

$$\begin{aligned}\prod_{i=1}^{I} f_{\mathbf{Y}_i}(\mathbf{y}_i;\boldsymbol{\theta}) &= \prod_{i=1}^{I}\int_{\mathcal{R}}\prod_{t=1}^{T} f_Y(y_{it}|b_i;\boldsymbol{\gamma}^{\mu})\varphi(b_i;0,\sigma^2)db_i \\ &= \prod_{i=1}^{I}\int_{\mathcal{R}}\prod_{t=1}^{T} f_{Beta}\left(y_{it};\phi\mu_{it},\phi(1-\mu_{it})\right)\ \varphi(b_i;0,\sigma^2)db_i, \end{aligned} \tag{2}$$

where $\mathbf{Y}_i = (Y_{i1},...,Y_{iT})'$.

2.2. Longitudinal inflated beta rRegression for modelling fractions

In this subsection, it is supposed that $Y_{it}|b_i$ comes from the zero and one inflated beta ($INFBE_{0,1}$) distribution. A longitudinal zero and one inflated beta regression model will be as follows:

$$\begin{aligned}&Y_{it}|b_i \sim INFBE_{0,1}(\alpha_{it},\lambda_{it},\mu_{it},\phi), \quad i=1,...,I, \quad t=1,...,T,\\ &logit(\alpha_{it}) = \mathbf{x}_{it}{}'\boldsymbol{\gamma}^{\alpha} + b_i,\ logit(\lambda_{it}) = \mathbf{x}_{it}{}'\boldsymbol{\gamma}^{\lambda} + b_i,\ logit(\mu_{it}) = \mathbf{x}_{it}{}'\boldsymbol{\gamma}^{\mu} + b_i.\end{aligned} \tag{3}$$

For $\boldsymbol{\theta} = (\boldsymbol{\gamma}^{\alpha\prime},\boldsymbol{\gamma}^{\lambda\prime},\boldsymbol{\gamma}^{\mu\prime},\phi,\sigma)'$, the joint density function of model (3) can be written as

$$\begin{aligned}\prod_{i=1}^{I} f_{\mathbf{Y}_i}(\mathbf{y}_i;\boldsymbol{\theta}) = \prod_{i=1}^{I}\int_{\mathcal{R}}\prod_{t=1}^{T}\Big[&\alpha_{it}\ \lambda_{it}{}^{y_{it}}\ (1-\lambda_{it})^{1-y_{it}}\ I_{\{0,1\}}(y_{it}) + (1-\alpha_{it})\times \\ &f_{Beta}\left(y_{it};\phi\mu_{it},\phi(1-\mu_{it})\right)\Big] I_{(0,1)}(y_{it})\ \varphi(b_i;0,\sigma^2)db_i,\end{aligned} \tag{4}$$

where $I_{\{0,1\}}(y_{it})$ is an indicator function of the subset $\{0,1\}$.

3. The Main Identification Results

Here, we will obtain the sufficient conditions for the identifiability of LB and LIB models in the form of two theorems.

Theorem 1 *If the vector of coefficients γ^{μ} does not contain an intercept and all the covariates X_{it}^{k} for $k=1,2,...,K$ take all values in $\mathcal{S}_{X^k}$, where $\mathcal{S}_{X^k}\subseteq\mathcal{R}$ contains at least one interval and zero, then the model* (1) *is identifiable.*

Proof: By Theorem 1 in **?**, it is enough to show that

$$f_{\mathbf{Y}_i}(\mathbf{y}_i;\boldsymbol{\theta}) = f_{\mathbf{Y}_i}(\mathbf{y}_i;\boldsymbol{\theta}^*), \tag{5}$$

in the joint density function (2) implies $\boldsymbol{\theta}=\boldsymbol{\theta}^*$. Without loss of generality, assume that there exists one covariate denoted by X_{it}, suppressing the index i for simplicity, multiplying both sides of (5) by y_1 and then integrating both of its sides with respect to y_1, y_2, ..., and y_T over $(0,1)^T$, we have

$$\int_{\mathcal{R}}\frac{exp(\gamma_1^{\mu}x_1+b)}{1+exp(\gamma_1^{\mu}x_1+b)}\ \varphi(b;0,\sigma^2)db = \int_{\mathcal{R}}\frac{exp(\gamma_1^{\mu*}x_1+b)}{1+exp(\gamma_1^{\mu*}x_1+b)}\ \varphi(b;0,\sigma^{*2})db. \tag{6}$$

Changing the order of integrations is allowed by Fubini's theorem. By the assumptions of Theorem 1, we are allowed to take the first derivative in both sides of (6) with respect to x_1 at x as an arbitrary point in $\mathcal{S}_X$. It yields

$$\gamma_1^{\mu}\int_{\mathcal{R}}\Big[\frac{exp(\gamma_1^{\mu}x+b)}{1+exp(\gamma_1^{\mu}x+b)} - \left(\frac{exp(\gamma_1^{\mu}x+b)}{1+exp(\gamma_1^{\mu}x+b)}\right)^2\Big]\ \varphi(b;0,\sigma^2)db =$$

$$\gamma_1^{\mu*} \int_{\mathcal{R}} \Big[\frac{exp(\gamma_1^{\mu*}x+b)}{1+exp(\gamma_1^{\mu*}x+b)} - \Big(\frac{exp(\gamma_1^{\mu*}x+b)}{1+exp(\gamma_1^{\mu*}x+b)}\Big)^2\Big]\,\varphi(b;0,\sigma^{*2})db. \tag{7}$$

Multiplying both sides of (5) by y_1y_2 and $y_1(1-y_1)$ and then integrating them with respect to y_1, y_2, ..., and y_T over $(0,1)^T$ gives us

$$\begin{aligned}&\int_{\mathcal{R}} \frac{exp(\gamma_1^{\mu}x_1+b)}{1+exp(\gamma_1^{\mu}x_1+b)}\,\frac{exp(\gamma_1^{\mu}x_2+b)}{1+exp(\gamma_1^{\mu}x_2+b)}\,\varphi(b;0,\sigma^2)db = \\ &\int_{\mathcal{R}} \frac{exp(\gamma_1^{\mu*}x_1+b)}{1+exp(\gamma_1^{\mu*}x_1+b)}\,\frac{exp(\gamma_1^{\mu*}x_2+b)}{1+exp(\gamma_1^{\mu*}x_2+b)}\,\varphi(b;0,\sigma^{*2})db,\end{aligned} \tag{8}$$

and

$$\begin{aligned}&\frac{\phi}{\phi+1}\int_{\mathcal{R}} \frac{exp(\gamma_1^{\mu}x_1+b)}{1+exp(\gamma_1^{\mu}x_1+b)}\Big[1-\frac{exp(\gamma_1^{\mu}x_1+b)}{1+exp(\gamma_1^{\mu}x_1+b)}\Big]\,\varphi(b;0,\sigma^2)db = \\ &\frac{\phi^*}{\phi^*+1}\int_{\mathcal{R}} \frac{exp(\gamma_1^{\mu*}x_1+b)}{1+exp(\gamma_1^{\mu*}x_1+b)}\Big[1-\frac{exp(\gamma_1^{\mu*}x_1+b)}{1+exp(\gamma_1^{\mu*}x_1+b)}\Big]\,\varphi(b;0,\sigma^{*2})db,\end{aligned} \tag{9}$$

respectively. Note that both functions μ_1 and $\partial\mu_1/\partial x_1$ are continuous with respect to x_1 and their integrals with respect to b are absolutely converge (since the absolute values of the integrands are at most 1). Thus, interchanging the differentiation and integration is allowed. Again, note that since $plogis(a) > plogis^2(a)$ for $\forall a \in \mathcal{R}$, the integrals in both sides of (7) cannot be zero. Thus, $\gamma_1^{\mu} = \gamma_1^{\mu*}$ and $\phi = \phi^*$ are yielded by substituting (6) and (8) at $x_1 = x_2 = x$ into (7) and (6) and (8) at $x_1 = x_2 = x$ into (9), respectively. Under $\phi = \phi^*$ and the assumptions of Theorem 1, make the change of variable $z = b/\sigma$ and $z = b/\sigma^*$ in the left and right sides of (9) at $x_1 = 0$, respectively. Since the function $e^{\sigma z}/(1+e^{\sigma z})^2$ is an even function, it yields

$$\int_0^{+\infty} \Big[\frac{e^{\sigma z}}{(1+e^{\sigma z})^2} - \frac{e^{\sigma^* z}}{(1+e^{\sigma^* z})^2}\Big]\,\varphi(z;0,1)dz = 0. \tag{10}$$

By contradiction, from (10), it can be proven that $\sigma = \sigma^*$. For more details, see **?**.

Theorem 2 *If the vector of coefficients* $\boldsymbol{\gamma}^{\alpha}$, $\boldsymbol{\gamma}^{\lambda}$, *and* $\boldsymbol{\gamma}^{\mu}$ *do not contain any intercept and all the covariates* X_{it}^k *for* $k = 1, 2, ..., K$ *take all values in* $\mathcal{S}_{X^k}$, *where* $\mathcal{S}_{X^k} \subseteq \mathcal{R}$ *contains at least one interval and zero, then the model* (3) *is identifiable.*

Proof: Again, without loss of generality, suppose that there exists one covariate denoted by X_{it}. We need to prove that

$$f_{\mathbf{Y}_i}(\mathbf{y}_i;\boldsymbol{\theta}) = f_{\mathbf{Y}_i}(\mathbf{y}_i;\boldsymbol{\theta}^*), \tag{11}$$

in (4) for all values of y_{it} implies $\boldsymbol{\theta} = \boldsymbol{\theta}^*$. Suppressing the index i for simplicity, Letting (1) $y_{i1} = y_{i2} = 0$, (2) $y_1 = y_2 = 1$, (3) $y_1 = 0$ and $y_2 = 1$, and (4) $y_1 = 1$ and $y_2 = 0$ in (11) and then summing over $\{0,1\}$ and integrating over $(0,1)$ both sides of this equation with respect to y_{it} for $t = 3, ..., T$, we have four equations. Adding these four equations together yields

$$\int_{\mathcal{R}} \alpha_1\,\alpha_2\,\varphi(b;0,\sigma^2)db = \int_{\mathcal{R}} \alpha_1^*\,\alpha_2^*\,\varphi(b;0,\sigma^{*2})db. \tag{12}$$

Now, let $y_1 \in \{0,1\}$ and $y_2 \in (0,1)$. Taking the sum over $\{0,1\}$ with respect to y_1, taking the integral over $(0,1)$ with respect to y_2, and finally taking sum over $\{0,1\}$ and integral over $(0,1)$ of both sides of the equation in (11) with respect to y_{it} for $t = 3, ..., T$, this equality reduces to

$$\int_{\mathcal{R}} \alpha_1 \, (1-\alpha_2) \, \varphi(b;0,\sigma^2)db = \int_{\mathcal{R}} \alpha_1^* \, (1-\alpha_2^*) \, \varphi(b;0,\sigma^{*2})db. \tag{13}$$

Substituting (12) into (13) yields

$$\int_{\mathcal{R}} \alpha_1 \, \varphi(b;0,\sigma^2)db = \int_{\mathcal{R}} \alpha_1^* \, \varphi(b;0,\sigma^{*2})db. \tag{14}$$

Substituting (13) at $x_1 = x_2 = x$ into the equation yielded by taking the first derivative in both sides of (14) with respect to x_1 at $x \in \mathcal{S}_X$ gives $\gamma_1^{\alpha} = \gamma_1^{\alpha *}$. By the assumptions of Theorem 2, letting $x_1 = x_2 = 0$ in (13), we face the same equation we used to prove $\sigma = \sigma^*$ in the proof of Theorem 1. Thus, by the same argument, it can be proven that $\sigma = \sigma^*$. Adding the equations yielded from the cases **??** and **??** together, we have

$$\int_{\mathcal{R}} \alpha_1 \, \alpha_2 \, \lambda_1 \, \varphi(b;0,\sigma^2)db = \int_{\mathcal{R}} \alpha_1^* \, \alpha_2^* \, \lambda_1^* \, \varphi(b;0,\sigma^{*2})db. \tag{15}$$

Now, subtracting the equation yielded by applying operation $\partial/\partial x_1$ in both sides of (15) at $x_1 = x_2 = x$ from the equation yielded by applying operation $\partial/\partial x_2$ in both sides of (15) at $x_1 = x_2 = x$ gives

$$\gamma_1^{\lambda} \int_{\mathcal{R}} \alpha^2 \, \lambda \, (1-\lambda) \, \varphi(b;0,\sigma^2)db = \gamma_1^{\lambda *} \int_{\mathcal{R}} \alpha^{*2} \, \lambda^* \, (1-\lambda^*) \, \varphi(b;0,\sigma^{*2})db, \tag{16}$$

where $\alpha = plogis(\gamma_1^{\alpha} x + b)$, $\alpha^* = plogis(\gamma_1^{\alpha *} x + b)$, $\lambda = plogis(\gamma_1^{\lambda} x + b)$, and $\lambda^* = plogis(\gamma_1^{\lambda *} x + b)$ in which $plogis(a) = exp(a)/(1 + exp(a))$. Substituting the equation yielded from the cases **??** or **??** at $x_1 = x_2 = x$ into (16) yields $\gamma_1^{\lambda} = \gamma_1^{\lambda *}$. Now, let $y_1, y_2 \in (0,1)$, and $y_t \in [0,1]$ for $t = 3, ..., T$. Multiplying both sides of (11) by y_1 and then integrating over $(0,1)$ with respect to y_1 and y_2, summing over $\{0,1\}$ with respect to y_t for $t = 3, ..., T$, and finally, integrating over $(0,1)$ with respect to y_t for $t = 3, ..., T$ yields

$$\int_{\mathcal{R}} (1-\alpha_1) \, (1-\alpha_2) \, \mu_1 \, \varphi(b;0,\sigma^2)db = \int_{\mathcal{R}} (1-\alpha_1^*) \, (1-\alpha_2^*) \, \mu_1^* \, \varphi(b;0,\sigma^{*2})db. \tag{17}$$

Adding the equations yielded by applying operations $\partial/\partial x_1$ and $\partial/\partial x_2$ in both sides of (17) at $x_1 = x_2 = x$ together gives

$$\begin{aligned} &\gamma_1^{\mu}\Bigg(\int_{\mathcal{R}} (1-\alpha)^2 \, \mu \, \varphi(b;0,\sigma^2)db - \int_{\mathcal{R}} (1-\alpha)^2 \, \mu^2 \, \varphi(b;0,\sigma^2)db\Bigg) = \\ &\gamma_1^{\mu *}\Bigg(\int_{\mathcal{R}} (1-\alpha^*)^2 \, \mu^* \, \varphi(b;0,\sigma^{*2})db - \int_{\mathcal{R}} (1-\alpha^*)^2 \, \mu^{*2} \, \varphi(b;0,\sigma^{*2})db\Bigg). \end{aligned} \tag{18}$$

To prove $\gamma_1^{\mu} = \gamma_1^{\mu *}$, it is enough to substitute (17) at $x_1 = x_2 = x$ into (18). Finally, let $y_1, y_2 \in (0,1)$, and $y_t \in [0,1]$ for $t = 3, ..., T$. If both sides of (11) are multiplied by $y_1(1-y_1)$ and then integrated over $(0,1)$ with respect to y_1 and y_2, summed over $\{0,1\}$ with respect to y_t for $t = 3, .., T$, and integrated over $(0,1)$ with respect to y_t for $t = 3, .., T$, we have

$$\frac{\phi}{\phi+1} \int_{\mathcal{R}} (1-\alpha_1) \, (1-\alpha_2) \, \mu_1 \, (1-\mu_1) \, \varphi(b;0,\sigma^2)db =$$

$$\frac{\phi^*}{\phi^*+1}\int_{\mathcal{R}}\left(1-\alpha_1^*\right)\left(1-\alpha_2^*\right)\mu_1^*\left(1-\mu_1^*\right)\varphi(b;0,\sigma^{*2})db. \tag{19}$$

Again, substituting (17) at $x_1 = x_2 = x$ into (19) yields $\phi = \phi^*$.

4. Discussion and Conclusion

The model identifiability is of great importance for the study of statistical models, since it is the necessary theoretical foundation for estimation and inference. We studied the problem of identifiability for the longitudinal beta regression model and the longitudinal zero and one inflated beta regression model. We indicated how some mild conditions provide a complete solution to this problem. Furthermore, the solution sheds some light on the identification problem from some other longitudinal statistical models. It will be interesting to further pursue several ideas brought out in this way: studying the identifiability of LB and LIB models including discrete covariates and studying the necessary and sufficient conditions for identifiability.

References

Cui X, Guo J, Yang G. On the identifiability and estimation of generalized linear models with parametric nonignorable missing data mechanism. Comput Stat Data Anal. 2017; 107: 64-80.

Culpepper SA. An exploratory diagnostic model for ordinal responses with binary attributes: Identifiability and estimation. Psychometrika. 2019; 84(4): 921-940.

Ferrari S, Cribari-Neto F. Beta regression for modelling rates and proportions. J Appl Stat. 2004; 31(7): 799-815.

Kieschnick R, McCullough BD. Regression analysis of variates observed on $(0, 1)$: percentages, proportions and fractions. Stat Model. 2003; 3(3): 193-213.

Miao W, Ding P, Geng Z. Identifiability of normal and normal mixture models with nonignorable missing data. J Am Stat Assoc. 2017; 111(516): 1673-1683.

Ospina R, Ferrari SL. Inflated beta distributions. Stat Pap. 2010; 51(1): 111-126.

Ospina R, Ferrari SL. A general class of zero-or-one inflated beta regression models. Comput Stat Data Anal. 2012; 56(6): 1609-1623.

Verkuilen J, Smithson M. Mixed and mixture regression models for continuous bounded responses using the beta distribution. J Educ Behav Stat. 2012; 37(1): 82-113.

Wang, W. Identifiability of linear mixed effects models. Electron J Stat. 2013; 7: 244-263.