IJER, Vol. 10, No. 2, July-December, 2013, pp. 475-486

ASSESSING THE PERFORMANCE OF BETA-LOGIT AND NORMAL-LOGIT SPECIFICATIONS ON GROUPED DEPENDENT VARIABLE LOGIT MODELS

Alfredo A. Romero

Department of Economics, North Carolina A&T State University, Greensboro, NC, USA

ABSTRACT

In applied work, the need for combining data from different sources is often times unavoidable, either for availability, practicality, or legality reasons. This is particularly true in agricultural and/or environmental research, where the information is often available on a national or a regional basis. One of these situations arises when the researcher has only access to aggregated information about the dependent variable (where the choices of n-individuals are averaged across groups) but possesses individual information for the independent variables. In this document, we assess the performance of two competing specifications for estimating this kind of models, a new beta-logit model, and the standard normal-logit model. We also propose a methodology for simulating the data that uses conditional and marginal distributions from which the true specifications are derived. This method of simulation allows for the direct comparison of the models without giving a computational advantage to either of them. We argue that a beta-distributed model provides more accurate estimators than those provided by the standard model.

JEL Codes: C25; C46; C52; C81

Keywords: Logit, aggregated choice data, combined data, beta regression.

Choice modeling endeavors require a set of characteristics of an economic agent, be a consumer, a household, a firm, or a farm; and a set of choices made by the economic agent. In a best case scenario, a researcher or policy maker has the ability to design a statistical study focused on specific goals using specific variables of interest, most of the time without taking into consideration the potential use and limitations that the data would pose to other researchers. For instance, a study on credit card decisions may include an array of sociodemographic characteristics of the individuals in a household, including race, but might omit a very important variable for a sociologist or a labor economist interested in phenotypical discrimination, such as skin tone. When situations like these arise, the researchers have no choice but to aggregate, consolidate, or combine information from different data sources to satisfactorily answer a policy question.

The need of combining data from different sources is particularly important on choice in agriculture and environmental settings, where the information is often available on a national or regional basis (Kurkalova and Wade, 2010). In studies of this nature, the sheer amount of data makes it difficult, expensive, or impractical to achieve a high level of specificity and the practitioner must conform tothe answers provided togeneral questions. The obvious consequence is the widespread lack of detailed information for state or smaller scale regions and the need to combine information from supplemental sources to complete a research project (see Lambert, Schaible, Johansson, Daberkow, 2006; Lambert, Schaible, Johansson, Vasavada, 2007; Soule, Tegene, Wiebe, 2000). Besides the issues of practicality or prohibitive costs, there might also exist a legal impediment from the researcher's side to fully or partially disclose detailed information about the data. In this situation, the use of aggregated and/or estimated values would be necessary.

It is that latter kind of models and the performance of two competing specifications that are the focal point of this document. In particular, we assess the performance of beta-logit and normal-logit specifications for a special case of data combination models where the dependent variable is agrouped average of undisclosed choices and the independent variables are observable individual characteristics (Grouped Dependent Variable Logit models). This scenario is pervasive in agricultural studies where the researcher might be able to obtain weighted averages of individual responses from published reports and would require pairing them to individual characteristics, usually taken from other datasets. As long as the response variable has been weighted to approximate the expected value of the geographic area (crop reporting district or county), it can be possible to match individual characteristics with group responses despite them not coming from the same survey without compromising the statistical integrity of the study (Kurkalova and Wade, 2010).

To assess the usefulness of GDVL models, we compare the estimation performance of two competing specifications, the normal-logit and the beta-logit. Both specifications use the sample average logit-link function to associate the averaged dependent variable to the individual independent variables. The key difference in the models is the specification of a normally distributed error, or a beta distributed error, and their corresponding likelihood functions. The a priori expectation is that the beta model will perform better in small samples. The rest of the paper is as follows; section 2 describes grouped dependent variable models using both the normal and the beta distribution. Section 3 provides the simulation and estimation strategy. Section 4 presents the simulation results while section 5 summarizes this research.

2. GROUPED DEPENDENT VARIABLE MODELS

Kurkalova and Rabotyagov (2005) introduced the specification and estimation of Group Dependent Variable Logit models (GDVL) as an alternative to grouping the independent variables when the dependent variable is grouped (see Miller and

Plantinga, 1999) or to disaggregating the dependent variable when information on individuals is available for the independent variables. Their rationale for not aggregating the independent variables is driven by the notion that when one is using aggregated explanatory variables the estimated coefficients are in general biased and inference conducted using this aggregated data may differ significantly from those retrieved from individual level data (Steel and Holt, 1996; Train, 2009). On the other hand, the GDVLavoids both the computational cumbersomeness and the creation of pseudo-data points that results from disaggregating the average choice so that a logistic regression can be run.

In its most general form, the GDVLconsiders a group of N agents, indexed by i, each making a binary choice. The choice outcome variable Y_i is either 1 or 0 depending on whether a certainalternative is chosen or not. If Y_i is observed, the exact relationship between Y_i and the set of K explanatoryvariables $x_i = (x_{i1},\ldots,x_{iK})'$ can be established via a logistic regression. So that $Pr[Y_i = 1] = \Pr[\eta_i < \beta' \boldsymbol{x}_i]$; where η_i is the *i.i.d.* logistic error term and $\boldsymbol{\beta} = (\beta_1,\ldots\ \beta_K)'$ is the vectorof unknown parameters of interest. If Y_i is not observed directly, the researcher uses non-empty subsets, G_j, $j = 1,\ldots,J$, of the set of all respondents $\{1,\ldots, N\}$, so that $\Sigma_j N^{G_j} = N$, where N^{G_j} is the number of individuals in subset G_j; and uses the expected value (or some other related central measure statistic) of the choice variables, i.e., $\overline{p}^{Gj} = \frac{1}{N^{G_j}} \Sigma_{i \in Gj}\ Y_i$, to perform the analysis.

The resulting general GDVL model, given $\overline{p}^{G_j}$, $j=1,\ldots,J$, and $x_i = (x_{i1},\ldots,\ x_{iK})'$ is specified as,

$$\overline{p}^{G_j} = \frac{1}{N^{G_j}} \Sigma_{i \in G_j} \frac{\exp(\beta' x_i)}{1 + \exp(\beta' x_i)} + \omega_j = \mu_t + \omega_j \tag{1}$$

where ω_j is a stochastic error drawn from either the normal or the beta distribution[1].

Assuming the normal distribution in Equation (1), the GDVL model leads to the following probabilistic specification for the j-th group of observations,

$$L_N\left(\beta, \sigma_\omega \middle| \overline{p}^{G_j}, x_i,\ (i \in G_j)\right) = -\frac{1}{2}\left[\ln\left(\sigma_\omega^2\right) + \ln(2\pi) + \frac{1}{\sigma_\omega^2}\left\{\overline{p}^{G_j} - \mu_t\right\}^2\right] \tag{2}$$

where $\mu_t = \frac{1}{N^{G_j}} S_{i \in G_j} \frac{\exp(\beta' x_i)}{1 + \exp(\beta' x_i)}$ and $\overline{p}^{G_j} \in (-\infty, +\infty)$. Analogously, assuming the beta distribution in Equation (1), the GDVLmodel leads to the following probabilistic specification for the j-th groups of observations[2] (see Romero, 2010),

$$L_B\left(\beta, \delta \middle| \overline{p}^{G_j},\ x_{i,}(i \in G_j)\right) = \frac{\Gamma(\delta)}{\Gamma(\delta\mu_t)\Gamma\left(\delta(1-\mu_t)\right)}\left(\overline{p}^{G_j}\right)^{\delta\mu_{t-1}}\left(1 - \overline{p}^{G_j}\right)^{\delta(1-\mu_t)-1} \tag{3}$$

where $\mu_t = 1\frac{1}{N^{G_j}} \Sigma_{i \in G_j} \frac{\exp(\beta' x_i)}{1 + \exp(\beta' x_i)}$ and $\overline{p}^{G_j} \in (0,1)$.

The likelihood functions in (2) and (3) are maximized using Matlab's nonlinearconstrainedminimization routine, allowing theestimates of the true underlying parameters β, σ_{ω}, and δ to be retrieved[3].

3. SIMULATION STRATEGY

Instead of simulating from Equation (1), we follow a probabilistic approach and simulate a logit specification directly fromconditional and marginal distributions. We then use the resulting simulated data to compute the needed averages for the dependent variable while keeping the individual information for the independent variables. This approach serves two purposes. For one, it allows for the direct comparison of the performance of the competing models without directly or indirectly giving a computational advantage to either; and two, it addresses the issue put forward by some researchers that simulating from and estimating (1) only attests to the ability of the model to estimate itself. Under the probabilistic simulation, we do not specify a priori the values of the parameters in (1) but rather we know the parameters of related specific distributions, from where the parameters in (1) are derived (see below).

The simulation strategy is described usingthe following methodology (see Bergtold, Spanos, Onukwugha, 2010 or Romero, 2012 for a more detailed derivation). Let $\{Y_i, i = 1,\ldots,N\}$ be a random variable distributed Bernoulli, with $E(Y_i) = p$ and $Var\ (Y_i) = p(1\text{-}p)$; For the moment, let $\{\boldsymbol{X}_i = (X_{1,i},\ldots,X_{K,i}),\ i=1,\ldots,N\}$ be a vector of K random variables of unspecified distribution but with a proper joint density function $f(\boldsymbol{X}_i, \boldsymbol{\Psi}_2)$. The related stochastic vector $\{Y_i, \boldsymbol{X}_i, i=1,\ldots,N\}$, whose joint density (probability) function takes the form $f(Y_1,\ldots,Y_N, \boldsymbol{X}_1,\ldots,\boldsymbol{X}_N;\varphi\}$ can be decomposed, following Bergtold *et al.* (2010), into

$$\frac{f(\mathbf{X}_i \mid Y_i = 1; \eta_1)\, f\,(Y_i = 1; p)}{f(\mathbf{X}_i \mid Y_i = 0; \eta_1)\, f\,(Y_i = 0; p)} = \frac{f(Y_i = 1 \mid \mathbf{X}_i\,; \psi_2)}{f(Y_i = 0 \mid \mathbf{X}_i\,; \psi_1)\, f\,(\mathbf{X}_i\,; \psi_2)} \tag{4}$$

Since $f(Y_i \mid \boldsymbol{X}_i; \psi_2)$ is Bernoulli distributed (see Chen and Liu, 1997; Spanos, 1999) with density function $f(Y_i \mid \boldsymbol{X}_i; \psi_1) = g\ (\boldsymbol{X}_i; \psi_1)^{Y_i}\,[1- g\ (\boldsymbol{X}_i; \psi_1)]^{1\text{-}Y_i}$, substituting this into equation (2) yields,

$$\frac{f(\mathbf{X}_i \mid Y_i = 1;\, \eta_1)}{f(\mathbf{X}_i \mid Y_i = 0;\, \eta_1)}\,\frac{\pi_1}{\pi_0} = \frac{g(X_i\,; \psi_1)}{1 - g\,(X_i\,; \psi_1)}, \tag{5}$$

where $\pi_j = p^j\,(1\text{-}p)^{1\text{-}j}$ for $j = 0{,}1$. Thus (see Kay and Little, 1987),

$$g(\mathbf{X}_i\,; \psi_1) = \frac{\pi_1 . f(\mathbf{X}_i \mid Y_i = 1; \eta_1)}{\pi_0 . f(\mathbf{X}_i \mid Y_i = 0; \eta_1) + \pi_1 . f(\mathbf{X}_i \mid Y_i = 1; \eta_1)} = \frac{\exp\{h(\mathbf{X}_i\,; \eta_1)\}}{1 + \exp\{h(\mathbf{X}_i\,; \eta_1)\}}, \tag{6}$$

where $h(\mathbf{X}_i\,; \eta_1) = \ln \dfrac{f(\mathbf{X}_i \mid Y_i = 1;\, \eta_1)}{f(\mathbf{X}_i \mid Y_i = 0; \eta_1)} + \kappa$, and $\kappa = \ln \pi_1 - \ln \pi_0$.

Hence, a proper statistical model in which the dependent variable is binary and the conditional relationship is Bernoulli naturally establishes the logistic cumulative density function as the transformation function and requires the use of logit specifications to model the statistical dependency of Y_i on $\boldsymbol{X}_i$. This result is tailored-made for our simulation analysis, for it implies that under perfect information, the researcher would conduct a thoroughlogistic analysis of the ungrouped data. Of course, this does not preclude us from adopting alternative simulation models, like the probit, it simply provides a window of direct comparison between the normal and the beta distribution models.

Given that the functional form of the index function $h(.)$ depends entirely on the conditional distribution of $\boldsymbol{X}_i$ given the two outcomes of Y_i; and to keep the analysis relatively simple, we simulated the data assuming conditional normal distributions for two uncorrelated[4] independent variables X_{1i} and X_{2i}. Specifying the conditional distributions of X_{1i} and X_{2i} given the two states of Y_i yields,

$$\left.\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\right|_{Y=j} \sim N\left[\begin{pmatrix} \mu_{x_1,Y=j} \\ \mu_{x2,Y=j} \end{pmatrix}, \begin{pmatrix} \sigma^2_{x_{1,Y=j}} & 0 \\ 0 & \sigma^2_{x_{2,Y=j}} \end{pmatrix}\right]$$

Thus, the previous probabilistic assumptions will give birth to the following unrestricted logit specification:

$$Y_i = \frac{\exp(\beta_0 + \beta_1 x_{1i} + \beta_1 x_{2i})}{1 + \exp(\beta_0 + \beta_1 x_{1i} + \beta_1 x_{2i})} + \omega_i \tag{7}$$

where the coefficients are directly related to the distributional assumptions about Y_i and $\boldsymbol{X}_i$ (see Appendix).

As mentioned above, a significant benefit of using the previous approach for developing the unconstrained logit specification is that it provides a mechanism for randomly generating the vector stochastic process, $\{(Y_i, \boldsymbol{X}_i),\ i=1,\ldots N\}$ using the relationship given by equation (6) for simulations involving binary choice models. The complete data simulation and estimation processescomprise performing the following steps:

Step 1: Assign p and generate a realization of the stochastic process $\{Y_i, i=1,\ldots, N\}$ using a binomial random number generator.

Step 2: Using the regression function associated with $f(\boldsymbol{X}_i \mid Y_i = j; \eta_1)$, generate a realization of the vector stochastic process, $\{\boldsymbol{X}_i,\ i=1,\ldots,N\}$ using appropriate random number generators for specified values of η_1.

Step 3: Calculate $\overline{p}^{Gj} = \frac{1}{N^{G_j}} \Sigma_{i\in Gj}\ Y_i$ for each of the J groups of size N^{Gj}.

Step 4: Estimate models (2) and (3).

Step 5: Repeat R times.

The simulation and estimation strategy focuses on the competing models' ability to accurately estimate the unknown parameters when the group sizes vary, when the average response within groups vary, and the when the researcher has a fixed numberof members in each group. Since a researcher's ability to obtain data in a geographic region may be limited, we will make the assumption that the investigator has access to state information and that she can examine three grouping sizes for any state (J=15, 50, 100), which can be interpreted as the number of counties surveyed; and three different sets of individual agents surveyed in each country, (N^{G_j}= 2, 5, 10). Additionally, we allow for three different sample average response probabilities (p = 15%, 30%, 50%). These assumptions are done simply to put the simulation analysis in an applied econometrics context. The narrative can equally apply to other domains.

4. SIMULATION RESULTS

Table 1-3 compare the performance of the models at estimating the true underlying parameters[5]. They include the true parameters, the sample mean of the estimated parameters, and the sample standard errors. We simulated and estimated three specifications, an unrestricted logit, a normal-logit, and a beta-logit for a total of R=1,000 times. To further contrast their accuracy, the tables also show the relative difference between the true parameter values and the parameter estimates as a percentage, calculated as $RD = 100 \times \left|\frac{\beta - \hat{\beta}}{\beta}\right|$, where β is the true underlying parameter and $\hat{\beta}$is the estimated coefficient. Since the lower the ARP the higher the variance of a Bernoullidistributed variable, we expect the results to increase in accuracy when the ARP approaches 50%, when the group size increases, and when the number of groups increases.

Table 1 shows the results for J=15; with group sizes 2, 5, and 10; and *ARP* of 15%, 30%, and 50%. Although both models do a relatively poor job at estimating the true values at this scale, the beta model seems to almost always outperform the normal distribution in this scenario when using the *RD* as the measure of reliability. Nevertheless, it seems that the relatively lower amount of information in this scenario would make either estimates too volatile to be reliably estimated. The only model which accuracy consistently increases with the group size and the ARP is the unrestricted logit, but these results should not be surprising.

In Table 2, we increased the number of groups to 50, significantly increasing the amount of statistical information for the models to estimate the true parameters. Once again, when the variance of the dependent variable is highest (*ARP* = 15%), the beta model appears to outperform the normal model when using the *RD* as the accuracy

Table 1
Parameter Estimates and Relative Difference from True Parameters as a Percentage when *J*=15 (Standard Errors in Parentheses)

		N^{Gj}=2			*N^{Gj}=5*			*N^{Gj}=10*		
ARP	*True Values*	*Logit*	*Normal*	*Beta*	*Logit*	*Normal*	*Beta*	*Logit*	*Normal*	*Beta*
0.15	β_0= -11.24	-49.962	-225.4	-4.8629	-14.212	-269.37	-8.4205	-12.383	-225.68	-94.989
		(154.19)	(525.63)	(1.68)	(8.39)	(542.71)	(7.55)	(3.13)	(504.52)	(280.25)
	RD	365.17	1938.5	56.715	38.179	2318.4	39.315	19.592	1932.1	782.04
	β_1= 1.00	4.2973	23.36	0.36221	1.2508	22.381	0.76713	1.096	18.823	13.202
		(15.62)	(90.60)	(0.41)	(1.01)	(68.52)	(1.64)	(0.46)	(52.83)	(55.08)
	RD	407.93	2808.7	67.526	59.042	2683.8	66.041	33.835	1974.8	1466.6
	β_2= 4.00	19.964	74.128	1.5876	5.1268	101.52	2.873	4.4239	85.594	24.372
		(65.19)	(150.31)	(0.83)	(2.94)	(214.25)	(2.66)	(1.16)	(194.39)	(98.69)
	RD	422.37	1892.2	60.375	42.105	2465	42.006	21.568	2069.6	752.81
0.3	β_0= -10.34	-21.515	-238.65	-6.1041	-12.757	-207.15	-37.031	-11.053	-152.23	-265.81
		(63.47)	(434.76)	(2.17)	(14.97)	(412.19)	(131.28)	(2.18)	(353.57)	(419.08)
	RD	121.5	2224.7	42.573	35.089	1928.3	295.26	15.94	1396.8	2487.9
	β_1= 1.00	2.0664	24.075	0.58953	1.2199	19.659	3.8732	1.0682	16.116	32.784
		(6.76)	(59.68)	(0.52)	(1.96)	(51.91)	(22.87)	(0.34)	(45.72)	(71.79)
	RD	153.26	2592.3	54.149	52.104	2128.1	435.38	26.761	1641.8	3467.4
	β_2= 4.00	8.7708	92.056	2.3284	4.9915	82.704	14.172	4.2696	56.78	89.366
		(29.57)	(171.89)	(1.02)	(5.19)	(167.00)	(64.04)	(0.87)	(133.22)	(166.47)
	RD	135.9	2231.1	44.23	37.009	1995.9	351.75	16.65	1348.8	2396.9
0.5	β_0= -9.50	-14.22	-203.23	-6.3652	-11.087	-162.78	-64.679	-10.09	-137.69	-161.43
		(13.49)	(349.74)	(3.05)	(4.54)	(330.52)	(188.54)	(1.94)	(301.33)	(313.54)
	RD	65.812	2061	39.686	28.873	1640.3	610.56	15.663	1374.2	1622.1
	β_1= 1.00	1.4731	22.792	0.66432	1.1535	17.505	7.613	1.064	15.315	16.603
		(1.65)	(56.54)	(0.61)	(0.64)	(43.86)	(28.71)	(0.31)	(41.12)	(43.62)
	RD	91.78	2551.5	54.599	43.253	1804.6	766.65	24.618	1571.6	1828.9
	β_2= 4.00	6.0334	82.293	2.6958	4.7032	67.796	25.771	4.2392	55.912	68.582
		(5.84)	(139.37)	(1.32)	(1.94)	(137.92)	(81.72)	(0.80)	(123.01)	(136.09)
	RD	66.17	1985.1	39.796	29.449	1622.5	614.64	15.645	1324.3	1665.4

metric. The sample statistics of each of the estimators is also closer in absolute value when using the beta model than when using the normal model. The accuracy of the parameters estimated with the beta model consistently increases when the *ARP* approaches 50%. Specifically, then the group size if *5* and *ARP* = 30%, the beta distribution seems spot on when estimating the true parameters. When the amount for information keeps increasing (N^{G_j} = 10) the performance of the beta model decreases while the accuracy of the normal model appears to increase at every *ARP* level.

Table 2
Parameter Estimates and Relative Difference from True Parameters as a Percentage when *J*=50 (Standard Errors in Parentheses)

		N^{Gj}=2			*N^{Gj}=5*			*N^{Gj}=10*		
ARP	*True Values*	*Logit*	*Normal*	*Beta*	*Logit*	*Normal*	*Beta*	*Logit*	*Normal*	*Beta*
0.15	β_0= -11.24	-13.41	-209.48	-4.8277	-11.824	-48.453	-7.7714	-11.469	-21.122	-10.8
		(5.28)	(481.72)	(0.84)	(2.00)	(221.45)	(1.31)	(1.21)	(86.71)	(7.67)
	RD	29.493	1776.3	57.029	13.537	346.58	31.026	8.4434	103.49	20.223
	β_1= 1.00	1.2108	19.642	0.38136	1.0449	4.4107	0.65554	1.0229	1.8983	1.168
		(0.71)	(53.34)	(0.19)	(0.31)	(22.14)	(0.28)	(0.21)	(8.39)	(6.32)
	RD	48.072	1952.8	61.913	24.299	373.58	38.228	16.455	118.66	57.295
	β_2= 4.00	4.7782	73.465	1.504	4.2283	17.227	2.6208	4.0815	7.6427	3.0873
		(1.97)	(168.18)	(0.39)	(0.78)	(77.91)	(0.58)	(0.48)	(33.15)	(18.44)
	RD	31.758	1751.1	62.401	14.739	348.08	34.864	9.3641	109.39	34.415
0.3	β_0= -10.34	-11.389	-87.746	-5.5426	-10.754	-24.449	-9.8549	-10.516	-23.33	-151.99
		(2.92)	(275.12)	(1.06)	(1.56)	(105.02)	(2.30)	(0.98)	(97.06)	(309.67)
	RD	21.324	761.75	46.435	11.516	151.98	17.417	7.4122	141.97	1381.2
	β_1= 1.00	1.1037	8.4473	0.51994	1.0354	2.4471	0.94316	1.0148	2.1956	18.077
		(0.47)	(31.01)	(0.23)	(0.26)	(11.36)	(0.44)	(0.17)	(9.54)	(46.51)
	RD	35.102	799.91	48.602	19.728	174.28	33.294	13.32	149.48	1780.2
	β_2= 4.00	4.3931	33.91	2.103	4.1635	9.3624	3.7609	4.0688	9.2889	53.209
		(1.20)	(106.26)	(0.48)	(0.60)	(39.86)	(1.03)	(0.41)	(39.64)	(116.94)
	RD	22.738	763.03	47.476	11.734	150.95	19.891	7.9468	150.38	1286.2
0.5	β_0= -9.50	-10.445	-56.307	-5.8253	-9.8299	-20.621	-15.237	-9.6483	-17.533	-28.957
		(2.68)	(189.92)	(1.19)	(1.39)	(82.03)	(48.38)	(0.88)	(67.51)	(108.35)
	RD	20.907	508.53	38.877	11.281	134.61	10.905	7.396	103.23	222.98
	β_1= 1.00	1.1038	5.9452	0.61226	1.0342	2.2163	1.5443	1.0154	1.8774	3.1228
		(0.44)	(22.45)	(0.26)	(0.25)	(10.54)	(4.15)	(0.15)	(7.84)	(13.33)
	RD	32.828	535.31	40.492	18.927	150.86	133.89	12.104	117.24	241.23
	β_2= 4.00	4.3725	23.689	2.448	4.1368	8.647	6.5277	4.0595	7.2957	12.026
		(1.11)	(79.15)	(0.50)	(0.56)	(34.05)	(22.90)	(0.40)	(27.79)	(45.87)
	RD	20.499	508.36	38.987	10.905	84.649	78.176	7.8586	101.92	220.14

Table 3 shows the simulation scenario that contains the largest amount of information. This is achieved when the number of groups equals 100. The same regularity found in Tables 1 and 2 is found, namely, when the size of the group is relative small, the beta model outperforms the normal model, when using *RD* for comparison. Also, the advantage of the beta model dwindles when the size of the group increases or when the variance of the dependent variable decreases.

Table 3
Parameter Estimates and Relative Difference from True Parameters as a Percentage when *J*=100 (Standard Errors in Parentheses)

		N^{Gj}=2			*N^{Gj}=5*			*N^{Gj}=10*		
ARP	*True Values*	*Logit*	*Normal*	*Beta*	*Logit*	*Normal*	*Beta*	*Logit*	*Normal*	*Beta*
0.15	β_0= -11.24	-12.139	-58.679	-4.6361	-11.469	-13.914	-7.7517	-11.399	-12.216	-10.841
		(2.44)	(238.29)	(0.58)	(1.21)	(39.71)	(0.92)	(0.90)	(6.28)	(1.62)
	RD	16.289	433.09	58.733	8.4434	35.992	31.002	6.331	20.604	11.83
	β_1= 1.00	1.0813	5.0732	0.36074	1.0229	1.3164	0.64872	1.0199	1.1067	0.96846
		(0.37)	(22.23)	(0.12)	(0.21)	(5.74)	(0.21)	(0.14)	(0.59)	(0.30)
	RD	28.362	437.3	63.926	16.455	56.292	36.114	11.47	31.385	23.863
	β_2= 4.00	4.3366	21.68	1.4347	4.0815	4.8582	2.591	4.0489	4.3334	3.7921
		(0.97)	(89.51)	(0.27)	(0.48)	(10.99)	(0.38)	(0.36)	(2.68)	(0.67)
	RD	18.192	454.86	64.133	9.3641	35.544	35.242	7.019	22.673	13.83
0.3	β_0= -10.34	-11.046	-24.251	-5.4404	-10.516	-11.663	-10.027	-10.474	-12.34	-41.197
		(1.83)	(106.42)	(0.70)	(0.98)	(5.12)	(1.60)	(0.72)	(30.33)	(147.13)
	RD	14.019	145.2	47.422	7.4122	25.417	12.55	5.5958	31.35	305.22
	β_1= 1.00	1.0786	2.3824	0.52282	1.0148	1.1185	0.95747	1.0177	1.19	4.7964
		(0.30)	(11.18)	(0.16)	(0.17)	(0.54)	(0.31)	(0.12)	(2.10)	(20.23)
	RD	23.144	158.75	47.759	13.32	34.059	24.304	9.6662	37.693	395.08
	β_2= 4.00	4.254	9.4627	2.0407	4.0688	4.5456	3.8054	4.0348	4.7937	14.684
		(0.75)	(42.18)	(0.30)	(0.41)	(2.22)	(0.70)	(0.29)	(13.78)	(53.98)
	RD	14.655	148.57	48.983	7.9468	27.355	14.374	5.7831	33.612	276.08
0.5	β_0= -9.50	-9.8733	-17.128	-5.6741	-9.6483	-9.0398	-11.402	-9.5912	-12.134	-16.342
		(1.55)	(69.01)	(0.76)	(0.88)	(33.51)	(2.98)	(0.64)	(41.91)	(59.71)
	RD	12.916	92.565	40.273	7.396	55.307	27.306	5.4057	54.635	84.348
	β_1= 1.00	1.045	1.9234	0.59734	1.0154	0.39564	1.1994	1.014	0.94716	1.7027
		(0.27)	(8.68)	(0.18)	(0.15)	(12.50)	(0.47)	(0.12)	(8.34)	(6.59)
	RD	20.929	115.03	40.594	12.104	163.3	37.828	9.2806	98.703	89.716
	β_2= 4.00	4.145	6.9649	2.3894	4.0595	6.4095	4.7979	4.0268	5.8853	6.9337
		(0.66)	(26.52)	(0.33)	(0.40)	(22.03)	(1.30)	(0.27)	(21.17)	(25.88)
	RD	12.858	86.687	40.267	7.8586	86.685	27.669	5.3683	66.032	86.083

5. CONCLUSION

In applied work, the need for combining data from different sources is often times unavoidable, either for availability, practicality, or legality reasons. This is particularly true in agricultural and/or environmental research, where the information is often available on a national or regional basis. One of these situations arises when the researcher has only access to aggregated information about the dependent variable (where the choices of *n*-individuals are averaged across groups) but possess individual information for the independent variables. These models are known in the literature as Grouped Dependent Logit Models, first introduced by Kurkalova and Rabotyagov (2005) as an alternative to averaging out the independent variables when the dependent variable is grouped or to disaggregating the dependent variable when information on individuals is available for the independent variables.

In this paper, we assessed the performance of two competing specifications for estimating this kind of models, a beta-logit model and a normal-logit model. We did so by using a sample average logit-link function to associate the averaged dependent variable to the individual independent variables but assuming two different distributions for the error term. We also proposed a methodology for simulating the data that does not use the proposed specification as the data generating mechanism but instead the data is generated using conditional and marginal distributions from which the true specifications are derived. This method of simulation allows for the direct comparison of the models without giving a computational advantage to either one of them.

The results of the simulations fall in line with standard statistical theory. The accuracy and reliability of the estimation increases with the amount of information but decreases with the dispersion of the dependent variable. Using three different grouping sizes (2,5,10) and three different number of groups (15,50,100), we propose the use of the beta distribution when estimating this kind of models given that the accuracy of the estimators is no worse that the estimation provided by the model using the normal distribution.

NOTES

1. This error term can be further decomposed as $w_j = \in_j + \alpha_j$. In this case, $\in_j$ and α_j refer to the errors related to the experts' opinion and model misspecification, respectively (see Kurkalova and Wade, 2010).
2. The use of the beta distribution seems tailor-made for our purposes, for it deals primarily with data with lower and upper bounds, such as rates and proportions.
3. In situations where the dependent variable takes the values 0 or 1, we rescaled the data following the procedure described in Smithson and Verkuilen, 2006.
4. The use of uncorrelated uniformly distributed regressors is standard in parameter simulation and estimation. In this situation, adding correlated regressors did not changes the resultsqualitatively.
5. The derivation of the true underlying parameters is described in the Appendix.

REFERENCES

Bergtold, Jason; Aris Spanos; Ebere Onukwugha (2010), "Bernoulli Regression Models: Revisiting the Specification of Statistical Models with Binary Dependent Variables," *Journal of Choice Modelling*, **3**(2), pp. 1-28.

Chen, Sean X. and Jun S. Liu (1997), "Statistical Applications of the Poisson-Binomial and Conditional Bernoulli Distributions," *Statistica Sinica*, 7, pp. 875-892.

Kay, R. and S. Little (1987), "Transformations of the Explanatory Variables in the Logistic Regression Model for Binary Data," *Biometrika*, **74**(3), pp. 495-501.

Kurkalova, Lyubov A. and Sergey Rabotyagov (2005), "Estimation of a Binary Choice Model with Grouped Choicedata,"*Economics Letters,* **90**(2), pp. 170-175.

Kurkalova, Lyubov A. and Tara Wade (2010), "Aggregated Choice Data and Logit Models," Working Paper, North Carolina A & T State University.

Lambert, Dayton; Glenn D. Schaible; Robert Johansson; Stan Daberkow (2006), *Working-land Conservation Structures: Evidence on Program and On-program Participants*, American Agricultural Economics Association. Annual Meeting, July 23-26, Long Beach, CA.

Lambert, Dayton; Glenn D. Schaible; Robert Johansson; Arwind U. Vasavada (2007), "The Value of Integrated CEAPARMS Survey Data in Conservation Program Analysis," *Journal of Soil and Water Conservation,* **62**(1), pp. 1-10.

Miller, Douglas J. and Andrew Plantinga (1999), "Modeling Land Use Decisions with Aggregate Data," *American Journal of Economics,* **81**(1), pp. 180-194.

Romero, Alfredo A. (2010), *Statistical Adequacy and Reliability of Inference in Regression-like Models*, Doctoral Dissertation, Virginia Polytechnic Institute and State University.

Romero, Alfredo A. (2012), Where Do Moderation Terms Come From in Binary Choice Models? Working Paper. North Carolina A&T State University.

Soule Meredith J., Abebayehu Tegene; Keith D. Wiebe (2000), "Land Tenure and the Adoption of Conservation Practices," *American Journal of Agricultural Economics,* **82**(4), pp. 993-1005.

Spanos, Aris (1999), *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*. Cambridge, UK: Cambridge University Press.

Smithson, Michael and Jay Verkuilen (2006), "A Better Lemon Squeezer? Maximum-Likelihood Regression With Beta-Distributed Dependent Variables," *Psychological Methods,* **11**(1), pp. 54-71.

Steel, David G. and D. Holt (1996), "Analyzing and Adjusting Aggregation Effects: The Ecological Fallacy," *International Statistical Review,* **64**(1), pp. 39-60.

Train, K. (2009), *Discrete Choice Methods with Simulation* (2nd ed). New York, NY: Cambridge University Press.

APPENDIX

A.1. Derivation of the True Parameters in Equation (7)

Let $\{Y_i, i=1,\ldots,N\}$ be a random variable distributed Bernoulli, with $E(Y_i)=p$ and $Var(Y_i)=p\ (1\text{-}p)$.

Let also $\left.\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\right|_{Y=j} \sim N\left[\begin{pmatrix} \mu_{x_1,j} \\ \mu_{x2,j} \end{pmatrix}, \begin{pmatrix} \sigma^2_{x_{1,j}} & 0 \\ 0 & \sigma^2_{x_{2,j}} \end{pmatrix}\right]$. Substituting this into equation (6), we get,

$$\beta_0 = \ln\frac{p}{1-p} + \ln\frac{\sigma_{x_1,0}}{\sigma_{x_1,1}} + \ln\frac{\sigma_{x_2,0}}{\sigma_{x_{2,1}}} + \frac{\mu^2_{x_{1,0}}}{2\sigma^2_{x_{1,0}}} - \frac{\mu^2_{x_{1,1}}}{2\sigma^2_{x_{1,1}}} + \frac{\mu^2_{x_{2,0}}}{2\sigma^2_{x_{2,0}}} - \frac{\mu^2_{x_{2,1}}}{2\sigma^2_{x_{2,1}}}$$

$$\beta_1 = \frac{\mu_{x_{1,1}}}{\sigma^2_{x_{1,1}}} - \frac{\mu_{x_{1,0}}}{\sigma^2_{x\,1,0}}$$

$$\beta_2 = \frac{\mu_{x_{2,1}}}{\sigma^2_{x_{2,1}}} - \frac{\mu_{x_{2,0}}}{\sigma^2_{x_{2,0}}}$$