Conference on Applied Statistics in Agriculture 2013 - 25th Annual Conference Proceedings

A SIMULATION STUDY OF THE SMALL SAMPLE PROPERTIES OF LIKELIHOOD BASED INFERENCE FOR THE BETA DISTRIBUTION

Kevin Thompson

Edward Gbur

Recommended Citation

Thompson, Kevin and Gbur, Edward (2013). "A SIMULATION STUDY OF THE SMALL SAMPLE PROPERTIES OF LIKELIHOOD BASED INFERENCE FOR THE BETA DISTRIBUTION," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1019

A SIMULATION STUDY OF THE SMALL SAMPLE PROPERTIES OF LIKELIHOOD BASED INFERENCE FOR THE BETA DISTRIBUTION

Kevin Thompson and Edward Gbur
Agricultural Statistics Laboratory, Arkansas Agricultural Experiment Station
University of Arkansas, Fayetteville AR 72701

Abstract

Researchers often collect proportion data that cannot be interpreted as arising from a set of Bernoulli trials. Analyses based on the beta distribution may be appropriate for such data. The SAS® GLIMMIX procedure provides a tool for these analyses using a likelihood based approach in the context of generalized linear mixed models. Since the t and F-distribution based inference employed in this approach relies on asymptotic properties, it is important to understand the sample sizes required to obtain reasonable approximate answers to inference questions. In addition, the complexity of the likelihood functions can lead to numerical issues for optimization algorithms that may or may not be related to sample size issues. This simulation study is based on a simple intercept-only model for known beta distributed responses. Convergence and estimation issues are investigated over a range of beta distributions and sample sizes.

Keywords Generalized linear mixed model, Beta distribution, GLIMMIX, Simulation

1. Introduction

Proportions that are measured on a continuum are often modeled by a beta distribution because of the wide range of possible shapes for its pdf. Our objective in this paper was to study the small sample likelihood based inference properties for the beta in the context of the one sample problem using PROC GLIMMIX in SAS®. This study represents the first step toward examining small sample inference for the beta in the context of generalized linear mixed models (GLMM).

2. Beta Distribution

The standard form of the pdf of a beta distribution is given by

$$f(y \mid \alpha, \beta) = y^{(\alpha - 1)}(1 - y)^{(\beta - 1)}/B \qquad \text{for } 0 < y < 1,$$

where B is the beta function with parameters $\alpha > 0$ and $\beta > 0$. For a GLMM, the alternative parameterization

$$f(y \mid \mu, \varphi) = y^{(\mu\varphi - 1)}(1 - y)^{(\varphi(\mu - 1) - 1)}/B,$$

with parameters $0 < \mu < 1$ and $\varphi > 0$ is often used, where $\mu = \alpha/(\alpha + \beta)$ and $\varphi = \alpha + \beta$. For this parameterization we have

$E(Y) = \mu$ and $Var(Y) = \mu(1 - \mu)/(1 + \varphi)$.

The parameter space for the beta can be divided into regions defined by α and β that determine the general shape of its pdf as shown in the Figure 1. Shapes of distributions with $\mu > 0.5$ are mirror images of the corresponding distributions having $\mu < 0.5$. Distributions with $\mu = 0.5$ are symmetric regardless of value of φ. For $\varphi > 10$, distributions have the same general shapes as those shown in Figure 1 for $\varphi = 10$.

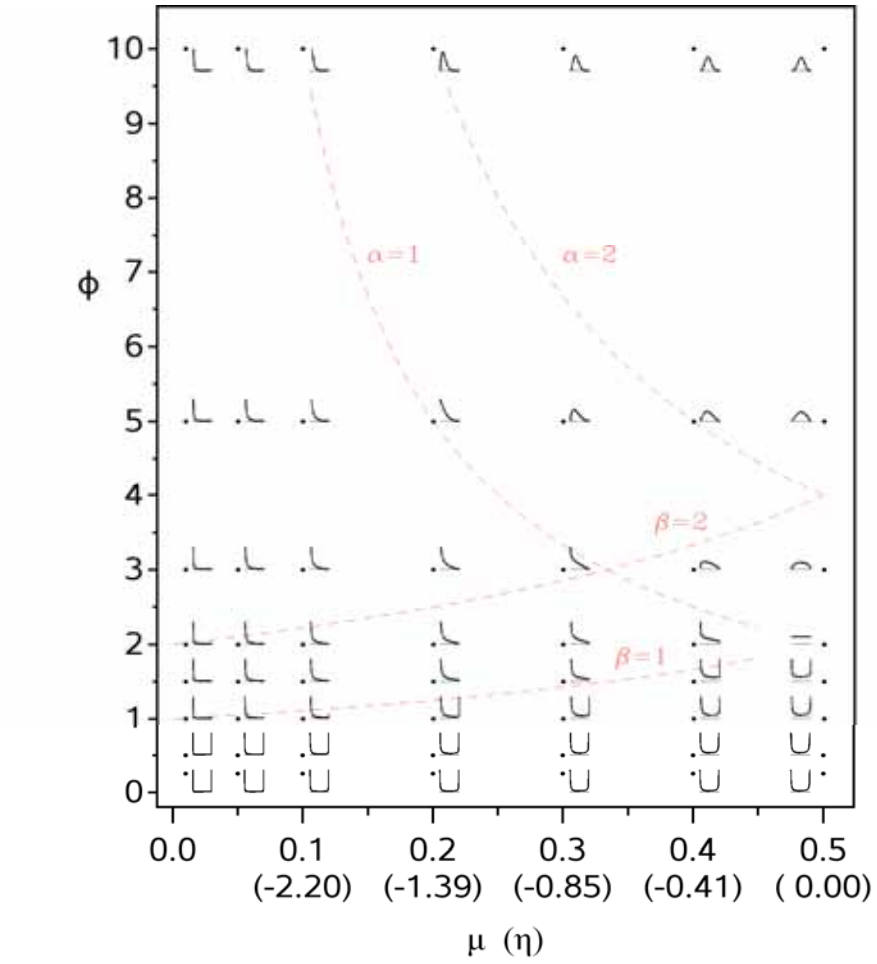

Figure 1. General shapes of the pdf for beta distributions when $\mu \leq 0.5$ and $\varphi \leq 10$.

3. Simulation and Analysis Details

The simulation was performed on a rectangular grid on (μ, φ) parameter space with

$\mu = 0.1, 0.2, 0.3, 0.4, 0.5$
$\varphi = 0.5, 1, 1.5, 2, 3, 5, 10, 15, 25, 50, 75, 100$

and sample sizes of N = 5, 10, 15, 20, 50, 100. Two thousand samples were generated for each (μ, φ, N) combination using the RAND("Beta") function in SAS. The software used was 32 bit SAS® for Windows, Version 9.3 (TS1M2, Analytics version 12.1).

Each sample was analyzed as a generalized linear model (GLM) using PROC GLIMMIX with a logit link function η = logit(μ) . Numerical techniques used were pseudo-likelihood (RSPL) and Laplace. The MODEL statement contained one factor having only one level. The no intercept (NOINT) and SOLUTION options provided estimates for η and φ. The LSMEANS statement provided 90%, 95% and 99% confidence intervals for η based on the t-distribution. Estimates and confidence intervals for μ were obtained by back-transforming using the ILINK option on the LSMEANS statement.

4. Convergence Issues

Convergence issues were divided into four categories:

(1) Convergence was declared and all estimates and standard errors were calculated,
(2) Convergence was declared but SE(φ-hat) was missing or zero,
(3) Convergence was declared but SE(η-hat) was missing or zero,
(4) Convergence was not attained.

Convergence results for RSLP are reported below. Results for Laplace were similar.

For sufficiently large μ and φ combinations, non-convergence was not an issue. For all $\varphi \geq 3$ and all (μ, N) combinations, at least 92.5% of the samples were in category 1. For $\varphi < 3$ and small values of μ, the percentage of samples in category 1 decreased dramatically to as small as 16% as N increased. Except for $\varphi = 0.5$ and $\mu = 0.1$, most of the non-category 1 samples were in category 2. For $\varphi = 0.5$ and $\mu = 0.1$, between 15% and 19% of the samples were in category 3 and inference on μ would not be possible for these samples.

In the results that follow, only samples from categories 1 and 2 combined were used in the construction of the tables.

5. Confidence Intervals for μ

Trends for empirical confidence interval coverage levels for nominal 90%, 95% and 99% confidence levels were similar. Only results for 95% are presented here. The results for RSPL and Laplace were similar as well and only numerical results for RSPL will be presented.

For all $\varphi \geq 3$ and all (μ, N) combinations, the empirical confidence interval coverage for nominal 95% confidence intervals for μ ranged from 92.4% to 96.1% for both RSPL and Laplace. Empirical coverage levels for $\varphi < 3$ using RSPL are presented in Table 1 with coverage levels less than 90% printed in red.

Table 1. Empirical coverage levels for nominal 95% confidence intervals for μ based on RSPL for selected (μ, φ, N) combinations. Trends for Laplace are similar.

φ	N	μ = 0.1	μ = 0.2	μ = 0.3	μ = 0.4	μ = 0.5
2.0	5	91.0	93.4	93.7	93.7	93.9
	10	91.4	94.5	94.0	93.9	93.9
	20	89.7	93.8	95.0	95.1	93.9
	50	83.9	94.7	94.3	95.8	95.1
	100	73.1	94.6	94.6	95.2	95.3
1.5	5	88.9	93.7	95.0	93.4	94.1
	10	85.6	92.9	94.9	95.8	94.2
	20	77.3	93.6	94.5	94.6	94.9
	50	56.1	94.8	95.7	94.7	94.8
	100	34.6	94.3	95.0	95.4	96.1
1.0	5	78.4	93.0	94.1	94.7	94.5
	10	65.8	92.8	93.9	94.7	95.2
	20	44.6	91.8	94.8	94.7	94.9
	50	18.2	85.4	94.3	93.9	95.5
	100	10.2	77.1	93.7	95.7	95.2
0.5	5	63.6	85.6	92.2	95.1	95.6
	10	38.8	70.4	92.4	94.9	94.8
	20	28.5	50.3	85.3	94.7	95.3
	50	22.4	17.8	67.0	90.3	94.5
	100	22.9	9.6	43.2	84.5	92.5

From Table 1 the following trends emerge.

- For each (φ, N) combination, as μ increases to 0.5 the empirical coverage levels tend to increase and then stabilize; i.e., as the shape of the distribution changes from highly skewed to symmetric, the coverage level tends to increase.

- For fixed N and small values of μ, as φ decreases, the coverage level decreases; i.e., as the distribution changes from a reverse J shape to a U shape, the coverage decreases.

- For each (μ, φ) combination where the empirical coverage level is consistently much smaller than the nominal level (entries in red), as the sample size N increases, the coverage levels tend to decrease which is contrary to what would be expected.

Issues that may contribute to poor empirical coverage levels include:
- Bias in the estimates of μ.
- Problems estimating φ, including potential effects of highly skewed distributions.
- Estimates of μ having large biases that are coupled with large overestimates of φ to produce narrow confidence intervals not centered near μ.

6. Estimation of μ and its effect on empirical coverage levels

The means of the estimates of μ obtained by back-transformation for both RSLP and Laplace were calculated. The results are presented in Table 2 for selected (μ, φ, N) combinations using RSPL. Entries in red correspond to empirical coverage levels for a nominal 95% level that are less than 90% in Table 1. The results for Laplace are similar.

For $\varphi \geq 3$ and all (μ, N) combinations, the means of the estimates were within 0.005 of μ for both RSPL and Laplace with most of the biases less than or equal to 0.001. For $\varphi < 3$ in Table 2, the means tend to be biased downward but trends or lack thereof appear to depend on the (μ, φ) combination. The largest biases correspond to empirical coverages in Table 1 that were well below the nominal 95% level.

The means of the lengths of the confidence intervals for μ obtained by back-transformation for both RSLP and Laplace were calculated. The means for RSPL are presented in Table 3 for selected (μ, φ, N) combinations, again with entries in red corresponding to poor empirical coverage levels in Table 1. The trends for Laplace are similar.

The following conclusions can be drawn from Table 3.
- For each (μ, φ) combination, as N increases, the mean confidence interval width for μ decreases. This would be expected and provides a potential explanation for decreasing patterns of poor empirical coverage levels in Table 1 despite the relatively small biases in the mean estimates of μ in Table 2.

- For each (μ, N) combination except N = 100, as φ decreases the mean width increases. Again this might be expected since $\text{Var}(Y) = \mu(1 - \mu)/(1 + \varphi)$. The exception for N = 100 may arise from samples which have more observations in the upper tail of the distribution that are far from μ.

- For each (φ, N) combination, as μ increases toward $\mu = 0.5$, the mean width increases; i.e., as the shape of the distribution changes from highly skewed to symmetric the mean width increases.

Table 2. Means of the estimates of μ based on RSPL obtained by back-transformation for selected (μ, φ, N) combinations. Trends for Laplace are similar.

φ	N	μ=0.1	μ=0.2	μ=0.3	μ=0.4	μ=0.5
2.0	5	0.095	0.195	0.296	0.394	0.504
	10	0.095	0.193	0.297	0.399	0.503
	20	0.097	0.198	0.299	0.402	0.500
	50	0.098	0.198	0.299	0.400	0.499
	100	0.099	0.200	0.300	0.399	0.499
1.5	5	0.094	0.191	0.300	0.399	0.502
	10	0.095	0.192	0.295	0.401	0.499
	20	0.097	0.196	0.300	0.402	0.500
	50	0.096	0.199	0.300	0.400	0.499
	100	0.100	0.200	0.299	0.399	0.501
1.0	5	0.096	0.191	0.287	0.391	0.502
	10	0.094	0.192	0.289	0.401	0.502
	20	0.087	0.195	0.297	0.398	0.500
	50	0.091	0.202	0.300	0.400	0.500
	100	0.119	0.203	0.302	0.399	0.500
0.5	5	0.120	0.189	0.289	0.393	0.501
	10	0.096	0.182	0.298	0.391	0.505
	20	0.085	0.167	0.298	0.398	0.503
	50	0.086	0.136	0.308	0.408	0.500
	100	0.097	0.138	0.319	0.416	0.501

Table 3. Mean confidence interval widths of the back-transformed nominal 95% confidence intervals for μ using RSPL for selected (μ, φ, N) combinations. Trends for Laplace are similar.

φ	N	μ=0.10	μ=0.20	μ=0.30	μ=0.40	μ=0.50
2.0	5	0.387	0.444	0.484	0.506	0.509
	10	0.226	0.286	0.321	0.341	0.348
	20	0.146	0.198	0.224	0.239	0.243
	50	0.085	0.123	0.141	0.150	0.153
	100	0.054	0.087	0.099	0.106	0.108
1.5	5	0.412	0.482	0.523	0.539	0.548
	10	0.231	0.305	0.345	0.368	0.372
	20	0.141	0.212	0.242	0.256	0.261
	50	0.067	0.132	0.151	0.161	0.164
	100	0.033	0.093	0.106	0.114	0.116
1.0	5	0.407	0.516	0.556	0.578	0.589
	10	0.214	0.330	0.370	0.395	0.404
	20	0.104	0.223	0.259	0.276	0.282
	50	0.033	0.135	0.163	0.174	0.177
	100	0.027	0.090	0.115	0.122	0.125
0.5	5	0.420	0.532	0.595	0.628	0.636
	10	0.188	0.305	0.399	0.426	0.434
	20	0.116	0.164	0.266	0.295	0.303
	50	0.067	0.055	0.150	0.184	0.190
	100	0.057	0.029	0.090	0.129	0.132

7. Estimation of φ and its effect on empirical coverage levels

The means of the estimates of φ for both RSLP and Laplace were calculated. The results are presented in Table 4 for selected (μ, φ, N) combinations using RSPL. Entries in red correspond to empirical coverage levels for a nominal 95% level that are less than 90% in Table 1. The results for Laplace are similar.

For $\varphi \geq 3$ and all μ, the means of the estimates of φ were biased upward with the bias decreasing as N increased. For N = 100, the biases in the means were small. For $\varphi < 3$ in Table 4, the behavior was similar to that for $\varphi \geq 3$ for those (μ, φ) combinations for the empirical coverage levels close to the nominal 95% level. For the remaining (μ, φ) combinations, the means were extremely large and biased upward. Samples producing very large estimates of φ would generate very narrow confidence intervals.

The medians of the estimates of φ were also calculated (data not shown). They were biased upward as well, indicating that φ was overestimated in more than 50% of the samples. The bias increased as N increased for (μ, φ) combinations having poor empirical coverage levels. However, the largest observed median (96.5) occurred for $\mu = 0.1$, $\varphi = 0.5$ and N = 100.

8. Conclusions

For a beta distribution that has been parameterized in terms of its mean μ and a scale parameter φ, when both μ and φ are not close to the boundary of the parameter space, there are no serious problems using both the pseudo-likelihood and Laplace methods in GLIMMIX for small sample inference in the one sample problem. Inference for samples sizes as small as 10 to 15 should produce reasonably good results in these situations. However, when μ or φ or both are near the boundary difficulties arise with convergence, estimation of both parameters, and confidence interval coverage for μ. It is known that when a parameter is actually on the boundary of the parameter space, the asymptotic theory of likelihood based inference is not the same as for non-boundary parameter values. Hence, it may not be unreasonable to expect poor performance in small sample inference when one or both parameters are near the parameter space boundary.

Table 4. Means of estimated scale parameter (phi) using RSPL for selected (μ, φ, N) combinations. Trends for Laplace are similar.

φ	N	μ = 0.1	μ = 0.2	μ = 0.3	μ = 0.4	μ = 0.5
2.0	5	46.0	6.1	4.9	4.5	4.8
	10	6.0	3.0	2.9	2.8	2.7
	20	7.0×10^{2}	2.4	2.3	2.3	2.3
	50	1.0×10^{4}	2.1	2.1	2.1	2.1
	100	6.4×10^{3}	2.1	2.1	2.1	2.0
1.5	5	1.5×10^{13}	5.3	3.5	3.6	3.6
	10	2.2×10^{3}	2.4	2.2	2.0	2.1
	20	8.8×10^{6}	1.8	1.7	1.7	1.7
	50	9.1×10^{6}	2.1	1.6	1.6	1.6
	100	4.6×10^{6}	6.9×10^{2}	1.7	1.5	1.5
1.0	5	3.4×10^{14}	7.0	3.4	2.7	2.1
	10	3.0×10^{6}	3.2	1.5	1.4	1.3
	20	5.9×10^{7}	1.3×10^{3}	1.2	1.2	1.1
	50	6.1×10^{7}	25.9	1.2	1.0	1.1
	100	1.9×10^{8}	28.1	1.1	1.0	1.0
0.5	5	4.2×10^{16}	2.7×10^{13}	9.5×10^{5}	2.7	1.7
	10	4.7×10^{8}	9.9×10^{5}	5.3×10^{6}	1.0	7.9
	20	2.3×10^{8}	2.4×10^{7}	8.1×10^{3}	1.2	0.6
	50	3.7×10^{9}	7.4×10^{7}	6.0×10^{2}	1.2	0.6
	100	1.3×10^{10}	5.5×10^{7}	3.6×10^{5}	1.7	1.0