The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*), *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

Subscriptions are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at

http://www.stata.com/bookstore/sj.html

Subscription rates listed below include both a printed and an electronic copy unless otherwise mentioned.

U.S. and Canada		Elsewhere	
Printed & electronic		**Printed & electronic**	
1-year subscription	$ 98	1-year subscription	$138
2-year subscription	$165	2-year subscription	$245
3-year subscription	$225	3-year subscription	$345
1-year student subscription	$ 75	1-year student subscription	$ 99
1-year institutional subscription	$245	1-year institutional subscription	$285
2-year institutional subscription	$445	2-year institutional subscription	$525
3-year institutional subscription	$645	3-year institutional subscription	$765
Electronic only		**Electronic only**	
1-year subscription	$ 75	1-year subscription	$ 75
2-year subscription	$125	2-year subscription	$125
3-year subscription	$165	3-year subscription	$165
1-year student subscription	$ 45	1-year student subscription	$ 45

Back issues of the *Stata Journal* may be ordered online at

http://www.stata.com/bookstore/sjj.html

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

http://www.stata-journal.com/archives.html

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.

The *Stata Journal* (ISSN 1536-867X) is a publication of Stata Press. Stata, Stata, Stata Press, Mata, Mata, and NetCourse are registered trademarks of StataCorp LP.

The Stata Journal (2014)
14, Number 2, pp. 292–303

Estimation and testing of binomial and beta-binomial regression models with and without zero inflation

James W. Hardin
Institute for Families in Society
Department of Epidemiology and Biostatistics
University of South Carolina
Columbia, SC
jhardin@sc.edu

Joseph M. Hilbe
School of Social and Family Dynamics
Arizona State University
Tempe, AZ
hilbe@asu.edu

Abstract. We present new Stata commands for carrying out several regression commands suitable for binomial outcomes. The `zib` command extends Stata's `binreg` command to allow zero inflation. The `betabin` command fits binomial regression models allowing for beta overdispersion, and the `zibbin` command fits a beta-binomial regression model with zero inflation. All the new commands allow the specification of links within the `glm` command's collection for both outcome and zero inflation. The zero-inflated commands optionally calculate a Vuong test comparing the zero-inflated model with the nonzero-inflated model, and the `zibbin` command optionally includes a likelihood-ratio test of the overdispersion parameter.

Keywords: st0337, betabin, zib, zibbin, binomial outcomes, Vuong test, zero inflation, beta overdispersion

1 Introduction

Regression modeling of binary outcomes is supported by several Stata commands. Missing from the official collection of commands is support for zero inflation and beta dispersion. We present Stata commands to evaluate zero-inflated binomial (ZIB) regression, beta-binomial regression, and zero-inflated beta-binomial regression. This article is organized as follows: in section 2, we review the regression models; in section 3, we present Stata syntax for the new commands; and in section 4, we present examples.

2 Binomial regression models

A binomial outcome is characterized by

$$P(Y=y) = \binom{n}{y}\mu^y(1-\mu)^{n-y}$$

where the expected value of the outcome is $n\mu$ for n independent trials, each with a probability of success given by μ. In binomial regression, we model the probability of success, μ, via a link function $g(\,)$, of a linear combination of covariates $X\beta$. The compound beta-binomial distribution results from assuming the probability of success, μ, is a random variable that follows a Beta(α, β) distribution. The resulting probability mass function can be written as

$$P(Y=y) = \binom{n}{y}\frac{B(y+\alpha, n-y+\beta)}{B(\alpha,\beta)}$$

with mean $n\alpha/(\alpha+\beta)$ and variance $n\alpha\beta(\alpha+\beta+n)/\{(\alpha+\beta)^2(\alpha+\beta+1)\}$. Substituting $p = \alpha/(\alpha+\beta)$ and $\sigma = 1/(\alpha\beta)$, the probability mass function is given by

$$P(Y=y) = \binom{n}{y}\frac{B(y+p/\sigma, n-y+(1-p)/\sigma)}{B(p/\sigma,(1-p)/\sigma)}$$

with mean np and variance $np(1-p)(1+n\sigma)/(1+\sigma)$ so that the variance exceeds that of the binomial model when $(1+n\sigma)/(1+\sigma) > 1$. Thus extra binomial dispersion is accounted for only when $n > 1$ and when $\sigma > 0$; the new commands support estimation using the same grouped binomial data that are addressed by Stata's `glm, family(binomial N)` for generalized linear models and by the `blogit` and `bprobit` commands for logistic and probit regression models, respectively.

In the following subsections, we review various approaches to regression modeling of binary outcomes.

2.1 General binomial regression

General binomial regression models are fit using Stata's `glm` command; one can also use the `binreg` command. Specific regression models (fitting binomial models for a specific link function) can also be fit using model-specific commands: `logit` and `logistic` fit logistic regression models; `probit` fits a probit regression model; and `cloglog` can be used to fit complementary log-log regression models. Using the `glm` command is the most convenient: it represents one command from which multiple models (via various link functions) can be fit.

2.2 General binomial regression with zero inflation

As with the zero-inflated Poisson and the zero-inflated negative binomial models, we can imagine two separate processes generating outcomes such that the outcome of the two

processes are partially visible. The two binary components of the model are programmed through the `glm` command in Stata and admit all supported `glm` link functions.

In the general (grouped) binomial regression model, each observation in the dataset contains information on the number of successes out of a number of trials where each trial has the same probability of success. This probability of success is parameterized via a user-specified link function of a linear predictor, $x\beta$. When we incorporate zero inflation, we consider a Bernoulli process that models the probability of zero successes. This probability of failure is parameterized via a user-specified link function of a linear predictor, $z\gamma$. Given that this part of the data-generating process is Bernoulli and models the probability of zero successes, it is modeling the probability of failure for one trial. This is backward from the binomial model with which it is combined, and users must take care interpreting the coefficients of the Bernoulli (inflation) process as being associated with a higher likelihood of failure (zero successes), while the coefficients of the binomial process are associated with a higher likelihood of success for each independent trial.

$$\begin{aligned} P(Y=0) &= P_{\text{Bernoulli}}(Y=0|z\gamma) \\ &\quad + \{1 - P_{\text{Bernoulli}}(Y=0|z\gamma)\}\, P_{\text{Binomial}}(Y=0|x\beta, n) \\ P(Y=y>0) &= \{1 - P_{\text{Bernoulli}}(Y=0|z\gamma)\}\, P_{\text{Binomial}}(Y=y|x\beta, n) \end{aligned}$$

A Vuong test (Vuong 1989) evaluates whether the binomial model with zero inflation or the binomial model without zero inflation is closer to the true model. This equation is equivalent to the interpretation of the Vuong test for the `zip` and `zinb` commands.

A random variable, ω, is defined as the vector $\log L_{\text{Z}} - \log L_{\text{S}}$, where L_{Z} is the likelihood of the zero-inflated model evaluated at its maximum likelihood estimates, and L_{S} is the likelihood of the standard (nonzero-inflated) model evaluated at its maximum likelihood estimates. The vector of differences over the N observations is then used to define the statistic

$$V = \frac{\sqrt{N}\overline{\omega}}{\sqrt{\sum_i (\omega_i - \overline{\omega})^2/(N-1)}}$$

which, asymptotically, is characterized by a standard normal distribution. A significant positive statistic indicates preference for the zero-inflated model, and a significant negative statistic indicates preference for the model without zero inflation. Nonsignificant Vuong statistics indicate no preference for either model. Results of this test are included in a footnote to the estimation of the model when the user includes the `vuong` option.

2.3 Beta-binomial regression

Griffiths (1973) and Prentice (1986) popularized a regression model for outcomes following the beta-binomial model. Subsequently, Guimarães (2005) provided readers of this journal an introduction to methods for estimating beta-binomial parameters. Herein, we provide a full regression command for grouped binomial data (with and without zero inflation) following this distribution. Similar to how the negative binomial model allows

for greater dispersion than the Poisson model, the beta-binomial regression model allows for greater dispersion than the binomial model. This extra variability originates from assuming that the mean parameter in the binomial model follows a beta distribution; the binomial component is programmed through the `glm` command and supports all the `glm` link functions.

Because the beta-binomial regression model differs from the general binomial regression model only when $\sigma > 0$, estimation of the model includes the appropriate test of the null hypothesis that $\sigma = 0$. Similar to testing the dispersion parameter in a negative binomial model, the hypothesis test evaluates the parameter at the boundary of the parameter space. As such, the resulting statistic is a distributed chi-bar with a single degree of freedom. Results of this test are included as a footnote to the output of the model estimation.

2.4 Beta-binomial regression with zero inflation

Analogous to how the ZIB model allows extra dispersion and zero inflation, the zero-inflated beta-binomial regression model allows extra-binomial dispersion and zero inflation. The two binary components of the model are programmed through the `glm` command in Stata and thus admit all supported `glm` link functions.

Because the zero-inflated beta-binomial regression model differs from the ZIB regression model only when $\sigma > 0$, estimation of the model includes the appropriate test. However, unlike with the `betabin` command, this test is included in the output only when the user requests the test via the `zib` option. This implementation mimics that of the `zip` option for the `zinb` command. Interpretation of the test results is the same as for the `zinb` command.

In addition, a Vuong test evaluates whether the beta-binomial model with zero inflation or the beta-binomial model without zero inflation is closer to the true model. As in the interpretation of the Vuong test for the `zip`, `zinb`, and `zib` commands, a significant positive statistic indicates preference for the zero-inflated model, and a significant negative statistic indicates preference for the model without zero inflation. Nonsignificant Vuong statistics indicate no preference for either model. Results of this test are included in a footnote to the estimation of the model when the user includes the `vuong` option.

3 Stata syntax

The software accompanying this article includes the command files and supporting files for dialogs and help. In all the following syntax diagrams, unspecified *options* include the usual collection of maximization and display options available to all estimation commands. In addition, all commands include the option `link(`*linkname*`)` to specify the link function for the binomial model, and the zero-inflated commands include the option `ilink(`*linkname*`)` to specify the link function for the inflation model. Supported *linkname*s include `logit`, `probit`, `loglog`, and `cloglog`.

Equivalent in syntax to the `zip` command, the basic syntax for the ZIB model is

`zib` *depvar* [*indepvars*] [*if*] [*in*] [*weight*] `,`
`inflate(`*varlist*[`, offset(`*varname*`)`] | `_cons)` `n(`*varname_n*`)` [`vuong` *options*]

Equivalent in syntax to the `nbreg` command, the basic syntax for the beta-binomial regression model is

`betabin` *depvar* [*indepvars*] [*if*] [*in*] [*weight*] `, n(`*varname_n*`)` [*options*]

Equivalent in syntax to the `zinb` command, the basic syntax for the zero-inflated beta-binomial regression model is

`zibbin` *depvar* [*indepvars*] [*if*] [*in*] [*weight*] `,`
`inflate(`*varlist*[`, offset(`*varname*`)`] | `_cons)` `n(`*varname_n*`)` [`vuong zib` *options*]

Help files are included for the estimation and postestimation specifications of these models. The help files include example specifications.

4 Example

Using data included in Hilbe (2009) on surviving passengers of the Titanic (see table 1), we highlight the use of the new commands and the interpretation of fitted coefficients. For pedagogical reasons, we have altered the data by changing some survivor numbers to facilitate examination of the data by the zero-inflated models.

The data are organized in 12 passenger types constituting the outcome for 1,316 passengers. Passenger types are defined by whether the members are adult, whether they are male, and whether they are first-, second-, or third-class passengers.

Table 1. Survivors among different categorizations of passengers on the Titanic

Survive	N	Adult	Male	Class	Survive	N	Adult	Male	Class
0	1	0	0	1	140	144	1	0	1
0	13	0	0	2	80	93	1	0	2
14	31	0	0	3	76	165	1	0	3
0	5	0	1	1	57	175	1	1	1
0	11	0	1	2	14	168	1	1	2
0	48	0	1	3	75	462	1	1	3

Because we are interested in the predictors of passenger survival, we first construct a regression with the outcome variable `survive` as the numerator of the binomial response and the variable `N` as the denominator of the binomial response. The `N` variable represents the number of passengers having the same pattern or values for the model predictors, that is, the same values for `adult`, `male`, and `class`. There are 12 separate sets of covariate patterns in the data. The explanatory predictors `adult` and `male` are binary; `class` is a categorical variable with three values or levels. We created three indicator variables for each level of `class`, designating `class1` (first-class passengers) as the reference level. Because we are treating these data as grouped data for which observations within a group might be more correlated than observations from different groups, we specify robust standard errors in generalized linear model specifications. Because the beta-binomial regression models incorporate an extra dispersion parameter, we specify model-based standard errors in those cases.

```
. glm survive adult male class2 class3, family(binomial N) nolog eform
> vce(robust)
Generalized linear models                         No. of obs      =        12
Optimization     : ML                             Residual df     =         7
                                                  Scale parameter =         1
Deviance         =  86.69204634                   (1/df) Deviance =  12.38458
Pearson          =  77.51997519                   (1/df) Pearson  =  11.07428
Variance function: V(u) = u*(1-u/N)               [Binomial]
Link function    : g(u) = ln(u/(N-u))             [Logit]
                                                  AIC             =  10.79288
Log pseudolikelihood = -59.75725742               BIC             =   69.2977
```

survive	Odds Ratio	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
adult	5.379405	4.653434	1.95	0.052	.987201	29.31318
male	.0741401	.0409807	-4.71	0.000	.0250931	.2190545
class2	.2766598	.1523033	-2.33	0.020	.0940488	.8138397
class3	.2041763	.106587	-3.04	0.002	.073392	.5680175
_cons	1.562367	1.383433	0.50	0.614	.2754655	8.861329

We note the Pearson dispersion statistic of 11.074, which clearly indicates overdispersion in the data.

```
. estat ic
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	12	.	-59.75726	5	129.5145	131.939

Note: N=Obs used in calculating BIC; see [R] BIC note

The Bayesian information criterion (BIC) statistic for this model is 131.94, and the Akaike's information criterion (AIC) is 129.51.

Subsequently, we specified the complementary log-log link in place of the canonical logit link:

```
. glm survive adult male class2 class3, family(binomial N) link(cloglog) nolog
> eform vce(robust)
Generalized linear models                         No. of obs      =        12
Optimization     : ML                             Residual df     =         7
                                                  Scale parameter =         1
Deviance         =  65.24144526                   (1/df) Deviance =  9.320206
Pearson          =   58.8251827                   (1/df) Pearson  =  8.403598
Variance function: V(u) = u*(1-u/N)               [Binomial]
Link function    : g(u) = ln(-ln(1-u/N))          [Complementary log-log]
                                                  AIC             =  9.005326
Log pseudolikelihood = -49.03195688               BIC             =   47.8471
```

survive	exp(b)	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
adult	3.663888	2.765667	1.72	0.085	.834482	16.08672
male	.1375246	.0394917	-6.91	0.000	.0783337	.2414418
class2	.4232986	.1475969	-2.47	0.014	.2137212	.83839
class3	.2782559	.0781055	-4.56	0.000	.1605147	.4823628
_cons	.860577	.6584041	-0.20	0.844	.1921138	3.854969

```
. estat ic
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	12	.	-49.03196	5	108.0639	110.4884

Note: N=Obs used in calculating BIC; see [R] BIC note

This model appears to be the best-fitted standard generalized linear model for the data. The Pearson dispersion statistic has reduced to 8.4, and the AIC and BIC statistics have respective values of 108.06 and 110.49, a substantial improvement over the logit model.

Ideally, an equidispersed binomial model has a dispersion statistic of 1.0. We seek to determine what may be causing the inflated dispersion statistic, which represents extra-binomial correlation in the data. We note that 5 of the 12 binomial numerators have 0

values. Like the zero-inflated Poisson model, a ZIB model can be used to accommodate binomial overdispersion, adjusting for Poisson overdispersion because of excessive zero counts.

We attempted both the logit and the complementary log-log links using the new `zib` command. The complementary log-log link provided better results.

```
. zib survive adult male class2 class3, n(N) inflate(adult male class2 class3)
> link(cloglog) nolog eform vce(robust)
Zero-inflated binomial regression               Number of obs   =         12
Regression link: cloglog                        Nonzero obs     =          7
Inflation link : logit                          Zero obs        =          5
                                                LR chi2(4)      =     480.98
Log pseudolikelihood = -35.96415                Prob > chi2     =     0.0000
```

survive	exp(b)	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
survive						
adult	1.383654	.3449018	1.30	0.193	.8488861	2.255306
male	.1412465	.0418428	-6.61	0.000	.0790347	.2524278
class2	.4484645	.1542485	-2.33	0.020	.2285382	.8800296
class3	.2641123	.0757093	-4.64	0.000	.1505868	.4632232
_cons	2.274691	.6520528	2.87	0.004	1.296942	3.989551
inflate						
adult	-79.48796	1.805339	-44.03	0.000	-83.02636	-75.94957
male	38.90549	1.477246	26.34	0.000	36.01014	41.80084
class2	-.2168801	1.061966	-0.20	0.838	-2.298296	1.864536
class3	-40.15798	1.397274	-28.74	0.000	-42.89659	-37.41937
_cons	20.52229	.9281515	22.11	0.000	18.70315	22.34144

```
. estat ic
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	12	-276.4523	-35.96415	9	89.92831	94.29247

```
Note:  N=Obs used in calculating BIC; see [R] BIC note
```

The initial part of the regression table reports usual binomial regression results. From this part of the model, we can see that males (compared with females), second-class (compared with first-class) passengers, and third-class (compared with first-class) passengers are more likely to be survivors of the disaster. The inflation part of the model reports the association of the covariates with the likelihood of a zero outcome. Compared with children, adults are far less likely to have zero survivors. Similarly, compared with males, females are far less likely to have zero survivors. Finally, third-class passengers, compared with first-class passengers, are far less likely to have zero survivors.

We do not have a Pearson dispersion statistic with this model, although it can be derived. However, the AIC statistic rather substantially decreased from 108.06 to 89.93, as did the BIC statistic (down to 94.29).

Next we fit a beta-binomial model for the data; the generalization from binomial to beta binomial is similar to that of Poisson to negative binomial. Though we do not illustrate results for the logit link, the complementary log-log link proved to be a better-fitting link than the logit link.

```
. betabin survive adult male class2 class3, n(N) nolog eform link(cloglog)
Beta-binomial regression                        Number of obs   =         12
Link           = cloglog                        LR chi2(4)      =      18.71
Dispersion     = beta-binomial                  Prob > chi2     =     0.0022
Log likelihood = -31.990675                     Pseudo R2       =     0.2263
```

survive	exp(b)	Std. Err.	z	P>|z|	[95% Conf.	Interval]
adult	9.762161	8.616098	2.58	0.010	1.730909	55.05765
male	.1380966	.0661646	-4.13	0.000	.0539954	.3531908
class2	.495203	.216559	-1.61	0.108	.2101558	1.166877
class3	.3748033	.1957181	-1.88	0.060	.1346839	1.043017
_cons	.2738493	.2611105	-1.36	0.174	.0422577	1.774669
/lnsigma	-2.164766	.7019774			-3.540616	-.7889153
sigma	.1147768	.0805707			.0289955	.4543374

```
Likelihood-ratio test of sigma=0:  chibar2(01) =   34.08 Prob>=chibar2 = 0.000
. estat ic
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	12	-41.34621	-31.99068	6	75.98135	78.89079

```
Note:  N=Obs used in calculating BIC; see [R] BIC note
```

The beta-binomial model results in a significantly better model than the ZIB. The AIC drops from 108.06 to 75.98, and the BIC from 94.29 to 78.89. Note that `adult` and `class3` (third class) are not significant at the 0.05 level. The likelihood-ratio test indicates that the beta-binomial model is preferred to the logit model.

Finally, we turn to the zero-inflated beta binomial model. There is still the problem of excessive zero counts in the data. Recall that the zero-inflated complementary log-log model was preferred to the standard complementary log-log model. Here the canonical logit link is preferred over the complementary log-log link. Note that we are using the logit link for the model's inflation component.

```
. zibbin survive adult male class2 class3, n(N)
> inflate(adult male class2 class3) link(cloglog) nolog eform
Zero-inflated beta-binomial regression           Number of obs   =         12
Regression link: logit                           Nonzero obs     =          7
Inflation link : logit                           Zero obs        =          5
                                                 LR chi2(4)      =      16.72
Log likelihood =  -26.1544                       Prob > chi2     =     0.0022
```

survive	exp(b)	Std. Err.	z	P>|z|	[95% Conf.	Interval]
survive						
adult	1.471449	.9772518	0.58	0.561	.4003373	5.408344
male	.0410846	.0223641	-5.86	0.000	.0141363	.1194054
class2	.2796984	.143201	-2.49	0.013	.102539	.7629415
class3	.0867963	.0593631	-3.57	0.000	.0227161	.3316411
_cons	9.555155	8.167465	2.64	0.008	1.789187	51.02932
inflate						
adult	-74.79802	21738.02	-0.00	0.997	-42680.53	42530.94
male	36.05695	13572.33	0.00	0.998	-26565.22	26637.34
class2	-.1523928	15692.19	-0.00	1.000	-30756.29	30755.98
class3	-36.94756	15063.43	-0.00	0.998	-29560.73	29486.83
_cons	19.01662	12869.68	0.00	0.999	-25205.09	25243.13
/lnsigma	-3.300808	.5982408			-4.473338	-2.128277
sigma	.0368534	.0220472			.0114092	.1190422

```
. estat ic
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	12	-34.51573	-26.1544	11	74.30881	79.64278

```
Note:  N=Obs used in calculating BIC; see [R] BIC note
```

The excess number of zero counts indeed appears to be a factor in the observed overdispersion in the data. The AIC drops to 74.31. Though the BIC increases slightly to 79.64, the difference between 79.64 and 78.89 is only 0.75. We affirm that there is no significant difference between the BIC measures for these models. Note that we have incorporated the `adult` variable in the main component of the model despite its not being a significant predictor of survival. We also point out that the `class2` variable is not significant in the logistic inflation component of the model.

```
. zibbin survive adult male class2 class3, n(N) inflate(adult male class3)
> link(cloglog) nolog eform zib vuong
Zero-inflated beta-binomial regression          Number of obs   =         12
Regression link: logit                          Nonzero obs     =          7
Inflation link : logit                          Zero obs        =          5
                                                LR chi2(4)      =      16.72
Log likelihood =  -26.1544                      Prob > chi2     =     0.0022
```

survive	exp(b)	Std. Err.	z	P>|z|	[95% Conf.	Interval]
survive						
adult	1.471449	.9772515	0.58	0.561	.4003373	5.408342
male	.0410846	.0223641	-5.86	0.000	.0141363	.1194055
class2	.2796986	.1432011	-2.49	0.013	.102539	.7629418
class3	.0867964	.0593631	-3.57	0.000	.0227162	.3316412
_cons	9.555147	8.167456	2.64	0.008	1.789186	51.02925
inflate						
adult	-74.75848	21851.03	-0.00	0.997	-42901.99	42752.48
male	36.06967	13642.47	0.00	0.998	-26702.68	26774.82
class3	-36.74312	11790	-0.00	0.998	-23144.72	23071.23
_cons	18.81744	8834.441	0.00	0.998	-17296.37	17334
/lnsigma	-3.300808	.5982408			-4.473338	-2.128277
sigma	.0368534	.0220472			.0114092	.1190422

```
Likelihood-ratio test of sigma=0:  chibar2(01) =   44.59 Prob>=chibar2 = 0.000
Vuong test of zibb vs. standard beta binomial:    z =     1.49   Pr>z = 0.0684
. estat ic
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	12	-34.51573	-26.1544	10	72.30881	77.15787

```
Note:  N=Obs used in calculating BIC; see [R] BIC note
```

For this model, the AIC is now 72.31, and the BIC is 77.16. Each of the predictors has significant p-values at the 0.05 level. The comparison of the zero-inflated beta-binomial regression model with the ZIB model for binomial outcome data is similar to the comparison of the zero-inflated negative binomial model with the zero-inflated Poisson model for count data. The requested Vuong statistic is not statistically significant. While the AIC and BIC statistics indicate a slight preference for this final model, the Vuong test does not indicate that either the beta-binomial model or the zero-inflated beta-binomial model is closer to the true model.

5 References

Griffiths, D. A. 1973. Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. *Biometrics* 29: 637–648.

Guimarães, P. 2005. A simple approach to fit the beta-binomial model. *Stata Journal* 5: 385–394.

Hilbe, J. M. 2009. *Logistic Regression Models*. Boca Raton, FL: Chapman & Hall/CRC.

Prentice, R. L. 1986. Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *Journal of the American Statistical Association* 81: 321–327.

Vuong, Q. H. 1989. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57: 307–333.

About the authors

James W. Hardin is an associate professor in the Department of Epidemiology and Biostatistics and an affiliated faculty member in the Institute for Families in Society at the University of South Carolina in Columbia, SC.

Joseph M. Hilbe is an emeritus professor at the University of Hawaii, an adjunct professor of statistics at Arizona State University in Tempe, AZ, and a Solar System Ambassador with NASA's Jet Propulsion Laboratory in Pasadena, CA.