October 15, 2013

Beta Binomial Regression

Joseph M Hilbe, *Arizona State University*

Beta Binomial Regression

Joseph M. Hilbe
(c) 2013, Joseph M Hilbe: hilbe@asu.edu

Modeling overdispersed binomial data can be developed by assuming that the binomial mean parameter is itself beta distributed. That is, we provide a prior beta distribution to μ, the logistic model probability of success (1). The beta distribution, unlike the binomial, is a doubly bounded two parameter distribution. This second parameter is employed in the model to adjust for any extra-binomial correlation found in the data. The two-parameter model, which is based on a mixture of beta and binomial distributions, is known as beta-binomial regression.

The binomial distribution below is expressed in terms of parameter μ. This is standard when the binomial distribution is being modeled as a generalized linear model (GLM), otherwise the parameter is typically symbolized as π. Since I will use the *glm* functions in Stata and R when modeling the binomial component of the beta binomial, we shall employ μ in place of π.

BINOMIAL PDF

$$f(y;\mu,n) = \binom{n}{y}\mu^{y}(1-\mu)^{n-y} \tag{1}$$

The $\binom{n}{y}$ choose function is the binomial coefficient, which is the normalization term of the binomial probability distribution function (PDF). It guarantees that the function sums to 1.0. This form of the function may also be expressed in terms of factorials,

$$\binom{n}{y} = \frac{n!}{y!\,(n-y)!} \tag{2}$$

which is easily recognized from basic algebra as a combination. Both terms can be interpreted as describing the number of ways that *y* successes can be distributed among *n* trials, or observations. Note though that the mean parameter, μ, is not a term in the coefficient,

Factorials may also be calculated in terms of factorial or gamma functions. In Stata the appropriate functions to use for calculating factorials are the log-factorial and log-gamma functions. For example, factorial 5 is 120; ie 1*2*3*4*5. We must exponentiate the natural log in both cases to obtain a factorial. In the case of the gamma function, 1 must be added to the number being factorialized. For example:

```
. di exp(lnfactorial(5))
120

. di exp(lngamma(5+1))
120
```

Using the Greek symbol Γ for a gamma function, Γ(), the binomial normalization term from (2) above may be expressed as:

$$\frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)} \tag{3}$$

The log-likelihood function for the binomial model can then be expressed as:

$$f(\mu; y, n) = ln\Gamma(n+1) - ln\Gamma(y+1) - ln\Gamma(n-y+1) + y\ln(\mu) + (n-y)\ln(1-\mu) \tag{4}$$

The beta distribution is used as the basis of modeling proportional data. That is, beta data is constrained between 0 and 1 - and can be thought of in this context as the proportion obtained by dividing the binomial numerator by the denominator. The Beta PDF is given below in terms of two shape parameters, *a* and *b*, although there are a number of different parameterizations.

BETA PDF

$$f(y; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, y^{a-1}(1-y)^{b-1} \tag{5}$$

where a is the number of successes and b the number of failures. The initial term in the function is the normalization constant, comprised of gamma functions.

The above function can also be parameterized in terms of μ. Since we plan on having the binomial parameter, μ, itself distributed as beta, we can parameterize the beta PDF as

$$f(\mu) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, \mu^{a-1}(1-\mu)^{b-1} \tag{6}$$

Notice that the kernal of the beta distribution is similar to that of the binomial kernal.

$$\mu^{y}(1-\mu)^{n-y} \sim \mu^{a-1}(1-\mu)^{b-1} \tag{7}$$

Even the coefficients of the beta and binomial are similar in structure. In probability theory such a relationship is termed *conjugate*. The beta distribution is conjugate to the binomial. This is a very useful property when mixing distributions, since it generally allows for easier estimation. Conjugacy plays a particularly important role in Bayesian modeling where a prior conjugate (beta) distribution of a model coefficient, which is considered to be a random variable, is mixed with the (binomial) likelihood to form a (beta-binomial) posterior distribution.

The mean and variance of the beta PDF may be given as:

$$E(y) = \frac{a}{a+b} = \mu \qquad V(y) = \frac{ab}{(a+b)^2(a+b+1)} \tag{8,9}$$

As mentioned before, the beta binomial distribution is a mixture of the binomial and beta distributions. The binomial parameter, μ, is distributed as beta, which adjusts for extra-binomial correlation in the data. Such overdispersion can be due to clustering effects; ie. that various sets of observations in the data are more similar to one another than than they are to other sets in the data, or to the data as whole. Overdispersion may also be due to proneness in the data, excessive zero counts in the binomial numerator, needed additional predictors, or a number of other reasons. In any case, the mixture can be obtained by multiplying the two distributions.

$$f(y; \mu, a, b) = f(y; \mu, n) f(y; \mu, a, b)$$

(10)

The result is the beta-binomial probability function.

BETA BINOMIAL

$$f(y;\mu,a,b) = \frac{\Gamma(a+b)\Gamma(n+1)}{\Gamma(a)\Gamma(b)\Gamma(y+1)\Gamma(n-y+1)}\pi^{y-a-1}(1-\mu)^{n-y+b-1}$$

(11)

The kernel of the distribution may also be expressed in terms of gamma functions, but this is not useful when developing a statistical model. I provide it here since it is many times found in beta binomial literature, particularly in Bayesian statistics.

$$f(y;\pi,a,b) = \frac{\Gamma(a+b)\Gamma(n+1)}{\Gamma(a)\Gamma(b)\Gamma(y+1)\Gamma(n-y+1)}\frac{\Gamma(a+y)\Gamma(n+y-b)}{\Gamma(n+a+b)}$$

(12)

The beta binomial mean and variance are

$$E(Y) = \frac{na}{a+b} \qquad V(Y) = \frac{nab(a+b+n)}{(a+b)^2(a+b+1)}$$

(13.14)

An alternative parameterization may be given in terms of μ and σ, with μ=a/(a+b).

$$f(y;\mu,\sigma) = \frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)}\frac{\Gamma\left(\frac{1}{\sigma}\right)\Gamma\left(y+\frac{\mu}{\sigma}\right)\Gamma\left(n-y+\frac{1-\mu}{\sigma}\right)}{\Gamma\left(n+\frac{1}{\sigma}\right)\Gamma\left(\frac{\mu}{\sigma}\right)\Gamma\left(\frac{1-\mu}{\sigma}\right)}$$

(15)

wiih y=0,1,2,...n, and 0<μ<1, and σ>0.

Under this parameterization, the mean and variance of the beta binomial are:

$$E(Y) = n\mu \qquad V(Y) = n\mu(1-\mu)\left[1+\frac{\sigma}{1+\sigma}(n-1)\right]$$

(16,17)

This is the parameterization that is used in the Stata *betabin* command (Hardin & Hilbe, 2013) and in R's *gamlss* function (Rigby & Stasinopoulos, 2005)

EXAMPLES

To begin we shall use a beta model to estimate the parameters of proportional data. The beta model is appropriate when the variable to be modeled has values between 0 and 1, representing the proportion of successes for a specific covariate pattern or counts per time period or area.

We shall use the grouped Titanic disaster data for an example of a beta regression. However, we must divide the number of passengers who survived the wreck by the number of passengers having the same pattern of covariates. For the data below, there was only one passenger who was a first class female child - and she survived. 14 passengers survived of the 31 female children 3rd class passengers. The data is stored in the **titanixgrp** file

R: *Modeling beta regression*

```
=========================================
library(Hmisc); library(foreign)
titanic <- read.dta("c://ado/titanicgrp.dta")
titanic ; attach(titanic) ; table(class)
y <- survive/cases              # create y as the proportion
cbind(y,survive,cases)          # list of y an binomial variables
y[y==1] <- .9999                # replace .9999 for 1's in y
class03 <- factor(titanic$class,
     levels=c("3rd class", "2nd class", "1st class")) # change reference
library(betareg)                # use of betareg model
summary(mymod <- betareg(y ~ age + sex + class03, data=titanic))
library(gamlss)                 # use of gamlss model
summary(mybeta <- gamlss(y ~ age + sex + class03, data=titanic,
           sigma.fo=~1, family=BEOI, method=RS()))
=========================================
```

```
. use titanicgrp,clear
. l
     +-----------------------------------------------+
     | survive   cases      age     sex        class |
     |-----------------------------------------------|
  1. |       1       1    child   women    1st class |
  2. |      13      13    child   women    2nd class |
  3. |      14      31    child   women    3rd class |
  4. |       5       5    child     man    1st class |
  5. |      11      11    child     man    2nd class |
     |-----------------------------------------------|
  6. |      13      48    child     man    3rd class |
  7. |     140     144   adults   women    1st class |
  8. |      80      93   adults   women    2nd class |
  9. |      76     165   adults   women    3rd class |
 10. |      57     175   adults     man    1st class |
     |-----------------------------------------------|
 11. |      14     168   adults     man    2nd class |
 12. |      75     462   adults     man    3rd class |
     +-----------------------------------------------+
```

Using Stata we observe the data without labels,

```
. list, nolab

     +--------------------------------------+
     | survive   cases   age   sex    class |
     |--------------------------------------|
  1. |       1       1     0     0        1 |
  2. |      13      13     0     0        2 |
  3. |      14      31     0     0        3 |
  4. |       5       5     0     1        1 |
  5. |      11      11     0     1        2 |
     |--------------------------------------|
  6. |      13      48     0     1        3 |
  7. |     140     144     1     0        1 |
  8. |      80      93     1     0        2 |
  9. |      76     165     1     0        3 |
 10. |      57     175     1     1        1 |
     |--------------------------------------|
 11. |      14     168     1     1        2 |
 12. |      75     462     1     1        3 |
     +--------------------------------------+
```

Since I would like to have third class passengers as the reference level for the categorical variable, *class*, we factor or level it to create separate dummy or indicator variables, each of which is formatted as 0,1.

```
. tab class, gen(class)

  passenger |
  class:1-3 |      Freq.     Percent        Cum.
------------+-----------------------------------
  1st class |          4       33.33       33.33
  2nd class |          4       33.33       66.67
  3rd class |          4       33.33      100.00
------------+-----------------------------------
      Total |         12      100.00
```

Now we can divide *survive* (the passengers who survived) by the number of observations sharing the identical covariate pattern (*cases*). Each observation in the Titanic data is a distinct covariate pattern. The result of the division, *y*, is a proportion - a division of the binomial numerator by its denominator.

```
. gen y=survive/cases
. list survive cases y

     +---------------------------+
     | survive   cases         y |
     |---------------------------|
  1. |       1       1         1 |
  2. |      13      13         1 |
  3. |      14      31   .4516129 |
  4. |       5       5         1 |
  5. |      11      11         1 |
     |---------------------------|
  6. |      13      48   .2708333 |
  7. |     140     144   .9722222 |
  8. |      80      93   .8602151 |
  9. |      76     165   .4606061 |
 10. |      57     175   .3257143 |
     |---------------------------|
 11. |      14     168   .0833333 |
 12. |      75     462   .1623377 |
     +---------------------------+
```

We spot a problem. It could have been identified when comparing *survive* and *cases*, but there are 4 *y*'s with a value of 1. The beta distribution requires that all response values are between 0 and 1. The solution is to recode these values as 0.999. Then we can model the data. The Stata *betafit* command was created by Maartin Buis (Univ of Tuebingen), Nicholas Cox (Durham Univ) and Stephen Jenkins (London School of Economics and Political Science), and is provided on the book's web site.

```
. replace y=.9999 if y==1

. betafit y, mu(age sex class2 class1) nolog

ML fit of beta (mu, phi)                          Number of obs   =         12
                                                  Wald chi2(4)    =      13.34
Log likelihood =  25.126578                       Prob > chi2     =     0.0097
```

```
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -2.287329   .7299636    -3.13   0.002    -3.718032   -.8566269
         sex |  -1.203256   .6961577    -1.73   0.084      -2.5677    .1611883
      class2 |   2.204885   .8739525     2.52   0.012     .4919693      3.9178
      class1 |   2.662907    .943594     2.82   0.005     .8134968    4.512317
       _cons |   1.094288   .8183261     1.34   0.181    -.5096019    2.698178
-------------+----------------------------------------------------------------
     /ln_phi |   1.108147   .4674121     2.37   0.018     .1920359    2.024258
-------------+----------------------------------------------------------------
         phi |   3.028741    1.41567                      1.211714     7.57049
------------------------------------------------------------------------------
```

The R output using the *betareg* function provides the same result as Stata.

```
> mymod <- betareg(y ~ age + sex + class03, data=titanic)
> summary(mymod)

Coefficients (mean model with logit link):
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)        1.0943     0.6576   1.664 0.096122 .
ageadults         -2.2873     0.6503  -3.517 0.000436 ***
sexman            -1.2032     0.5997  -2.006 0.044813 *
class032nd class   2.2049     0.7711   2.859 0.004246 **
class031st class   2.6629     0.7921   3.362 0.000774 ***

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)    3.029      1.318   2.299   0.0215 *

Type of estimator: ML (maximum likelihood)
Log-likelihood: 25.13 on 6 Df
Pseudo R-squared: 0.7598
```

Using the *gamlss* package (Rigby and Stasinopoulos, 2005) we can duplicate the results of Stata and *betareg*, acknowledging that rounding errors give us slightly different - but statistically identical - output.

```
> summary(mybeta)
*******************************************************************
Family:  c("BEOI", "One Inflated Beta")

Call:  gamlss(formula = y ~ age + sex + class03, sigma.formula = ~1,
    family = BEOI, data = titanic, method = RS())

Fitting method: RS()
-------------------------------------------------------------------
Mu link function:  logit
Mu Coefficients:
                  Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)          1.090      0.8178    1.332   0.24023
ageadults           -2.275      0.7296   -3.118   0.02631
sexman              -1.190      0.6960   -1.709   0.14808
class032nd class     2.194      0.8731    2.513   0.05366
class031st class     2.647      0.9426    2.808   0.03762
```

```
------------------------------------------------------------------
Sigma link function:  log
Sigma Coefficients:
  Estimate  Std. Error     t value    Pr(>|t|)
   1.10111     0.46787     2.35347     0.06528

------------------------------------------------------------------
Nu link function:  logit
Nu Coefficients:
  Estimate  Std. Error     t value    Pr(>|t|)
-1.843e+01   2.899e+03  -6.359e-03   9.952e-01

------------------------------------------------------------------
No. of observations in the fit:  12
Degrees of Freedom for the fit:  7
      Residual Deg. of Freedom:  5
                      at cycle:  8

Global Deviance:     -50.25368
            AIC:     -36.25368
            SBC:     -32.85933
```

Note that the *gamlss* dispersion parameter value of 1.10111 is statistically the same as *betareg* and *betafit* output. The estimate is given in log form. Exponentiating 1.10111 produces a value of 3.01

The greater the value of the proportion the more likely a passenger is a female child, and the more likely they are of a superior passenger class. This information is not particularly helpful for understanding criteria of survival in this case, although there are data situations for which beta model make good sense. If we only have proportional data the beta model may be the only regression model we can use.

Prior to concluding our look at beta models, we should determine the extent of extra correlation in the data by employing a robust or sandwich variances adjustment, which results in the output below. Note that the adjustment does little to the values of the standard errors. The *p*-values (Wald statistics) are therefore nearly the same.

```
. betafit y, mu(age sex class2 class1) nolog vce(robust)

------------------------------------------------------------------------------
             |               Robust
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -2.287329   .6682538    -3.42   0.001    -3.597083   -.9775758
         sex |  -1.203256   .6802613    -1.77   0.077    -2.536543    .1300319
      class2 |   2.204885   .9155048     2.41   0.016     .4105283    3.999241
      class1 |   2.662907   .9157482     2.91   0.004     .8680736     4.45774
       _cons |   1.094288   .9632493     1.14   0.256     -.793646    2.982222
-------------+----------------------------------------------------------------
     /ln_phi |   1.108147   .2682577     4.13   0.000     .5823714    1.633922
-------------+----------------------------------------------------------------
         phi |   3.028741    .812483                      1.790279    5.123933
------------------------------------------------------------------------------
```

Robust variance and scaling do not work well with *betareg* and *gamlss*. Since beta regression is itself not our foremost concern in this monograph, I'll pass on additional discussion.

Before considering the use of a beta binomial model on grouped binomial data it is necessary to determine if the binomial data is overdispersed. In particular, we shall model the Titanic data using a grouped logistic model to determine if the data is overdispersed. If it is, we cannot in general trust the coefficients or standard errors of the resulting model. Other links can be used, and we can adjust the standard errors to account for the excess variability. We'll look at these alternatives first.

We model the data using a grouped logistic model. The *glm* command provides accurate estimates. R's *glm* function is also the function of choice for grouped logistic models. The *eform* option provides odds ratios to be displayed, and *nolog* depresses a print out of the iteration log that is by default displayed when a model is estimated.

R *Logistic regression models*

```
===================================================================
died <- cases - survive
summary(jhlogit <- glm(cbind(survive,died) ~ age + sex + class03,
                    data=titanic, family=binomial))
exp(coef(jhlogit))    # Odds ratios
library(COUNT)
modelfit(jhlogit)     # same as Stata abic command
summary(sclogit <- glm(cbind(survive,died) ~ age + sex + class03,
        data=titanic, family=quasibinomial))    # scaled SEs
OR <- exp(coef(sclogit)); OR
library(sandwich)
rse <- sqrt(diag(vcovHC(jhlogit, type = "HC0"))) # robust SEs
ORrse <- OR*rse; ORrse     # robust SE for odds ratios
===================================================================
```

```
. glm survive age sex class2 class1, fam(bin cases)eform nolog

Generalized linear models                          No. of obs      =        12
Optimization     : ML                              Residual df     =         7
                                                   Scale parameter =         1
Deviance         =  110.8437538                    (1/df) Deviance =  15.83482
Pearson          =  100.8828206                    (1/df) Pearson  =  14.41183 <=

Variance function: V(u) = u*(1-u/cases)            [Binomial]
Link function    : g(u) = ln(u/(cases-u))          [Logit]

                                                   AIC             =  13.14728
Log likelihood   = -73.88365169                    BIC             =  93.44941
------------------------------------------------------------------------------
             |                 OIM
     survive | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .3479809   .0844397    -4.35   0.000     .2162749    .5598924
         sex |   .0935308   .0135855   -16.31   0.000     .0703585    .1243347
      class2 |   2.129343   .3731801     4.31   0.000     1.510315    3.002091
      class1 |    5.84959   .9986265    10.35   0.000     4.186109    8.174107
       _cons |   3.652859   .9053449     5.23   0.000     2.247327    5.937442
------------------------------------------------------------------------------

. abic

AIC Statistic   =   13.14727          AIC*n      = 157.7673
BIC Statistic   =   13.65514          BIC(Stata) = 160.19183
```

The dispersion statistic has a value of 14.4, indicating extensive overdispersion. An equi-dispersed model -- one meeting the distributional assumptions of the model -- would have a dispersion statistic of approximately 1.0. The AIC statistic is 157.77, which we can later compare with a model of the data using beta binomial regression.

Note that the *p*-values appear to indicate that all of the predictors are significant contributors to understanding the response; ie predictors for survival. As we have shown earlier though, we need to at first scale the standard errors by the dispersion to determine the extent of the extra-binomial correlation in the data. The R *quasi-binomial* family provides the same results. The *nohead* option suppresses a display of header statistics, which are identical to the statistics above.

```
. glm survive age sex class2 class1, fam(bin cases) nolog eform scale(x2) nohead

------------------------------------------------------------------------------
             |                 OIM
     survive | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .3479809   .3205576    -1.15   0.252      .057205    2.116784
         sex |   .0935308   .0515744    -4.30   0.000     .0317386    .2756263
      class2 |   2.129343     1.4167     1.14   0.256     .5779923    7.844574
      class1 |    5.84959   3.791078     2.73   0.006     1.642358     20.8345
       _cons |   3.652859   3.436953     1.38   0.169     .5777532    23.09529
------------------------------------------------------------------------------
(Standard errors scaled using square root of Pearson X2-based dispersion.)
```

We find that age is no longer a significant predictor, nor is *class2*. this means that *class1* and *class2* are togther the reference level for *class1*. There is no statistical difference in 2nd and 3rd classs passengers with respect to survival. First class passengers, however, have nearly 6 times greater odds of survival than 3rd class passengers -- as well as 2nd and 3rd class passengers together. It appears from the output that first class females survived the Titanic accident significantly higher than other passengers. We know from independent sources that this is indeed what happened.

To confirm what we found using scaled standard errors, we'll employ a robust variance adjustment to the standard errors.

```
. glm survive age sex class2 class1, fam(bin cases) nolog eform vce(robust) nohead

------------------------------------------------------------------------------
             |               Robust
     survive | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .3479809   .2592807    -1.42   0.157     .0807839    1.498946
         sex |   .0935308   .0461042    -4.81   0.000     .0355935     .245775
      class2 |   2.129343   1.320289     1.22   0.223     .6316292    7.178426
      class1 |    5.84959   3.189964     3.24   0.001     2.008809    17.03383
       _cons |   3.652859   2.905024     1.63   0.103     .7685894    17.36087
------------------------------------------------------------------------------
```

The results are consistent with the "*quasibinomial*" or scaled model. We can be sure that based on a logistic regression model, *age* and *class2* are not significant contributors to an understanding of survival. It appears that female 1st class passengers had a significantly greater survival odds than did other passenger. Age was not a determinant.

BETA BINOMIAL

We use the *betabin* command from Hardin and Hilbe (2013). The command is a two parameter model with the scale parameter, *sigma*, serving as dispersion parameter, adjusting the model for any extra-binomial correlation. The origin of the model was outlined earlier in this section.

R *Beta binomial*

```
=================================================================
summary(mybb <- gamlss(cbind(survive,died) ~ age + sex + class03,
            data=titanic, family=BB))
exp(coef(mybb))
======================================
```

```
. betabin survive age sex class2 class1, n(cases) nolog eform

Beta-binomial regression                          Number of obs   =         12
Link           = logit                            LR chi2(4)      =      14.46
Dispersion     = beta-binomial                    Prob > chi2     =     0.0129
Log likelihood = -36.901181                       Pseudo R2       =     0.1639
------------------------------------------------------------------------------
     survive |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .1093592   .0862822    -2.81   0.005     .0232957    .5133763
         sex |   .1123649    .085515    -2.87   0.004     .0252829    .4993845
      class2 |   7.608313   5.994373     2.58   0.010     1.624243    35.63903
      class1 |   15.94947   13.97772     3.16   0.002     2.862691    88.86238
       _cons |   4.504546   3.740425     1.81   0.070     .8847928    22.93298
-------------+----------------------------------------------------------------
    /lnsigma |  -1.791043   .6072598                      -2.98125   -.6008352
-------------+----------------------------------------------------------------
       sigma |   .1667862   .1012825                      .0507294    .5483534
------------------------------------------------------------------------------
Likelihood-ratio test of sigma=0:  chibar2(01) =   73.96 Prob>=chibar2 = 0.000

. abic
AIC Statistic   =   7.150197          AIC*n      = 85.802361
BIC Statistic   =   7.941956          BIC(Stata) = 88.7118
```

R output using *gamlss* is given as displayed below. Note again that there is a slight difference in estimates. Note that the estimate for sigma is statistically the same as what is displayed in the Stata output for the log of sigma, */lnsigma*. Sigma is the dispersion parameter, and can itself be parameterized, having predictors like the mean or location parameter, *mu*. The dispersion estimates inform the analyst which predictors significantly influence the extra correlation in the data, therefore influencing the value of sigma. In this form below it is only the intercept of sigma that is displayed. In this respect, the beta binomial is analagous to the heterogeneous negative binomial count model (Hilbe, 2011, 2014), and the binomial logistic regression function is a analagous to Poisson model.

```
> summary(mybb)
                  Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)          1.498      0.6814    2.199  0.063855
ageadults           -2.202      0.8205   -2.684  0.031375
sexman              -2.177      0.6137   -3.547  0.009377
class032nd class     2.018      0.8222    2.455  0.043800
class031st class     2.760      0.8558    3.225  0.014547
```

```
---------------------------------------------------------------------
Sigma link function:  log
Sigma Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)    -1.801      0.7508   -2.399   0.03528

---------------------------------------------------------------------

> exp(coef(mybb))
     (Intercept)        ageadults          sexman class032nd class
       4.4738797        0.1105858       0.1133972        7.5253615
class031st class
      15.8044343
```

The dispersion parameter is significant in that its confidence interval does not include one (1).. Moreover, the AIC and BIC statistics are substantially lower than the binomial logistic model,. 85.8 and 88.7 respectively to the logistic model values of 157.7 and 160.2. The Beta binomial model is preferred to the single parameter logistic model. However, extra correlation still needs to be checked and adjusted. A robust or sandwich variance adjustment is applied to the model.

```
. betabin survive age sex class2 class1, n(cases) nolog eform vce(robust)

Beta-binomial regression                          Number of obs   =         12
Link           = logit
Dispersion     = beta-binomial                    Wald chi2(4)    =      17.01
Log likelihood = -36.901181                       Prob > chi2     =     0.0019
------------------------------------------------------------------------------
             |               Robust
     survive |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .1093592   .1220964    -1.98   0.047     .0122607    .9754317
         sex |   .1123649   .1235181    -1.99   0.047     .0130299    .9689943
      class2 |   7.608313   6.195038     2.49   0.013     1.542414    37.52976
      class1 |   15.94947   12.42715     3.55   0.000     3.463588    73.44568
       _cons |   4.504546   6.227455     1.09   0.276     .2998333    67.67405
-------------+----------------------------------------------------------------
    /lnsigma |  -1.791043   .6542925                     -3.073432   -.5086528
-------------+----------------------------------------------------------------
       sigma |   .1667862    .109127                      .0462621    .6013051
------------------------------------------------------------------------------
Likelihood-ratio test of sigma=0:  chibar2(01) =   73.96 Prob>=chibar2 = 0.000
```

Note that even after applying a robust variance adjustment, all of the main effect predictors are significant. In addition, the likelihood ratio test of the value of sigma shows us that the beta binomial model is preferred to the logistic model. We should check for an interactive effect between age and sex, and between both *age* and sex and *class1*. I shall leave that as an exercise for the reader. It appears, though, from looking at the main effects, only that 1st class female children stood the best chance of survival on the Titanic,

Zero-Inflated binomial

At times, even grouped data consists of more observations having a no successes for given covariates than acceptable based on model assumptions. Data having excessive zero values for the response are commonly referred to as zero-inflated models. Most zero-inflated models are count models, but grouped binomial models are also subject to having excess zeros. For a complete analysis of zero-inflation for count models, see Hilbe, (2011) or Hilbe (2014). For a discussion of beta binomial

$$f(y; n, \mu, \sigma) = \sigma + (1-\sigma)(1-\mu)^n \qquad if\ y == 0$$

$$f(y; n, \mu, \sigma) = \frac{(1-\sigma)n!\, \mu^y (1-\mu)^{n-y}}{y!\,(n-y)!} \qquad if\ y > 0$$

(18)

where $0<\mu<1$ and $0<\sigma<1$. The mean and variance or Y are given as:

$$E(Y) = (1-\sigma)n\mu \qquad V(Y) = n\mu(1-\sigma)[1-\mu+n\mu\sigma]$$

(19)

I use the **titanicgrp0** data set, which is nearly the same as **titanicgrp**, except that four of the observations have zero successes rather than what they had in **titanicgrp**..

R *Zero inflated Binomial*

```
=================================================================
library(Hmisc); library(foreign)
titanic0 <- read.dta("c://ado/titanicgrp0.dta")
attach(titanic0); head(titanic0)
class03 <- factor(titanic0$class,
     levels=c("3rd class", "2nd class", "1st class")) # change reference
summary(mylogit <- glm(cbind(survive,died) ~ age + sex + class03,
            data=titanic0, family=binomial))
summary(mybb <- gamlss(cbind(survive,died) ~ age + sex + class03,
            data=titanic0, family=BB))
summary(binBI0 <- gamlss(cbind(survive,died) ~ age + sex + class03,
                       nu.fo =~ age + sex,
                       data=titanic0, family=ZIBI))
exp(coef(binBI0))
summary(binBB0 <- gamlss(cbind(survive,died) ~ age + sex + class03,
                       nu.fo =~ age + sex,
                       data=titanic0, family=ZIBI))
exp(coef(binBB0))
# EXTRA: zero-inflated beta
y <- survive/cases            # create y as the proportion
cbind(y,survive,cases)        # list of y an binomial variables
y[y==0] <- .0001              # replace .0001 for 0's in y
summary(beta0 <- gamlss(y ~ age + sex + class03, data=titanic0,
          sigma.fo=~1, family=BEOI, method=RS()))
=================================================================
```

```
. titanicgrp0

. l survive-class, nolab

     +-------------------------------------+
     | survive   cases   age   sex   class |
     |-------------------------------------|
  1. |       0       1     0     0       1 |
  2. |       0      13     0     0       2 |
  3. |      14      31     0     0       3 |
  4. |       0       5     0     1       1 |
  5. |       0      11     0     1       2 |
```

```
     |-----------------------------------|
  6. |       0      48     0     1       3 |
  7. |     140     144     1     0       1 |
  8. |      80      93     1     0       2 |
  9. |      76     165     1     0       3 |
 10. |      57     175     1     1       1 |
     |-----------------------------------|
 11. |      14     168     1     1       2 |
 12. |      75     462     1     1       3 |
     +-----------------------------------+
```

Model the data as a standard logistic regression, without concern for the excess zero response values. The model standard errors have been adjusted by robust variance estimates. Recall that the model without zero responses was highly overdispersed; we therefore employ robust variance adjustment.

```
. glm survive age sex class2 class1, fam(bin cases) eform nolog 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/cases)            [Binomial]
Link function    : g(u) = ln(u/(cases-u))          [Logit]

                                                   AIC             =  10.79288
Log pseudolikelihood = -59.75725742                BIC             =   69.2977
------------------------------------------------------------------------------
             |               Robust
     survive | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   5.379405   4.653434     1.95   0.052      .987201    29.31318
         sex |   .0741401   .0409807    -4.71   0.000     .0250931    .2190545
      class2 |   1.355005   .8377221     0.49   0.623     .4033586    4.551873
      class1 |   4.897728   2.556782     3.04   0.002     1.760509    13.62546
       _cons |   .3189983   .2577471    -1.41   0.157     .0654677    1.554354
------------------------------------------------------------------------------

. abic
AIC Statistic   =   10.79288          AIC*n      = 129.51451
BIC Statistic   =   11.30074          BIC(Stata) = 131.93904
```

The zero-inflated binomial model has exponentiated coefficients that are close to the logistic model, except for *age*, which is not significant.

```
. zib survive age sex class2 class1, n(cases) eform nolog vce(robust) inflate(age sex
class2 class1)

Zero-inflated binomial regression                Number of obs   =         12
Regression link: logit                           Nonzero obs     =          7
Inflation link : logit                           Zero obs        =          5
                                                 LR chi2(4)      =     456.01
Log pseudolikelihood =  -48.4482                 Prob > chi2     =     0.0000
------------------------------------------------------------------------------
             |               Robust
     survive |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
survive      |
         age |   1.910453   1.032149     1.20   0.231     .6626171    5.508204
```

```
         sex |   .0773431   .0427009    -4.64   0.000     .0262105    .2282275
      class2 |   1.602383    1.03357     0.73   0.465     .4526093     5.67295
      class1 |   5.281635   2.930272     3.00   0.003     1.780403    15.66817
       _cons |   .8235294   6.93e-11 -2.3e+09   0.000     .8235294    .8235294
-------------+----------------------------------------------------------------
inflate      |
         age |  -74.92587   1.798567   -41.66   0.000      -78.451   -71.40074
         sex |   36.61934   1.479621    24.75   0.000     33.71933    39.51934
      class2 |   37.67849   1.380459    27.29   0.000     34.97284    40.38414
      class1 |   37.85468   1.404743    26.95   0.000     35.10144    40.60793
       _cons |  -18.51198   1.044466   -17.72   0.000     -20.5591   -16.46487
------------------------------------------------------------------------------

. abic
AIC Statistic   =   9.741367          AIC*n      = 116.89641
BIC Statistic   =   11.91234          BIC(Stata) = 121.74547
```

The AIC value drops from 129.5 to 116.9, a substantial reduction. The BIC drop a little over 10 points, which is also considered to be substantial. Inflated values inform us that all of the predictors significantly influence zero survival values. The zero-inflated binomial model appears to fit the data better than standard grouped logistic regression.

In order to display a Vuong test in Stata, standard errors cannot be adjusted. We drop the *vce(robust)* option and rerun the model. Model standard errors are displayed, as is the Vuong test results. For space purposes only the Vuong test results are shown.

```
. zib survive age sex class2 class1, n(cases) eform nolog inflate(age sex class2
class1) vuong

Vuong test of zib vs. standard binomial:          z =     1.85   Pr>z = 0.0322
```

The zero-inflated binomial fits the data better than does the logistic model, which we had earlier concluded. The Vuong test reconfirms our finding.

Zero-Inflated beta binomial

The zero-inflated beta binomial PDF can be expressed as,

$$f(y;n,\mu,\sigma,\nu) = \nu + (1-\nu)f(0;\mu,a,b) \quad if\ y == 0$$
$$f(y;n,\mu,\sigma,\nu) = (1-\nu)f(y;\mu,a,b) \qquad if\ y > 0 \tag{20}$$

where 0<μ<1, 0< <1,ν and σ>0

The mean and variance can be defined as:

$$E(Y) = (1-\nu)n\mu \qquad V(Y) = (1-\nu)n\mu(1-\mu)\left[1-\frac{\sigma}{1-\sigma}(n-1)\right] + \nu(1-\nu)n^2\mu^2 \tag{21}$$

We shall compare results of the beta binomial and zero-inflated beta binomial models using the **titanicgrp0** data,

```
. betabin survive age sex class2 class1, n(cases) eform nolog vce(robust)

Beta-binomial regression                         Number of obs   =         12
Link           = logit
Dispersion     = beta-binomial                   Wald chi2(4)    =      51.53
Log likelihood = -32.111548                      Prob > chi2     =     0.0000
------------------------------------------------------------------------------
             |               Robust
     survive |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   17.05713   23.53551     2.06   0.040     1.141338    254.9164
         sex |   .0647876   .0453762    -3.91   0.000     .0164181    .2556582
      class2 |    .856555   1.212357    -0.11   0.913     .0534535     13.7257
      class1 |   2.932857   3.637627     0.87   0.386     .2579554     33.3455
       _cons |   .1678694   .3710879    -0.81   0.420     .0022046    12.78251
-------------+----------------------------------------------------------------
    /lnsigma |  -1.980238   .5428028                     -3.044112   -.9163638
-------------+----------------------------------------------------------------
       sigma |   .1380364   .0749266                      .0476386    .3999708
------------------------------------------------------------------------------
Likelihood-ratio test of sigma=0:  chibar2(01) =   55.29 Prob>=chibar2 = 0.000

. abic
AIC Statistic   =   6.351924          AIC*n      = 76.223099
BIC Statistic   =   7.143684          BIC(Stata) = 79.132538
```

The AIC and BIC statistics have substantially dropped. For the AIC statistic, the logistic model has an AIC of approximately 130, for the zero-inflated binomial (ZIB) it reduced to 117, and now for the beta binomial it is 76.22. Next we model the zero-inflated beta binomial.

```
. zibbin survive age sex class2 class1, n(cases) eform nolog vce(robust) inflate(age
sex class2 class1)

Zero-inflated beta-binomial regression           Number of obs   =         12
Regression link: logit                           Nonzero obs     =          7
Inflation link : logit                           Zero obs        =          5
                                                 Wald chi2(4)    =      43.66
Log pseudolikelihood =  -26.1544                 Prob > chi2     =     0.0000
------------------------------------------------------------------------------
             |               Robust
     survive |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
survive      |
         age |   1.471143   .8602602     0.66   0.509     .4676376    4.628075
         sex |   .0410867   .0348104    -3.77   0.000     .0078078    .2162096
      class2 |   3.222762   3.011053     1.25   0.210     .5163495    20.11467
      class1 |   11.52215   13.05589     2.16   0.031     1.250335    106.1795
       _cons |   .8293798   .0037798   -41.05   0.000     .8220046    .8368212
-------------+----------------------------------------------------------------
inflate      |
         age |  -67.61926   1.806085   -37.44   0.000    -71.15913    -64.0794
         sex |   32.54278   1.527725    21.30   0.000     29.54849    35.53706
      class2 |   33.30773   1.392911    23.91   0.000     30.57767    36.03778
      class1 |   33.15821   1.401135    23.67   0.000     30.41204    35.90439
       _cons |  -16.08771   1.044466   -15.40   0.000    -18.13483   -14.04059
-------------+----------------------------------------------------------------
    /lnsigma |  -3.300706   .6550799                     -4.584639   -2.016773
-------------+----------------------------------------------------------------
       sigma |   .0368572   .0241444                      .0102074    .1330843
------------------------------------------------------------------------------
```

```
. abic
AIC Statistic   =     6.1924          AIC*n      = 74.308807
BIC Statistic   =   8.755208          BIC(Stata) = 79.642776
```

The AIC drops from 76.2 to 74.3, and the BIC drops from 79.1 to 79.6. *age* and *class2* are still the only non-significant predictors. The change in information statistics is not great, and indicates that the use of a zero-inflated beta binomial model may not be warranted. We drop the adjustment to the standard errors and remodel the data, calling for a Vuong test.

```
. zibbin survive age sex class2 class1, n(cases) eform nolog inflate(age sex class2
class1) vuong zib

Vuong test of zibb vs. standard beta binomial:    z =     1.49   Pr>z = 0.0684
```

Our suspicions appear to be correct. the Vuong test informs us that the standard beta binomial model is preferred for this data.

The beta binomial is an important model, and should be considered for all overdispersed logistic models. In addition, for binomial models with *probit* and *complementary loglog* links, the *betabin* and *zibbin* commands also have options for these models. We may therefore test to determine if the data is better modeled as a beta binomial with a complementary loglog link. Information criterion tests can be used to determine fit, as well as a goodness-of-link test.

REFERENCES

Hardin, J.W and J.W Hilbe (2012), *Generalized Linear Model and Extensions, 3rd edition*, Stata Press/Chapman & Hall/CRC

Hardin, J.W .and J.M.Hilbe (2013), Estimation and Testing of Binomial and Beta Binomial Regression Models with and without Zero Inflation *Stata Journal* Vol- ,--<forthcoming>

Hilbe, J.M (2011), *Negative Binomial Regression, 2nd edition*, Cambridge: Cambridge University Press

Hilbe, J.M. (2014), *Modeling Count Data*, Cambridge:Cambridge University Press

Hilbe, J.M and A.P Robinson (2012). *COUNT* package, CRAN

Hilbe, JM and A.P Robinson (2013) *msme* package, CRAN,

Rigby, B and M Stasinopoulos (2012), *Gamlss package*, 2nd edition, CRAN,

Zuur, A.F, J.M Hilbe, and E.N Zuur (2013), *A Beginner's Guide to GLM and GLMM with R: A frequentist and Bayesian perspective for ecologists.* Highland
http://www.highstat.com/BGGLM.htm