Paper 325-2012

Inflated Beta Regression: Zero, One, and Everything in Between

Christopher J. Swearingen[1], Maria S. Melguizo Castro[1], and Zoran Bursac[2]
[1]Biostatistics Program, Department of Pediatrics
[2]Biostatistics, College of Public Health
University of Arkansas for Medical Sciences, Little Rock, AR

ABSTRACT

Beta Regression, an extension of generalized linear models, can estimate the effect of explanatory variables on data falling within the (0,1) interval. Recent developments in Beta Regression theory extend the support interval to now include 0 and 1. The %Beta_Regression macro is updated to now allow for Zero-One Inflated Beta Regression.

KEY WORDS: Beta Regression, Inflated Beta Regression, PROC NLMIXED, Macro, Generalized Linear Model

INTRODUCTION

Beta Regression, an extension of generalized linear model (GLM) theory, primary assumption holds that a continuous, percentage-scaled dependent variable (i.e. ranging from 0 to 1, non-inclusive) can be characterized by the Beta distribution [1-4]. Unimodal and bimodal densities with varying severity of skewness can be characterized by the Beta distribution, a fact that gives Beta Regression incredible flexibility in modeling dependent variables for which normalizing transformations are impossible. Beta Regression further assumes that changes in the dependent variable's mean, precision (scaling factor related to variance), or both can be associated with changes in explanatory variables.

While Beta Regression is a very flexible and useful regression model, one limitation of the regression model is that true observations existing at either 0 or 1 must be scaled away from these values. Recent theoretical work on Beta Regression now incorporates a mixture model to estimate observations existing at either 0 or 1 [5], and through a slight modification, we introduce a general model of Beta Regression that simultaneously estimates probability masses at both 0 and 1. Inflated Beta Regression incorporates the existing Beta distribution with degenerate distributions to model the extreme values, thereby allowing for complete modeling of the entire continuous percentage space.

Based upon the existing Beta Regression macro[6], we introduce Zero-Inflated, One-Inflated and Zero-One-Inflated Beta Regression macros using SAS® PROC NLMIXED. Moreover, we further develop the macro call to execute the appropriate Beta Regression model based upon the parameterization submitted. For each new inflated class, an example of the macro usage and brief description of the data analysis is presented.

REGRESSION ON A BETA DISTRIBUTED DEPENDENT VARIABLE

Beta Regression assumes the dependent variable can be assumed to follow a Beta distribution with two parameters μ and ϕ:

$$f(y;\mu,\phi)=\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1},\qquad 0<y<1$$

where $0\le\mu\le1$, $\phi>0$ and Γ is the gamma function[5-6]. This parameterization dictates that $E(y)=\mu$ and $Var(y)=\mu(1-\mu)/(\phi+1)$, in which the variance of the dependent variable is defined as a function of the distribution mean μ and the precision parameter ϕ. Extending GLM theory to accommodate this distribution, parameter estimates obtained in Beta Regression associate changes in the dependent variable's mean and/or precision as a function of explanatory variables [1,5-6].

Inflated beta distributions incorporate degenerate probability statements producing a mixture density. For Zero-Inflation, a new parameter π_0 is added to account for the probability of observations at zero. The subsequent mixture density is:

$$f(y;\pi_0,\mu,\phi)=\begin{cases}\pi_0, & \text{if } y=0\\ (1-\pi_0)f(y;\mu,\phi), & \text{if } 0<y<1\end{cases}$$

One-Inflation follows the same logical though, although here the new parameter π_1 is added to account for the probability of observations at one. The subsequent mixture density is:

$$f(y;\pi_1,\mu,\phi)=\begin{cases}(1-\pi_1)f(y;\mu,\phi), & \text{if } 0<y<1\\ \pi_1, & \text{if y = 1}\end{cases}$$

Finally, Zero-One-Inflated Beta Regression combines the two prior inflated densities into one density:

$$f(y;\pi_0,\pi_1,\mu,\phi)=\begin{cases}\pi_0, & \text{if y = 0}\\ (1-\pi_0)(1-\pi_1)f(y;\mu,\phi), & \text{if } 0<y<1\\ \pi_1, & \text{if y = 1}\end{cases}$$

It is important to note that this Zero-One-Inflated Beta Regression model differs slightly from previously published models [5]. However, our inflated density allows for distinct covariate and parameter estimate design matrices to be linked to each of the four parameters.

DEFINING THE MACRO CALL

```
%Macro Beta_Regression(Dataset,tech,details,mu_vars,phi_vars,zero_vars,one_vars,
        depvar);
   %if &zero_vars ne and &one_vars ne %then
      %Beta_Regression_Zero_One(&Dataset,&tech,&details,&mu_vars,&phi_vars,
         &zero_vars,&one_vars,&depvar);
   %else %if &zero_vars ne %then
      %Beta_Regression_Zero(&Dataset,&tech,&details,&mu_vars,&phi_vars,
         &zero_vars,&depvar);
   %else %if &one_vars ne %then
      %Beta_Regression_One(&Dataset,&tech,&details,&mu_vars,&phi_vars,
         &one_vars,&depvar);
   %else
      %Beta_Regression_Only(&Dataset,&tech,&details,&mu_vars,&phi_vars,&depvar);
%mend;
```

The macro *Beta_Regression* is specified in the following manner:

- Dataset – the LIBNAME.DATA file
- tech – allows for different optimization schemes to be used
- details – allows for other options to be specified
- mu_vars – variables modeling changes in mean
- phi_vars – variables modeling changes in precision
- zero_vars – variables predicting a response of zero
- one_vars – variables predicting a response of one
- depvar – the dependent variable scaled to a [0,1] interval

CREATING LABELS FOR PREDICTOR VARIABLES

```
%Macro Preprocessing(Vars,b0,xb2,b);

data HPG;
   %global &xb2;
   length  &xb2 $200.;
   &xb2=&b0;

   %if &Vars ne '' %then %do;
         %let n=1;
         var&n="%scan(&Vars,&n,' ')";
         %do %while( %scan(&Vars,&n,' ') ne  );
               %let n=%eval(&n+1);
               var&n="%scan(&Vars,&n,' ')";
         %end;
         %let n_1=%eval(&n-1);

         array xbv {*} $ var1--var&n_1;

         %do j=1 %to &n_1;
               &b&j= "&b&j";
         %end;
         %let one=1;
         array &b{*} $8 &b&one--&b&n_1 ;
         array p{1} $ 8 ('+');
         array m{1} $ 8 ('*');

         do i=1 to dim(xbv) while (xbv{i} ne '');
               &xb2= cats(of &xb2 p{1} &b{i} m{1} xbv{i});
         end;
   %end;
   call symput("&xb2",&xb2);
run;

%mend;
```

PROC NLMIXED requires the specification of parameter estimate labels for each variable regressed on the dependent variable. For example, “b1 *gender” identifies the parameter estimate “b1” quantifying the effect of “gender” in the model. For ease of use, a nested macro *Preprocessing* creates labels for the intercept and each predictor variable; this macro is unchanged from our previous publication [6].

If no variables are specified, the *Preprocessing* macro will return intercept labels. Mean covariates are specified as “b” parameter estimates, precision covariates are specified as “d” parameter estimates, zero-inflated covariates are specified as “zero” parameter estimates and one-inflated covarates are specified as “one” parameter estimates. . Each parameter estimate is labeled in ascending order corresponding to the listing order of the variables in the macro statement.

CREATION OF MODEL PREDICTION DATASET

```
%Macro Postprocessing(hat, predict, depvar);
   data &predict;
      retain record &depvar &hat;
      set &predict;
      if &depvar = . then &hat = .;
         else &hat = pred;
      record=_n_;
      keep record &depvar &hat;
   run;
%mend;
```

PROC NLMIXED utilizes the PREDICT statement to output the estimated linear prediction. However, prediction must be performed for each parameter specified in the model; for example, a Zero-Inflated model produces linear prediction estimates for the probability of zero, the mean and the precision parameters. For ease of use, each Beta Regression macro automatically predicts linear fits for each parameter and a nested macro *Postprocessing* creates one combined temporary datasets of all predictions.

IMPLEMENTATION OF ZERO-INFLATED BETA REGRESSION

```
%Macro Beta_Regression_Zero(Dataset,tech,details,mu_vars,phi_vars,zero_vars,depvar);

   %Preprocessing(&mu_vars,'b0',xb,b);
   %Preprocessing(&phi_vars,'d0',wd,d);
   %Preprocessing(&zero_vars,'zero0',zeroxb,zero);

   proc nlmixed data = &Dataset tech = &tech &details;
      pizero = exp(&zeroxb)/(1 + exp(&zeroxb));
      mu = exp(&xb)/(1 + exp(&xb));
      phi = exp(&wd);
      w = mu*phi;
      t = phi - mu*phi;
      if (&depvar = 0) then
            ll = log(pizero);
      else ll = lgamma(w+t) - lgamma(w) - lgamma(t) + ((w-1)*log(&depvar)) +
            ((t-1)*log(1 - &depvar)) + log(1-pizero);
      model &depvar ~ general(ll);
      predict mu out=mu_results (keep=&depvar pred);
      predict phi out=phi_results (keep=&depvar pred);
      predict pizero out=pizero_results (keep=&depvar pred);
   run;
```

```
   %Postprocessing(mu_hat,mu_results,&depvar);
   %Postprocessing(phi_hat,phi_results,&depvar);
   %Postprocessing(pizero_hat,pizero_results,&depvar);

   data prediction;
      merge mu_results phi_results pizero_results;
            by record;
   run;

%mend;
```

Since PROC NLMIXED can maximize any programmable likelihood, the Zero-Inflated Beta Regression model is programmed exactly as the density function would suggest. Once called, the macro passes covariates to the *Preprocessing* macro for labeling. The dependent variable is subsequently modeled as a degenerate function of the constructed design matrices through the usage of the canonical link functions for each parameter. Once maximized, the linear prediction for each parameter is passed to the *Postprocessing* macro and then merged into the model's *Prediction* dataset.

EXAMPLE – ANALYSIS OF BRAZILIAN TRAFFIC ACCIDENT MORTALITY

```
%Beta_Regression(zoib.traffic,trureg, ,lnpop prop2029 idhe,lnpop prop2029 idhe,lnpop
         prop2029 idhe,, pro00);
```

Ospina and Ferrari illustrate the utility of Zero-Inflated Beta Regression in an analysis of traffic accident mortality from 200 randomly selected cities in Brazil [5]. The dependent variable is the proportion of deaths caused in 2002 by traffic accidents (*pro00*); 39% of reported deaths were not caused by traffic accidents (i.e. zero percent). Explanatory variables used to predict the percentage of traffic accident mortality included log-transformed population (*lnpop*), proportion of residents between 20 and 29 years of age (*prop2029*), and an index measure of education within each city (*idhe*).

Table 1. Results of Brazilian Traffic Accident Mortality by Software Package

	Results from Ospina & Ferrari [5]				SAS v9.3 *%Beta_Regression* Macro			
		Estimate	Standard Error	P		Estimate	Standard Error	P
π_0	intercept	27.27	4.63	1.73e−08	zero0	27.27	4.35	<.0001
	lnpop	−1.17	0.27	2.07e−05	zero1	-1.17	0.26	<.0001
	prop2029	−48.06	17.46	6.49e−03	zero2	-48.06	17.20	0.0057
	idhe	−11.34	4.01	5.13e−03	zero3	-11.35	3.93	0.0043
		Estimate	Standard Error	P		Estimate	Standard Error	P
μ	intercept	−4.72	1.20	1.11e−04	b0	-4.73	1.20	0.0001
	lnpop	−0.53	0.05	7.14e−17	b1	-0.53	0.06	<.0001
	prop2029	27.68	6.37	2.33e−05	b2	27.73	6.38	<.0001
	idhe	3.1	1.44	3.28e−02	b3	3.10	1.44	0.0332
		Estimate	Standard Error	P		Estimate	Standard Error	P
ϕ	intercept	9.46	3.25	4.08e−03	d0	9.48	3.27	0.0042
	lnpop	0.47	0.10	8.54e−06	d1	0.48	0.10	<.0001
	prop2029	−28.34	16.63	9.01e−02	d2	-28.46	16.60	0.0880
	idhe	−6.70	3.90	8.78e−02	d3	-6.69	3.91	0.0883

As can be seen in the macro call above, all three design matrices contain the same explanatory variables. Published results from Ospina and Ferrari's *R* function are summarized next to the SAS results (**Table 1**), indicating that the same inference can be made using the *%Beta_Regression* macro in SAS.

IMPLEMENTATION OF ONE-INFLATED BETA REGRESSION

```
%Macro Beta_Regression_One(Dataset,tech,details,mu_vars,phi_vars,one_vars,depvar);

   %Preprocessing(&mu_vars,'b0',xb,b);
   %Preprocessing(&phi_vars,'d0',wd,d);
   %Preprocessing(&one_vars,'one0',onexb,one);

   proc nlmixed data = &Dataset tech = &tech &details;
      pione = exp(&onexb)/(1 + exp(&onexb));
      mu = exp(&xb)/(1 + exp(&xb));
      phi = exp(&wd);
      w = mu*phi;
      t = phi - mu*phi;
      if (&depvar = 1) then
            ll = log(pione);
      else ll = lgamma(w+t) - lgamma(w) - lgamma(t) + ((w-1)*log(&depvar)) +
            ((t-1)*log(1 - &depvar)) + log(1-pione);
      model &depvar ~ general(ll);
      predict mu out=mu_results (keep=&depvar pred);
      predict phi out=phi_results (keep=&depvar pred);
      predict pione out=pione_results (keep=&depvar pred);
   run;

   %Postprocessing(mu_hat,mu_results,&depvar);
   %Postprocessing(phi_hat,phi_results,&depvar);
   %Postprocessing(pione_hat,pione_results,&depvar);

   data prediction;
      merge mu_results phi_results pione_results;
            by record;
   run;

%mend;
```

One-Inflated Beta Regression differs from Zero-Inflated Beta Regression only in the partitioning of the degenerate probability mass.

EXAMPLE – ANALYSIS OF BARTHEL INDEX IN NINDS RT-PA CLINICAL TRIAL

```
%Beta_Regression(zoib.tpa_data,trureg, ,tpa decade,decade, ,tpa decade,
          oi_barthel12);
```

As detailed in our previous paper [6], our data example comes from the two National Institute of Neurological Diseases and Stroke (NINDS) "recombinant tissue-type plasminogen activator" (rt-PA) trials [7]. The primary aim of the double-blind trials was to assess the effectiveness of in treating cerebral stroke secondary to artery thrombosis (clot restricting or stopping blood flow) within three hours of symptom

onset. The trials were designed to collect the same data using the same procedures, but each was powered to test a different primary endpoint. Additional details of the NINDS rt-PA clinical trials are available [7].

The "Part 2" trial assessed the combined functional outcomes at three months as measured by the Barthel Index [8], a clinical outcome scale ranging from [0-100] that assesses various activities of daily living achieved by an individual post-stroke, as well as three clinical outcomes [7]. Data was also collected at twelve months post-stroke, in which a majority of study participants achieved functional independence as measured by the raw Barthel Index scores (**Figure 1**). The resulting distribution is severely negatively skewed with signficant mass at the maximum score.

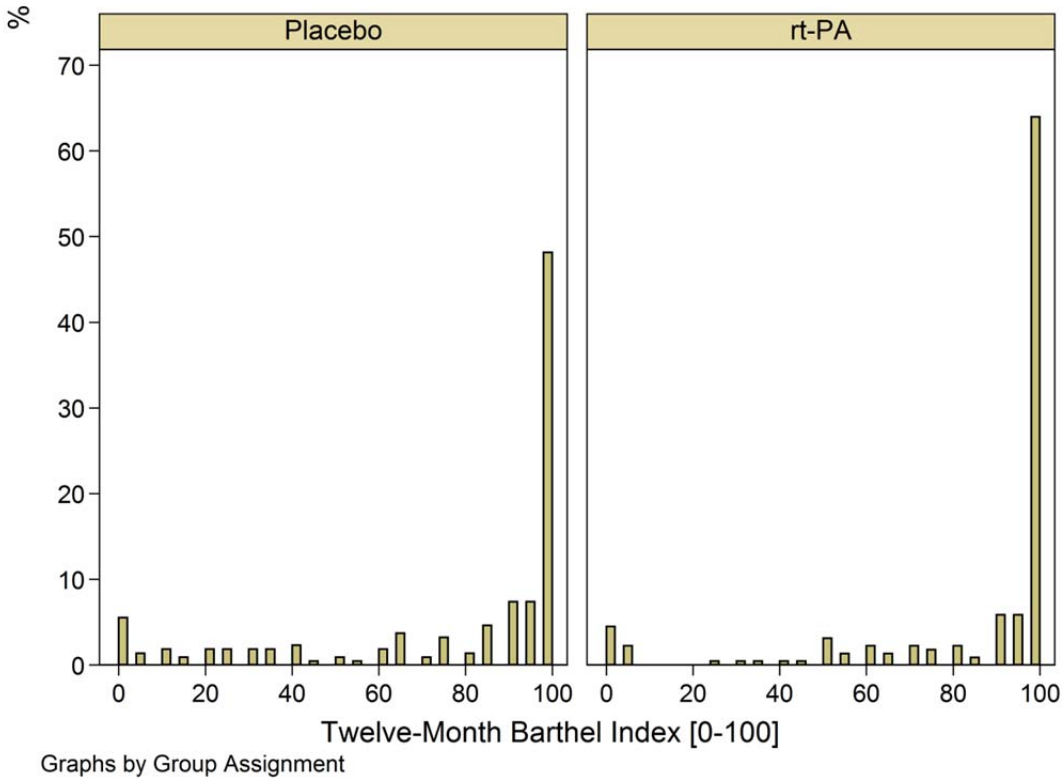

Figure 1. Histogram of Barthel Index Outcome at Twelve Months by Treatment Group.

Dividing the raw Barthel Index scores by 100 would result in a distribution that could be supported by the Beta domain. The standard transformation Z = [Y'(N-1) + 0.5] / N where Y' = [Y - Minimum(Y)] / [Maximum(Y) - Minimum(Y)], N is the number of total observations and Y is the original scale measurement [9] was used for all responses less than 100; responses that were 100 were scaled to 1.

Results from the analysis are summarized in **Table 2**. Examining first the one inflated results, the parameter estimate "*one1*" relates to the indicator variable for treatment group (rt-PA = 1, placebo = 0). These results indicate that the odds of achieving a Barthel Index score of 100 in rt-PA group was 106 times more than the odds of the placebo group (OR = exp(0.72) = 2.06, 95% CI (1.39, 3.04), $p < 0.001$). The treatment group is not associated with any further significant findings when examining the mean or precision results, although age is significantly associated with all three parameters. In terms of the mean and probability parameters, increasing age is associated with decreased likelihood of successful treatment. In terms of precision, increasing age is associated with decreased precision (i.e. increased variability).

Table 2. Barthel Index Analysis using One-Inflated Beta Regression

		Estimate	Standard Error	P
π_1	one0	1.85	0.60	0.0021
	one1	0.72	0.20	0.0003
	one2	-0.30	0.09	0.0010

		Estimate	Standard Error	P
μ	b0	3.10	0.53	<.0001
	b1	0.17	0.17	0.3355
	b2	-0.44	0.08	<.0001

		Estimate	Standard Error	P
ϕ	d0	2.42	0.64	0.0002
	d1	-0.29	0.09	0.0018

IMPLEMENTATION OF ZERO-ONE INFLATED BETA REGRESSION

```
%Macro Beta_Regression_Zero_One(Dataset,tech,details,mu_vars,phi_vars,zero_vars,
          one_vars,depvar);

   %Preprocessing(&mu_vars,'b0',xb,b);
   %Preprocessing(&phi_vars,'d0',wd,d);
   %Preprocessing(&zero_vars,'zero0',zeroxb,zero);
   %Preprocessing(&one_vars,'one0',onexb,one);

   proc nlmixed data = &Dataset tech = &tech &details;
      pizero = exp(&zeroxb)/(1 + exp(&zeroxb));
      pione = exp(&onexb)/(1 + exp(&onexb));
      mu = exp(&xb)/(1 + exp(&xb));
      phi = exp(&wd);
      w = mu*phi;
      t = phi - mu*phi;
      if (&depvar = 0) then
      ll = log(pizero);
            else if (&depvar = 1) then
            ll = log(pione);
      else ll = lgamma(w+t) - lgamma(w) - lgamma(t) + ((w-1)*log(&depvar)) +
            ((t-1)*log(1 - &depvar)) + log(1-pizero) + log(1-pione);
      model &depvar ~ general(ll);
      predict mu out=mu_results (keep=&depvar pred);
      predict phi out=phi_results (keep=&depvar pred);
      predict pizero out=pizero_results (keep=&depvar pred);
      predict pione out=pione_results (keep=&depvar pred);
   run;

   %Postprocessing(mu_hat,mu_results,&depvar);
   %Postprocessing(phi_hat,phi_results,&depvar);
   %Postprocessing(pizero_hat,pizero_results,&depvar);
   %Postprocessing(pione_hat,pione_results,&depvar);
```

```
   data prediction;
      merge mu_results phi_results pizero_results pione_results;
            by record;
   run;

%mend;
```

Zero-One-Inflated Beta Regression combines the probability mass estimation for both zero and one into one likelihood.

EXAMPLE – ANALYSIS OF BARTHEL INDEX IN NINDS RT-PA CLINICAL TRIAL

```
%Beta_Regression(zoib.tpa_data,trureg, ,tpa decade,decade,tpa decade,tpa decade
         ,zoi_barthel12);
```

It should be noted that all analysis of the rt-PA trial present to date has only examined the "per protocol" observations; those study participants that died prior to the twelve month observation are **not** included in the analysis. An "intention-to-treat" analysis was also performed, imputing the worst observation (i.e. zero) for those participants who died during the trial [7]. Subsequently, this imputation led to the Barthel Index having two separate probability masses as zero and one (**Figure 2**). Fortunately, a simple transformation (dividing by 100) yields a dependent variable supported by the zero-one-inflated Beta.

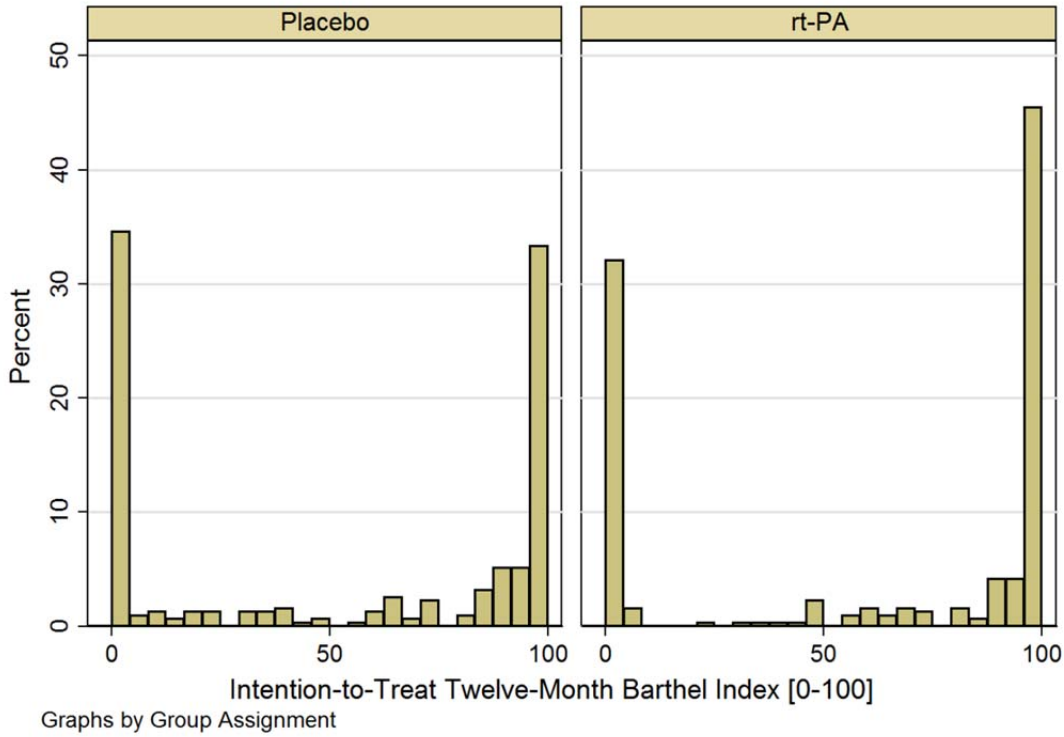

Figure 2. Histogram of Intention-to-Treat Barthel Index Outcome at Twelve Months by Treatment Group.

Table 3. Intention-to-Treat Barthel Index Analysis using Zero-One-Inflated Beta Regression

	Estimate	Standard Error	P
zero0	-2.67	0.70	0.0001
zero1	0.16	0.22	0.4483
zero2	0.41	0.10	<.0001

	Estimate	Standard Error	P
one0	1.48	0.61	0.0147
one1	0.72	0.21	0.0005
one2	-0.23	0.09	0.0145

	Estimate	Standard Error	P
b0	2.96	0.54	<.0001
b1	0.17	0.17	0.3098
b2	-0.36	0.08	<.0001

	Estimate	Standard Error	P
d0	2.8262	0.80	0.0004
d1	-0.2805	0.12	0.0170

Results from the intention-to-treat analysis are summarized in **Table 3**. As was seen in the previous analysis, age is significantly associated with each estimated parameter. Here, however, the inference is different for the zero-inflated parameter, as increasing age (*zero2*) is associated with increased odds of death. Specifically, for every unit increase in a participant's age (in decades), the odds of death increase 50% (OR = exp(0.41) = 1.50, 95% CI (1.23, 1.83), p<0.001). Inference for the other parameters remains the same.

The only treatment parameter estimate significantly associated with the outcome is again in the one-inflation part of the model. These results are very similar to the one-inflated model of the per-protocol analysis, indicating that the odds of achieving a Barthel Index score of 100 in rt-PA group was 105 times more than the odds of the placebo group (OR = exp(0.72) = 2.05, 95% CI (1.37, 3.07), p < 0.001).

CONCLUSION

It has been shown how Beta Regression can provide utility in modeling continuous percentage dependent variables. Current development in Beta Regression now allows for the accounting of observations existing at zero or one. PROC NLMIXED can be utilized to the degenerate inflated functions, and the %Beta_Regression macro has been updated to implement all of the inflated Beta Regression models in a straightforward, easy-to-use manner..

REFERENCES

1. Paolino P. Maximum likelihood estimation of models with beta-distributed dependent variables. *Political Analysis,* 2001; 9:325-346.
2. Kieschnick R, McCullough BD. Regression analysis of variates observed on (0,1): percentages, proportions and fractions. *Statistical Modelling*, 2003; 3:193-213.
3. Ferrari SLP, Cribari-Neto F. Beta regression for modelling rates and proportions. *J Applied Statistics*, 2004; 31:799-815.
4. Ospina R, Ferrari SLP. Inflated beta distributions. *Stat Papers*, 2010; 51: 111-126.
5. Ospina R, Ferrari SLP. A general class of zero-or-one inflated beta regression models. *Comp Stat Data Analysis*, 2012; 56:1609-1623.
6. Swearingen CJ, Melguizo castro MS, and Bursac Z. Modeling percentage outcomes: The %Beta_Regression macro. SAS® Global Forum Proceedings 2011; Paper 335:1–12. http://goo.gl/n2M0V
7. NINDS rt-PA Stroke Study Group. Tissue plasminogen activator for acute ischemic stroke. *New England J Med*. 1995; 333:1581-1587.
8. Mahoney FI, Barthel DW. Functional evaluation: the Barthel Index. *Maryland State Med J*, 1965; 14:61-65.
9. Smithson M, Verkuilen J. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psych Methods*, 2006; 11:54-71.

ACKNOWLEDGEMENTS

The NINDS rt-PA dataset is a public use dataset and can be obtained through the National Technical Information Service (http://www.ntis.gov/search/product.aspx?ABBR=PB2006500032). The authors appreciate the willingness of Profs. Ospina and Ferrari in sharing their data.

RECOMMENDED READING

For more details on PROC NLMIXED, consult SAS.com

http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#nlmixed_toc.htm

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Christopher J. Swearingen PhD
UAMS Department of Pediatrics
1 Children's Way, Biostatistics Slot 512-43
Little Rock, AR 72202
Phone: 501-364-6639
Fax: 501-364-1431
E-mail: cswearingen@uams.edu
Web: www.arpediatrics.org/research/biostatistics

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.
Other brand and product names are trademarks of their respective companies.