Paper 335-2011

Modeling Percentage Outcomes: The *%Beta_Regression* Macro

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

Data are often encountered that fail the normality assumption, creating a dilemma between implementing the most appropriate analysis while maintaining appropriate inference. A recent extension of generalized linear models includes support for the Beta distribution, a flexible and accommodating distribution. SAS® can implement "Beta Regression" through PROC NLMIXED, allowing the model's likelihood to be specified in terms of its mean and a variance component, providing a flexible modeling tool and providing intuitive inference. We provide a brief theoretical introduction to Beta Regression as well as a macro that implements Beta Regression and provides residuals plots for model fit diagnostics.

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

INTRODUCTION

The Beta distribution is a highly flexible parametric distribution, able to accommodate both unimodal and bimodal densities with varying severity of skewness (**Figure 1**). It is this flexibility of the Beta distribution that is of particular importance as it can be used estimating distributions with intractable skew, making normalizing transformations impossible. Further, the support for the Beta distribution lies within the [0, 1] closed interval, leading Beta distributions to be a natural choice for characterizing percentages.

Expanding the generalized linear model (GLM) to regress predictor variables on a dependent variable that is considered to be marginally distributed following a Beta distribution is referred to as Beta Regression [1-4]. Beta Regression has been shown to provide more accurate and efficient parameter estimates than ordinary least squares regression when the dependent variable follows a skewed underlying distribution [1] or when there is underlying heteroskedasticity [2]. Recent developments have provided for standardized residuals to be calculated when a Beta Regression model is specified with both mean and precision covariates [4].

While the Beta distribution is characterized by two shape parameters, a simple algebraic transformation of these parameters defines the Beta distribution in terms of its mean and a scaling, or precision, parameter [1,4]. In this manner, Beta Regression can provide parameter estimates associated with, and thereby allowing inference to be made with respect to changes in the dependent variable's mean and precision. To our knowledge, previous discussion of regression on a Beta distributed dependent variable in this forum has been limited to models utilizing the shape parameters, not the mean and precision parameters [5-6]. In this paper, we present the Beta Regression model using the mean and precision parameterization as well as introduce a custom macro implementing Beta Regression within PROC NLMIXED based upon an existing example template [7]. Finally, we conclude with an example of an analysis using Beta Regression.

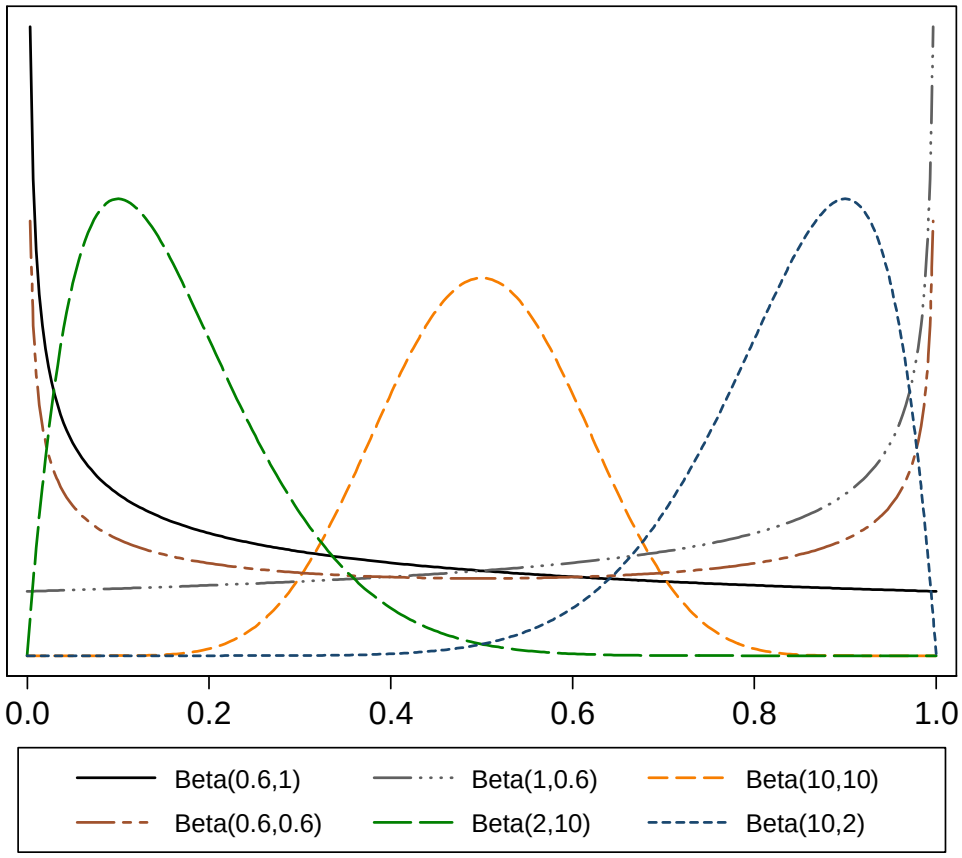

Figure 1. Examples of Beta Distribution by Varying Shape Parameters.

REGRESSION ON A BETA DISTRIBUTED DEPENDENT VARIABLE

The standard, two-parameter Beta distribution is defined with two parameters, ω and τ:

$$f(y \mid \omega, \tau) = \frac{\Gamma(\omega+\tau)}{\Gamma(\omega)\Gamma(\tau)} y^{\omega-1}(1-y)^{\tau-1}$$

where $\omega, \tau > 0$, $0 \le y \le 1$, and Γ is the gamma function[6]. The mean of the Beta distribution is given as:

$$E(Y) = \frac{\omega}{\omega+\tau} \equiv \mu$$

and its variance:

$$Var(Y) = \frac{\omega\tau}{(\omega+\tau)^2(\omega+\tau+1)} = \frac{\mu(1-\mu)}{\phi+1}$$

where $\phi \equiv \omega+\tau$ and given the algebraic substitution $\tau(\omega+\tau)^{-1} = 1-\mu$. This parameterization dictates that $0 < \mu < 1$ and $\phi > 0$ as it can be easily shown that $\omega = \mu\phi > 0$ and $\tau = \phi(1-\mu) > 0$. Assuming

this mean and precision parameterization, the setup of the GLM is relatively straightforward as predicting variables will be regressed upon each parameter.

Let $\mathbf{X}$ be a matrix of size $n \times (p+1)$ independent predicting variables to regress upon the mean parameter μ, with $\boldsymbol{\beta}$ the corresponding regression coefficient vector of length $p+1$ to be estimated. Considering that the mean parameter is restricted to the open interval (0,1), a link function that maps the parameter from the interval into the real number space is needed. While there are several choices that can be used for this link function, the logit function is the canonical link function and returns parameter estimates in terms of log-odds. Subsequently, the function $h(\mu)$ is defined:

$$h(\mu) = \text{logit}(\mu) = ln\left(\frac{\mu}{1-\mu}\right) = \mathbf{X\beta} \longrightarrow \mu = \frac{exp(\mathbf{X\beta})}{1+exp(\mathbf{X\beta})}.$$

In a similar manner, let $\mathbf{W}$ be a matrix of size $n \times (p+1)$ independent predicting variables to regress upon the precision parameter ϕ, with $\boldsymbol{\delta}$ the corresponding regression coefficient vector. $\mathbf{X}$ and $\mathbf{W}$ are separate design matrices with respect to the linear model algebra, but do not necessarily have to be distinct. Since the precision parameter must be strictly positive, again a link function that maps the parameter from the restricted space into the real number space is needed. Here, the canonical link is the natural logarithmic link function, leading to the function $g(\phi)$ as:

$$g(\phi) = ln(\phi) = \mathbf{W\delta} \longrightarrow \phi = exp(\mathbf{W\delta}).$$

Relating the two parameter functions to the dependent variable, $\mathbf{Y} = f\left(h'(\mathbf{X\beta}), g'(\mathbf{W\delta})\right)$, leads to the likelihood of $\mathbf{Y}$ in terms of the regressors:

$$L(\boldsymbol{\beta}, \boldsymbol{\delta}; \mathbf{Y}, \mathbf{X}, \mathbf{W}) = \frac{\Gamma(\exp(\mathbf{W\delta}))}{\Gamma(s)\Gamma(t)} \mathbf{Y}^{s-1}(1-\mathbf{Y})^{t-1},$$

$$s = \frac{\exp(\mathbf{X\beta} + \mathbf{W\delta})}{1+\exp(\mathbf{X\beta})} \quad t = \frac{\exp(\mathbf{W\delta})}{1+\exp(\mathbf{X\beta})}.$$

Understanding the GLM setup of Beta Regression allows its specification within PROC NLMIXED. Based upon an existing code template [7], we developed the following macro to automate the execution of Beta Regression as well as produce residual plots.

DEFINING THE MACRO CALL

```
%Macro Beta_Regression(Dataset,tech, details, Vars, Vars2, depvar,
   residuals, gpath );
```

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
- Vars – list of mean covariates or double single quotes [‘’] if none

- Vars2 – list of precision covariates or double single quotes [‘’] if none
- depvar – the dependent variable scaled into the (0,1) open interval
- residuals – specify “1” to generate Pearson and Standardized Residual plots
- gpath – specify output directory for residual plots.

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

IMPLEMENTING BETA REGRESSION

PROC NLMIXED provides maximum likelihood estimation for any programmable likelihood. The Beta Regression likelihood, following the derivation above, can be specified as a GENERAL likelihood. The piecewise specification of the mean and precision parameters is not required, but creates easy to read and interpret code [6].

```
%Macro Beta_Regression(Dataset,tech,details,Vars,Vars2,depvar,
    residuals, gpath);

ODS LISTING;
%Preprocessing(&Vars,'b0',xb,b);
%Preprocessing(&Vars2,'d0',wd,d);

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

%if &residuals = 1 %then %Beta_Regression_Residuals;

%mend;
```

Moreover, using the PREDICT statement to output the mean and precision parameter estimates for residual examination is made easier with the piecewise specification of the Beta Regression likelihood. Prediction must be performed in two statements, one for each parameter, leading to the generation of two temporary datasets.

GENERATING AND PLOTTING THE RESIDUALS

Pearson residuals are calculated by standard definition, taking the difference between actual and fitted values and dividing by the estimated standard deviation. In terms of Beta Regression, this residual is:

$$r_i = \frac{y_i - \widehat{\mu_i}}{\sqrt{\widehat{\mu_i}(1-\widehat{\mu_i})(\widehat{\phi_i}+1)^{-1}}}.$$

Considering that the Beta distribution is supported within the (0,1) interval, the Pearson residuals are not necessarily expected to be normal and centered around zero. The standardized residuals [4] account for the (0,1) mean parameter space by calculating the difference between the logit-transformed dependent

```
%Macro Beta_Regression_Residuals;

data mu_results;
    set mu_results;
    if &depvar = . then mu_hat = .;
          else mu_hat = pred;
    record=_n_;
    keep record &depvar mu_hat;
run;

data phi_results;
    set phi_results;
    if &depvar = . then phi_hat = .;
          else phi_hat = pred;
    record=_n_;
    keep record &depvar phi_hat;
run;

data residual;
    merge mu_results phi_results;
    by record;
    pearson = (&depvar - mu_hat) /
          sqrt( (mu_hat * (1-mu_hat)) / (phi_hat + 1) );
    mustar = digamma(mu_hat * phi_hat) –
          digamma( (1-mu_hat) * phi_hat);
    astar = trigamma(mu_hat * phi_hat) +
          trigamma( (1-mu_hat) * phi_hat);
    logit_y = log(&depvar / (1 - &depvar) );
    scoreres = ( logit_y - mustar) / sqrt(astar);
run;

proc rank data=residual out=qqplot normal=blom;
    var pearson scoreres;
    ranks pearsonrank scorerank;
run;

proc sql;
    Create table Range As
    select      ceil(abs(max(pearsonrank))) as Max1,
                ceil(abs(min(pearsonrank))) as Min1,
                ceil(abs(max(scorerank))) as Max2,
                ceil(abs(min(scorerank))) as Min2
    from qqplot ;

Data Range;
    set Range;
    Range1=Max(Max1,Min1);
    Range2=Max(Max2,Min2);
    call symput("Range1",Range1);
    call symput("Range2",Range2);
run;
```

```
ods html style= Journal;
ods graphics on / reset scale=off;
ods listing gpath=&gpath;

goptions
    reset=all
    device=png
    ftext='Arial' htext=4 gunit=pct
    hsize = 5
    vsize = 5
    cback=white
    noborder
    gsfname=outgraph
    gsfmode=replace;

symbol1 interpol=none cv=black value=circle h=3;
symbol2 interpol=join cv=black value=none h=3;

axis1 label=("Inverse Normal")
    order=(-&Range1 to &Range1 by 1)
    minor=none;

axis2 label=("Inverse Normal")
    order=(-&Range2 to &Range2 by 1)
    minor=none;

axis3 label=(angle=90 "Pearson Residual")
    order=(-&Range1 to &Range1 by 1)
    minor=none;

axis4 label=(angle=90 "Standardized Residual")
    order=(-&Range2 to &Range2 by 1)
    minor=none;

filename outgraph '.\pearson_qqplot.png';
proc gplot data=qqplot;
    plot pearson*pearsonrank  pearsonrank*pearsonrank / overlay
                haxis = axis vaxis = axis3;
run;

filename outgraph '.\standardized_qqplot.png';
proc gplot data=qqplot;
    plot scoreres*scorerank  scorerank*scorerank / overlay
                haxis = axis2 vaxis = axis4;
run;

ods _all_ close;
ods graphics off;

%mend;    /* End of Macro Beta_Regression_Residuals*/
```

variable and the mean and precision parameter estimates. Once calculated, PROC RANK is used to give the inverse quantiles associated with each residual. The ordered quantiles are passed to PROC SQL to determine the absolute minimum and maximum value for subsequent usage in setting the axis range in the quantile-quantile (QQ) plots. With all of the calculations complete, PROC GPLOT is used to create square QQ plots with an overlaid linear reference line. The GRAPHIC options are given as example only, and can be adjusted to suit the user's specific style.

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

Two concurrent randomized, placebo-controlled, double-blind clinical trials sponsored by the National Institute of Neurological Diseases and Stroke (NINDS) were conducted between January 1991 and October 1994 [8]. The primary aim of the trials was to assess the effectiveness of "recombinant tissue-type plasminogen activator" (rt-PA) in treating cerebral artery thrombosis (clot restricting or stopping blood flow) given within three hours of ischemic stroke onset , although safety of rt-PA administration was also of great importance [8]. 624 individuals with ischemic stroke participated in these trials.

The trials were designed to collect the same data using the same procedures, but each was powered to test a different primary endpoint. The Part 1 trial assessed twenty-four hour functional improvement as measured by the National Institutes of Health Stroke Scale (NIHSS) as its primary endpoint; the Part 2 trial assessed the combined functional outcomes at three months as measured by the NIHSS and other clinical outcomes [8]. Additional details of the NINDS rt-PA clinical trials are available [8].

A secondary clinical outcome in the NINDS rt-PA trials was the Barthel Index [9], a clinical outcome scale ranging from [0-100] that assesses various activities of daily living achieved by an individual post-stroke. If an individual can achieve independence in performing activities, the highest score is assigned. Since a majority of study participants achieved functional independence twelve months post-stroke, the resulting distribution of raw Barthel Index scores was severely skewed (**Figure 2**).

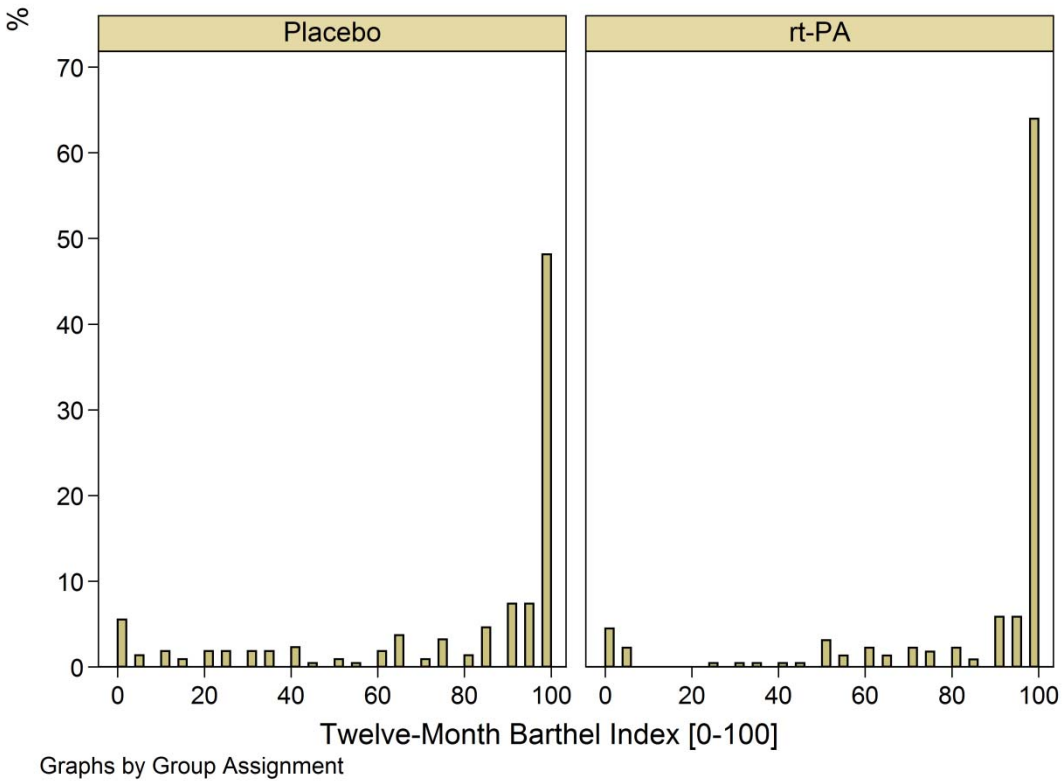

Figure 2. Histogram of Barthel Index Outcome at Twelve Months by Treatment Group.

The underlying distribution of the raw Barthel Index scores could be characterized as a Beta distribution provided the support interval was transformed. Given that the absolute minimum and maximum of the Barthel Index include 0 and 1, the transformation Z = [Y'(N-1) + 0.5] / N was used where Y' = [Y - Minimum(Y)] / [Maximum(Y) - Minimum(Y)], N is the number of total observations and Y is the original scale measurement [10]. This transformation shifted the scores into the open interval while maintaining the overall distributional shape.

MEAN-ONLY COVARIATE MODEL

```
%Beta_Regression(tpa.dataset,trureg,hess cov itdetails,tpa decade, '',
   barthel12m_raw, 1, "\\tpa\analysis\" );
```

A model estimating differences in twelve month Barthel Index scores between treatment groups while adjusting for age (in decades) was initially fit with the predictor variables only in the mean function, a mean-only covariate Beta Regression. The iteration history (ITDETAILS), optimized Hessian matrix (HESS), and covariance matrix of the parameter estimates (COV) were requested in the macro call, but are not discussed in detail here. Convergence criterion was satisfied in 10 iterations and covariance between the parameter estimates was negligible.

Parameter estimates for the mean-only model are summarized in **Table 1**. The mean function estimates intercept (b0), treatment group indicator (b1), and age (b2) are the log-odds of an increase in Barthel Index per unit change in the parameter. For example, the odds of having a higher Barthel Index score in the rt-PA group are 1.3 times those in the placebo group given this model (exp(b1)=1.326). The precision function intercept (d0) is in actuality the log-transformed precision parameter.

Table 1. Parameter estimates from Mean-Only Covariate Beta Regression

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha
b0	2.1319	0.3440	438	6.20	<.0001	0.05
b1	0.2824	0.1171	438	2.41	0.0163	0.05
b2	-0.1775	0.05137	438	-3.46	0.0006	0.05
d0	-0.2456	0.06347	438	-3.87	0.0001	0.05

Examining the residuals for the mean-only covariate model indicates the model may not be fitting the data well (**Figure 3**). While it is common for a Pearson Residual plot from a Beta Regression model to not be normal, the Standardized Residual plot should fit the QQ plot. The disjoint Standardized Residual plot further indicates that the mean-only covariate model should be improved.

MEAN AND PRECISION COVARIATE MODEL

```
%Beta_Regression(tpa.dataset,trureg,hess cov itdetails,tpa decade, decade,
   barthel12m_raw, 1, "\\tpa\analysis\" );
```

A mean and precision model was fit adding age to the precision function. Convergence criterion was satisfied in 21 iterations due to the added parameter. Parameter estimates for the mean and precision model are given in **Table 2**. While the parameter estimate for the treatment group (b1) does not change

dramatically, the age parameter (b2) increases almost three-fold. The age parameter in the precision function (d1) is also statistically significant, indicating that the log-precision decreases and concordantly overall variance increases as age increases.

Table 2. Parameter estimates from Mean and Precision Covariate Beta Regression

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha
b0	4.4509	0.4743	438	9.38	<.0001	0.05
b1	0.2902	0.1141	438	2.54	0.0113	0.05
b2	-0.5048	0.06958	438	-7.26	<.0001	0.05
d0	2.9000	0.4865	438	5.96	<.0001	0.05
d1	-0.4505	0.07002	438	-6.43	<.0001	0.05

The residual plots for the mean and precision model indicate an improved fit of the data as compared to the mean-only covariate model (**Figure 4**). Improved fit is also seen in comparing the Akaike Information Criteria between the two models (mean-only AIC = -1877, mean and precision AIC = -1917). While model fit has improved, the residual plots for the mean and precision model indicate that more improvement is still desired.

CONCLUSION

Beta Regression combines the strength of generalized linear models with the flexibility of the Beta distribution, providing utility in modeling skewed and/or bounded variables. While the Beta distribution can be fit in other procedures, PROC NLMIXED can be utilized to model changes in the dependent variable's mean and precision. We provide a set of macros that allow the user to implement Beta Regression in a straightforward manner, providing flexibility in using procedure options, maintaining ease of variable selection and assessing model fit through residual plots.

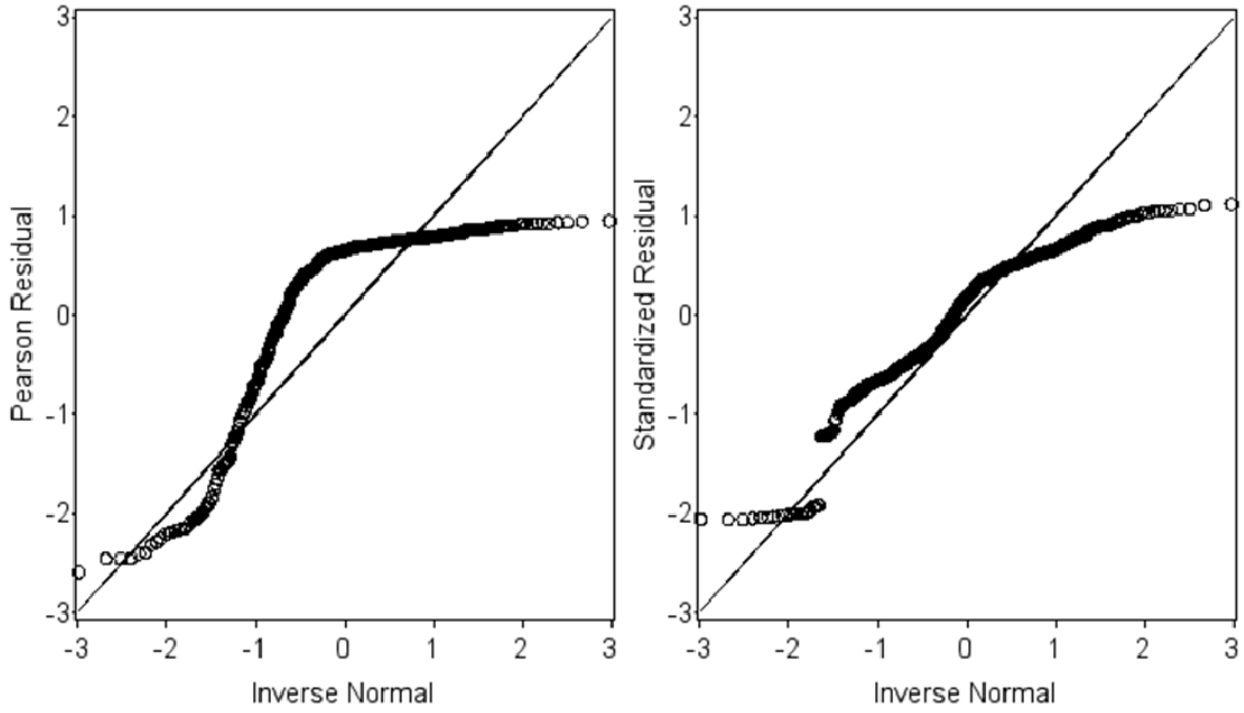

Figure 3. Residual Plots after Fitting Mean-Only Covariate Model

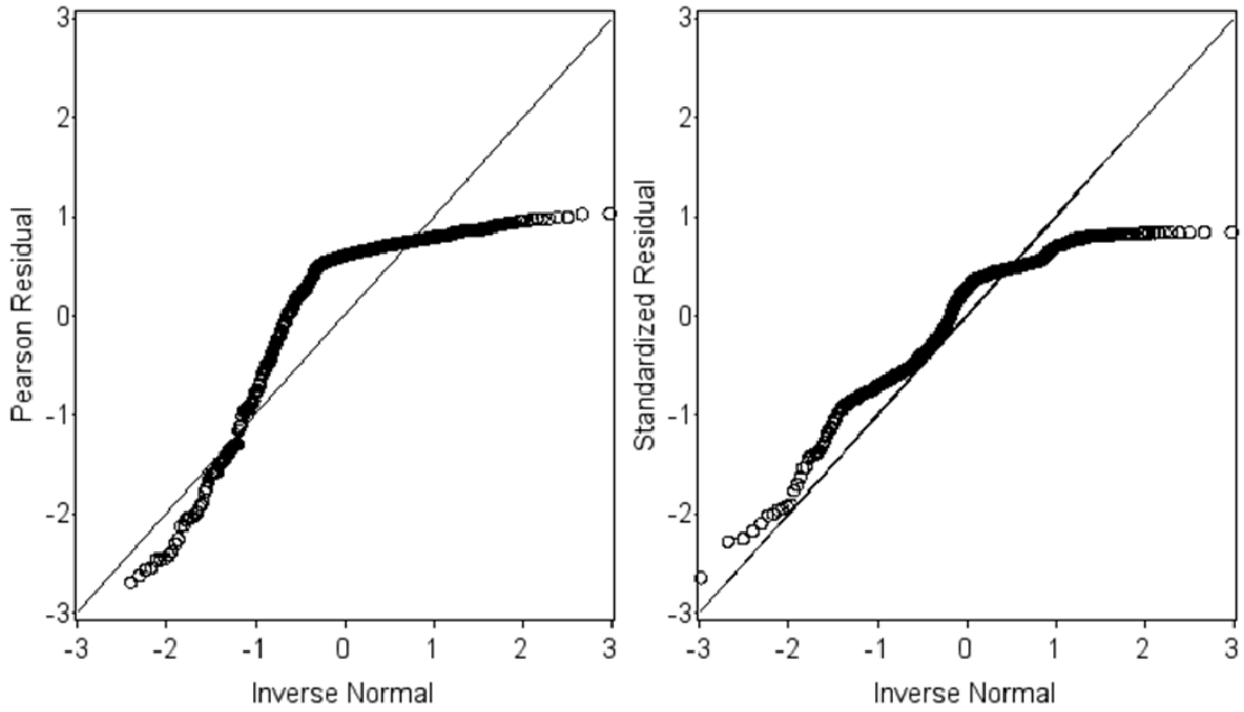

Figure 4. Residual Plots after Fitting Mean and Precision Covariate Model

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. Rocha AV, Simas AB. Influence diagnostics in a general class of beta regression models. *Test*, 2010; epub 23 March.
5. Dickey DA. Ideas and examples in generalized linear mixed models. *SAS Global Forum Proceedings*, 2010; 263:1-12.
6. Osborne JA. Estimating the false discovery rate using SAS. *SAS Users Group International Proceedings*, 2006; 190:1-10.
7. Verkuilen J. Examples from Smithson & Verkuilen (2005), "A Better Lemon Squeezer". http://psychology3.anu.edu.au/people/smithson/details/betareg/SAS_beta_regression.sas. Accessed Oct 22, 2010.
8. NINDS rt-PA Stroke Study Group. Tissue plasminogen activator for acute ischemic stroke. *New England J Med*. 1995; 333:1581-1587.
9. Mahoney FI, Barthel DW. Functional evaluation: the Barthel Index. *Maryland State Med J*, 1965; 14:61-65.
10. 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).

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