The Beta Regression Model: An Alternative Method for Political Science Research

Kaylee Hodgson

Brigham Young University

Abstract. This paper proposes beta regression as a tool to regress a dependent variable with the same bounds and shape as the beta distribution in political science research. I begin by introducing the concepts and advantages of beta regression. I then estimate the parameters of the beta regression model and employ hypothesis testing to test the significance of the estimated coefficients. A simulation study is performed to demonstrate the accuracy of these estimates. I compare the beta model to a linear model to show the advantages that beta regression has over the traditional linear regression model. Lastly, these methods are used to analyze a specific dataset on women in national government and the environment. I find that the beta regression model is a better fit for datasets with beta-distributed response variables. This can become a useful tool for political science researchers faced with beta-distributed data that does not meet the assumptions of a linear model.

Keywords: beta regression, rates, proportion, maximum likelihood estimators, hypothesis testing, R

1 Introduction

Political science research often evaluates data given in proportions and rates, such as voter demographic statistics or percentage of recipients of a program or policy. In political science research, linear regression has commonly been used to estimate regression models for datasets with numerical and categorical response variables that range between 0 and 1. The binomial and quasibinomial generalized linear regression model (using a logit link function) is commonly employed to predict this type of response variable. This linear model is useful for many types of data, such as categorical data coded as dichotomous variables. However, the assumptions of both the linear model and the binomial generalized linear model are not satisfied by rate and proportion data. In reference to the linear model, Cribari-Neto and Zeileis (2009) find that "Gaussian-based approximations for interval estimation and hypothesis testing [using the linear model] can be quite inaccurate in small samples." Rate and proportion data tends to be skewed and heteroskedastic. As Cribari-Neto and Zeileis (2009) discuss, proportion and rates data often do not satisfy the normality assumptions of linear regression models because "distributions of rates and proportions are typically asymmetric." Additionally, as Cribari-Neto and Zeileis (2009) discuss, this type of data "display[s] more variation around the mean and less variation as we approach the lower and upper limits of the standard unit interval." This heteroskedasticity of the data is a problem because the linear regression model assumes homoskedasticity, which leads to overestimation of the Pearson regression coefficient. While the the logistic model with a binomial distribution is often employed in these cases, the assumptions of proportions and rates data are still not satisfied because they tend to resemble a beta distribution, not a binomial. Additionally, this binomial-logistic regression model specifically does not allow for direct inferences regarding the expected value of the response variable from the regression parameters because its response variable is written as an odds ratio function in terms of the cumulative distribution function, $\tilde{y} = \log(F(y)/(1 - F(y)))$.

These inconsistencies in the assumptions of both the linear and the logistic-binomial model with the actual properties of proportions and rates data could result in inacurrate analyses and conclusions, specifically in small datasets. The common use of these models in political science to analyze rates and proportions is therefore a serious issue. This paper discusses the use of the beta regression model in place of the linear regression model, largely based on Ferrari and Zeileis's (2004) paper, which proposed that the beta regression model replace the linear regression in these cases. I specifically apply their proposal to political science data on women in government positions and the environment, and compare that analysis to a logit linear regression model.

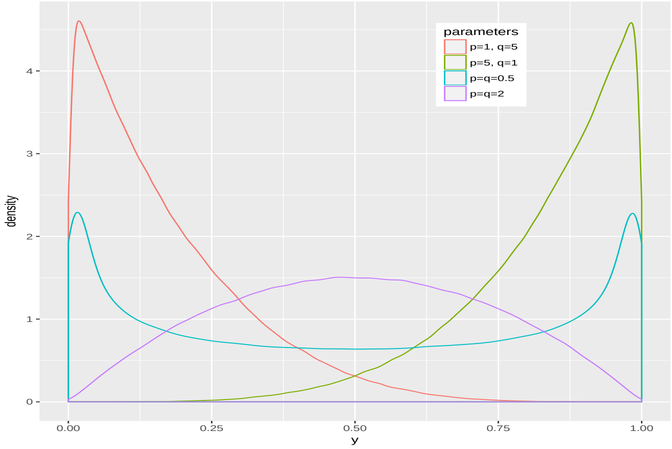
This paper will proceed as follows: section 1 will introduce and estimate the parameters of the beta regression model, discuss a method of hypothesis testing on these parameters, use these to perform probability testing. Section 2 will perform a simulation study to test the accuracy of the estimation method introduced in section 1, and section 3 will apply these methods to the political science dataset. Finally, the results and conclusion will be summarized in section 4.

2 Methodology

In 2004 Ferrari and Cribari-Neto introduced the idea of a beta regression model, "based on the assumption that the response is beta-distributed," to predict continuous response variables that fall in the same interval as the beta distribution (0,1), specifically rate and proportion data. This model assumes the response data is beta-distributed and heteroskedastic. The model is particularly appealing because its assumptions of the response data's shape adjust as the parameters adjust, consistent with the behavior of beta-distributed variables. This is demonstrated in Figure 1, which shows how the shape of the density for the beta distribution changes at different levels of parameters. These properties solve the issues inherent in the use of the linear regression model for beta-distributed response variables. The density function for the beta distribution is

$$f(y|p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}, 0 < y < 1. \tag{1}$$

Fig. 1. Beta Density with Different Parameters



Because direct inferences cannot be made regarding the mean of y from either of the parameters in the original beta distribution, Ferrari and Cribari-Neto (2004) introduced a new parameterization of the distribution with a location parameter, μ, and a precision parameter, ϕ. This new parameterization of the beta distribution allows for more direct inferences regarding the mean and variance of y in the regression model. The new parameters can be calculated from the original parameters as follows:

$$\mu = \frac{p}{(p+q)} \text{ and } \phi = p + q$$

The reparameterized model then becomes

$$f(y|\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, 0 < y < 1 \tag{2}$$

where $\Gamma(.)$ is the Gamma function, $0 < \mu < 1$ and $\phi > 0$. With this new parameterization, the expected values and variance of y can be written more directly as

$$E(y) = \mu \text{ and } Var(y) = \frac{\mu(1-\mu)}{1+\phi}$$

In this reparameterized model, Cribari-Neto and Zeileis (2009) “[n]ote that the variance of y is a function of μ which renders the regression model based on this parameterization naturally heteroskedastic.” The assumptions of heteroskedasticity in this beta distribution match the performance of the rates and proportions data. The location parameter, μ, is estimated by the beta regression model and ϕ is a fixed parameter. It is worth noting that ϕ can also be considered a non-constant parameter, determined by a different function of regressors. This can be easily adjusted for using the R packages I will discuss in this paper. However, for the purpose of this study, I will assume that ϕ is a constant.

The beta regression model can be written as $g(\mu_t) = x_t^T\beta$ where $g(.)$ is the link function and $t = 1, 2, ...n$. In this paper, the logit regression function $g(\mu) = \log(\mu/(1-\mu))$ is used as the link function to define the beta regression model. This link function is partially chosen for the purpose of comparing the results of thise model to the commonly used logistic-binomial model. The beta regression model with the logit link function can be written as

$$\mu_t = g^{-1}(\hat{y}_t) = \frac{e^{x_t^T\beta}}{1+e^{x_t^T\beta}}. \tag{3}$$

This regression model now directly predicts the expected value of y for each combination of values for the explanatory variables. The density function for each predicted value of y_t can then be calculated, as mentioned above, as

$$f(y|\mu_t, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu_t\phi)\Gamma((1-\mu_t)\phi)} y^{\mu_t\phi-1}(1-y)^{(1-\mu_t)\phi-1}, 0 < y < 1 \tag{4}$$

with a cumulative distribution function of

$$F(y|\mu_t, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu_t\phi)\Gamma((1-\mu_t)\phi)} \int_0^k y^{\mu_t\phi-1}(1-y)^{(1-\mu_t)\phi-1} dy \tag{5}$$

With this reparameterized beta distribution and regression model, the regression parameters can be estimated and tested.

2.1 Parameter Estimation

The parameters of the regression model are $\boldsymbol{\beta}$, which is a vector of β_i (and from which we can directly predict the expected value of y_t, $\hat{\mu_t}$), and ϕ, which is the precision parameter. Ferrari and Cribari-Neto (2004) introduced parameter estimation for the beta regression model using maximum likelihood estimation (MLE). Because the beta distribution's parameters do not have closed-form solutions, two methods have been proposed to find the maximum likelihood estimation: the Newton-Raphson method and the optimization method. I first set up the necessary functions for finding the MLE's, then describe the two methods.

MLE Functions The log-likelihood function for the beta regression model is

$$\begin{aligned} L(\theta) = \sum_{t=1}^{n} &\log\Gamma(\phi) - \log\Gamma(\mu_t\phi) - \log\Gamma((1-\mu_t)\phi) \\ &+ (\mu_t - 1)\log(y_t) + \{(1-\mu_t)\phi - 1\}\log(1-y_t) \end{aligned} \tag{6}$$

The gradient vector, which contains the partial derivatives of the log-likelihood function with respect to each of the parameters, is

$$\mathbf{U} = \begin{bmatrix} U_\beta(\beta,\phi)^T \\ U_\phi(\beta,\phi) \end{bmatrix}$$

$$U_\beta(\beta,\phi) = \phi X^T T(y* - \mu*) \tag{7}$$

$$U_\phi(\beta,\phi) = \sum_{t=1}^{n} \{\mu_t(y_t * - \mu_t *) + \log(1-y_t) - \psi((1-\mu_t)\phi) + \psi(\phi)\}, \tag{8}$$

where $y* = (y_1*, ..., y_n*)^T$ with $y_t* = \log\{y_t/(1-y_t)\}$ and $\mu* = (\mu_1*, ..., \mu_n*)^T$ with $\mu_t* = \psi((1-\mu_t)\phi)$, where $T = diag\{1/g'(\mu_1), ..., 1/g'(\mu_n)\}$, and where $\psi(.)$ is the digamma function. Finally, the hessian matrix, which contains the partial derivatives of each of the gradient functions with respect to each of the parameters, is

$$\mathbf{K} = \begin{bmatrix} K_{\beta\beta} & K_{\beta\phi} \\ K_{\phi\beta} & K_{\phi\phi} \end{bmatrix}$$

$$K_{\beta\beta} = \phi X^T W X \tag{9}$$

$$K_{\beta\phi} = K_{\phi\beta}^T = X^T T c \tag{10}$$

$$K_{\phi\phi} = tr(D) \tag{11}$$

where $W = diag\{w_1, ..., w_n\}$ with $w_t = \phi\{\psi'(\mu_t\phi) + \psi'((1-mu_t)\phi)\}/\{g'(mu_t)\}^2$, and where $c = (c_1, ..., c_n)^T$ with $c_t = \phi\{\psi'(\mu_t\phi)\mu_t - \psi'((1-\mu_t)\phi)(1-\mu_t)\}$.

Methods for MLE Ferrari and Cribari-Neto proposed the Newton-Raphson method in 2004, which requires solving for both the gradient vector and the hessian matrix, given in the previous section. The next step is to employ a Newton-Raphson simulation to solve for the MLE's.

The initial values selected for the Newton-Raphson method for approximating the maximum likelihood estimators of the parameters are based on the logit link function in the linear regression model. This method for selecting initial values was introduced and proven effective by Ferrari and Cribari-Neto (2004). To select the initial values for $\boldsymbol{\beta}$, the linear regression method for the logit model is used, such that $\boldsymbol{\beta} = X^T(X^TX)^{-1}z$, where $z = [g(y_1), ..., g(y_n)]^T$. The initial value suggested for ϕ is based on the variance of y for the reparameterized beta distribution, where $var(y_t) = \mu_t(1-\mu_t)/(1+\phi)$. Solving for ϕ, I obtain $\phi = \frac{1}{n}\sum_{t=1}^{n}\frac{\mu_t(1-\mu_t)}{\sigma_t^2} - 1$, where, using the linear regression method to solve for μ, $\mu = g^{-1}(x_t^T X^T (X^T X)^{-1} z)$ and $\sigma^2 = \tilde{e}^T\tilde{e}/[(n-k)g'(\mu_t)^2]$, with $\tilde{e} = z - X(X^T X)^{-1} X^T z$, which is the vector of residuals calculated from the linear regression model.

The Newton-Raphson algorithm to derive these likelihoods is as follows:

1. the 1x(k+1) vector θ contains the initial guesses for the estimates of each β_i and of ϕ.
2. initial count = 0 and ϵ, or the tolerated error, is set
3. while the absolute value of the gradient is greater than ϵ, count = count + 1 and $\theta = \theta$ - (gradient x hessian)$^{-1}$
4. θ = vector of maximum likelihood estimates for the parameters

While this is an effective method, the hessian matrix can become quite large, depending on the number of $\boldsymbol{\beta}$ parameters in the regression model. This can make the Newton-Raphson algorithm complicated and unstable, especially with the inclusion of the ϕ parameter, which this algorithm specifically struggles to estimate.

Cribari-Neto and Zeileis discussed a more direct and simple method in 2010: the optimization method. Using this method, we need only find the log-likelihood function and the gradient vector. The optim() function in R is then employed to estimate the MLE's for the parameters by maximizing the likelihood function. The optim() function is based on multiple algorithmic methods. The most common, and what is believed to be the most accurate, algorithm is the quasi-Newton, which uses the gradient to optimize the values of the parameters.

In this paper, I utilize this second method to estimate regression parameters, by using **betareg** package in R which relies on this optimization method. This parameter estimation method gives an accurate approximation of parameters for proportion and rates response variables that follow a beta distribution, which will be shown in the simulation study. These estimated values can be used to directly predict the expected value of the response variable when plugged into the regression function introduced in Equation 3.

2.2 Hypothesis Testing

In a research design, after approximating the regression coefficients, the next step is to employ hypothesis testing to find which variables are significant to the model. One way to perform these tests is to use the Wald test. The Wald test is useful because both the overall significance of the model and the significance of individual regression coefficients can be tested using this method.

To test the null hypothesis regarding the overall significance of the model, $H_o : \beta_0 = \beta_1 = ... = \beta_k = 0$, the Wald test is used to find an F statistic and a p-value. If the p-value is less than $\alpha = 0.05$, I can conclude that at least one regression parameter is significant to the model. The equation to used to find this F statistic is:

$$F = \frac{(C\hat{\beta})'[C(X'X)^{-1}C](C\hat{\beta})}{qs^2} \tag{12}$$

where C is the qx(k+1) contrast matrix,$\boldsymbol{\beta}$ is the vector of regression coefficients, β_i, and X is the nx(k+1) matrix of explanatory variables. If the test statistic is significant, further tests can be performed to find which coefficients are significant to the model. These test of the null hypotheses, $H_o : \beta_0 = 0, ..., H_o : \beta_k = 0$, require finding a a t statistic for each individual regression estimate. The equation to find this t statistic is:

$$t = \frac{a\hat{\beta}}{\sqrt{s^2 a(X'X)^{-1}a')}} \tag{13}$$

where a is the 1x(k+1) constrast vector, with a 1 that correponds with the coefficient being tested, and 0's everywhere else. This t test allows us to make specific conclusions about which variables are significant to the model. Cribari-Neto and Zeileis (2009) suggest the use of the waldtest() function from the **lmtest** R package to perform this test, and this is what is used in this article to perform the hypothesis testing.

2.3 Probability Testing

Finding the probability that the next observation will be greater than a certain value k is useful for regression analysis. This allows us to make inferences regarding the data and to use the model to predict probabilities of future observations. As discussed in the previous section, in the linear regression model with the logit link function, while direct inferences cannot be made regarding the expected value of the response variable, it does directly infer a probability function. However, the ability to make a probability inference is not lost with the beta regression model. Once the beta regression model has been set up and the parameters specified, we can find the expected value, μ_t, for each set of observations for the explanatory variables. The cumulative distribution function for a specified distribution estimates the probability that an observation will be less than a certain value. In order to find the probability that the observation will be greater than or equal to that value, we subtract the reparameterized beta CDF from one:

$$1 - F(y|\mu_t, \phi) = 1 - \frac{\Gamma(\phi)}{\Gamma(\mu_t\phi)\Gamma((1-\mu_t)\phi)} \int_0^k y^{\mu_t\phi - 1}(1-y)^{(1-\mu_t)\phi-1} dy \tag{14}$$

Using the function for the regression model, $\mu_t = e^{x_t^T\beta}/(1+e^{x_t^T\beta})$, we can replace μ_t with its regression function. The function for estimating the probability then becomes

$$1 - F(y|\frac{e^{x_i^T\beta}}{(1+e^{x_i^T\beta})},\phi) = 1 - \frac{\Gamma(\phi)}{\Gamma(\frac{e^{x_i^T\beta}}{(1+e^{x_i^T\beta})}\phi)\Gamma((1-\frac{e^{x_i^T\beta}}{(1+e^{x_i^T\beta})})\phi)} \int_0^k y^{\frac{e^{x_i^T\beta}}{(1+e^{x_i^T\beta})}\phi-1}(1-y)^{(1-\frac{e^{x_i^T\beta}}{(1+e^{x_i^T\beta})})\phi-1}dy \tag{15}$$

This beta regression model can now be used to estimate the probability that the next observation will be greater than some value k, dependent on the values of the explanatory variables for the observation. To estimate this numerical integration for the CDF, I employ Simpson's rule for integration. Simpson's rule is uses a computational algorithm to approximate the numerical integration value by dividing the integration interval (in this case [0,k]) into n subintervals. Jones, Maillardet, and Robinson (2014) provide the following function for this rule;

$$S = \frac{h}{3}(f(y_0) + 4f(y_1) + 2f(y_2) + 4f(y_3) + ... + 4f(y_{n-1}) + f(y_n)), \tag{16}$$

and describe the algorithm: "[O]n each consecutive pair of subintervals, it approximates the behavior of $f(x)$ by a parabola." This should estimate closely the integral of the pdf, and allow us to predict the probability that the next observation will be greater than a certain value k.

2.4 Comparison to Other Regression Models

As previously discussed, there are three main reasons that a beta regression model provides better conclusions for proportions and rates data: the beta regression model allows for direct inference regarding the expected value of the response variable, the beta regression model assumes that the response variable beta-distributed, and the beta regression model assumes that the response variable is heteroskedastic.

In logit regression analysis, the model is written in terms of the logistic cumulative distribution function (CDF):

$$F(y) = \frac{e^{x^T\beta}}{1+e^{x^T\beta}}. \tag{17}$$

The estimation of regression coefficients for this model provides information regarding how fast the CDF, which is the probability that y will be below a certain value, changes as an explanatory variable changes by 1. This can be useful for some types of analysis; however, the regression model does not give direct information regarding the expected value of y. Because the beta regression function is written as

$$\mu = \frac{e^{x^T\beta}}{1+e^{x^T\beta}}, \tag{18}$$

we can directly estimate the expected values of the response variable from the model. This is very useful to an inferential analysis, and, as discussed above, does not prevent us from being able to estimate probabilities from the CDF.

Additionally, if the data is beta-distributed, then its CDF does not meet the normality assumptions for a linear model. However, using the beta regression model, we only have to assume that the estimated values of μ_t, or $\bar{y}_t$, are normally-distributed, which is satisfied by the Central Limit Theorem. The Central Limit Theorem concludes that the distribution of an expected value converges to a normal distribution as n approaches infinity. This expected value of μ_t is the parameter estimate for the reparameterized beta regression model, which, of course, assumes that y is beta-distributed.

Lastly, the linear regression model assumes homoskedasticity because of its normality assumption. Homoscedasticity is the assumption that the variance of the residuals is the same or similar regardless of the values of the explanatory variables. This assumption in the linear model holds because the variance is calculated without taking in as arguments any values of X. However, the beta regression model's variance is calculated using a function of μ, shown in Equation 3. As given in Equation 18, μ is a function of the observations of the explanatory variables. This makes the model's assumptions heteroskedastic.

All of these updated assumptions in the beta regression model are a better fit for beta-distributed y variables than the linear regression or logit-binomial regression models, and make inferential conclusions easier to interpret. This beta regression model is used in the following simulation study.

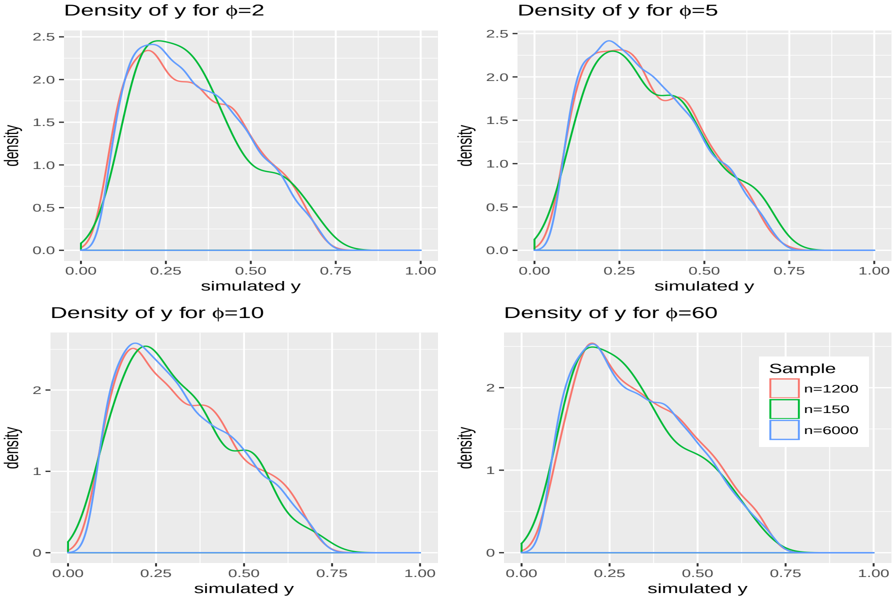
3 Simulation Study

Data is drawn from a simulation study to test the parameter estimation proposed in the previous section. The values of the simulated data closely resemble the political science data selected in the following section. The purpose of this simulation study is to ensure that the estimators for the beta regression model perform well before implementing that and the other methods with an actual dataset.

To simulate the data for the model, I first randomly draw values for the $\boldsymbol{\beta}$ coefficients in a $kx1$ vector. I then randomly draw values for the variables in the nxk X matrix from different uniform distributions that match the bounds of the actual explanatory variables. Next, to calculate the μ_t parameters, I employ the regression function by plugging $\boldsymbol{\beta}$ and X into $g^{-1}(\hat{y}_t)$. I then plug the estimated values of μ_t and a selected value of ϕ into the functions to find the parameters p and q from the original parameterization of the beta distribution. This step is implemented because R uses the original parameterization in its rbeta() function. Once I have found these parameter values, I plug each combination into rbeta() to simulate values of y.

I simulate data for the model with $\phi = 2$, $\phi = 5$, $\phi = 10$ and $\phi = 60$. These data are simulated for $n = 150$, $n = 1200$, and $n = 6000$. I repeat the n draws of y 1000 times, and plug each of the 1000 y vectors into betareg() to obtain the coefficients. Lastly, I take the average of each of the coefficient estimates for the 1000 replications of the simulation. The density plots in Figure 2 show the beta distributions of y from the different simulations. These all follow a fairly similar right-skewed shape, similar to the shape of the response variables in the data. However, for lower values of ϕ, the distribution becomes bimodal. The y values do not follow a normal distribution, and are better-fitted for a beta regression model.

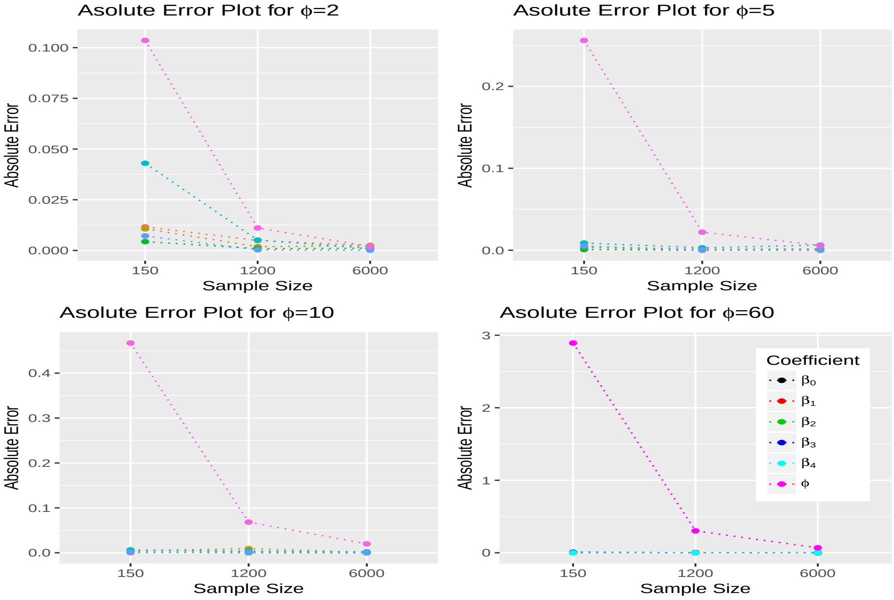
Fig. 2. Density Plots of Simulated y Values



Once I find the beta regression coefficients for the simulated data, I estimate the absolute error for these estimates using the randomly selected $\boldsymbol{\beta}$ vector, which is: $[-0.33696, -0.15116, -0.45989, -0.024487, 0.46680]$. The results can be found in Tables 5, 6, 7, and 8 in the Appendix.

Notice that as the sample size increases, the general trend is for the absolute error to decrease. This is demonstrated in the plots in Figure 3. This is generally the trend with parameter estimation- that as the sample size increases, the estimates become more accurate. Although, generally, especially for higher values of ϕ, the estimates for the β values remain fairly consistent and very close to the actual value. The estimate accuracy that changes the most as n increases is ϕ. This is likely due to the fact that the values for ϕ are much larger than the values for the β coefficients, so there's more room for estimation error.

Fig. 3. Absolute Error PLots



The simulation study has shown that the **betareg** package generally performs very well with estimating the regression coefficients. Although the estimation is slightly less accurate for small sample sizes, this is not a concern for the dataset below, which has a sample size of 150. The parameter estimates for $n = 150$, while not quite as accurate as for larger simulated datasets, still only have very small errors. The following section, therefore, applies this method to the dataset and compares this to regression model estimations to demonstrate the advantages of the beta regression model. I also employ the other beta regression methods proposed in the previous section: hypothesis and probability testing.

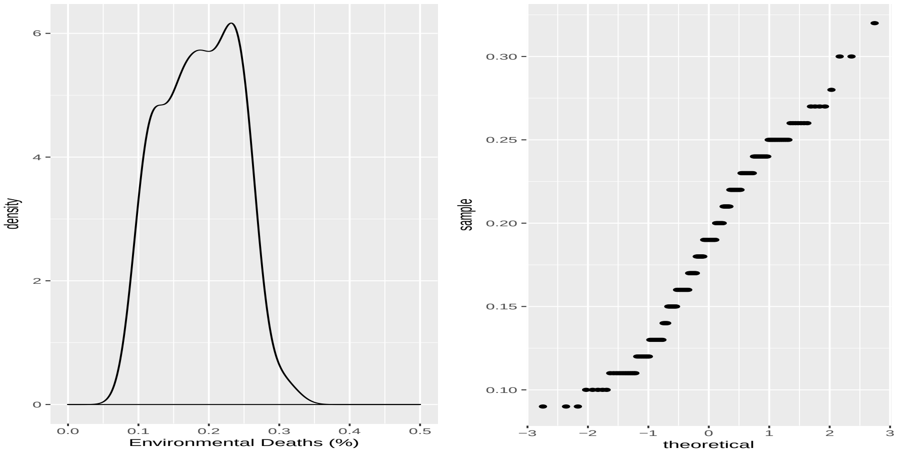
4 Results

Political scientists and women's rights scholars have begun to evaluate the intersection between women's rights and the environment. The dataset selected for the beta regression analysis evaluates how women's presence in national government affects a country's performance on an environment measure. The measures of women's presence in national governments include the percentage of seats held by women in parliament from The World Bank (2016), the percentage of women in ministerial positions from the United Nations (2016), a government participation scale from The WomanStats Project (2005-2009), and a multivariate scale on government framework for gender equality from The WomanStats

Project (2015). The government participation scale combines two subscales on representation of women in parliament and ministerial positions, and gives countries a rating according to the overall representation of women in government. Countries with excellent representation receive a 0, and those with very poor representation of women receive a 4 (with intermittent scores in between). The WomanStats multivariate scale on government framework for gender equality is a 0-5 scale, where countries with the best government framework for gender equality receive a 0, and those with the worst receive a 5. The environmental measure comes from the World Health Organization and is a percentage of deaths attributable to environmental factors in each country. If a country's environmental wellbeing is good, we expect that the percentage of deaths caused by environmental impacts will be low.

This data was chosen because of the nature of the response variable's distribution. The Environmental Deaths variable is measured as proportion data, which does not fit a normal distribution. This data's density is given in the first plot in Figure 4, and its normal QQ plot is given in the second. Although the plots look similar to data with a normal distribution, the data is slightly right-skewed, and may be better explained under the assumptions of the beta regression model.

Fig. 4. Density and QQ-plot of National Data Environmental Deaths



Using the beta regression model, I implement maximum likelihood estimation to estimate the regression parameters. I also perform hypothesis testing (using the Wald test) on each of the coefficients to determine whether each of the variables is significant to the model. As is shown in Table 1, the only variable that is statistically significant at an $\alpha = 0.05$ level is the women in ministerial positions variable.

Table 1. Parameter Estimates of Women and Environment Data

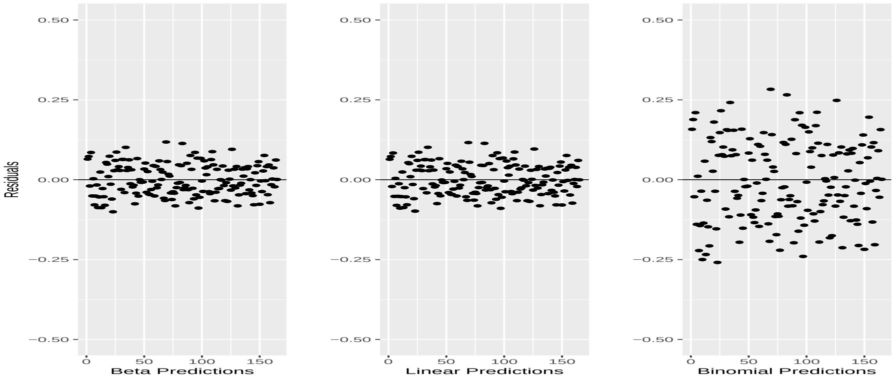
Parameter	**Estimates**	**Standard Error**	**test statistic**	**p-value**
(Intercept)	-1.366	0.147	-9.318	0.000**
womeninparliament	0.031	0.289	0.108	0.914
genderequalityfw	-0.011	0.016	-0.666	0.505
govtparticipationscale	0.036	0.032	1.223	0.261
womenministerial	-0.863	0.306	-2.819	0.005**
ϕ	60.822	6.668	9.121	0.000**

I next run both a linear regression model and a binomial regression model with a logit link function on the data to compare their results to the beta regression model. The regression coefficients, test statistics, and p-values can be found in the Table 2. I find that the same variables that were significant in the beta regression model are also significant in these other two models, although there are deviations in the values. When the standardized residuals for the three models, which measure how well the model fits the data, are compared, the beta regression model performs the same as the linear model, but much better than the binomial model. The Residual Sum of Squares, which measure the accuracy of the predictions from the model, is 0.016 for the binomial model and 0.002 for both the beta and the linear models. Figure 5 compares the residual plots for these three models. The residuals for the beta and linear models are much closer to 0, which indicates better-fit models.

Table 2. Alternative Parameter Estimates of Women and Environment Data

Linear Regression Model				
Parameter	**Estimates**	**Standard Error**	**test statistic**	**p-value**
(Intercept)	0.201	0.023	8.889	0.000 **
womeninparliament	0.006	0.045	0.126	0.900
genderequalityfw	-0.001	0.002	-0.497	0.620
govtparticipationscale	0.005	0.005	1.113	0.267
womenministerial	-0.123	0.046	-2.666	0.008 **
Binomial Regression Model (with Logit Link)				
Parameter	**Estimates**	**Standard Error**	**test statistic**	**p-value**
(Intercept)	-1.373	0.150	-9.153	0.000**
womeninparliament	0.032	0.296	0.107	0.915
genderequalityfw	-0.008	0.016	-0.491	0.624
govtparticipationscale	0.034	0.032	1.060	0.291
womenministerial	-0.843	0.313	-2.689	0.008 **

Fig. 5. Comparison of Standardized Residual Plots



The similarities in the results of the linear and the beta regression models can be explained by the distribution of the response variable plotted in Figure 4. Because there is only a slight right-skew in the data, it may approximate closely enough to a normal curve that the linear model performs just as well in this study. However, given a curve that deviates further from a linear model the beta regression model would be a clearer choice. Even with similar results in these two models, the assumptions of the beta regression model are more flexible with the shape and distribution of data, making it a safer model to use when the choice isn't obvious.

I also perform the tests on this data to find the probability that the next observation of the response variable will be greater than or equal to a certain value, given specific observations of the measures of women's presence in national governments. I use Simpson's rule approximation to find this probability, as discussed in the Methodology section. This is a useful prediction specifically for this type of data. If I want to know the probability that a country will have 20% or more deaths caused by environmental factors, given the percentage of women they have in parliament and ministerial positions, along with their performance on the women representation and framework scales, this method can inform that research question. This is a very useful inferential method because it can help determine whether a country's inclusion of women in national government increases their chances of having very high numbers of deaths caused by environmental factors. These probabilities are recorded in Table 3, along with the absolute error of the Simpson's integration estimate. Notice that as women's involvement in national government improves, the probability that a country will have 20% or more deaths caused by environmental factors decreases. Also notice that the absolute error for the Simpson's integration method is very low, which indicates that this is a very good estimator of probability.

Table 3. Probability $y > k = 0.20$

New X Values	**Poor Values**	**Mediocre Values**	**Good Values**
womeninparliament	0.10	0.25	0.39
genderequalityfw	5	3	0
govtparticipationscale	4	2	0
womenministerial	0.10	0.25	0.39
Probability	0.50854	0.30233	0.16700
Absolute Error	0.0000	0.0000	0.0000

This probability testing can also be done when predicting the probability that observations will be less than k by only finding the CDF, and not subtracting that from 1. I demonstrate this by plugging in the same new values as above and finding the probability that a new observation y will be less than 10%. The results in Table 4 show that as a country's involvement of women in national government increases, the probability that the proportion of deaths caused by environmental factors is less than 10% also increases.

Table 4. Probability $y < k = 0.10$

New X Values	**Poor Values**	**Mediocre Values**	**Good Values**
womeninparliament	0.10	0.25	0.39
genderequalityfw	5	3	0
govtparticipationscale	4	2	0
womenministerial	0.10	0.25	0.39
Probability	0.00960	0.03973	0.10337
Absolute Error	0.0000	0.0000	0.0000

These methods allow more accurate and clear inferential conclusions regarding the data, and the prediction ability of the logit linear model is not lost because probability can be estimated using the beta cumulative distribution function. I can conclude from this analysis that the beta regression model is an effective way to measure the percentage of deaths resulting from environmental causes in a country dependent on the presence of women in national government. I find that women in ministerial positions is a significant variable, and that as the percentage of women in these ministerial positions increases, the percentage of deaths caused by environmental factors decreases. I can find the expected value of this environmental variable given any set of explanatory variable observations, and can easily predict probabilities.

5 Conclusion

Political science research could significantly benefit from implementing this alternative regression model in studies with proportion and rate response variables. Political scientists would be able to create better-fit models for this type of response variable if the beta regression model were implemented in more studies. The simulation study indicated that the beta regression model provides accurate estimates of the parameters of the regression model.

Additionally, the data application indicated that the beta regression model is a more reliable alternative to the logit binomial regression model for this type of data, and allows for clearer and simpler inferential conclusions. This model also has the added benefit of having the expected value, which we obtain directly from the model, as one of the parameters of the distribution. With this property, inferential abilities are not lost with the use of this model because the easily-approximated expected value can be plugged into the beta distribution's CDF to also estimate probabilities. Although the linear model performed similarly to the beta regression model in this study, the flexibility of the assumptions in the beta regression model regarding shape and consistency of spread make this a safer choice for analyses on proportions and rates data because the beta regression model can more easily adjust to skewed and heteroskedastic data.

6 Appendix

Table 5. Estimates of Simulated Data for $\boldsymbol{\phi = 2}$

Sample Size	**Coefficient**	**Estimated Coefficients**	**Absolute Error**
n=150	β_0	-0.325	0.0117
	β_1	-0.1618	0.0106
	β_2	-0.4642	0.0043
	β_3	-0.2879	0.0430
	β_4	0.4740	0.0072
	ϕ	2.1036	0.1036
n=1200	β_0	-0.3320	0.0049
	β_1	-0.1530	0.0019
	β_2	-0.4608	0.0009
	β_3	-0.2500	0.0051
	β_4	0.4665	0.0003
	ϕ	2.0111	0.0111
n=6000	β_0	-0.3345	0.0025
	β_1	-0.1535	0.0023
	β_2	-0.4610	0.0011
	β_3	-0.2439	0.0010
	β_4	0.4669	0.0001
	ϕ	2.0020	0.0020

Table 6. Estimates of Simulated Data for $\boldsymbol{\phi = 5}$

Sample Size	**Coefficient**	**Estimated Coefficients**	**Absolute Error**
n=150	β_0	-0.3406	0.0037
	β_1	-0.1547	0.0035
	β_2	-0.4607	0.0008
	β_3	-0.2536	0.0088
	β_4	0.4718	0.0050
	ϕ	5.2560	0.2560
n=1200	β_0	-0.3384	0.0014
	β_1	-0.1512	0.0001
	β_2	-0.4598	0.0001
	β_3	-0.2419	0.0030
	β_4	0.4672	0.0003
	ϕ	5.0220	0.0220
n=6000	β_0	-0.3385	0.0015
	β_1	-0.1527	0.0015
	β_2	-0.4599	0.0000
	β_3	-0.2386	0.0062
	β_4	0.4669	0.0001
	ϕ	5.0058	0.0058

Table 7. Estimates of Simulated Data for $\boldsymbol{\phi = 10}$

Sample Size	**Coefficient**	**Estimated Coefficients**	**Absolute Error**
n=150	β_0	-0.3371	0.0002
	β_1	-0.1558	0.0047
	β_2	-0.4620	0.0021
	β_3	0.2515	0.0066
	β_4	0.4681	0.0013
	ϕ	10.4670	0.4670
n=1200	β_0	-0.3346	0.0024
	β_1	-0.1605	0.0093
	β_2	-0.4600	0.0001
	β_3	-0.2486	0.0037
	β_4	0.4673	0.0005
	ϕ	10.0683	0.0683
n=6000	β_0	-0.3372	0.0002
	β_1	-0.1523	0.0012
	β_2	-0.4599	0.0000
	β_3	-0.2471	0.0023
	β_4	0.4673	0.0005
	ϕ	10.0200	0.0200

Table 8. Estimates of Simulated Data for $\boldsymbol{\phi = 60}$

Sample Size	**Coefficient**	**Estimated Coefficients**	**Absolute Error**
n=150	β_0	-0.3306	0.0063
	β_1	-0.1517	0.0005
	β_2	-0.4615	0.0016
	β_3	-0.2503	0.0054
	β_4	0.4671	0.0003
	ϕ	62.8925	2.8925
n=1200	β_0	-0.3350	0.0020
	β_1	-0.1526	0.0014
	β_2	-0.4601	0.0002
	β_3	-0.2469	0.0020
	β_4	0.4666	0.0002
	ϕ	60.3027	0.3027
n=6000	β_0	-0.3372	0.0003
	β_1	-0.1507	0.0005
	β_2	-0.4598	0.0001
	β_3	-0.2451	0.0002
	β_4	0.4667	0.0001
	ϕ	60.0693	0.0693

Bibliography

1. Cribari-Neto, Francisco, and Achim Zeileis. “Beta regression in R.” (2009).
2. Ferrari, Silvia, and Francisco Cribari-Neto. “Beta regression for modelling rates and proportions.” Journal of Applied Statistics 31, no. 7 (2004): 799-815.
3. Geneva, World Health Organization, `http://www.who.int`, 2016.
4. Jones, Owen, Robert Maillardet, and Andrew Robinson. Introduction to scientific programming and simulation using R. CRC Press, 2014.
5. The World Bank, `http://www.worldbank.org`, 2016
6. United Nations, `http://www.un.org`, 2016
7. WomanStats Project Database, `http://www.womanstats.org`, 2016