Modeling Social Attitudes with Beta Regression Hurdle Models

Mark H. White II[1]

[1] University of Kansas

Author Note

Author note will go here.

Correspondence concerning this article should be addressed to Mark H. White II.
E-mail: markhwhiteii@gmail.com

Abstract

Abstract will go here.

Keywords: beta regression, hurdle models, norms, social attitudes

Modeling Social Attitudes with Beta Regression Hurdle Models

Statistical Background

The Beta Distribution

The beta distribution can be used to model the residuals in a generalized linear model when the values of the dependent variable are bounded $0 < y < 1$ (Coxe, West, & Aiken, 2013). The probability density function (pdf) of the beta distribution is determined by two parameters, α and β, that are called "shape" parameters:

$$f(y; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}$$

where $\Gamma(.)$ is the gamma function. The two shape parameters pull the mean toward zero and one, respectively. If α is larger than β, the mean leans toward zero; if the reverse is true, the mean leans toward one. One of the benefits of the beta distribution is that it is flexible and can take a number of distributional shapes. ADD FIGURE 1 HERE TO SHOW VARIOUS SHAPES.

These parameters are not intuitive to applied researchers used to using regression and analysis of variance, however. Rigby, Stasinopoulos, Heller, and De Bastiani (2017) "reparameterized" the beta distribution to make it easier understand in a regression framework (but see Ferrari & Cribari-Neto, 2004 for a different parameterization). Instead of predicting α and β, they parameterize the beta distribution with two different parameters: μ (called the "location" parameter) and σ (called the "scale" parameter), where $\mu = \frac{\alpha}{(\alpha+\beta)}$ and $\sigma = \sqrt{\frac{1}{(\alpha+\beta+1)}}$. μ is equivalent to the mean, and σ is related positively to the variance. (Note that σ is *not* the standard deviation, even though σ refers to the standard deviation in many other contexts.) The variance is equivalent to $\sigma^2\mu(1-\mu)$. There are two important things to note from this equation: First, the greater the σ, the greater the variance; Second, the variance depends on the mean. This means that beta regression, covered shortly, will be naturally heteroskedastic.

But how can we model dependent variables that equal zero and/or one?

Zero-One Inflated Beta Distribution

Rigby et al. (2017) show that the beta distribution can be "inflated" at zero or one—that is, zeros and ones can be modeled. This distribution is a mixture distribution. When the dependent variable contains zeroes and ones (i.e., $0 \leq y \leq 1$), the pdf for this beta mixture, beinf, is:

$$\text{beinf}(y; \mu, \sigma, \nu, \tau) = \begin{cases} p_0 & \text{if } y = 0 \\ (1 - p_0 - p_1) f(y; \mu, \sigma) & \text{if } 0 < y < 1 \\ p_1 & \text{if } y = 1 \end{cases}$$

for $0 \leq y \leq 1$, where $f(y; \mu, \sigma)$ is the beta pdf with μ and σ bounded *between* zero and one. The two additional parameters, ν and τ, are mixture parameters. p_0 is the probability that a case equals zero, p_1 is the probability that a case equals one, and p_2 (i.e., $1 - p_0 - p_1$) is the probability that the case comes from the beta distribution. In terms of these two additional parameters, $p_0 = \frac{\nu}{(1+\nu+\tau)}$ and $p_1 = \frac{\tau}{(1+\nu+\tau)}$. Rearranging these algebraically, ν is the odds that a case is zero to being from the beta distribution, $\nu = p_0/p_2$, and τ is the odds that a case is a one to being from the beta distribution, $\tau = p_1/p_2$.

Beta Regression Models

The goal now is to predict these four parameters, μ, σ, ν, τ, from any number of predictor variables. All four parameters can be predicted by an identical set of predictors, none of the same predictors, or anywhere in between. Both μ and σ have to be between zero and one, so we can use the logistic link function to fit predicted values in this range; both ν and τ have to be greater than zero, so we can use the log link function to fit predicted values in this range. Imagine we have one predictor, x. We could include this variable as a predictor of all four variables with the equations:

$$\log(\tfrac{\mu}{1-\mu}) = \beta_{10} + \beta_{11}X$$
$$\log(\tfrac{\sigma}{1-\sigma}) = \beta_{20} + \beta_{21}X$$
$$\log(\nu) = \beta_{30} + \beta_{31}X$$
$$\log(\tau) = \beta_{40} + \beta_{41}X$$

Or, equivalently stated:

$$\mu = \tfrac{1}{1+e^{-(\beta_{10}+\beta_{11}X)}}$$
$$\sigma = \tfrac{1}{1+e^{-(\beta_{20}+\beta_{21}X)}}$$
$$\nu = e^{\beta_{30}+\beta_{31}X}$$
$$\tau = e^{\beta_{40}+\beta_{41}X}$$

where e^x is the natural exponential function. A regression model can also be used when the dependent variable contains zeros but no ones (e.g., $0 \leq y < 1$) or when it contains ones but no zeros (e.g., $0 < y \leq 1$). Let c be the value—0 or 1—that is inflated. The pdf is:

$$\text{beinf}_c(y; \mu, \sigma, \nu) = \begin{cases} p_c & \text{if } y = c \\ (1-p_c)f(y; \mu, \sigma) & \text{if } 0 < y < 1 \end{cases}$$

where $\nu = p_c/(1-p_c)$ and everything else is the same as above. The same link functions are used as above, but the fourth parameter, τ, is not included, as the dependent variable can only take on values of c or from the beta distribution. Lastly, if no zeros or ones are observed, the beta distribution alone can be used as the pdf. This results in a beta regression where the researcher is only predicting the location, μ, and shape, σ. Since there is no mixture with values of being 0 or 1, the latter two mixture parameters, ν and τ, are not included.

Applications to Social and Personality Psychology (NEW NAME?)

These beta regression models are often used when dealing with rates and proportions, given that these are naturally bounded $0 \leq y \leq 1$ (Buntaine, 2011; Eskelson, Madsen, Hagar,

& Temesgen, 2011; Gallardo, Bovea, Colomer, & Prades, 2012; Hubben et al., 2008; Peplonska et al., 2012). The beta distribution is doubly bounded continuous distribution, meaning that, although it is on a continuous scale, values cannot be greater than the upper bound, u, or lesser than the lower bound, l. Although researchers generally model variables on Likert scales and sliding scales (e.g., feeling thermometers) as being conditionally normally distributed, these variables are, by definition, not strictly normal.

Observations from a normal distribution can take on any value on the real number line. Observations from a standard, 1 (Strongly Disagree) to 7 (Strongly Agree) Likert scale can only take on values between 1 and 7. There are no meaningful negative consequences of using the normal distribution to model a variable with observed $M = 4.0$ and $SD = 0.8$, given that virtually all of the distribution falls between 1 and 7 anyways. However, measures in prejudice and politics are often skewed toward one of the poles, such as $M = 2.0$ and $SD = 1.5$. In this case, one might observe predicted values outside the bounds, and the model is likely to violate the assumption of homoskedasticity, as well. SHOW EXAMPLES IN FIGURE 2. While some advocate for using robust estimators for standard errors to address this problem [CITE], one can explicitly take into account that the response is doubly bounded by using the beta regression models described above.

If one observes a dependent variable limited between two bounds, there is a straightforward way to rescale the variable to the $0 \leq y \leq 1$ range:

$$y_i' = (y_i - l)/(u - l)$$

where y is the variable on the original scale, y' is the transformed variable, u is the upper bound (i.e., the largest possible value on the scale), l is the lower bound (i.e., the smallest possible value on the scale), and the i subscript denotes an individual's score. On a standard 7-point Likert scale, $l = 1$ and $u = 7$. This transformation allows a researcher to explicitly model conditional variance, floor effects, and ceiling effects.

Conditional variance

Ceiling and floor effects

Variance as a Proxy for Norms

Ceiling and Floor Effects as Hurdles

Implementation in GAMLSS

References

Buntaine, M. T. (2011). Does the Asian Development Bank respond to past environmental performance when allocating environmentally risky financing? *World Development*, *39*(3), 336–350.

Coxe, S., West, S. G., & Aiken, L. S. (2013). Generalized linear models. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods, volume 2* (pp. 26–51). New York, NY: Oxford University Press.

Eskelson, B. N., Madsen, L., Hagar, J. C., & Temesgen, H. (2011). Estimating riparian understory vegetation cover with beta regression and copula models. *Forest Science*, *57*(3), 212–221.

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, *31*(7), 799–815.

Gallardo, A., Bovea, M. D., Colomer, F. J., & Prades, M. (2012). Analysis of collection systems for sorted household waste in Spain. *Waste Management*, *32*(9), 1623–1633.

Hubben, G. A. A., Bishai, D., Pechlivanoglou, P., Cattelan, A. M., Grisetti, R., Facchin, C., … Tramarin, A. (2008). The societal burden of HIV/AIDS in Northern Italy: An analysis of costs and quality of life. *AIDS Care*, *20*(4), 449–455.

Peplonska, B., Bukowska, A., Sobala, W., Reszka, E., Gromadzińska, J., Wasowicz, W., … Ursin, G. (2012). Rotating night shift work and mammographic density. *Cancer, Epidemiology, Biomarkers, & Prevention*, *21*(7), 1028–1037.

Rigby, R., Stasinopoulos, M., Heller, G., & De Bastiani, F. (2017). *Distributions for modelling location, scale, and shape: Using GAMLSS in R. Unpublished draft.* Retrieved from gamlss.com/books-articles