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

The Beta Distribution

The beta distribution can be used to model the residuals in a generalized linear model when the dependent values are bounded $0 < y < 1$. 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. These parameters are not intuitive to applied researchers used to using regression and analysis of variance.

Stasinopoulos and colleagues "reparameterized" the beta distribution to make it easier understand in a regression framework. 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 to the variance. σ is not the standard deviation (even though σ refers to the standard deviation in 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 σ is, the greater the variance is; 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?

Mixture Beta Distribution

Stasinopoulos and colleagues show that the beta distribution can be "inflated" at zero or one. However, I will not use the language of "inflation," as this distribution is more intuitively thought of as a mixture distribution. When the dependent variable contains zeroes and ones (i.e., $0 \leq y \leq 1$), the pdf for this beta mixture, bmix, is:

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

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 compared to being from the beta distribution, $\nu = p_0/p_2$, and τ is the odds that a case is a one compared to being from the beta distribution, $\tau = p_1/p_2$.

Beta Regression Hurdle Models

The goal now is to predict these four parameters, μ, σ, ν, τ. 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, for the dependent variable y. We could include this variable as a predictor of all four variables with the equations:

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

Or, equivalently stated:

$$\begin{aligned} \mu &= \tfrac{1}{1+e^{-(\beta_{10}+\beta_{11}X)}} \\ \sigma &= \tfrac{1}{1+e^{-(\beta_{20}+\beta_{21}X)}} \end{aligned}$$

$$\nu = e^{\beta_{30}+\beta_{31}X}$$

$$\tau = e^{\beta_{40}+\beta_{41}X}$$

Where e^x is the natural exponential function.

References