Discretised Beta Regression for Analysis of Rating Data: The R Package DBR

Mansour T.A. Sharabiani
School of Public Health
Imperial College London, UK

Cathy M. Price
Solent NHS Trust
Southampton, UK

Alex Bottle
School of Public Health
Imperial College London, UK

Alireza S. Mahani
Davison Kempner Capital Management
New York, USA

Abstract

The question of whether to treat rating data - often generated from survey responses - as ordinal or numeric has received considerable attention over the years. Theoretical arguments notwithstanding, when the number of response levels is high, practitioners often seek a numeric interpretation of the response variable and the effect of predictors on 'mean' response, thus using linear regression. In this paper, we introduce Discretised Beta Regression (DBR) - mathematical framework and open-source software implementation - as a more suitable alternative to linear regression for numerical interpretation and analysis of rating data. DBR is an adaptation of beta regression with several features: First, it handles the forward and backward mapping between the observed range of responses and the standard range of the beta distribution. Secondly, it properly takes into account the discrete nature of observations, including the use of cumulative-density terms in constructing the likelihood function. Thirdly, DBR properly accounts for extreme-value count inflation, often seen in survey responses, both in estimation and prediction steps. Finally, by adopting a Bayesian framework using Markov Chain Monte Carlo sampling for estimation, DBR benefits from robust estimation, credible interval calculation and prediction functionalities. Unlike standard linear regression which is homoscedastic, DBR successfully replicates the variability of slope and dispersion observed in ratings data, making it a more realistic framework for analysis of such datasets.

Keywords: Discretised Beta Regression, Ordinal Regression, Likert, Bayesian, Markov Chain Monte Carlo.

1. Introduction

When analysing survey-response data, a key decision is whether the data should be treated as nominal, ordinal or numeric. When there is no natural order in responses, the data should clearly be treated as nominal. An example would be the type of lung cancer detected in a patient (Table **??**, first question). Choice models such as multinomial logit Hasan, Wang, and Mahani (2016) and probit are suitable for regression analysis of nominal response variables. If responses present a natural order but do not carry a clear numeric interpretation (ordinal data), one can use ordered logit and probit regression models Goodrich, Gabry, Ali, and Brilleman (2018). An example would be a patient's degree of happiness in sending their child to school after a prolonged period of remote learning (Table **??**, second question).

The third type of survey response - referred to as ratings data - is similar to ordinal data, but contains more levels, with levels often associated with numbers. When it comes to ratings data, there has been considerable debate about whether the responses should be treated as ordinal or numeric Harpe (2015); Liddell and Kruschke (2018); Jamieson (2004); Norman (2010); Kuzon, Urbanchek, and McCabe (1996); Armstrong (1981); Knapp (1990); Pell (2005); Carifio and Perla (2007, 2008). Examples of rating scales - used to elicit rating responses - are Likert, numerical, fully-anchored and adjectival Harpe (2015). Numeric treatment of ratings data allows for easier interpretation of regression coefficients, but has been shown to lead to inconsistent results Liddell and Kruschke (2018) when there are few levels. When dealing with many levels, the numeric treatment has the advantage of consuming significantly fewer degrees of freedom compared with ordinal regression, but the underlying assumptions of unboundedness and homoscedasticity remain at odds with the nature of ratings data.

In this paper, we offer a new mathematical framework - called Discretised Beta Regression (DBR) - for regression analysis of ratings data, along with an open-source software implementation, the `DBR` R package. DBR offers a middle ground between linear regression - built on a strict equidistant interpretation of the response scale - and ordinal regression with full flexibility in partitioning a latent variable into sub-regions that are mapped to the observed, discrete levels.

Discretised Beta Regression (DBR) is an adaptation of beta regression, following the specification of Ferrari and Cribari-Neto (2004); Zeileis, Cribari-Neto, Grün, and Kos-midis (2010). It is similar to ordinal regression, especially the ordered probit model, in that it maps a continuous, latent variable to the observed discrete response by partitioning the range of the latent variable. However, DBR has two important differences from ordered probit: 1- the underlying distribution is assumed to be beta (with proper shift and scale factors applied) rather than normal, 2- cutoff points in DBR are assumed to be halfway points between the observed values. (However, see the discussion of left and right buffers in Section 2.4). This setup allows DBR to create a numeric yet realistic interpretation of ratings data.

DBR is similar to beta-binomial regression, which has also been recommended for analysis of ratings data Najera-Zuloaga, Lee, and Arostegui (2018). There are differences, however: first, rather than directly mapping responses to a discrete distribution (binomial or beta-binomial), DBR follows the latent-variable approach in ordinal regression, which is more in line with our intuition about the process of response selection by survey respondents. Secondly, the `DBR` software accounts for extreme-value count inflation using cumulative-density terms in the log-likelihood function.

The rest of this paper is organized as follows. In Section **??** we describe the detailed math-

ematical framework underlying DBR. In Section **??**, we introduce the datasets used in the paper, namely the National Pain Audit (NPA) data and the SOLO data. In Section **??**, we present the results of applying DBR to NPA and SOLO datasets, and compare the performance of DBR to linear regression. Section **??** includes a discussion of future work. Supplementary material - available online - include the `DBR` package along with a tutorial for using the package.

add a table?

2. Discretised Beta Regression (DBR)

We begin this section with a brief review of beta regression. This is followed by changes made to beta regression in DBR for adapting it to rating responses.

2.1. Overview of Beta Regression

The probability density function (PDF) for beta distribution is given by:

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

$$\Gamma(z) \equiv \int_0^\infty u^{z-1}e^{-u}du. \tag{1b}$$

where the random variable y is restricted to the interval $[0,1]$, $\alpha,\beta>0$ are the so-called shape parameters of the distribution, and $\Gamma(.)$ is the Gamma function, which is a generalisation of the factorial function to real (and complex) numbers. For beta regression, we follow Ferrari and Cribari-Neto (2004); Zeileis *et al.* (2010) by using the alternative, mean-precision parameterisation of beta distribution:

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\,\phi)\,\Gamma\left((1-\mu)\,\phi\right)}\, y^{\mu\,\phi-1}\,(1-y)^{(1-\mu)\,\phi-1} \tag{2}$$

where the parameters μ (mean) and ϕ (precision) are linked to the shape parameters as follows:

$$\begin{cases} \mu &= \frac{\alpha}{\alpha+\beta} & \text{(3a)} \\ \phi &= \alpha+\beta & \text{(3b)} \end{cases}$$

and reversely:

$$\begin{cases} \alpha &= \mu\,\phi & \text{(4a)} \\ \beta &= (1-\mu) & \text{(4b)} \end{cases}$$

We also require that $0<\mu<1$ and $\phi>0$. The first and second moments of the distribution can be expressed in terms of mean and precision parameters:

$$\mathrm{E}[y] = \mu \tag{5a}$$

$$\mathrm{VAR}[y] = \frac{\mu\,(1-\mu)}{1+\phi} \tag{5b}$$

We can see from the above that the model is heteroscedastic, i.e., the response variance is reduced (approaching zero) - for a fixed precision parameter - as the mean approaches either end of the $(0, 1)$ range. This dispersion-compression at extreme ends of the response range is consistent with our expectation.

With the above mean-precision specification in hand, we can set up beta regression by assuming that the mean parameter - via a link function - is a linear function of model predictors, $\mathbf{x}$:

$$g(\mu) = \mathbf{x}^\top \beta, \tag{6}$$

where $g(.)$ could be a suitable function that maps $(0, 1)$ to real line, e.g., the logit function, $g(u) = \log(u/(1-u))$. Further flexibility can be achieved by making the precision parameter a function of predictors, also via a suitable link function such as log. Note that the nonlinear link function causes the first derivative of mean response with respect to any explanatory variable (or predictor), x_k, to be non-constant, i.e.:

$$\frac{\partial \mathrm{E}[y]}{\partial x_k} = \beta_k \, \frac{dg^{-1}(z)}{dz}\Big|_{z=\mathbf{x}^\top \beta} \tag{7}$$

2.2. Forward and Backward Transformation of Response Variable

Most statistical software packages use the Maximum-Likelihood (ML) technique for parameter estimation, which typically involves maximising the 'logarithm' of the likelihood function. Note that setting $x = 0$ or $x = 1$ causes the beta-distribution PDF to become zero (Eq. 2), and hence its logarithm to become infinite. For this reason, software packages only allow an open-ended interval for x, i.e., they require $x \in (0, 1)$. Therefore, the first step towards adapting beta regression for DBR is to map the raw data to the $(0, 1)$ range.

Consider K unique response values, sorted in increasing order: $y_1 < \cdots < y_K$. A naive transformation could be:

$$z_k = \frac{y_k - y_1}{y_K - y_1}. \tag{8}$$

But the above would map to $[0, 1]$, rather than to $(0, 1)$. Instead, we introduce left (b_l) and right (b_r) buffers:

$$b_l \equiv (y_2 - y_1)/2 \tag{9a}$$

$$b_r \equiv (y_K - y_{K-1})/2 \tag{9b}$$

We have essentially extended the 'latent' range of the data to $y_1 - b_l$ on the left, and $y_K + b_r$ on the right. This leads to the revised linear transformation:

$$y \longrightarrow z = u(y) \equiv \frac{y - \delta}{r}. \tag{10}$$

where $y \in \{y_1, \ldots, y_K\}$, and we have defined

$$\delta \equiv y_1 - b_l \tag{11a}$$

$$r \equiv y_K - y_1 + b_l + b_r \tag{11b}$$

It can be easily verified that the above transformation would map the data to the following range:

$$b_l/(y_K - y_1 + b_l + b_r) \leq u(y) \leq (y_K - y_1 + b_l)/(y_K - y_1 + b_l + b_r) \tag{12}$$

The above is what is needed for model training (i.e., regression). For prediction, we differentiate between two modes. For 'point' prediction, we simply apply the reverse of $u(.)$ defined in Eq. 10. We refer to this reverse transformation as $u_p^{-1}(.)$, formally defined as

$$z \longrightarrow y = u_p^{-1}(z) \equiv r\, z + \delta \tag{13}$$

On the other hand, we can also generate 'samples' during prediction, in which case we must add a discretisation step, where we report y_k that is closest to the sample drawn from beta distribution according to mean and dispersion parameters provided by the regression model. Referring to this transformation as $u_s^{-1}(.)$, we formally define it as

$$z \longrightarrow y = u_s^{-1}(z) \equiv y_k \;\; s.t. \; |r\,\hat{x} + \delta - y_k| \leq |r\,\hat{x} + \delta - y_{k'}|, \;\; \forall k' \in \{1, \ldots, K\}. \tag{14}$$

(Ties are theoretically possible given finite resolution of floating-point math on digital computers, but rare cases can be handled by choosing the smallest of the (at most two) k's.)

2.3. Discretisation Correction

The discretisation process must be reflected in the likelihood function for estimation. In other words, if we observe the value z_k, we cannot be certain that the latent sample drawn from the beta distribution - before discretisation - was z_k, but only that it was between $\frac{z_{k-1}+z_k}{2}$ and $\frac{z_k+z_{k+1}}{2}$, when $1 < k < K$. When $k = 1$, the left boundary is 0, and when $k = K$, the right boundary is 1. We summarise the above by introducing boundary functions $z_l(.)$ and $z_r(.)$:

$$z_l(y_k) = \begin{cases} 0 & k = 1 \\ \frac{u(y_{k-1})+u(y_k)}{2} & 1 < k \leq K \end{cases} \tag{15}$$

and

$$z_r(y_k) = \begin{cases} \frac{u(y_k)+u(y_{k+1})}{2} & 1 \leq k < K \\ 1 & k = K \end{cases} \tag{16}$$

Given the above, we assert that the contribution of a data point with response y_k to the likelihood is

$$P(y = y_k) = F\left(z_r(y_k)\right) - F\left(z_l(y_k)\right) \tag{17}$$

where $F(.)$ is the cumulative density function for beta distribution (Eq. 1a or 2), defined as:

$$F(x) = \int_0^x f(u)\, du. \tag{18}$$

2.4. Handling Extreme Responses

Extreme response to survey questions is one of several known types of bias in survey data Furnham (1986). For example, in Likert scales, the proportion of 0's and 10's for a 0-10 scale may be higher than 1 and 9, respectively. Researchers have discussed reasons for, impact of, and ways to handle this bias Meisenberg and Williams (2008); Greenleaf (1992).

Aside from study/question-design approaches, one method for analysis of extreme responses is a mixture model, similar to zero-inflated Poisson distribution Lambert (1992). In the case

of beta distribution, we can modify Eq.17 as follows

$$P(y = y_k) = (1 - \pi_l - \pi_r) \left\{ F\left(z_r(y_k)\right) - F\left(z_l(y_k)\right) \right\} + \begin{cases} \pi_l & k = 1 \\ 0 & 1 < k < K \\ \pi_r & k = K \end{cases} \tag{19}$$

The new parameters π_l, π_r are both probabilities, and thus must be between 0 and 1. In a regression context, they can be both made to be linear functions of predictors, via a suitable link function.

We take a different approach in DBR, however, and utilise the existing framework for handling discretisation by allowing the left and right buffers, b_l, b_r to be estimated from the data, rather than being fixed according to Eqs. 9a and 9b.

Besides boundary values, extreme response can also be observed for midpoint/neutral points on a Likert scale. While the inflation/mixture-density approach of Eq. 19 can be deployed for this case as well, we refrain from including it in our implementation of DBR due to increase in parameter count and hence risk of overfitting. Including a neutral point on the Likert scale may encourage the respondent to take an easy way out, thus providing more noise than information. Hence some have argued in favor of removing the neutral options, e.g., by using an even number of levels instead of an odd number Allen and Seaman (2007).

2.5. Bayesian Estimation

Due to the complexity of likelihood function, especially when including left and right buffers in estimation, we opt for a Bayesian framework, which allows for consistent estimation of credible intervals. The conditional probability of observed responses is given by:

$$P(\mathbf{y}|\mathbf{X}; \phi, \beta, b_l, b_r) = \prod_{n=1}^{N} \left\{ F\left(z_r(y_{k[n]}; b_l, b_r); g^{-1}(\beta^\top \mathbf{x_n}), \phi\right) - F\left(z_l(y_{k[n]}; b_l, b_r); g^{-1}(\beta^\top \mathbf{x_n}), \phi\right) \right\} \tag{20}$$

In the above, $z_l(y; b_l, b_r)$ and $z_r(y; b_l, b_r)$ are functions that map each observed response to its left and right intervals over the $(0, 1)$ scale that is the domain of the beta distribution, $g^{-1}(\beta^\top \mathbf{x})$ is the 'mean function', i.e., function that calculates the mean of the beta distribution by forming the linear predictor $\beta^T \mathbf{x}$, followed by the logistic function. From the above, we obtain the following log-posterior:

$$L(\phi, \beta, b_l, b_r) = \log\left(F\left(z_r(y_{k[n]}; b_l, b_r); g^{-1}(\beta^\top \mathbf{x_n}), \phi\right) - F\left(z_l(y_{k[n]}; b_l, b_r); g^{-1}(\beta^\top \mathbf{x_n}), \phi\right)\right) + \Phi(\phi) + \mathbf{B}(\beta) + B_l(\tag{21}$$

where $\Phi(\phi)$, $\mathbf{B}(\beta)$, $B_l(b_l)$ and $B_r(b_r)$ are the log-prior functions specified for precision parameter of beta distribution (ϕ), coefficients for the mean parameters (β) and the left and right buffers (b_l, b_r), respectively. For results shown in this work, we use non-informative, flat priors for all parameters (with conservative boundaries).

For parameter estimation, we use the Markov Chain Monte Carlo (MCMC) sampling technique, using our `MfUSampler` R package Mahani and Sharabiani (2017). This software relies on a Gibbs wrapper around the univariate slice sampler Neal (2003). MCMC has the inherent advantage of being able to escape local optima and finding the true global optimum, which is highly desirable for complex functions such as 21. (However, this is not guaranteed to happen

in every problem.) In addition, the fact that slice sampler is derivative-free provides further convenience.

3. DBR Implementation and Features

4. Using DBR

We begin by loading the necessary libraries and the dataset:

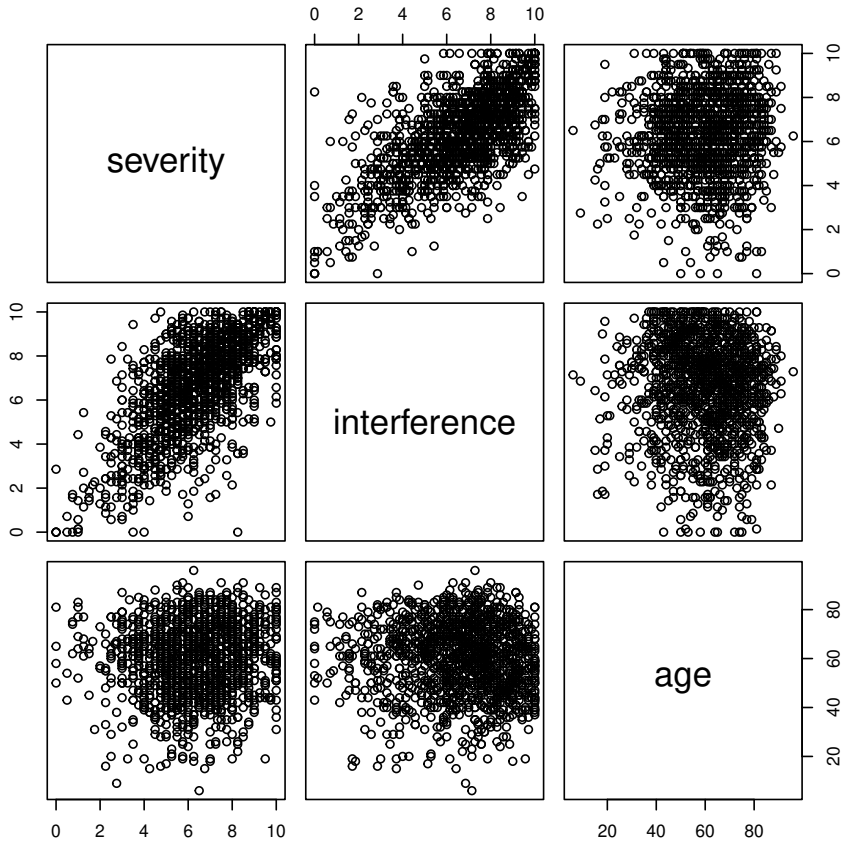
```
R> library("DBR")
R> data("pain")
R> df <- pain
R> df$age <- as.integer(df$age)
```

Pain severity and interference are two aggregate scores calculated from patient survey responses, each between 0 and 10:

```
R> summary(df)
```

```
    severity       interference         age
 Min.   : 0.000   Min.   : 0.000   Min.   : 6.0
 1st Qu.: 5.000   1st Qu.: 5.286   1st Qu.:50.0
 Median : 6.500   Median : 7.000   Median :61.0
 Mean   : 6.318   Mean   : 6.584   Mean   :59.6
 3rd Qu.: 7.750   3rd Qu.: 8.286   3rd Qu.:71.0
 Max.   :10.000   Max.   :10.000   Max.   :96.0
```

We can also examine the scatterplots:



We observe a clear positive correlation between pain severity and pain interference scores, but the impact of age on pain interference is less clear, The spearman test below indicated a statistically-significant negative correlation between age and pain interference.

```
R> ret <- with(df, {
+    print(cor.test(severity, interference, method = "spearman"))
+    print(cor.test(age, interference, method = "spearman"))
+  })
```

```
        Spearman's rank correlation rho

data:  severity and interference
S = 136306075, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.6427926
```

```
        Spearman's rank correlation rho
```

```
data:  age and interference
S = 417536805, p-value = 0.0006158
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.09420837
```

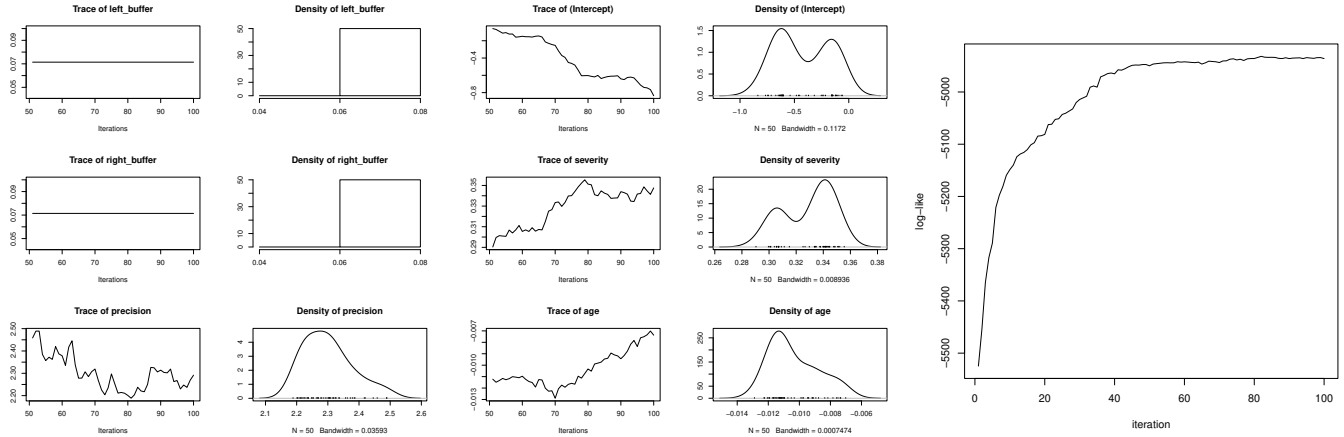
5. Discretized Beta Regression - First Attempt

Using the DBR is as simple as a one-line call to the `dbr` function. The following call specifies only the two required parameters, `formula` and `data`, relying on the default values for the others:

```
R> est.1 <- dbr(
+    formula = interference ~ severity + age
+    , data = df
+  )
```

We can review the estimated model via the `summary` function, which produces MCMC diagnostic plots as well as credible intervals for estimated model parameters:

```
R> summary(est.1)
```



	left_buffer	right_buffer	precision	(Intercept)	severity	age
2.5%	0.07142857	0.07142857	2.201476	-0.76076143	0.2997927	-0.012266314
50%	0.07142857	0.07142857	2.294415	-0.47196006	0.3361209	-0.011025895
97.5%	0.07142857	0.07142857	2.481968	-0.07421716	0.3519739	-0.007347752

A couple of observations can be made upon examining the above figures and table:

1. The left and right buffer trace plots show constant values. This is because, by default, `dbr` does not estimate their values; instead it uses fixed values according to the details described earlier in the paper. To instruct the software to estimate these buffers from data, we need to set the flags `estimate_left_buffer` and `estimate_right_buffer` to `TRUE` in the call to `dbr.control`.

2. By default, MCMC runs for 100 iterations and discards the first 50 as the burn-in period. The MCMC diagnostics plots suggest that we need more iterations: 1) Trace plots for model parameters have not stabilized, 2) Log-likelihood trace plot also has not reached a stable, and is still rising at the end of iterations.

In addition to above issues, we check on another point: Do we expect the unique levels of the response variable to be all represented in the training data? As we said in description of the data before, the pain interference score is an average of 7 individual responses, each an integer between 0 and 10. Therefore, the average ratings form a sequence with increments of 1/7. However, as seen below, there are a couple of gaps:

```
R> setdiff(0:70, round(7 * sort(unique(df$interference))))
```

```
[1] 2 3
```

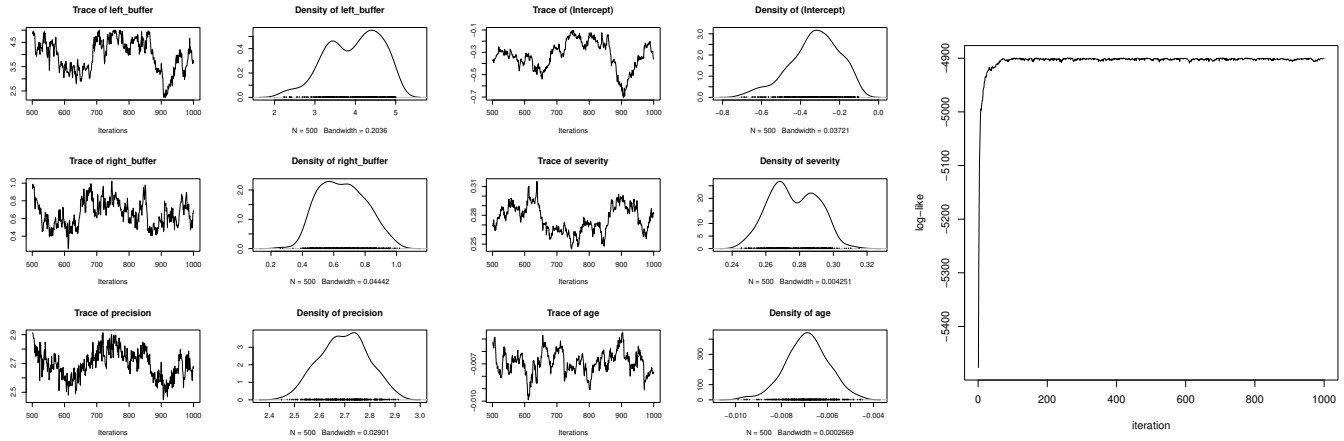
In other words, values of 2/7 and 3/7 have not occurred in the training data. This could distort the DBR algorithm's calculation of cutpoints. To correct this, we can explicitly override the `yunique` argument in the call to `dbr`.

6. Revised Model

Given above observations, we re-estimate the model, but with several function arguments overridden:

```
R> est.2 <- dbr(
+    formula = interference ~ severity + age
+    , data = df
+    , control = dbr.control(
+      nsmp = 1000
+      , nburnin = 500
+      , estimate_left_buffer = T
+      , estimate_right_buffer = T
+    ), yunique = 0:70 / 7)
```

```
R> summary(est.2)
```



	left_buffer	right_buffer	precision	(Intercept)	severity	age
2.5%	2.484779	0.4212212	2.512354	-0.6170400	0.2506109	-0.008801291
50%	4.059479	0.6505089	2.697751	-0.3197953	0.2742341	-0.006879292
97.5%	4.954448	0.9459980	2.876430	-0.1281612	0.2993229	-0.005126241

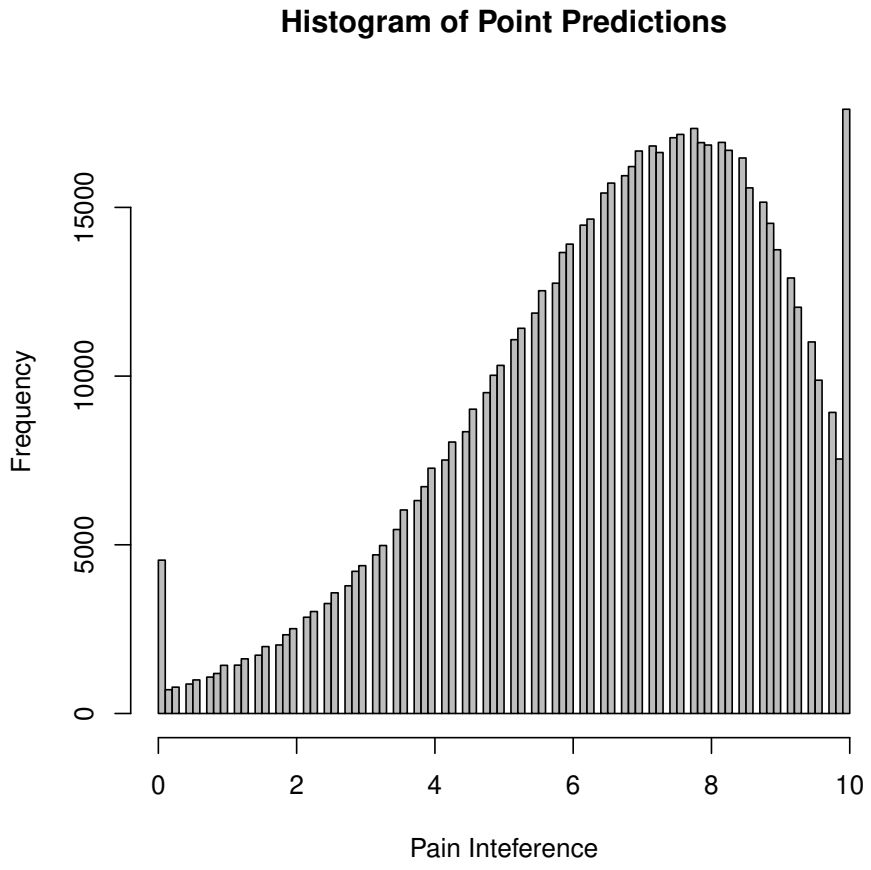
We can see that the MCMC chains show relative stability, and so does the log-posterior trace plot. Also, examining the credible intervals for severity and age coefficients indicates that they are both significant at the 95% level. Let's discuss interpretation of DBR coefficients.

Firstly, the left and right buffers represent the scale of the latent, continuous variable. For example, a left buffer of 4.5 means that when the latent variable is anywhere between -4.5 and $+1/70$, it is mapped to the observed response 1. Similarly, a right buffer of 0.7 means that when the latent variable is between $10-1/70$ and 10.7, it is mapped to an observed value of 10. The coefficients of severity and age are interpreted on the standard beta-distribution scale. For example, a severity coefficient of 0.27 means that, every unit increase in pain severity increases the logit of the mean of the beta distribution that generates pain interference score (before reverse scaling) by $+0.27$.

7. Model Prediction

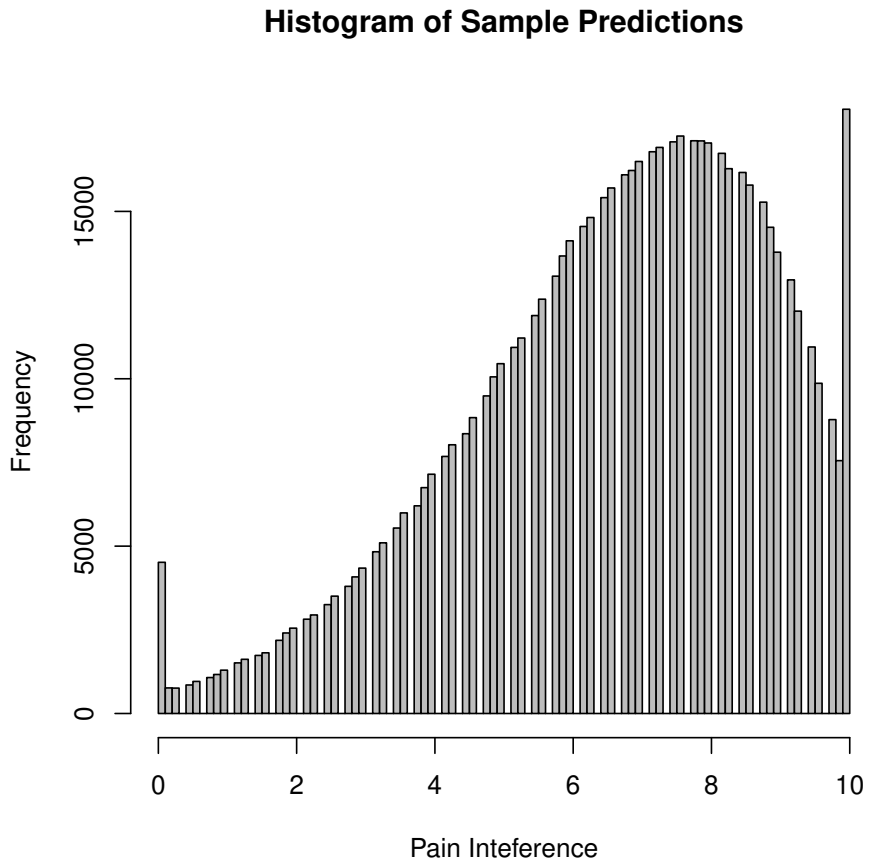
DBR offers two prediction modes, `point` prediction and `sample` prediction. The `point` prediction returns expected response value, and is thus a continuous value. This is achieved via a call to the `predict` function and by setting the `type` argument to `point` (which is also the default value) .

```
R> pred_point <- predict(est.2, newdata = df, type = "point")
R> hist(pred_point, breaks = 100, col = "grey"
+       , xlab = "Pain Inteference"
+       , main = "Histogram of Point Predictions"
+       )
```



The reader may wonder why the extreme-value inflation is not seen in the histogram of predicted values. This is because these are mean responses. To see the full dispersion of predictions, we have to switch to `sample` mode:

```
R> pred_sample <- predict(est.2, newdata = df, type = "sample")
R> hist(pred_sample, breaks = 100, col = "grey"
+       , xlab = "Pain Inteference"
+       , main = "Histogram of Sample Predictions"
+       )
```



8. Discussion

We have presented the DBR mathematical framework for analysing ratings data (preferably with many levels of response) using a discretised version of beta distribution. DBR allows for quantifying the impact of predictors on the response while relaxing the unrealistic assumptions of fixed slope and variance, which are present in linear regression. We have also prepared an open-source implementation of the DBR framework as an R package, also called `DBR`, available in Supplementary Material for this paper. A tutorial for how to use `DBR` has also been provided in Supplementary Material, with more help available as part of package documentation.

The Bayesian framework, along with Markov Chain Monte Carlo (MCMC) sampling technique offers several advantagesKruschke and Liddell (2018); Liddell and Kruschke (2018), including robust credible-interval calculation without resorting to unrealistic assumptions about the asymptotic behavior of the log-likelihood function. While MCMC can be time-consuming for large datasets, there are several techniques proposed in the literature for speeding it up Mahani and Sharabiani (2015).

One future step in taking full advantage of the Bayesian framework is to allow for users of the DBR R package to supply or select non-uniform (non-informative) priors for regression parameters. Another direction of future work is to add support in the software for inflated

midpoint values using the mixture framework described in Section 2.4. Another direction for future work is to embed DBR in composite settings such as multi-level and mixture models. The Bayesian framework adopted for DBR would facilitate such extensions. Finally, another direction for future research is to systematically compare DBR and beta-binomial regression, e.g., using the `PROreg` R package Najera-Zuloaga, Lee, and Arostegui (2020).

References

Allen IE, Seaman CA (2007). "Likert scales and data analyses." *Quality progress*, **40**(7), 64–65.

Armstrong GD (1981). "Parametric statistics and ordinal data: A pervasive misconception." *Nursing Research*, **30**(1), 60–62.

Carifio J, Perla R (2008). "Resolving the 50-year debate around using and misusing Likert scales." *Medical education*, **42**(12), 1150–1152.

Carifio J, Perla RJ (2007). "Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes." *Journal of social sciences*, **3**(3), 106–116.

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

Furnham A (1986). "Response bias, social desirability and dissimulation." *Personality and individual differences*, **7**(3), 385–400.

Goodrich B, Gabry J, Ali I, Brilleman S (2018). "rstanarm: Bayesian applied regression modeling via Stan." *R package version*, **2**(4), 1758.

Greenleaf EA (1992). "Measuring extreme response style." *Public Opinion Quarterly*, **56**(3), 328–351.

Harpe SE (2015). "How to analyze Likert and other rating scale data." *Currents in Pharmacy Teaching and Learning*, **7**(6), 836–850.

Hasan A, Wang Z, Mahani AS (2016). "Fast Estimation of Multinomial Logit Models: R Package mnlogit." *Journal of Statistical Software*, **75**(3), 1–24. doi:10.18637/jss.v075.i03.

Jamieson S (2004). "Likert scales: How to (ab) use them?" *Medical education*, **38**(12), 1217–1218.

Knapp TR (1990). "Treating ordinal scales as interval scales: an attempt to resolve the controversy." *Nursing research*, **39**(2), 121–123.

Kruschke JK, Liddell TM (2018). "The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective." *Psychonomic Bulletin & Review*, **25**(1), 178–206.

Kuzon W, Urbanchek M, McCabe S (1996). "The seven deadly sins of statistical analysis." *Annals of plastic surgery*, **37**, 265–272.

Lambert D (1992). "Zero-inflated Poisson regression, with an application to defects in manufacturing." *Technometrics*, **34**(1), 1–14.

Liddell TM, Kruschke JK (2018). "Analyzing ordinal data with metric models: What could possibly go wrong?" *Journal of Experimental Social Psychology*, **79**, 328–348.

Mahani AS, Sharabiani MT (2015). "SIMD parallel MCMC sampling with applications for big-data Bayesian analytics." *Computational Statistics & Data Analysis*, **88**, 75–99.

Mahani AS, Sharabiani MTA (2017). "Multivariate-From-Univariate MCMC Sampler: The R Package MfUSampler." *Journal of Statistical Software, Code Snippets*, **78**(1), 1–22. doi:10.18637/jss.v078.c01.

Meisenberg G, Williams A (2008). "Are acquiescent and extreme response styles related to low intelligence and education?" *Personality and individual differences*, **44**(7), 1539–1550.

Najera-Zuloaga J, Lee DJ, Arostegui I (2018). "Comparison of beta-binomial regression model approaches to analyze health-related quality of life data." *Statistical methods in medical research*, **27**(10), 2989–3009.

Najera-Zuloaga J, Lee DJ, Arostegui I (2020). "PROreg: Patient Reported Outcomes Regression Analysis." R package version 1.1, URL https://CRAN.R-project.org/package=PROreg.

Neal RM (2003). "Slice sampling." *Annals of statistics*, pp. 705–741.

Norman G (2010). "Likert scales, levels of measurement and the "laws" of statistics." *Advances in health sciences education*, **15**(5), 625–632.

Pell G (2005). "Use and misuse of Likert scales."

Zeileis A, Cribari-Neto F, Grün B, Kos-midis I (2010). "Beta regression in R." *Journal of statistical software*, **34**(2), 1–24.

A. Setup

Below is the corresponding R session information.

```
R> sessionInfo()

R version 4.1.1 (2021-08-10)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 20.04.3 LTS

Matrix products: default
BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.9.0
LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.9.0
```

```
locale:
 [1] LC_CTYPE=C.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=C.UTF-8        LC_COLLATE=C
 [5] LC_MONETARY=C.UTF-8    LC_MESSAGES=C.UTF-8
 [7] LC_PAPER=C.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C           LC_TELEPHONE=C
[11] LC_MEASUREMENT=C.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods
[7] base

other attached packages:
[1] DBR_1.2.2

loaded via a namespace (and not attached):
[1] compiler_4.1.1   ars_0.6          tools_4.1.1
[4] HI_0.4           coda_0.19-4      grid_4.1.1
[7] MfUSampler_1.0.6 lattice_0.20-44
```

Affiliation:

Alireza S. Mahani
Quantitative Research Group
Davidson Kempner Capital Management
New York, NY
US
E-mail: alireza.s.mahani@gmail.com