MODEL SELECTION UNCERTAINTY AND STABILITY IN BETA
REGRESSION MODELS: A STUDY OF BOOTSTRAP-BASED
MODEL AVERAGING WITH AN EMPIRICAL APPLICATION TO
CLICKSTREAM DATA

BY CORBAN ALLENBRAND\textsuperscript{1,*} AND BEN SHERWOOD\textsuperscript{2}

\textsuperscript{1}SCHOOL OF BUSINESS, UNIVERSITY OF KANSAS, *CALLENBRAND@KU.EDU

\textsuperscript{2}SCHOOL OF BUSINESS, UNIVERSITY OF KANSAS BEN.SHERWOOD@KU.EDU

Statistical model development is a central feature of many scientiﬁc investigations with a vast methodological landscape. However, uncertainty in the model development process has received less attention and is frequently resolved non-rigorously through beliefs about generalizability, practical usefulness, and computational ease. This is particularly problematic in settings of abundant data, such as clickstream data, as model selection routinely admits multiple models and imposes a source of uncertainty, unacknowledged and unknown by many, on all post-selection conclusions. Regression models based on the beta distribution are class of non-linear models, attractive because of their great ﬂexibility and potential explanatory power, but have not been investigated from the standpoint of multi-model uncertainty and model averaging. For this reason, a formalized tool that can combine model selection uncertainty and beta regression modeling is presented in this work. The tool combines bootstrap model averaging, model selection, and asymptotic theory to yield a procedure that can perform joint modeling of the mean and precision parameters, capture sources of variability in the data, and achieve more accurate claims of estimate precision, variable importance, generalization performance, and model stability. Practical utility of the tool is demonstrated through a study of model selection consistency and variable importance in average exit and bounce rate statistical models. This work emphasizes the necessity of a departure from the all-too-common practice of ignoring model selection uncertainty and introduces an accessible technique to handle frequently neglected aspects of the modeling pipeline.

1. Introduction.

1.1. Motivation. In this generation of big data, the issue of intelligent data analysis with stable and robust statistical models must receive a greater emphasis especially when issues of high dimensionality, spurious correlations, and heterogeneity of the data can confound the reliability of insights (Fan, Han and Liu, 2014). Online platforms, such as eCommerce platforms, generate a steady stream of a particular type of big data known as clickstream data which is derived from the second-by-second online behavior of website visitors. Although this source of information receives considerable commercial attention, it has not attracted the attention of the broader statistical community. As a result, rigorous exploration and validation of the many clickstream variable relationships mentioned commercially have heretofore not been subjected to thorough analysis. In particular, specific clickstream variables, including exit and bounce rate, are bounded and proportional, thus complicating statistical model development. A relatively newer type of regression model based on the beta distribution represents a candidate model...
structure, capable of dealing with the numerical properties of the clickstream variables, that has not been used to analyze clickstream data. On top of this model structure problem exists a model selection problem. The velocity, volume, and variety of clickstream data promotes the selection of several models during fitting with potentially differing conclusions on aspects like variable inclusion, significance, and estimate precision. This is part of the broader issue of model selection uncertainty and how to properly handle it. Any decision to model proportional and bounded variables with beta regression must then also include a systematic method to deal with multi-model selection and its associated uncertainty.

1.2. Background and Literature Review. Response variables that can be regarded as continuous, unbounded, and on an interval-scale are routinely analyzed with normal-theory models using ordinary least squares (OLS) (Kieschnick and McCullough, 2003). Not all variables fit this categorization with bounded, proportional data being one exception. Work by Ferrari and Cribari-Neto (2004), Smithson and Verkuilen (2006), and Simas, Barreto-Souza and Rocha (2010) established a comprehensive framework for beta regression models which can flexibly handle these variety of variables. Further development of this framework with greater understanding of beta regression residuals was done by Espinheira, Ferrari and Cribari-Neto (2008a). Work by Rocha and Simas (2010) and Espinheira, Ferrari and Cribari-Neto (2008b) yielded results that provided methods to measure and diagnose influence in beta models. Beta regression has been extended into the time domain realm by Rocha and Cribari-Neto (2009) to model moving average processes. Computational intensity or analytical challenges associated with likelihood-based approaches may be mitigated under a Bayesian approach to estimating a Beta regression model. A Bayesian methodology for modeling the parameters in a beta distribution was proposed by Cepeda-Cuervo and Gamerman (2005). Simultaneous modeling of the mean and dispersion in beta regression models was presented by Cepeda (2015) and is implemented in the Bayesianbetareg R package by Marin, Rojas and Jaimes (2014). Incorporation of random effects, latent variables, and regression coefficient specific priors expanded the Bayesian treatment of beta regression models. (Liu and Kong, 2015a; Liu and Li, 2016; Liu and Kong, 2015b)

Given its relative novelty compared to alternatives, many scientists and practitioners may opt for non-beta models; each of the possible alternatives raises its own set of pitfalls. Modeling proportions with an assumption of normally distributed errors is less tenable given a few characteristics of proportional data: their domain is not $\mathbb{R}$, a bounded domain for a response variable implies that the conditional mean function for the response will be a nonlinear function of the predictors, and the conditional variance must be a function of the mean since it will converge to zero as the mean converges to either 0 or 1. Hence, a normal theory OLS regression on proportional data will yield incorrect and unreliable estimates (Paolino, 2001). Other regression schemes exist as possible alternatives to OLS. Use of a censored normal distribution model (Tobit) may be considered, but this assumes the observed response can be modeled as a censored latent variable with normally distributed errors and the actual response variable in (0,1) is the observed part of the latent variable (Tobin, 1958). This can be a poor assumption as a restriction to (0,1) interval may originate from the physics of the process, not because of censoring. Also, any assumption of normality, where it does not hold, will suffer the same criticisms as for OLS. Data transformations to the response such as the logit, log, or arcsine may be used followed by OLS or nonlinear least squares (NLS) estimation (Warton and Hui, 2011). Dual interpretations of model coefficients on the original and non-transformed scale also constitutes a potential hazard for the
applied scientist or practitioner. Proceeding with OLS after a data transformation also assumes errors are additive and normally distributed on the new scale, an assumption not assured to be true. For instance, using a logit transformation on the response and running OLS assumes the original response is sampled from a conditional logistic normal distribution.

Explicitly non-linear alternatives to beta regression are not without fault. Although NLS does not make similarly restrictive assumptions on the errors as OLS does, it does require the specification of a nonlinear functional form and good starting values for all parameters being estimated, both of these may not be available to the modeler. Also, given that the relationship between the predictors and mean response is non-linear, nonlinear transformations will introduce greater bias towards the endpoints of the response’s domain than at the mean. Binomial regression with a logit link might also appear as a capable tool for modeling proportional data; however, this option presumes the proportions are weighted counts of multiple trials of a Bernoulli process with the counts and total number of trials known for each observation. This choice of logistic regression is clearly flawed if the proportions are not aggregated Bernoulli outcomes. An appropriate modeling of the dispersion, without assumption on dispersion being unity, is also crucial if sources of data variability are too be understood. In light of the aforementioned shortcomings of alternatives and the asymmetry, skewness, and heteroscedasticity of proportional data, a more fitting model is needed. A maximum-likelihood based regression technique, based on the beta distribution, that can model both location and dispersion jointly is a superior way to accommodate varying amount of skew and heteroscedasticity while respecting the bounded nature of the proportional data (Ferrari and Cribari-Neto, 2004; Smithson and Verkuilen, 2006; Simas, Barreto-Souza and Rocha, 2010).

Uncertainty about the model remains extensive throughout the entire fitting process and originates from a decision on which predictors to incorporate, what data to use for model construction and evaluation, and the functional form of the model (Chatfield, 1995). Parallel growth in model selection uncertainty and model complexity is a highly pertinent point with the replacement of simpler regression alternatives with beta regression models. All estimates returned by any model - beta regression or not - have, in reality, two sources of uncertainty. The first source is the uncertainty in the estimates, conditional on a given model, and is the inexactness commonly reported in inference done with the model. A second, more elusive, source of uncertainty comes from the model selection process itself. In practice, a model is selected wherefrom inferential conclusions are drawn with the unstated and, likely unrecognized, assumption that this is the true model. This yields an overstatement of the precision of the estimates and overly liberal conclusions. This is common in empirical work where a baseline model with certain structural assumptions, such as the functional form of the model, may be taken as mostly known with an exploration of slight modifications to it (Claeskens and Hjort, 2008; Burnham and Anderson, 2002). Variation in the selected model over data sets contributes to variance in estimates in a similar way as does the variation in parameter estimates from data set to data set. Hence, the possibility that different models may be supported by independent data sets from the same phenomena indicates that model uncertainty should be included in measures of precision about parameter estimates.

One alternative to ignoring model uncertainty is to estimate all candidate models using frequentist or Bayesian methodology, compute a weighted average of all the estimates, and base inference on the space of all models and not a singly selected one (Buckland, Burnham and Augustin, 1997). A Bayesian framework provides an alternative to frequentist non-Bayesian techniques, like resampling or information criteria, for
model averaging (MA) with Hoeting et al. (1999) providing a thorough discussion on the topic. Although Bayesian model averaging is a well-characterized approach to MA, it entails several difficulties not easily amended in all application domains such as the need to specify priors for competing models, priors for parameters under each model, and numerical difficulties with summation and integration of likelihoods (Hoeting et al., 1999). Non-Bayesian MA techniques, like the one employed in this work, escape the limitations of Bayesian techniques but still involve the key element to any MA method, the selection of model weights (Moral-Benito, 2015). Different model weighting schemes including information criteria based ones have been explored and discussed for use in model averaging (MA) efforts (Burnham and Anderson, 2002). Practical usefulness and a comparison of Bayesian and non-Bayesian MA methods was evaluated by Augustin, Sauerbrei and Schumacher (2005) for prognostic factors models for survival data. They found that using bootstrap resampling in the model averaging yielded more parsimonious models. Properties of predictions from linear regression models fitted under a bootstrap MA approach were explored by Bucholz, Hollander and Sauerbrei (2008).

This paper’s authors extended the methodology to the broader non-linear model setting, particularly for the beta regression model.

1.3. Overview. In this work, a bootstrap model averaging approach is introduced as a method to incorporate model selection uncertainty in the estimation of beta regression models. A combined program for joint mean-dispersion modeling and model
selection in a class of models for proportional data that has received considerably less attention than some of its misguided alternatives is the primary contribution of this paper. Problems of model selection are amplified in the beta regression model which is made of two submodels, one for the mean and the other for the dispersion. The program elaborated on and shown in this work is a step-forward in addressing issues of multi-model inference, selection consistency, and model stability in the beta regression setting. A simulation study was conducted to illustrate how precision in estimates from beta regression models change when the variance component from model selection uncertainty is factored into estimation. A narrowness bias in confidence intervals is investigated in a Monte Carlo study wherein the coverage probability of different estimation techniques is computed. In the presence of several competing models equally supported by the data, predictions may be expected to be more accurate if each model is allowed to contribute. To assess this, a separate Monte Carlo study was performed where models were estimated on independent training data sets and predictions were gathered on a test set independent from each of the training sets. Expected prediction accuracy was appraised through three measures of out-of-sample predictive performance on independently generated test data. Practical usefulness of beta regression and the bootstrap mode averaged estimators was depicted by analyzing the relationship between clickstream variables contained in data from an online eCommerce retailer. Two such variables, exit rate and bounce rate values, were shown to be amenable to analysis with the beta regression MA technique. Model selection and predictor variable inclusion frequencies were developed as robust quantifiers of model stability and relative variable importance.

This paper is further organized as follows. In Section 2, mathematical and theoretical details of the beta distribution, beta regression model, and estimation with model combinations are provided. In Section 3, a description of the algorithm for the beta regression bootstrap-based model averaging procedure, model selection criteria employed, and the metrics used to assess out-of-sample performance of the techniques are presented. In Section 4, formulation of and results from Monte Carlo numerical simulations of the proposed model averaging procedure and other alternatives over several different beta regressions settings are provided. In Section 5, background for the setting of the empirical application, a description of the clickstream data used, and results from model selection, variable selection, and out-of-sample prediction are offered. Discussion of results from both the simulated and empirical work are located in the same section that their respective results appear.

2. Theoretical Background.

2.1. Beta Regression Model. A beta regression model assumes the conditional distribution of response variable \( Y \) follows \( \text{Beta}(\mu, \phi) \). A more comprehensive survey of the beta distribution and some of its theoretical attributes can be found in Section A of the Supplementary Material (Allenbrand and Sherwood, 2021). In its original formulation, the precision parameter \( \phi \), was taken to be constant across the observations (Ferrari and Cribari-Neto, 2004). An argument for loss of efficiency with an assumed constant \( \phi \) and the need for joint modeling of the response’s mean and variance prompted an extension to the original beta regression model where \( \phi \) could be assigned a dependence on covariates (Smithson and Verkuilen, 2006; Simas, Barreto-Souza and Rocha, 2010). Let \( \mathbf{y} = (y_1, \ldots, y_n)^T \) denote a vector of \( n \) random variables such that \( y_i \sim \text{Beta}(\mu_i, \phi_i) \). By definition, strictly monotonic and twice differentiable functions of \( \mu_i, g(\mu_i), \) and \( \phi_i, h(\phi_i) \), that map \((0, 1)\) and \( \mathbb{R}^{++} \) (positive open orthant in \( \mathbb{R} \)), respectively, onto \( \mathbb{R} \) are
required in a beta regression model. Similar to a generalized linear model (GLM), let \( \eta^i = x_i^T \beta \) denote the linear combination of \( p \) predictors where \( \beta = (\beta_0, \ldots, \beta_p) \in \mathbb{R}^{p+1} \) is a vector of regression coefficients and \( x_i = (x_{i0}, \ldots, x_{ip}) \in \mathbb{R}^{p+1} \) is the vector of predictors for observation \( i \) in the mean submodel with the first element equating to one to allow for an intercept. Setting \( g(\mu_i) = \eta^i, \forall i = 1, \ldots, n \), let \( g : (0, 1)^n \rightarrow \mathbb{R}^n \) be a vector-valued link function that is applied component-wise to its argument so that we have,

\[
g(\mu) = X\beta
\]

or equivalently, \( \mu = g^{-1}(X\beta) \), where \( \mu = E[y; X, \beta] \in (0, 1)^n \) is the vector of conditional means and \( X = (x_1, \ldots, x_n)^T \in \mathbb{R}^{n \times (p+1)} \) is the design matrix corresponding to \( \beta \). A standard formulation of \( Y|X \sim Beta(\mu, \phi) \) assumes \( \phi \) is constant across all observations. However, a predictor’s influence on a response may not always be strictly limited to the response’s mean and direct modeling of a predictor’s effect on \( \phi \) may be warranted.

As was done for the mean submodel, define \( \eta^i_\phi = z_i^T \gamma \) where \( \gamma = (\gamma_0, \ldots, \gamma_q)^T \in \mathbb{R}^{q+1} \) is a vector of regression coefficients and \( z_i = (z_{i0}, \ldots, z_{ir})^T \in \mathbb{R}^{r+1} \) is the vector of predictors for the precision submodel with the first element equating to one to allow for an intercept. It should be noted that the predictors that appear in \( \eta^i_\phi \) are not required to appear in \( \eta^i \) although in practice and by theory they can. Recasting the specification with a similarly defined vector-valued link gives,

\[
h(\phi) = Z\gamma
\]

or equivalently, \( \phi = h^{-1}(Z\gamma) \), where \( \phi \in \mathbb{R}^{++} \) is the vector of precision parameters and \( Z \in \mathbb{R}^{n \times (r+1)} \) denotes the design matrix corresponding to \( \gamma \). Equation (2) is a model for the precision parameter but by inverting the sign on the predictors, it becomes possible to model the dispersion, \( \phi^{-1} \), which may make interpretation of signs for \( \gamma \) easier.

Similar to the GLM framework, various transformation (link) functions for \( g(.) \) and \( h(.) \) can be employed in order to model various structures in the data. Readers interested in a deeper examination of the relationship between beta regression models and GLMs should consult Section B or the Supplementary Material (Allenbrand and Sheridan, 2021). In this article, attention will be on the logit link for the mean submodel \( \log[\mu_i/(1 - \mu_i)] = \eta^i_\mu \) and the log link for the precision submodel \( \log(\phi_i) = \eta^i_\phi \). The logit satisfies the necessary properties of \( g(.) \) and confers a benefit with respect to the interpretation of \( \beta \) as estimates of the influence on the log odds of the mean. Alternatives to the logit include other links based on inverse cumulative distribution functions such as the probit, cauchit, complementary log-log (cloglog). Each demonstrates different scaling behaviors with respect to their corresponding parameters; this scaling prominently affects the tails of corresponding distributions and can generate large difference on estimation of fitted values (Smithson and Verkuilen, 2006). A logit link is the focus of this work because most practitioners would likely have a preference for it and it eases interpretation of the estimates. It was chosen over a probit as they both exhibit similar numerical behavior except the logit has slightly fatter tails than the probit and allows for a more gradual approach to 0 or 1. Further, the probit would be preferred to the logit when the response variable is suspected to normally distributed in the population, an assumption not easily justified. A logit was chosen over the cloglog because the asymmetrical property of the cloglog, i.e. response approaches 0 or 1 more or less
quickly, suggests understanding of a population property of the response variable that is not always available. For the precision submodel, \( h(.) \), the log has the property that it only permits strictly positive values for its domain, a restriction shared by \( \phi \). Maximum likelihood estimators of \( \beta \) and \( \gamma \) are obtained from maximizing the log likelihood objective function

\[
\sum_{i=1}^{n} \log L(\beta, \gamma; y_i) = \sum_{i=1}^{n} \log \frac{\Gamma \left( z_i^T \gamma \right)}{\Gamma \left( (1-x_i^T \beta)z_i^T \gamma \right) \Gamma \left( (1-x_i^T \beta)z_i^T \gamma - 1 \right)} \cdot \frac{\Gamma \left( (1-x_i^T \beta)z_i^T \gamma \right)}{\Gamma \left( (1-x_i^T \beta)z_i^T \gamma - 1 \right)}
\]

Solving for the maximum cannot be done in closed-form and so numerical optimization through a quasi-Newton method such as the Broyden-Fletcher-Goldfarb-Shanno algorithm must be completed with suitable starting conditions for \( \beta \) and \( \gamma \) (Ferrari and Cribari-Neto, 2004; Smithson and Verkuilen, 2006).

2.2. Model Combination. Consider a situation in which a researcher has gathered or observed data on a theorized relationship between several variables. Further, suppose the theoretical description of the relationship is thought to occupy some universe of models, \( M \), with \( |M| \) denoting the number of candidate models in this space. After some preferred model selection technique, model \( m_1 \in M \) is found to fit the data well with the directions and magnitude of coefficient estimates consistent with theory. This selected model, along with its associated within model variance, is then used to compute estimates, standard errors, make judgments about significance, and make predictions. Basing inferences solely upon \( m_1 \) is insecure when the possibility exists that a second model \( m_2 \in M \), or any number of alternatives, exists that may support different conclusion on the relationship under investigation. Information contained within the data concerning the theorized relationship is partly consumed during model selection and so ambiguity regarding any estimates should be higher post-selection (Hodges, 1987). Disregard some subset of \( M \), and the resultant model may misrepresent the stochastic behavior of the phenomena under study.

2.3. Beta Regression Model Averaging. Model uncertainty is factored into selection when using a model averaging estimator which combines estimates from competing models instead of relying on only one. Using the philosophy of Buckland, Burnham and Augustin (1997), let \( \theta \) be a parameter from the beta regression model to be estimated and assumed to be common to all models. Given our space of contending models \( M \), we can form a model averaging estimator of \( \theta \) by

\[
\hat{\theta} = \sum_{k=1}^{|M|} w_k \hat{\theta}_k,
\]

where \( w_1, \ldots, w_k \) are model specific weights for model \( m_k \) with \( w_k \geq 0 \), \( \sum_k w_k = 1 \), and \( \hat{\theta}_k \) represents the estimate of \( \theta \) from model \( m_k \). A particular realization of \( \hat{\theta} \) is attained by using the estimates \( \hat{\theta}_k \) of \( \theta \) for each \( m_k \). In this formalization, how one approaches sampling from \( M \) may impact interpretations of estimates of \( \hat{\theta}_k \). A reason for this is because exclusion of some subset of \( M \) precludes the use of information from that subset to estimate \( \theta \) requiring a conditioning on the subset of \( M \) actually used. In this work, random sampling from \( M \) is adopted, under which the interpretation of \( \hat{\theta}_k \) as that of an estimate of \( \theta \) is consistent with classical sampling theory. From here on, \( \hat{\theta} = g(\hat{\mu}) \) where \( \hat{\mu} \) is the estimate of the conditional mean of the beta regression model
and $g(.)$ is its link function. Expression (4) is a generic statement and depends on a methodology to determine $w_k$, $k = 1, \ldots, M$.

Estimation of $\text{Var}[g(\hat{\mu})]$ is as important as the value of $g(\hat{\mu})$ if inference is to be pursued. In order for model selection uncertainty to be more accurately accounted for, $\text{Var}[g(\hat{\mu})]$ must incorporate a variance component proportional to a multi-model estimation bias $\delta_k$ for model $m_k$, defined as $\delta_k = g(\mu_k) - g(\mu)$. The reason this estimation bias must be incorporated into the variance term is because its distribution across different models is unknown to the researcher before model selection and has a positive contribution to uncertainty in the estimates. Assume the expectation of $\delta_k$ over all competing models is equal to zero with notation $E_M(.)$ to signal expectation over the space of models $M$. With this supposition, we can write

$$E_M\left[g(\hat{\mu}_k)\right] = g(\mu) + E_M[\delta_k] = g(\mu_k)$$

(5) with the conditional variance of $g(\mu_k)$ given model $m_k$

$$\text{Var}\left[g(\hat{\mu}_k)\right] = E_M\left\{[g(\hat{\mu}_k) - g(\mu_k)]^2\right\}$$

(6) which can be seen as a general mean-squared error of using $g(\hat{\mu}_k)$ as an estimate for $g(\mu)$. An unconditional sampling variance of $g(\mu_k)$ can be determined using $\delta_k = g(\mu_k) - g(\mu)$ as the representation of the across model bias by

$$\text{Var}\left[g(\hat{\mu}_k)\right] = E_M\left\{[g(\hat{\mu}_k) - g(\mu)]^2\right\} = \text{Var}\left[g(\hat{\mu}_k)\right] + (\delta_k)^2.$$  

(7) Thus, variance of the model averaging estimator given in (4) under known constant weights and identically distributed $g(\hat{\mu}_k)$ is

$$\text{Var}\left[g(\hat{\mu})\right] = \sum_k w_k^2 \text{Var}\left[g(\hat{\mu}_k)\right] + \sum_k \sum_{l \neq k} w_k w_l \text{Cov}\left[g(\hat{\mu}_k), g(\hat{\mu}_l)\right].$$

(8) With the substitution, $\text{Cov}\left[g(\hat{\mu}_k), g(\hat{\mu}_l)\right] = \rho_{kl} \sqrt{\text{Var}\left[g(\hat{\mu}_k)\right] \text{Var}\left[g(\hat{\mu}_l)\right]}$, where $\rho_{kl}$ denotes the correlation coefficient between estimates $g(\hat{\mu}_k)$ and $g(\hat{\mu}_l)$ from models $m_k$ and $m_l$ and if it is further assumed $\rho_{kl} = \rho$ for all $k \neq l$, $k, l = 1, \ldots, M$ it can be established from (8) that

$$\sum_k \sum_{l \neq k} w_k w_l \text{Cov}\left[g(\hat{\mu}_k), g(\hat{\mu}_l)\right]$$

(9) $$= (1 - \rho) \sum_k w_k^2 \text{Var}\left[g(\hat{\mu}_k)\right] + \rho \left\{ \sum_k w_k \sqrt{\text{Var}\left[g(\hat{\mu}_k)\right]} \right\}^2.$$

Now, if we set $\rho = 1$ then

$$\text{Var}\left[g(\hat{\mu})\right] = \left\{ \sum_k w_k \sqrt{\text{Var}\left[g(\hat{\mu}_k)\right]} + \delta_k^2 \right\}^2.$$ 

(10) The above expression makes the simplifying assumption that $\delta_k$ and $\delta_l$ are perfectly correlated across data sets. This may appear to be a strong assumption but consider for $k = 2$ we have
we have that
totic properties of MLE estimates so that under standard MLE regularity conditions
is a consistent estimator of
penalize for uncertainty when
is performed by taking the variance term from (13) and substituting it in for the model
It then follows that inference on the model average estimators for beta regression models
observation,
(12) Var \( \hat{\Sigma} = \hat{\Sigma} \)

\[ \sum w_k^2 \text{Var} \left[ g(\hat{\mu}_k) \right] = \left\{ \sum w_k \sqrt{\text{Var} \left[ g(\hat{\mu}_k) \right]} \right\}^2 \]

and thus for any \( \rho < 1 \) the model-averaging variance in (10) will conservatively over-
penalize for uncertainty when \( \rho = 1 \) does not hold. Although shown for \( k = 2 \), by in-
duction, this results holds for any \( k \). On the other hand, for estimation on completely
independent data, where \( \rho = 0 \) the expression in (9) becomes

\[ \text{Var} \left[ g(\hat{\mu}) \right] = \sum w_k^2 \left\{ \text{Var} \left[ g(\hat{\mu}_k)|\hat{\delta}_k + \hat{\delta}_k^2 \right] \right\} \]

where \( \text{Var} \left[ g(\hat{\mu}_k)|\hat{\delta}_k \right] \) and \( \hat{\delta}_k = g(\hat{\mu}_k) - g(\hat{\mu}) \) are estimated for each model using valid
statistical estimation which is detailed in the next section. In practice, the extreme cases
of \( \rho = 0 \) or \( \rho = 1 \) do not likely hold perfectly but adopting either assumption enables
the derivation of closed-form expressions for the variance of model average estimates.

2.4. Inference with Beta Regression Model Averaging. Suppose a beta model estimated
on training data \( D_{\text{train}} \), with \( x_i \) and \( z_i \) available and fixed for each observation
\( i \) exists. Solving (3) yields MLE derived consistent estimates of \( \beta \) and \( \gamma \) represented
by \( \hat{\beta} \) and \( \hat{\gamma} \), respectively. Now, consider the subsequent calculation of some function of
interest, \( f_i(x_i, \hat{\beta}) \), calculated from the mean submodel predictors and estimates for the
\( i^{th} \) observation. In this work, \( f_i(x_i, \hat{\beta}) \) is the conditional expected value of the response
variable but is left in a more general form to emphasize \( f_i(.) \) could be replaced with the
expected value for \( \phi_i \) for which case it would depend on \( z_i \) and \( \gamma_i \). If it is assumed
that \( f_i(x_i, \hat{\beta}) \) is linear and differentiable in a local neighborhood of \( x_i \) then a linear
approximation given by first-order Taylor series expansion around \( f_i(x_i, \hat{\beta}) \) evaluated at
\( x_i \) is acceptably accurate asymptotically, provided \( \hat{\beta} \) displays appropriate convergence
behavior.

Let \( \Sigma = \text{Cov}(\hat{\beta}) \) be the variance-covariance matrix of the coefficients for the beta
regression mean submodel and \( \hat{\Sigma} = \text{Cov}(\hat{\beta}) \) denote its consistent estimator. Since \( \hat{\beta} \)
is a consistent estimator of \( \beta \), for large enough samples we can rely on the asymptotic
properties of MLE estimates so that under standard MLE regularity conditions
we have that \( \sqrt{n} \left[ f_i(x_i, \hat{\beta}) - f_i(x_i, \beta) \right] \sim \mathcal{N}(0, \nabla f_i(x_i, \beta)^T \Sigma \nabla f_i(x_i, \beta)) \). Further, the
variance of \( f_i(x_i, \beta) \) for observation \( i \) can be approximated asymptotically as,

\[ \text{Var} \left\{ f_i(x_i, \hat{\beta}) \right\} \approx \frac{1}{n} \nabla f_i(x_i, \hat{\beta})^T \hat{\Sigma} \nabla f_i(x_i, \hat{\beta}). \]

For large \( n \), the above asymptotic results indicates that for each model, \( k \), and
observation, \( i \), a normal sampling distribution for \( f_i(x_i^T \hat{\beta}_k) \) is a justified approximation.
It then follows that inference on the model average estimators for beta regression models
is performed by taking the variance term from (13) and substituting it in for the model
specific variance variance term \( \text{Var} \left[ g(\hat{\mu}_k)|\hat{\delta}_k \right] \) in (10) or (12) (depending on the assumption for
\( \rho \)) for each candidate model, \( m_k \). The multi-model estimation bias term \( \hat{\delta}_k \) for model
\( m_k \) and observation \( i \), on the scale of the link function, is given by

\[ \hat{\delta}_k = x_i^T \hat{\beta}_k - \sum_{k=1}^{M} w_k \left( x_i^T \hat{\beta}_k \right) \]
where the second term denotes the model average estimator. Therefore, the variance of the model average estimate of the conditional mean of the response on perfectly correlated data would be

\[
\text{Var}[g(\hat{\mu}_i)] \approx \left\{ \sum_k w_k \sqrt{\text{Var}[g(\hat{\mu}_{ik})|\delta_{ik}]} + \delta_{ik}^2 \right\}^2
\]

and on completely uncorrelated data

\[
\text{Var}[g(\hat{\mu}_i)] \approx \sum_k w_k^2 \left\{ \text{Var}[g(\hat{\mu}_{ik})|\delta_{ik}] + \delta_{ik}^2 \right\}.
\]

Hence, an approximate \((1 - \alpha)\) confidence interval for the conditional mean of the response is defined by,

\[
\text{lower limit} = f \left[ x_i^T \hat{\beta} - \Phi^{-1}(1 - \alpha/2) \frac{1}{\sqrt{n}} \sqrt{\nabla f_i(x_i, \hat{\beta})^T \hat{\Sigma} \nabla f_i(x_i, \hat{\beta})} \right]
\]

\[
\text{upper limit} = f \left[ x_i^T \hat{\beta} + \Phi^{-1}(1 - \alpha/2) \frac{1}{\sqrt{n}} \sqrt{\nabla f_i(x_i, \hat{\beta})^T \hat{\Sigma} \nabla f_i(x_i, \hat{\beta})} \right]
\]

where \(f(.)\) represents the logistic function for choice of the logit for \(g(.)\) and \(\Phi^{-1}\) is the inverse CDF for the standard normal distribution. This is expected to work well when the finite-sample distribution of \(g(y_i)\) is well approximated by the normal distribution and (13) is a reliable estimator of the variance of \(g(y_i)\). A similar derivation can be done to deduce an approximate confidence interval for conditional expectation of \(\phi_i\) where \(g^{-1}\) would be replaced with the inverse of the link function used for the precision submodel and \(x_i^T \hat{\beta}\) would be replaced with \(z_i^T \hat{\gamma}\). Asymptotic existence, consistency, and normality are used in the previous expressions to accomplish inference with the model average estimator. Resampling based techniques and the profile likelihood are possible alternatives where inference on model average estimators is desired but asymptotic assumptions are not met (Murphy and Vaart, 2000; Chalmers, 2017; DiCiccio and Efron, 1996).

3. Beta Regression Model Averaging Methodology.

3.1. Model Averaging Weighting Scheme. The weights in (4) can be understood to represent the confidence in the corresponding model’s closeness to the truth or its degree of reliability. Determination of these weights entails either scientific justification of priors (when using Bayesian methodology) or allowing the data to reveal the information. From a conceptual standpoint, an infinite amount of independent and identically distributed data could enable an experiment where model selection over the space of all beta regression models is performed a very large number of times. If our space of models includes one that is sufficiently close to the truth, then we might expect the probability of selecting this model to become close to one, given the sufficiently abundant data. If a best approximation is not singular, that is, our space admits several equally good models (failure to include the true, lets say) then the probability of selecting one of the best approximators may still exhibit some limiting behavior but not with any single model’s selection probability tending to one.
This idealized experiment is not available and so the next best thing when the true generating process is unknown and data is not of infinite extent is to mobilize bootstrap resampling. Bootstrapping provides access to sampling properties of the selection frequencies of each competing model, which will represent weights in the estimation of the model averaging estimator. Model-averaging based on the bootstrap was proposed by Buckland, Burnham and Augustin (1997), discussed by Burnham and Anderson (2002), extended by Augustin, Sauerbrei and Schumacher (2005), applied by Bucholz, Hollander and Sauerbrei (2008) to the linear regression model setting, and interrogated under different resampling schemes by Bin et al. (2016). Model selection uncertainty is not restricted to linear models and are even more warranted with non-linear two-parameter models capable of modeling mean and dispersion behavior simultaneously, such as a beta regression model. Because of this, development and verification of a bootstrap-based model averaging scheme for beta models is highly warranted and of significant importance to proper use of it in applied settings.

3.2. Model Selection Criteria. The choice of model selection criterion is crucial in the computation of bootstrap selection frequency weights that encode the relative quality of the competing models. Penalized likelihood-based criteria, including AIC and BIC, are widely used and readily recognized for their information-theoretical properties (Kuha, 2004). Both AIC and BIC are evaluated as 

\[
-2 \log L + q
\]

where \( L \) is the likelihood function evaluated at the maximum likelihood estimates and \( q \) is a criterion-specific penalty term. If we denote the number of predictors in the mean submodel as \( p \) and the number of predictors in the dispersion submodel as \( r \) then \( q = 2(p + r + 2) \) for AIC and \( q = \log(n)(p + r + 2) \) for BIC. It is possible to correct for small sample size by adjusting AIC to a bias-adjusted form called CAIC according to (Bin et al., 2016)

\[
CAIC = AIC + \frac{2(q)(q + 1)}{n - q - 1}.
\]

Although the CAIC is derived under Gaussian assumptions, it has been found to be useful in non-Gaussian models (Anderson, Burnham and White, 1994). Alternatively, a penalized pseudo \( R^2 \) criterion that represents a global measure of explained variation can be used. In the case of a beta regression model, \( R^2_p \) can be defined as the square sample correlation coefficient between the estimated linear predictor \( \hat{\eta} \) and the link transformed response \( g(y) \) penalized by the number of predictors appearing in the two submodels (Cribari-Neto and Zeileis, 2010)

\[
R^2_p = 1 - \left\{ \frac{(1 - R^2_c)(n - 1)}{n - (p + r + 2)} \right\}
\]

where \( R^2_c \) is the squared sample correlation coefficient between \( \hat{\eta} \) and \( g(y) \) given by

\[
R^2_c = \frac{\left\{ \sum_{i=1}^{n} [\hat{\eta}_i - \overline{\eta}] \left[g(y_i) - \overline{g(y)}\right] \right\}^2}{\sum_{i=1}^{n} [\hat{\eta}_i - \overline{\eta}]^2 \sum_{i=1}^{n} \left[g(y_i) - \overline{g(y)}\right]^2}, \quad 0 \leq R^2_c \leq 1.
\]
3.3. Bootstrap Procedure. With a choice made for model selection criteria, a bootstrap-based model averaging estimation for beta regression models is guided by the following procedure:

1. Generate \( B \) bootstrap samples from training data.
2. Apply a method of model selection to each bootstrap sample from step 1.
   - CAIC based backward elimination was enacted in the simulations.
3. Build record of if a model is selected from at least one bootstrap sample from step 2.
   - After \( B \) iterations, a vector of length \( C \) where \( C \) is the number of candidate models entering step 2 will result after step 3 with its \( i^{th} \) component corresponding to the selection frequency of model \( m_k \).
   - Model selection weights \( w_k \) correspond to the number of times model \( k \) is chosen in the \( B \) bootstrap iterations.
4. Fit all models with non-zero selection frequency to entire training data.
5. Use fitted models from 4 to obtain estimate \( \hat{\theta}_{ik} \) where \( i \) indicates the observation and \( k = 1, ..., M \) is the model index where \( |M| \) is the number of competing models returned in step 4.
   - Estimate \( \hat{\theta}_{ik} = \hat{\mu}_{ik} \) for the simulations but any identifiable parameter common to a certain subset of competing models could be estimated.
6. Compute the estimate of model averaging estimator \( \hat{\theta}_i \) for all observations \( i = 1, ..., n \) to get a vector of model averaging estimator realizations \( \hat{\theta}_i \).
   - \( \hat{\theta}_i = \hat{\mu}_i \) when estimation of conditional mean of the response is conducted whereas \( \hat{\theta}_i = \hat{\phi}_i \) when estimation of the average precision parameter is of interest.

4. Numerical Simulations. Simulations presented in this section were done using varying and fixed dispersion beta regression models as the underlying data generating mechanism. Mean submodels were comprised of \( p = 10 \) predictors whereas precision submodels had only \( q = 1 \) predictor. The mean and precision models did not share predictors. Link functions were restricted to the logit for the mean and log for the precision submodels. Both the regression structures for the mean and precision allow for an intercept in their descriptions. Coverage probabilities below the nominal level of 95% would indicate an underestimation of variance whereas coverage probability above the nominal level would indicate overestimation of variance. Improvement in the consistency of estimates and a smaller variance of those estimates from data set to data set should correspond to generally higher predictive accuracy of MA estimators. Three measures of out-of-sample performance were investigated - penalized pseudo \( R^2 \) (\( R^2_p \)), logarithmic score (\( \log S \)), and continuous ranked probability score (\( CRPS \)) - with definitions and explanations in Section C of the Supplementary Material. Constant and non-constant dispersion beta regression models were investigated with the setup and results for the constant dispersion case provided in Section E of the Supplementary Material. Moreover, Logarithmic scoring is presented in the main work with results for the other two metrics, \( CRPS \) and \( R^2_p \), contained in Sections G of the Supplementary Material (Allenbrand and Sherwood, 2021).

4.1. Data Generation. The data generating mechanisms were comprised of a univariate response variable \( Y \) where \( Y_i | X_i, Z_i \sim Beta(\mu_i, \phi_i) \) with \( \mu_i = 1/(1 + e^{x_i^T \beta}) \) and

\[
\phi_i = \begin{cases} 
  e^{z_i^T \gamma}, & \text{if non-constant dispersion} \\
  \phi, & \text{if constant dispersion}
\end{cases}
\]
with vectors of mean predictors given by $x_i$ where $x_{ip} \sim U(0, 1)$ for all $i = 1, \ldots, n$, $p = 1, \ldots, 10$ and precision predictors given by $z_i \sim U(0, 1)$ for all $i = 1, \ldots, n$. Different values of parameters for $\beta$ and $\gamma$ were investigated, as displayed in Table 1. Values for the parameters were chosen to interrogate the influence of different patterns in predictors on precision and accuracy of estimates returned by the models. Specific features of the data generating process are surmised to contribute to greater uncertainty in the estimates and are examined in the simulations: a mix of positive and negative predictor influence with range in magnitude, the presence of noise variables with no influence, predictors whose influence is close in magnitude with gradual increases or decreases between them. Another feature explored by the various settings was the impact of variability, controlled by magnitude of $\phi$, on model uncertainty and prediction accuracy. It is thought that data contaminated with random variation should confer greater uncertainty about model quantities.

| Setting | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | $\beta_5$ | $\beta_6$ | $\beta_7$ | $\beta_8$ | $\beta_9$ | $\beta_{10}$ | $\gamma_0$ | $\gamma_1$
|---------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| 1       | 1         | -1.50     | -1.50     | 0.00      | 0.00      | 0.00      | 0.00      | 0.00      | 1.50      | 1.50      | 1.50      | 1          | 5.0
| 2       | 1         | -1.50     | -1.00     | -0.50     | 0.00      | 0.00      | 0.00      | 1.50      | 1.00      | 0.50      | 1          | 2.5
| 3       | 1         | -3.00     | -0.75     | -0.50     | -0.25     | 0.00      | 3.00      | 0.75      | 0.50      | 0.25      | 1          | 5.0
| 4       | 1         | 1.50      | 0.75      | 0.50      | 0.25      | 0.00      | 0.00      | 0.00      | 4.00      | 2.00      | 1          | 2.5
| 5       | 1         | -1.50     | -0.75     | -0.50     | -0.25     | -0.10     | 0.00      | 0.00      | 0.50      | 2.00      | 1          | 5.0
| 6       | 1         | 0.75      | 0.50      | 0.25      | 0.10      | 0.00      | 0.00      | -0.75     | -0.50     | -0.25     | -0.10     | 1          | 10.0
| 7       | 1         | 0.75      | 0.60      | 0.45      | 0.30      | 0.15      | 0.10      | 0.00      | 0.00      | -4.00     | 1          | 5.0

4.2. Simulations Setup. The model averaging procedure described in Section 3.3 was followed in order to estimate the model averaging estimator of the conditional mean, $\mu_i$, for the beta-distributed response variable. Simulations were calibrated with the following values: $D_{train}$ and $D_{test}$ were of sample size $n = 100$, the number of bootstrap samples for step one of the MAP was set to $B = 500$, second-order corrected AIC (CAIC) was used in the backward elimination variable selection in step two. In addition to the MA estimator, estimates of $\mu_i$ from non-model averaging based techniques including AIC, BIC, CAIC, $R^2_p$, and the true model were computed. All models from the different model selection techniques were fitted on the same $D_{train}$ and had accuracy evaluated on the same testing data $D_{test}$. Accuracy and quality of fit to out-of-sample data was appraised with $R^2_p$, log S, and CRPS. All metrics were computed from 100 Monte Carlo simulations using $n = 100$ sample size.

Coverage probability of each model selection technique was computed through a Monte Carlo process by taking the models selected on $D_{train}$ and generating estimates of $\mu_i$ and $\phi_i$ for each observation $i$ on an independent data set $D_{conf}$, of size $n = 500$, created by the same process that established $D_{train}$. From these estimates, 95% confidence intervals were formed according to equation (17). A larger sample size was operated with so that the asymptotic arguments used to derive the 95% CI’s were more likely to be valid. This was repeated 100 times on new and independent $D_{conf}$. At the end of each of the 100 iterations, 500 CIs were determined for each of the 500 estimates of $\mu$. An arithmetic mean of the coverage probabilities for the 500 CIs was computed at each Monte Carlo iteration to attain within-data averages so as to avoid results that were reliant on only a small subset of the data. This was done because bias in
the estimates are not random across the data for beta regression models and a focus on only a small fraction of observations would introduce selection bias into the results and the statistical quality of a technique’s inference should not be assessed narrowly. Within-data averages were then averaged across the 100 Monte Carlo iterations. In each of the 100 Monte Carlo iterations, all model selection techniques had coverage assessed on the same $D_{con}$. In effect, the coverage probabilities computed this way fit into the paradigm argued for in this paper that statements about findings must not circumvent major sources of uncertainty.

4.3. Simulation Results and Discussion.

4.3.1. Coverage Rate Probability. Model averaged (MA) estimates of $\mu$ yielded coverage probabilities closer to the nominal level of 95% for the MA-based selection process when compared to single model selection based on AIC, BIC, CAIC, or $R^2_p$ for both constant and non-constant precision models across all seven settings. Readers are directed to Section D of the Supplementary Material for elaboration on the importance of this finding. (Allenbrand and Sherwood, 2021). For non-constant precision models, the average coverage rate across the settings ranged from 93% for setting seven to 97% for setting five as displayed in Figure 1. The true model with a fully correct specification always achieved coverage beyond that of the MA approach and met or exceeded the nominal level in all settings. Conditional CIs from the single model selection techniques were below 86% in all settings with each technique exhibiting coverage below 70% in at least one setting. Setting seven included predictors with small to moderately positive influence, a single predictor with a large negative influence, and three noise predictors with no effect. Findings were similar for the constant precision models. Mechanisms with mean effects that span positive/negative values and that show variable magnitudes, contain noise, and exhibit gradual changes in predictor effects would benefit the most from a MA-based approach to inferences about estimator precision. For the non-constant precision beta regression models, the average number of models supported by the Monte Carlo simulations across the seven settings was 52 whereas the average for the constant precision was 27. More models indicate more uncertainty that would have been ignored in inference had model averaging been denied. Overall, these findings support the necessity of directly factoring model selection uncertainty and bias into model inference if assertions about precision and reliability are sought.

4.3.2. Out-of-Sample Performance. Results from the previous section illustrated the consequences of ignoring model selection uncertainty on precision of estimates. However, a more precise estimate is not equivalent to a more accurate one and obtaining improved precision at the expense of poorer out-of-sample behavior is not a favorable tradeoff. Results depicted in Figure 2 and the results found in Section G of the Supplementary Material provide evidence for MA-based estimation’s capability of achieving out-of-sample quality similar to, if not better than, single-model estimation while possessing greater estimator precision (Allenbrand and Sherwood, 2021). Also of note is that MA-based prediction performance is more stable across the settings as opposed to non-MA techniques that alternate more in rankings on the metrics.
The logarithmic scoring results in Figure 2 demonstrates that MA-based models obtained values for $\log S$ comparable to those obtained from the other techniques. Interestingly, BIC-based models ranked the worst on out-of-sample fit when assessed by $\log S$ on non-constant precision models which completely inverts BIC’s rank when assessed by $R^2_p$ (Allenbrand and Sherwood, 2021). MA-based log scores were in the top three for all settings besides setting 1 for the non-precision model and settings four and five for the constant precision models. The MA-based model selection achieved greater reliability in prediction performance by always achieving a value of the metrics near the top performing technique and never being the worst performing technique except for setting 1. All other model selection techniques demonstrated worst-performance
in multiple settings. Setting 1 included the greatest number of noise variables (60%) with the other four predictors being split into subsets, one positive and one negative, with identical magnitudes of influence in each. In the presence of several irrelevant predictors, many of the potential fitted models returned by the MA approach included these irrelevant predictors. From the simulation results, the model averaging method improves precision of estimates and yields models that are competitive or superior to single model selection techniques in terms of prediction accuracy. In the next section, MA estimation with beta regression models is assessed on an applied problem using clickstream data from an online retailer.

5. Empirical Application. When it comes to an electronic commerce (eCommerce) platforms, keeping visitors on the site long enough to induce a sale or some other type of conversion can be understood as a fundamental requirement for success of the site. Possession of a model that faithfully describes and can predict a visitor’s path through the site, referred to as the clickstream, and subsequent exit point would strongly benefit website and business managers. Knowing visitors may differ markedly with respect to their characteristics, the belief that a single model would remain stable across replicates of clickstream data is wishful thinking. Since the rate by which visitors leave a page is bounded with a non-symmetric variance, a beta model represents a natural choice to model dependencies in the clickstream data. Hence, the beta regression model averaging framework introduced earlier is applied to clickstream data from an online retailer to examine the advantages and applicability of this method to an empirical case. Before presenting results of this application, a brief introduction to clickstream variables and their analysis is presented.

5.1. Clickstream Data and Web Analytics. Digitally mediated transactions are becoming increasingly more ubiquitous as the advantages of speed, convenience, access to personalized options, and quality of online systems increase. In particular, the purchasing and selling of goods electronically, known as eCommerce, continues to demonstrate an increasing share of total sales (Ramanathan et al., 2018). Digitalization of economic interactions was stimulated by the COVID-19 pandemic and will not likely recede into the future. Extraction of a benefit from this continually growing technological trend depends crucially on tools that can identify and analyze patterns contained in data generated by the online systems and yield actionable insights (Ansari et al., 2002). In the setting of online retail, web analytics (WA), sometimes referred to as webometrics, represents a family of methods and tools that can be employed to deliver insights about the relationships between online visitor behavior, visitor characteristics, and decisions made on the site (Chaffey and Patron, 2012; Thelwall, 2010). Efficacy and validity of insights from WA are based on a plethora of supposed relationships between numerous metrics and user behavior which are collectively referred to as clickstream variables (Chaterjee, Hoffman and Novak, 2003; Moore and Fader, 2004).
Fig 2: Out-of-sample quality as measured by the average \( \log S \) for the non-constant precision beta regression model. Panel (a) shows results for setting 1, (b) for setting 2, (c) for setting 3, (d) for setting 4, (e) for setting 5, (f) for setting 6, and (g) for setting 7 where details of settings are found in Table 1. Average \( \log S \) values were computed over 100 Monte Carlo simulations with independent testing data.

Standard categories of clickstream variables include site usage, referral and web traffic, site content analysis, and quality assurance. Site usage encompasses measures such as geographic information and number and type of user visits. Referral and web traffic comprises measures such as source or web traffic onto a site and search terms used to locate the site. Site content analysis contains measures for effectiveness of site content and top pages with respect to exits and value. Quality assurance measures the presence and impact of broken pages and visitor errors (Hansen, 2009).

An important web analytic quantity to track and model is traffic quality. (Booth and Jansen, 2008). Connected to this idea of traffic quality are the bounce and exit rates of the pages that constitute the website. The essence of these two metrics is that bounce rate quantifies how many arrivals to a page leave immediately whereas exit rate provides
information on which pages of a site are contributing, expectedly or unexpectedly, to customer departure (Editor Google Analytics, 2020). Many factors determine bounce and exit decisions but bounces are theorized to be generated by customers with no interest in the content of the site and is largely related to factors external to the website. Exit rate values, on the other hand, are generated by the interaction between internal factors of the website and the user. In other words, the human-computer interactions that occur at the user interface of a website are theorized to dictate exit rate decisions. Exits are inevitable but the page location of the exit and whether or not this location is suitable for an exit becomes a question of optimization.

5.2. Clickstream Relationships. An interpretable understanding of a visitor’s behavior on an eCommerce platform equips a firm to rationally design certain features of the platform. A classic model of the purchasing behavior posits that the purchase process begins with an intention to buy but can become activated such as when a customer makes contact with informative content and a new impulse to purchase is created (Howard and Sheith, 1969; Cialdini, 2001; Baumeister, 2002). This intent to purchase can easily be understood to be time dependent such that variation in purchase intention can be expected to occur between days or months (Vohs and Faber, 2007). Similar to an experience in a physical store, insufficient choice or information are known to be significant predictors of shopping cart abandonment (Cho, Kang and Cheon, 2006; Park and Kim, 2010). Not only insufficient information, but any force that raises customer irritation with the online environment predisposes abandonment and departure from the site. Negative correlations between navigational and informational aspects of website design with irritation were found with consumer irritation being particularly sensitive to navigational design (Hasan, 2016). In addition to irritation, any form of mismatch between customer expectation and experience will engender disappointment and an increase probability of abandonment of the site (Swinyard and Smith, 2003). In the setting of eCommerce, the compatibility between the shopping platform and the customer’s prior experiences, expectations, or lifestyle becomes a relevant factor. It is generally accepted that decisions regarding the use of a technology follow from arguments based on compatibility and usefulness. Components that may influence the compatibility of an eCommerce site with a user include demography such as region of purchase, technological preferences such as use of a specific browser, or temporal factors such as that day of the week.

A representation of the complex relationships possibly discoverable between clickstream variables and the two outcomes (exit and bounce rate) are presented in Figure 3. Formation of Figure 3 was partly guided by the literature but includes exploratory elements. Elucidation of and commentary on Figure 3 is provided in Section H of the Supplementary Material (Allenbrand and Sherwood, 2021). As an owner, manager, or web designer, a data-driven approach to design, redesign, and maintenance is appealing for many of the same reasons data-driven approaches to other problems is, understanding of the existence and strength of relationships between clickstream variables. A disciplined and principled methodology for extracting insight from clickstream data is needed.
Fig 3: A conceptual model of the possible interactions between clickstream variables. Directed arrows indicate the proposition that the variable which is the origin of the arrow is an independent variable whereas variables that are terminal are considered dependent, response variables. Solid lines signify direct relationships between clickstream variables and the decision to bounce or exit from the platform. Dashed lines represent direct relationships between clickstream variables and the decision to buy on the platform. The nature of the relationships are most likely non-linear and complex but the image shows an abstraction of them.

The sheer number of variables and their complex interrelationships in Figure 3 could result in a substantial amount of noise permeating the signal. Failures to choose proper models for the outcome variable and rigorously incorporate multi-model selection uncertainty has the potential to generate dynamically inconsistent and unstable modeling findings and post-modeling decisions. Dynamic inconsistency and instability is used here to denote modeling results that change, sometimes dramatically, from dataset to dataset. Conclusions about clickstream relationships should not be excessively dependent on chance variation, otherwise, the insights are fragile. With these considerations in mind, the bootstrap-based beta regression multi-model selection procedure was applied to clickstream data to ascertain stability and consistency of the relationships between exit and bounce rate value with other clickstream variables. Also, a more robust indicator of predictor variable importance is developed, where robustness is achieved with respect to inter-data variability and model selection uncertainty. Before presenting the results, a description of the clickstream data and beta regression models are provided.
5.3. Empirical Data. The dataset is available on UC Irvine’s machine learning repository for public use and is comprised of clickstream data gathered from 12,172 sessions from an online retailer’s web platform. It was collected and used by Sakar et al. (2019) to predict purchasing intention of platform users using multilayer perceptron and long-short term memory recurrent neural networks. Formation of the dataset was done in such a way so that each session corresponded to a different user over a one year time frame through a process called sessionization. Sessionization ultimately allows the sequence of actions taken by a visitor to be ordered and pairs online actions with user identification. Basically, session IDs generated when a user visits a web page are linked to the actions taken while that user is on the page, such as clicking a button or filling in a data field, and these tuples are then linked to a unique user ID (Mehrzadi and Feitelson, 2012). A primary reason for this sessionization was to avoid any tendency towards a specific advertisement campaign, special day, user profile, or time period (Sakar et al., 2019). Table 2 provides a description of the variables.

Each clickstream variable used was a member of one of the four classes of clickstream variables depicted in Figure 3. It will be shown in later sections that the model-averaging of beta regression models procedure provides a method to test the relationships in Figure 3 by way of model selection and variable inclusion frequencies. Exit and bounce rates vary between pages of a website visited by a user with values determined by previous arrival and departure events of other visitors. Observed values for exit and bounce rate were not strictly in $(0, 1)$ interval because of the nature of exit and bounce values. For instance, a user who chooses not to bounce following arrival on the site would have a bounce rate value of zero. Existence of zeros and ones violates the open interval requirement for the beta regression model parameterization used in this paper and hence a data transformation was employed to shift the the zero and ones slightly towards the center of the distribution. Formally, for response variable $y \in [0, 1]$ the transformation $[y \times (n - 1) + 0.5]/n$, where $n$ indicates the sample size, makes it so $y \in (0, 1)$. This data transformation was first introduced in Smithson and Verkuilen (2006).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Class of Variable</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Administrative</td>
<td>Number of pages visited about account management</td>
<td>Site Content and Quality</td>
<td>Integer</td>
</tr>
<tr>
<td>Informational</td>
<td>Number of pages visited about platform</td>
<td>Site Content and Quality</td>
<td>Integer</td>
</tr>
<tr>
<td>Product related</td>
<td>Number of pages visited about products</td>
<td>Site Content and Quality</td>
<td>Integer</td>
</tr>
<tr>
<td>Administrative duration</td>
<td>Total time spent on administrative pages (seconds)</td>
<td>Site Content and Quality</td>
<td>Integer</td>
</tr>
<tr>
<td>Informational duration</td>
<td>Total time spent on informational pages (seconds)</td>
<td>Site Content and Quality</td>
<td>Integer</td>
</tr>
<tr>
<td>ProductRelated duration</td>
<td>Total time spent on product related pages (seconds)</td>
<td>Site Content and Quality</td>
<td>Integer</td>
</tr>
<tr>
<td>Bounce rates</td>
<td>Average bounce rate of pages visited by user</td>
<td>Outcome</td>
<td>Proportional</td>
</tr>
<tr>
<td>Exit rates</td>
<td>Average exit rate of pages visited by user</td>
<td>Outcome</td>
<td>Proportional</td>
</tr>
<tr>
<td>Page values</td>
<td>Average page values of pages visited by user</td>
<td>Site Content and Quality</td>
<td>Integer</td>
</tr>
<tr>
<td>SpecialDay</td>
<td>Closeness of the session to a holiday</td>
<td>Temporal</td>
<td>Ratio</td>
</tr>
<tr>
<td>Month</td>
<td>Calendar Month of a Session</td>
<td>Temporal</td>
<td>Categorical</td>
</tr>
<tr>
<td>Operating Systems</td>
<td>Operating system of user</td>
<td>Technological</td>
<td>Categorical</td>
</tr>
<tr>
<td>Browser</td>
<td>Web browser of the user</td>
<td>Technological</td>
<td>Categorical</td>
</tr>
<tr>
<td>Region</td>
<td>Geographic region of session</td>
<td>Demographics and Site Usage</td>
<td>Categorical</td>
</tr>
<tr>
<td>Traffic type</td>
<td>Traffic source for arrival of user to site</td>
<td>Demographics and Site Usage</td>
<td>Categorical</td>
</tr>
<tr>
<td>Weekend</td>
<td>Site visited on a weekend or not</td>
<td>Demographics and Site Usage</td>
<td>Categorical</td>
</tr>
<tr>
<td>Visitor type</td>
<td>User is a new or returning visitor</td>
<td>Session finalized with a transaction or not</td>
<td>Dichotomous</td>
</tr>
<tr>
<td>Revenue</td>
<td></td>
<td>Outcome</td>
<td>Dichotomous</td>
</tr>
</tbody>
</table>

5.4. Exit and Bounce Rate Models. Average exit and bounce rate values, being proportions bounded in [0, 1] and with no a priori justification for symmetry or homoscedasticity, are candidates for being modeled as beta-distributed response variables.
whose conditional distributions are induced by the other clickstream variables from Table 2. Support for the validity of a beta distribution model for the exit and bounce rate dependent variables was gathered from a comparison of the data distribution to a theoretical beta distribution displayed in Figure 4. Two additional sources of support for the use of a beta distribution model are included. First, an appraisal of scaled third and fourth moments (skewness and kurtosis, respectively) for a set of different theoretical distributions revealed observed bounce and exit rate data had tailedness and skewness consistent with a beta distribution, see Section I of the Supplementary Material (Allenbrand and Sherwood, 2021). Second, use of the Kolmogorov-Smirnov (KS) statistic yielded evidence for good agreement between empirical values and a theoretical beta distribution. Both shape parameters used to specify the theoretical beta CDF were the MLEs in Figure 4 obtained after fitting a beta distributions to the exit and bounce rate data. Exit rate had a KS statistic of 0.106 with p-value of 0.08 whereas bounce rate had a KS statistic of 0.031 with p-value of 0.4. The question on the reasonableness of a statistical model for data is not answerable solely from a formal hypothesis test, so it was decided to conduct KS tests for both exit and bounce rate using normal, log normal, and logistic distributions as alternatives to the beta distribution. Distributional parameters for all three alternatives \((\mu, \sigma)\) for the normal, \((\mu_{\log}, \sigma_{\log})\) for log normal, and (location, scale) for logistic were determined by ML estimation from the exit and bounce rate data. It can be seen from Table 3 that KS statistics were smallest for the beta distribution and had the largest p-values.

<table>
<thead>
<tr>
<th>Response</th>
<th>Distribution</th>
<th>KS Statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ExitRate</td>
<td>Beta</td>
<td>0.106</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>Normal</td>
<td>0.52</td>
<td>&lt;&lt;&lt;&lt; 0.05</td>
</tr>
<tr>
<td></td>
<td>Lognormal</td>
<td>0.14</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>Logistic</td>
<td>0.13</td>
<td>0.06</td>
</tr>
<tr>
<td>BounceRate</td>
<td>Beta</td>
<td>0.031</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>Normal</td>
<td>0.67</td>
<td>&lt;&lt;&lt;&lt; 0.05</td>
</tr>
<tr>
<td></td>
<td>Lognormal</td>
<td>0.22</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>Logistic</td>
<td>0.11</td>
<td>0.08</td>
</tr>
</tbody>
</table>

Table 3

Values of KS-test statistics for the beta, normal, lognormal, and logistic reference distributions. The magnitude of the KS statistic is interpreted as a ranking for the suitability of the corresponding reference distribution for modeling the probabilistic behavior of exit and bounce rate.
Fig 4: Plot comparing empirical exit (left) and bounce (right) rate data to a maximum likelihood estimation based best fit theoretical beta distribution. MLE based estimates for the first and second shape parameters of the beta distribution are 0.862 and 4.28 for exit rate and 0.206 and 3.36 for bounce rate. Red line indicates the theoretical distribution whereas black indicates the observed.

Overall, both data-based and theoretical considerations promote the use of a conditional beta distributional models for the exit and bounce rate, ExitRate$_i \sim \text{Beta}(\mu_{ER}^{iBR}, \phi_{ER})$ and BounceRate$_i \sim \text{Beta}(\mu_{BR}^{iBR}, \phi_{BR})$ with $\mu_{ER} = 1/(1 + e^{-\eta_{ER}^i})$, $\mu_{BR} = 1/(1 + e^{-\eta_{BR}^i})$. Using a slight abuse of notation, $\eta_{ER}^i = x_i^T \beta_{ER}$ and $\eta_{BR}^i = x_i^T \beta_{BR}$ where $x_i$ is a vector of both categorical and non-categorical predictor variables for user $i$, $\beta_{ER}$ is the vector of regression coefficients for the mean submodel for ExitRate, and $\beta_{BR}$ is vector of coefficients for BounceRate. Predictors included all variables from Table 2 excluding exit rate, bounce rate, page value, and revenue.

Both $\phi_{ER}$ and $\phi_{BR}$ are modeled as constants. The decision to limit analysis to the constant precision setting was empirically motivated as it is not clear if variability in either exit rate or bounce rate value were influenced by a small set of regressors or a complex interplay between them. Moreover, as was revealed in the simulation results, the addition of a regression structure on $\phi$ amplified the results in favor of a need for model averaging. Hence, a restriction to constant dispersion would underestimate the benefits of model averaging in the presence of non-constant dispersion. If dispersion is truly constant, results regarding the value of model-averaging would be closer to the truth. Conclusions regarding variable importance may suffer instability issues and
may not be robust to random variations in different clickstream datasets. To interrogate these points, an assessment on model selection consistency, variable inclusion frequency, and out-of-sample prediction accuracy were conducted on the bounce and exit rate models fitted using both multi-model bootstrap model averaging and single model selection based on AIC, BIC, and $R_p^2$.

5.5. Empirical Results and Discussion. Empirical beta regression analysis of average exit and bounce rate values was conducted with bootstrap model averaging (MA) over the entire clickstream data set with $B = 500$ resampled data sets using AIC as the selection criterion for each iteration of the procedure. The non-adjusted AIC was utilized in the empirical work, instead of CAIC, because CAIC shows high agreement with AIC on large samples and its more recognizable among practitioners (Hurvich and Tsai, 1989). The decision to implement AIC over BIC selection comes down to the a priori belief that online behavior can be approximated to some degree of accuracy but never fully captured. Single-model selection using AIC, BIC, and $R_p^2$ was conducted to compare results for the two different methodologies. Results for average bounce rate are presented in the main work, results for average exit rate are located in Section J of the Supplementary Material (Allenbrand and Sherwood, 2021).

5.5.1. Model Selection Frequency and Consistency. Model selection frequencies from the bootstrap replications quantify how likely a particular set of predictors is selected. With constant $\phi$ and $p = 13$ predictors considered, there were $2^{13} = 8192$ competing models in the candidate model space for both exit and bounce rate. The effect of random data variability can be seen in the model selection frequency results across the bootstrap resamples for bounce rate as shown in Figure 5. For the bounce rate outcome, nine models were found to have non-zero support on the data with selection frequency ranging maximum support for model 1 (m1) at 64% support and minimum support for model 6 (m6) with only 0.5% support. About 56% of the models received negligible support, defined here as a selection frequency below 0.05, and could reasonably have been arrived at due to stochasticity in a particular data sets which would not be expected to be found in a different set of data. Of the remaining models, three received support over 5%, indicating a possibly true yet weak signal in the data. Existence of more than one model supported by the data would prove to be inauspicious for managers whose responsibility it is to modify attributes of the web system in response to findings from model development. This follows from the recognition that any of the nine models could be selected depending on which random sample of the data is used for analysis. Low model selection consistency generates greater model selection uncertainty with reduced confidence in the judgement made from any single model. Expensive decisions could quickly become nullified as soon as new clickstream data supports a new model. Instability coming from the model fitting technique translates to instability in business decisions.

Information contained in Figure 5 would be valuable to a manager, owner, web designer, or any other stakeholder in an eCommerce platform for multiple reasons. First, the models that did not receive any support during the bootstrap-MA procedure would be deemed unreliable and not be used to make design, marketing, or other structural decisions to the eCommerce platform. For example, solely increasing the number of pages on your site when models with site content independent variables included received low support should be avoided by the logic of the bootstrap-MA procedure. Effectively, the bootstrap-MA gave each model the ability to find support over the data; if the data and bootstrap simulations were large enough, no support is evidence
that the phenomena producing the clickstream data does not abide by the unsupported model. Second, the selection frequencies in Figure 5 can be interpreted as rankings of support from the data. If the sample size and bootstrap simulations were large enough, the frequencies could be considered as rankings of the degree to which the model approximates the data generating phenomena. A model with high ranking, such as model m1, would be endowed with greater decisions support influence. According to Table 3, model m1 largely selected non-site specific variables such as TrafficType, VisitorType, and Region while not selecting site specific variables like Informational, ProductRelated, and Administrative Duration. This is highly consistent with the directed relationships posited in Figure 3 which indicates that site content and quality clickstream variables should not have a consistent influence on bounce decisions. According to this, a focus on appealing to visitors through initiatives that target how the visitor reaches the site (TrafficType), improve retention (VisitorType), and tailor the experience to geographical information (Region) would be more effective than initiatives that try to increase the time a visitor spends on pages (Administrative Duration, Informational Duration, ProductRelated Duration) and if the visitor comes during a weekend or not (Weekend). When the mechanism of bouncing from a platform is considered, the internal details of the platform, including the number of product related pages and time spent on them, are irrelevant. Instead, factors external to the platform, such as the source from which traffic is coming, are dominant. Model averaging the beta regression models consistently chose the model closely aligned with this phenomenological understanding of bouncing from a platform.

5.5.2. Variable Inclusion and Relative Importance. Model selection bias may appear when a predictor variable has little to no effect on the observed average exit or bounce rate value but may appear so by random distortion of the signal carried by inter-data noise (Miller, 1990). Problems like this are expected when identical data is used to both select and make inferences from a model (Leamer, 1969; Copas, 1983; Lehmann, 1983; Breiman and Freedman, 1983; Breiman, 1992). So what are the negative consequences of including marginally to barely important variables? In data sets where they are selected, estimates of coefficients would be biased away from zero and inference to the population will exaggerate the influence this variable has on the response (Ye, 1998). Searches for patterns in noise can be costly. In light of these points, model averaging can serve as a more robust method of disentangling variable importance. Predictor variable inclusion frequency and variable make-up for bounce rate of the MA selected models, AIC, BIC, and $R^2_p$ singly selected models are exhibited in Table 4.

Individuals responsible for the qualities of a eCommerce platform face the need to know how to allocate resources to different aspects of the platform. Each explanatory clickstream variable signifies certain features of the platform. A simple question looms large, which variables should be acted on and in what order. It can be seen that the variable inclusion frequencies in Table 4 provides a succinct answer to this question, you act in the order of the inclusion frequencies. For instance, the source of the visitor (traffic type) being included in 100% of the bootstrapped models ranks it as a priority feature to focus on. If web traffic coming from social media sites is the dominant source then it would be clear that keeping a presence on these sites would be important. Conversely, if web traffic from other private web sites was found to be small, then efforts to form partnerships with other companies with the sole objective of getting a presence on their website could be avoided. On the other hand, finding that the day a visitor traffics to a site being a weekend or not (Weekend) is chosen in only 33% of bootstrapped models means it should still be investigated further but does not take precedence like those

...
ranking above it. In fact, the inclusion frequencies could be categorized into tiers - Tier 1, Tier 2, Tier 3 - where Tier 1 factors are those whose inclusion frequency is in the top quartile, Tier 2 factors are in the middle two quartiles, and Tier 3 factors are in the bottom quartile. A company protocol could then be introduced where Tier 1 factors are acted on first followed by Tier 2 then Tier 3 factors. The methodology presented can serve as a project organization and prioritization tool.

Inclusion frequencies can also serve as a test of the relationships conjectured in Figure 3. Inclusion frequencies that are high for certain clickstream variables operate as evidential support for the directed edge from that clickstream variable to the outcome under consideration. Once a visitor traffics onto a site and makes the decision not to bounce, the path traversed by this visitor through the site and the location of departure is largely governed by the human-system interaction. If this were true, one would expect to see variables that correspond to factors intrinsic to the website or the visitor to be repeatedly selected in models for average exit rate behavior over independent data. Results for the exit rate models confirmed that site content related variables were the most frequently included variables whereas those unrelated to site content received considerably less selection frequency. Conversely, the decision to bounce occurs before exploration of more than one page occurs and is considered a more extrinsically determined process. Hence, repeated analysis of a beta regression model over independent data should yield indications of importance for explanatory variables not directly linked to the site content and quality. Upon examining the results in Table 4, variables with 100% inclusion in all of the nine different bounce rate models selected over the 500 bootstrap replications are reasonably independent of the website besides the number of administrative pages viewed. Interestingly, the five variables that do correspond to site content and quality - informational, informational duration, product related, product related duration, administrative duration - received less than 45% inclusion in the models. These multi-model and resampling-based results provide evidence that the connections and non-connections surmised in Figure 3 are supported in this data.

5.5.3. Out-of-Sample Prediction Accuracy. The stabilization effect of model averaging (MA) should be exhibited in the problem of prediction on unseen data. In order to determine if the MA approach contributes positively to prediction error, 10-fold cross-validation was performed over the entire clickstream data set. Prediction of exit and bounce rate values on out-of-sample data can encounter the difficulty of imbalanced categorical variables. To control this, observations that did not have the same levels as the data used to fit the models were dropped. Only traffic type and browser variables experienced this issue with no more than three observations dropped in any iteration of cross-validation. Results for the three out-of-sample accuracy measures used in the simulation study are given in Table 5.

For both responses, the MA procedure achieved values for all three measures of out-of-sample performance greater than or equal to its single-model counterparts. Out of the single-model techniques, BIC-based backward elimination fared the worst for all three metrics. Basing selection on AIC or $R^2_p$ yielded larger models which would be expected to perform slightly better when predicting complex phenomena. From the standpoint of prediction error, MA acts as a sort of voting function with smoothing properties that up-votes predictions from models with more pronounced support on the data and down-votes those predictions from models that have lesser support. Looking at the results in Table 5 it is noticed that MA obtained comparatively better results than its competitors when predicting exit rate value than when predicting bounce rate value. It was found that no model that made up the exit rate MA estimator commanded
a disproportionate amount of support whereas support prepotency did exist for one model in the bounce rate MA estimator. Accordingly, one would expect the smoothing action of MA to be more pronounced in the exit rate prediction setting and less so in the bounce rate prediction setting.

One implication of these findings is in the pursuit of tailoring a visitor’s experience in real-time on the eCommerce platform, a problem of predicting how tuning a feature of the platform impacts resultant outcomes. Suppose knowledge of a visitor’s properties and behavior thus far on the platform is to be used to tune his future behavior on the platform, in particular, his exit and bounce rate behavior. What is the best way to go about achieving this? The most frequently selected explanatory variables can be regarded as direct votes for what features of the platform should be tuned. For instance, Table 4 shows that model averaging regarded the number of product related pages visited, the time spend on informational pages, and if visitation occurred on a weekend to be some of the least important factors for understanding bounce behavior. However, using a single-model selected via AIC would result in a conclusion that these factor are important. If a visitor comes on to the site, employing strategies to keep in on informational pages longer would be supported by a single AIC based model. However, when it comes to bounce behavior, the visitor does not go beyond the arrival page and decides to leave before exploring. In other words, factors other than the information contained elsewhere on the site were operative in the decision. Hence, not using the model-averaged results would amount to manipulating the visitor using unstable and inconsistent advice. However, using the model averaging approach, would be equivalent to first stabilizing the advice before using it to influence the visitor, a safer method. Similar limitations are incurred if using either BIC or $R^2_p$ based single-models. For example, under the BIC metric, the geographical region a visitor is in is considered unimportant when it comes to predicting bounce rate but is considered highly important according to the model-averaging model. A little thought reveals that many economic decisions are made according to regional patterns and this dependence should not be ignored. If visitors from certain regions are more likely to bounce, then discounting this dependence by using the BIC-based single model is a violation of intelligence use of the data.

In general, if the action to change the visitor’s experience is poorly tuned, the platform risks losing customer goodwill. Even if the model averaging approach may have only a small accuracy advantage over the single model approach on any single finite instance of data, it offers a more stable prediction of how to tune a visitor’s experience. A major risk to not using the model-averaged approach is that as more data and more relationships are entertained, the assumption of single-model adequacy will become increasingly untenable with the ramifications that unreliable and unstable predictions are reached. Decisions to alter visitors’ experience on the platform, when adjusted according to predictions with unknown and unreliable precision, becomes a gamble in the success of the platform.
Fig 5: Beta regression model inclusion frequencies for bounce rate value in bootstrap-based model averaging approach using 500 bootstrap replications of training data. A total of 8192 candidate models were explored with 9 being selected at least once in one of the 500 bootstrap replicates. AIC was used as the model selection criteria for the model averaging procedure.
Table 4
Variable inclusion result for the 9 different bounce rate value beta regression models selected in the bootstrap-based model averaging approach. Inclusion percentages are given as the proportion of times an explanatory variable was included within one of the 9 models. AIC, BIC, and $R_p^2$ columns correspond to variable inclusion results for the beta regression models returned by single model selection on the training data with AIC, BIC, and $R_p^2$. 

<table>
<thead>
<tr>
<th>Variable</th>
<th>Inclusion</th>
<th>m1</th>
<th>m2</th>
<th>m3</th>
<th>m4</th>
<th>m5</th>
<th>m6</th>
<th>m7</th>
<th>m8</th>
<th>m9</th>
<th>AIC</th>
<th>BIC</th>
<th>$R_p^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Administrative</td>
<td>100%</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SpecialDay</td>
<td>100%</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Month</td>
<td>100%</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Browser</td>
<td>100%</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Region</td>
<td>100%</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>TrafficType</td>
<td>100%</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>VisitorType</td>
<td>100%</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Informational</td>
<td>44%</td>
<td>x</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
<td>x</td>
<td>✓</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td>✓</td>
</tr>
<tr>
<td>ProductRelated</td>
<td>44%</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td>✓</td>
</tr>
<tr>
<td>Duration</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weekend</td>
<td>33%</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ProductRelated</td>
<td>22%</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>✓</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Administrative</td>
<td>11%</td>
<td>x</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Duration</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Informational</td>
<td>11%</td>
<td>x</td>
<td>x</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Duration</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5
Out-of-sample performance metrics for exit and bounce rate beta regression models acquired through 10-fold cross validation over the entire clickstream data set.

<table>
<thead>
<tr>
<th>Response</th>
<th>Technique</th>
<th>$R_p^2$</th>
<th>log S</th>
<th>CRPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>ExitRate</td>
<td>MA</td>
<td>0.29</td>
<td>2.27</td>
<td>0.017</td>
</tr>
<tr>
<td></td>
<td>AIC</td>
<td>0.24</td>
<td>2.26</td>
<td>0.019</td>
</tr>
<tr>
<td></td>
<td>BIC</td>
<td>0.23</td>
<td>2.21</td>
<td>0.021</td>
</tr>
<tr>
<td></td>
<td>$R_p^2$</td>
<td>0.25</td>
<td>2.26</td>
<td>0.019</td>
</tr>
<tr>
<td>BounceRate</td>
<td>MA</td>
<td>0.19</td>
<td>3.69</td>
<td>0.017</td>
</tr>
<tr>
<td></td>
<td>AIC</td>
<td>0.17</td>
<td>3.69</td>
<td>0.017</td>
</tr>
<tr>
<td></td>
<td>BIC</td>
<td>0.16</td>
<td>3.68</td>
<td>0.018</td>
</tr>
<tr>
<td></td>
<td>$R_p^2$</td>
<td>0.18</td>
<td>3.69</td>
<td>0.018</td>
</tr>
</tbody>
</table>
5.5.4. Comparison to Machine Learning Models. Results from the previous section given in Table 5 highlight the benefit of model-averaging against single-model approaches. However, this comparison may be disregarded by some because the single-model beta regression approach is not considered sophisticated enough. Machine learning (ML) models would seemingly be a stronger comparison to make against the MA method. Two learning methods are chosen, deep feedforward neural networks (NN) and random forest regression (RF). Readers are directed to Biau and Scornet (2016) for a review of random forests and Emmert-Streib et al. (2020) for a review of deep learning with neural networks. Random forest regression was chosen as the underlying algorithm is similar to model-averaging; a prediction from RF is the mean of the predictions from many regression trees trained on a random subset of the data and features. This is designed to both improve the accuracy and precision of predictions. Deep neural networks were chosen as they are lauded as universal function approximators and can powerfully capture relationships in the data that otherwise weaker methods could not. It should be emphasized that both of the ML models chosen can achieve higher predictive performance but surrender interpretability. It is this loss of interpretability that makes the ML models immediately less appealing than the MA method.

The architecture of the NNs consisted of two hidden layers with the first layer comprised of 14 neurons and the second layer had 5. The choice of the number of nodes within the layers should be between the number of input and output, in this case, 14 possible input predictors with one output. Optimization of the number of nodes was not pursued as the moderate size of the training dataset would make optimization a task in overusing the data. The loss function was set to the sum of squared errors and the activation function was the logistic function. The output from the last hidden layer was transformed with the activation function to ensure (0,1) predictions were given. The architecture of the RFs consisted of 500 regression trees where five predictors were randomly sampled as candidate predictors at each split within the training of the individual trees. During training of each tree, 63.2% of the training data was sampled for direct training with the other 36% used as out-of-bag data to test the single tree. Evaluation of predictive performance was assessed by evaluating the mean-squared prediction error (MSPE) and mean absolute prediction error (MAPE) through 10-fold cross validation over all observations from the clickstream dataset. Each of the three methods received the same train-test subsets for each of the 10 folds. Results are given in Table 6.

It is clear that the beta regression bootstrap-based model averaging achieved MSPE and MAPE values smaller than the NNs for both exit and bounce rate prediction. Unlike bootstrap-MA, neural networks do no sort of averaging of predictions and cannot control for excessive variance from noise. This lack of averaging out the noise can explain why the NNs underperformed MA. When compared to RF, MA showed better performance with respect to MSPE but worst performance with respect to MAPE. It is not easy to explain this reversal without uncertain speculation but MAPE is known to weight individual errors equally and hence will not penalize large errors any differently. On the other hand, MSPE places a larger penalty on larger errors and is useful in the presence of outlying values. These authors conclude that MA is competitive with RF but the choice of how the methods are evaluated should be thoughtfully reached. Overall, the results in Table 6 indicate that the method is competitive to machine learning models.
Table 6
Estimated predictive performances’ as measured by mean-squared prediction error (MSPE) and mean-absolute prediction error (MAPE) of beta regression bootstrap-MA (MA), feedforward deep neural network (NN), and random forest regression (RF) methods. Estimation was conducted through 10-fold cross-validation over the entire clickstream dataset with all three methods receiving the same train-test splits. Values for explanatory variables, variables excluding BounceRate, ExitRate, and Revenue, were normalized prior to the training of the neural networks. This scaling is important for neural networks to enhance learning.

<table>
<thead>
<tr>
<th>Response</th>
<th>Method</th>
<th>MSPE</th>
<th>MAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ExitRate</td>
<td>MA</td>
<td>0.0018</td>
<td>0.027</td>
</tr>
<tr>
<td></td>
<td>NN</td>
<td>0.030</td>
<td>0.028</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.0026</td>
<td>0.0018</td>
</tr>
<tr>
<td>BounceRate</td>
<td>MA</td>
<td>0.0026</td>
<td>0.022</td>
</tr>
<tr>
<td></td>
<td>NN</td>
<td>0.028</td>
<td>0.025</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.0035</td>
<td>0.0021</td>
</tr>
</tbody>
</table>

5.5.5. Comparison to Alternative Regression Models. When faced with proportional exit and bounce rate data, it is reasonable that a modeller may choose a more familiar model type instead of the less well known beta model. Recognizing this possibility, three alternative single-model structures were fitted and assessed for suitability to the clickstream data. It was important to highlight shortcomings of these alternatives to the beta regression model. Results from OLS regression, binomial logistic regression, and a non-linear logit transformation of the response variable followed by OLS regression, are provided in Section K of the Supplementary Material (Allenbrand and Sherwood, 2021). At the single model level, fitting constant dispersion beta regression models to the exit and bounce rate data represented a model structure more flexible and capable of capturing distributional regularities of the proportional and bounded data better than the other alternatives considered. Evaluation of prediction accuracy for the alternative models was not conducted. All three alternatives to a beta model are known to generate biased and inconsistent estimates because of direct violations of model assumptions. Unreliable predictions follow from unreliable models, so comparisons of the predictions would be unavailing.

6. Additional Policy Indications and Future Work. In any empirical setting where random variables demonstrate a beta distribution conditional on other variables and where model selection is pursued, the bootstrap model averaging (bootstrap-MA) method presented can establish notions of model stability and dynamic consistency, rankings of variable importance, and more reliable predictions. Consider in manufacturing and other process industries where the objective of quality which might be measured as the proportion of total occurrences that result in failure. What if a manufacturing firm has many plant locations with different types of employees at those locations, different protocols, different technologies, different suppliers, and so on, which are regarded as predictors of failure rate. Based on beta regression modeling results derived from one month worth of data, corporate decision makers promote one plant, defined by its set of predictor values, as the combination to be copied. This conversion process incurs costs such as the time and money to switch out a technology. A problem manifests when, on the next month, results over the new data instance suggests a change to the predictor values. If this dynamic instability and inconsistency continues, the firm would either alternate between investments or face indecision. Instead, the bootstrap-MA could have been executed with its capability to explore multiple models, rank the
models by selection frequency, rank predictors by inclusion frequencies in the models, and allow model-averaging of estimates for the mean and precision of the failure rate. This would be done in the first month, before costly investments are made.

The merits of bootstrap-MA are not confined to the domain of physically produced goods but is expected to hold for services as well. Take the success rate of immunization, a complex function of myriad explanatory factors such as recipient specific features such as age or immunization specific features such as the mode of delivery. Choosing to model the success rate in a beta regression framework invites issues of inter-data reliability and stability of estimates from the possible existence of many supported models. The bootstrap-MA procedure can be generalized to a diverse plethora of settings: the availability of a service, what proportion of packages are received by a given delivery time, customer approval, turnover rate of employees, the rate of bankruptcy for firms, labor force participation rate, and the proportion of countries that assent to climate change mitigation plans.

Future research includes an examination of the role that the selection of different model structures may have on fitted model conclusions. Model selection uncertainty was shown to strongly influence precision and model conclusions but this was after a certain structure was elected. In reality, a modeler may not know which structure to use and would prefer a data-driven approach to assist. What are the implications of this beyond adding extra uncertainty would be worth pursuing. In this work, a data transformation was applied to make the interval of response variables open and bounded instead of closed and bounded. If the number of observations of the variable whose values are on the extremes of the interval are numerous, and these end-points are meaningful, i.e. mapping to an open-interval is thought to result in the loss of information, then a zero/one inflated beta regression model approach needs to be sought. In particular, the problem of how to incorporate zero/one inflated beta regression models into the model averaging arena warrants attention. The work presented in this paper used a frequentist approach to beta regression model averaging which imposes limitations on interpreting model-averaged estimates of regression coefficients. A Bayesian approach to beta regression model averaging, with its powerful hierarchical formulations, would allow a direct estimation of effect sizes (regression coefficients) thus permitting an interpretation of the model-averaged coefficients in terms of posterior distributions. One must handle the specification of priors for each model and the parameters for the mean and precision parameter regression sub-models to be able to perform a full Bayesian beta regression model-averaging.

7. Conclusion. Statistical model development encompasses a collection of decisions including what structure to use, how to select between competing models, and how to make inferences from fitted models. When a quantity to be modeled is proportional and bounded, the beta regression structure becomes a natural candidate to explore. Embedded in the entire model development pipeline is the uncertainty originating from the existence of multiple models supported by the data. This work presented a formalized tool that can combine model selection uncertainty and beta regression modeling. A combination of bootstrap model averaging, model selection, and asymptotic theory yielded a procedure (bootstrap-MA) that can perform joint modeling of the mean and precision parameters of a beta model, capture multi-model selection uncertainty, and achieve more accurate claims of estimate precision, variable importance, and model stability and consistency. A careful simulation study revealed that confidence intervals for model-averaged estimates of the mean for a constant and non-constant precision beta-distributed variable achieved coverage much closer to the chosen level. It was shown
that using single model intervals resulted in severe undercoverage, a result of not accounting for multi-model selection uncertainty. Further, the bootstrap-MA procedure produced more reliable out-of-sample prediction performance when compared to single model selection techniques. Empirical utility of the bootstrap-MA procedure was further highlighted by applying it to clickstream data. It was found that bootstrap-MA, by directly factoring in the existence of competing models for the average exit and bounce rate on an eCommerce platform, could serve as a multi-tool that is capable of expressing model selection stability and dynamic consistency, variable importance, and inter-data estimate reliability in a rigorous and interpretable way. It was further shown that the prediction power of the MA-based models were competitive with both deep neural networks and random regression forests. Replacement of the beta model component of bootstrap-MA with alternative regression structures, such as the normal linear model, was shown to have major limitations. Overall, the beta regression bootstrap-MA procedure would be incredibly useful when the relationships under study are numerous and complex and when random noise contained in the data can lead to unstable and constantly changing modeling conclusions.

SUPPLEMENTARY MATERIAL


REFERENCES


Ansari, S. et al. (2002). Integrating E-Commerce and Data Mining: Architecture and Challenges. In IEEE International Conference on Data Mining.


