The Annals of Applied Statistics
2023, Vol. 17, No. 1, 680–710
https://doi.org/10.1214/22-AOAS1647

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[a] AND BEN SHERWOOD[b]

School of Business, University of Kansas, [a]*callenbrand@ku.edu,* [b]*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 nonrigorously 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 postselection conclusions. Regression models, based on the beta distribution, are a class of nonlinear models, attractive because of their great flexibility and potential explanatory power, but have not been investigated from the standpoint of multimodel 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

Received January 2022; revised May 2022.

Key words and phrases. Beta distribution, generalized linear model, variable selection, model combination, corrected AIC, rates and proportions.