Predicting Rare Events by Shrinking Towards Proportional Odds

Gregory Faletto & Jacob Bien

University of Southern California

September 8, 2023

Presented by Quinn Lanners

Introduction

Goal: To predict the probability of an ordinal outcome, $k \in \{1, ..., K-1\}$, where the later outcomes are rare.

The paper presents a method, **PRESTO**, as a relaxation of the proportional odds model for ordinal regression. The authors are focused on estimating probabilities of *rare events*.

Real-world examples

- In online marketing, a customer is first served an ad, then may click on it, then may indicate interest in making a purchase (by "liking" the product, for example), and finally may make a purchase.
- In health and medicine, many outcomes can be encoded as ordered categorical variables, like reported quality of life and disease progression (Norris et al., 2006).
- Sales of high-price durable goods typically follow a sales funnel (Duncan Elkan, 2015). For example, when buying a car often a potential buyer first comes in to see a car, may take a test drive, and finally may buy the car.

Background

The authors draw largely from the proportional odds model (McCullagh, 1980).

For ordinal outcomes $k \in \{1, \ldots, K-1\}$, given a feature vector $\boldsymbol{x} \in \mathbb{R}^p$ and the corresponding class label $y \in \{1, \ldots, K-1\}$,

$$\log \left(\frac{\mathbb{P}\left(y \leq k \mid \boldsymbol{x}\right)}{\mathbb{P}\left(y > k \mid \boldsymbol{x}\right)} \right) = \alpha_k + \boldsymbol{\beta}^\top \boldsymbol{x}, \tag{1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of weights.

Motivating Theory

Problem: Logistic regression struggles at estimating rare event probabilities as class-imbalance worsens.

Theory:

- *Theorem 2.2*: A logistic regression model is better able to estimate probabilities (of rare events) when the parameters $\boldsymbol{\beta}$ are known.
- *Theorem 2.3* The proportional odds model can precisely estimate $\boldsymbol{\beta}$ as long as two adjacent classes are reasonably common, even if the remaining classes are arbitrarily rare.

Conclusion: The authors conclude that the proportional odds model can better estimate probabilities of rare events because it leverages data from decision boundaries between abundant classes to better estimate decision boundaries near rare classes.

Current Problem

Note that the proportional odds model assumes that all the decision boundaries are parallel.

The authors claim that the proportional odds model may be too rigid to be realistic as the different decision boundaries may not be parallel.

Real-world examples

- In online marketing, users may click on an ad only to realize that the product is not what they were expecting, resulting in a particularly low probability of purchase.
- For expensive goods like a home or car, potential buyers may express interest by going on a tour or taking a test drive purely out of curiosity; this may be distinct from their level of interest in actually making a purchase.
- Students may place weights on different factors when deciding whether to apply to graduate school than they did when deciding whether to apply to an undergraduate program—they may have more appealing alternatives to additional schooling, they may face new financial or personal constraints because they are older, etc.

Proposed Solution

The authors propose casting the problem as $K - 1$ binary classification problems for adjacent classes

$$\log \left(\frac{\mathbb{P}\left(y \leq k \mid \boldsymbol{x}\right)}{\mathbb{P}\left(y > k \mid \boldsymbol{x}\right)} \right) = \alpha_k + \boldsymbol{\beta}_k^\top \boldsymbol{x}, \qquad k \in \{1, \ldots, K-1\}. \tag{2}$$

and imposing ℓ_1 penalties on adjacent $\boldsymbol{\beta}_k$ vectors to limit how much they vary.

This is reminiscent of the fused lasso (Tibshirani et al., 2005).

Method

The authors call their method PRESTO: Predicting Rare Events by Shrinking Towards proportional Odds.

They solve the following optimization problem for data $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$:

$$\begin{aligned} \arg\min_{\boldsymbol{\beta}, \boldsymbol{\alpha}} \Bigg\{ & -\frac{1}{n} \sum_{i=1}^{n} \log \Bigg[F\left(\alpha_{y_i} + \boldsymbol{\beta}_{y_i}^\top \boldsymbol{x}_i\right) \\ & - F\left(\alpha_{y_i - 1} + \boldsymbol{\beta}_{y_i - 1}^\top \boldsymbol{x}_i\right) \Bigg] \\ & + \lambda_n \left(\sum_{j=1}^{p} |\beta_{j1}| + \sum_{j=1}^{p} \sum_{k=2}^{K-1} |\beta_{jk} - \beta_{j,k-1}| \right) \Bigg\}, \end{aligned} \tag{3}$$

Consistency Results

Assumptions:

- An assumption that the observed data, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$, can be modeled by a linear model with flexible portions for each $\boldsymbol{\beta}_k$.
- An assumption that none of the decision boundaries cross in the support of $\boldsymbol{X}$.

With these assumptions, the author's prove that PRESTO is a consistent estimator of $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{K-1}$.

Experimental Results

The authors include results for two synthetic and two real-world experiments. They compare to logistic regression and the proportional odds model. Since they are interested in estimating the *probability* of a rare event, they use the mean squared error as their metric of choice.

Synthetic Experiments: 2500 samples with ten covariates and four ordinal classes. The final class is the rare class of interest and varies in rate.

Real-world Experiments:

- `soup` data set from the R `ordinal` package.
- `PreDiabetes` data set from the R `MLDataR` package.

Experimental Results

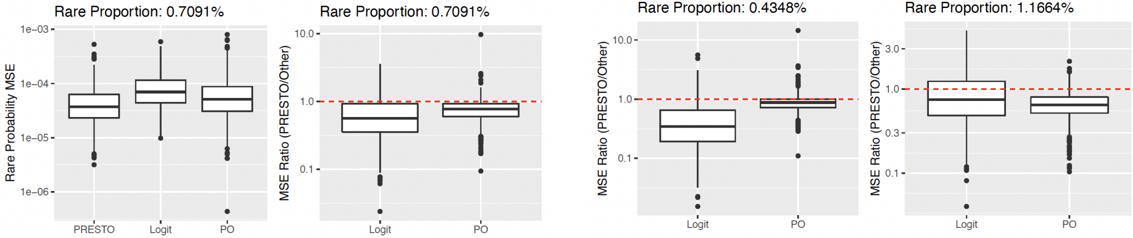


Figure: Experiment 1: Synthetic DGP favorable for PRESTO.

Experimental Results

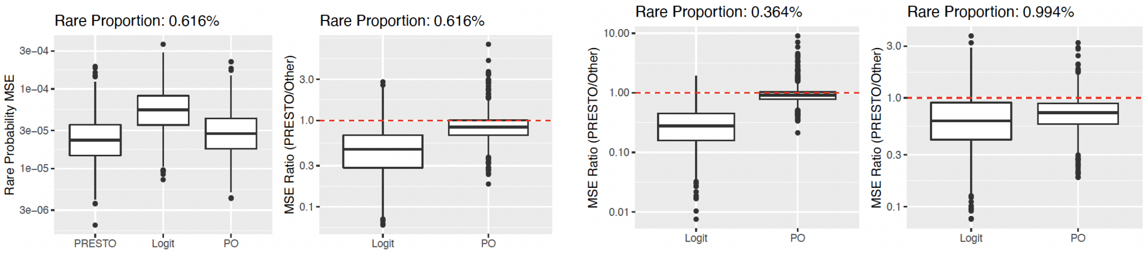
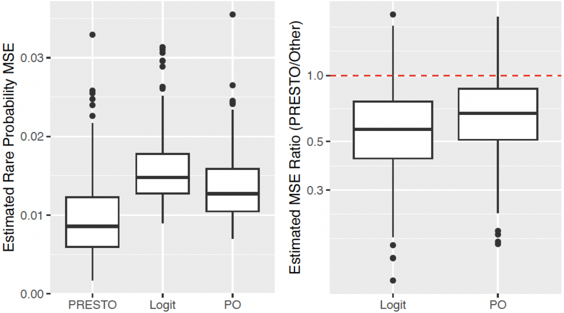


Figure: Experiment 2: More difficult synthetic DGP.

Experimental Results

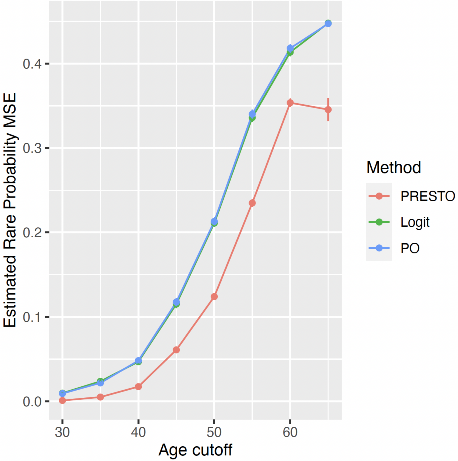
Data from study where 1,857 participants tasted soups and said if they thought each soup was an existing product or a new test product (Christensen2011). The respondents also stated how sure they were in their response on a three-level scale, yielding a total of $K = 6$ possible ordered outcomes.



Note: True probabilities are not observed so authors use a method similar to expected calibration error (Naeini et al., 2015) to estimate the probabilties and compute MSE.

Experimental Results

Data of 3,059 patients who were diagnosed with diabetes at various ages. The authors treat the disease state as ordinal classes (no diabetes, prediabetes, diabetes) and evaluate the methods at different age cutoffs.



Conclusions

Summary: PRESTO is an accurate method for probability estimation of rare ordinal events. Like the proportional odds model, it leverages more plentiful data from earlier decision boundaries to help predict later events that are more rare. But by allowing the coefficient vectors to vary, it is more accurate for probability estimation when decision boundaries are not parallel.

Things I Liked:

- Author's writing style is clean and they provide good examples and intuition for their theory.
- Idea is a simple extension and improves performance in certain settings.

Things I Found Lacking:

- Experiments (especially real-world) were not overly convincing.
- Supplemental section was extraordinarily long. The author's seemed to convulate their idea at times by just putting more stuff in the paper.

Recommendations:

- **Worth reading?** Perhaps. The paper is theory-heavy but is well written.
- **Worth implementing?** Yes. If you have a use case, they have a very clean GitHub repo that would be worth exploring.